NONLINEAR INTERACTIVE SOURCE-FILTER MODEL FOR VOICED SPEECH

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
MIDDLE EAST TECHNICAL UNIVERSITY

BY

TURGAY KOÇ

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF DOCTOR OF PHILOSOPHY
IN
ELECTRICAL-ELECTRONICS ENGINEERING

SEPTEMBER 2012

Approval of the thesis:

NONLINEAR INTERACTIVE SOURCE-FILTER MODEL FOR VOICED SPEECH

submitted by **TURGAY KOÇ** in partial fulfillment of the requirements for the degree of **Doctor of Philosophy in Electrical-Electronics Engineering Department, Middle East Technical University** by,

Prof. Dr. Canan Özgen
Dean, Graduate School of **Natural and Applied Sciences**

Prof. Dr. İsmet Erkmen
Head of Department, **Electrical-Electronics Engineering**

Prof. Dr. Tolga Çiloğlu
Supervisor, **Electrical - Electronics Engineering Department, METU**

Examining Committee Members:

Prof. Dr. Mübeccel Demirekler
Electrical - Electronics Engineering, METU

Prof. Dr. Tolga Çiloğlu
Electrical - Electronics Engineering, METU

Prof. Dr. Aydın Alatan
Electrical - Electronics Engineering, METU

Assoc. Prof. Dr. Çağatay Candan
Electrical - Electronics Engineering, METU

Assist. Prof. Dr. Yakup Özkazanç
Electrical - Electronics Engineering, Hacettepe University

Date:

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name: TURGAY KOÇ

Signature :

ABSTRACT

NONLINEAR INTERACTIVE SOURCE-FILTER MODEL FOR VOICED SPEECH

Koç, Turgay
Ph.D., Department of Electrical-Electronics Engineering
Supervisor : Prof. Dr. Tolga Çiloğlu

September 2012, 93 pages

The linear source-filter model (LSFM) has been used as a primary model for speech processing since 1960 when G. Fant presented acoustic speech production theory. It assumes that the source of voiced speech sounds, glottal flow, is independent of the filter, vocal tract. However, acoustic simulations based on the physical speech production models show that, especially when the fundamental frequency (F0) of source harmonics approaches to the first formant frequency (F1) of vocal tract filter, the filter has significant effects on the source due to the nonlinear coupling between them. In this thesis, as an alternative to linear source-filter model, nonlinear interactive source-filter models are proposed for voiced speech.

This thesis has two parts, in the first part, a framework for the coupling of the source and the filter is presented. Then, two interactive system models are proposed assuming that glottal flow is a quasi-steady Bernoulli flow and acoustics in vocal tract is linear. In these models, instead of glottal flow, glottal area is used as a source for voiced speech. In the proposed interactive models, the relation between the glottal flow, glottal area and vocal tract is determined by the quasi-steady Bernoulli flow equation. It is theoretically shown that linear source-filter model is an approximation of the nonlinear models. Estimation of ISFM's parameters from only speech signal is a nonlinear blind deconvolution problem. The problem is solved by a robust method developed based on the acoustical interpretation of the systems. Experimental results show that ISFMs produce source-filter coupling effects seen in the physical simulations and the parameter estimation method produce always stable and better performing models than LSFM model. In addition, a framework for the incorporation of the source-filter interaction into classical source-filter model is presented. The Rosenberg source model is extended to an interactive source for voiced speech and its performance is evaluated on a large speech database. The results of the experiments conducted on vowels in the database show that the interactive Rosenberg model is always better than its noninteractive version.

In the second part of the thesis, LSFM and ISFMs are compared by using not only the speech signal but also HSV (High Speed Endocopic Video) of vocal folds in a system identification approach. In this case, HSV and speech are used as a reference input-output data for the analysis and comparison of the models. First, a new robust HSV processing algorithm is developed and applied on HSV images to extract the glottal area. Then, system parameters are estimated by using a modified version of the method proposed in the first part. The experimental results show that speech signal can contain some harmonics of the fundamental frequency of the glottal area other than those contained in the glottal area signal. Proposed nonlinear interactive source-filter models can generate harmonics components in speech and produce more realistic speech sounds than LSFM.

Keywords: Speech Modeling, Source-Filter Interaction, Nonlinear System Modeling, Nonlinear System Identification, High Speed Endoscopic Video Processing

ÖZ

SESLİ SESLER İÇİN DOĞRUSAL OLMAYAN ETKİLEŞİMLİ KAYNAK-SÜZGEÇ MODELİ

Koç, Turgay
Doktora, Elektrik - Elektronik Mühendisliği Bölümü
Tez Yöneticisi : Prof. Dr. Tolga Çiloğlu

Eylül 2012, 93 sayfa

Doğrusal kaynak-süzgeç modeli (DKSM) Gunnar Fant tarafından önerildiği 1960'dan beri konuşma işlemede birincil model olarak kullanılmaktadır. Bu model de kaynak glottal akım ve süzgeç vokal boşluk olup bunların birbirlerinden bağımsız olduğu kabul edilir. Fakat, konuşma sisteminin fiziksel modelleri ile yapılan simülasyonlar, özellikle ses tellerinin titreşim frekansı vokal boşluğun birinci rezonans frekansına yaklaştığında, vokal boşluğun glottal akım üzerinde önemli etkileri olduğunu göstermiştir. Bu tezde klasik doğrusal kaynak-süzgeç modeline alternatif olarak sesli sesler için kaynak ve süzgecin birbirini bağlı olduğu yeni doğrusal olmayan etkileşimli kaynak-süzgeç modeli önerilmektedir.

Bu tez iki kısımdan oluşmaktadır. Birinci kısımda kaynak ve süzgeçin etkileşimi için bir platform sunulmaktadır. Ardından vokal boşluktaki akustiğin doğrusal ve ses tellerindeki hava akışının durağan-benzeri Bernoulli akımı olduğu kabul edilerek, iki adet doğrusal olmayan etkileşimli kaynak-süzgeç modeli (EKSM) önerilmiştir. Bu modelde klasik modelden farklı olarak glottal alan kaynak olarak kabul edilmiş, glottal akım, glottal alan ve vokal boşluk Bernoulli denklemi ile ilişkilendirilmiştir. Teorik olarak DKSM'nin bu modellerin bir yaklaşımı olduğu gösterilmiştir.

Önerilen modellerin parametrelerinin sadece konuşma sinyalinden bulunması doğrusal olmayan bir ters evrişim problemidir. Bu problem konuşmanın akustik teorisinden faydalanılarak geliştirilen güçlü bir algoritma ile çözülmüştür. Yapılan deney sonuçları EKSM'lerinin fiziksel sistemlerde gözlenen kaynak-süzgeç etkileşimini üretebildiğini göstermiş ve aynı zamanda önerilen parametre tahmin algoritması her zaman kararlı, doğrusal modelden daha iyi doğrusal olmayan modeller üretmiştir. Ayrıca, klasik kaynak modellerini etkileşimli kaynak modeline dönüştürmek için bir çerçeve sunulmuştur. Konuşma literatüründe yaygın olarak kullanılan Rosenberg kaynak modeli etkileşimli hale getirilerek büyük bir konuşma veri tabanı üzerinde etkileşimsiz model ile karşılaştırılmıştır. Ötümlü sesler üzerinde yapılan karşılaştırma sonuçları etkileşimli kaynak modelinin her zaman etkileşimsiz modelden daha iyi olduğunu göstermiştir.

Tezin ikinci kısmında DKSM ve EKSM modelleri hem konuşma sinyali hemde ses tellerinin yüksek hızlı endoskopik görüntüleri (YHEG) kullanılarak bir sistem tanıma işlevi biçiminde karşılaştırılmıştır. Bu bölümde YHEG ve konuşma sinyalleri modellerin giriş ve çıkış verilerini oluşturup analiz ve karşılaştırma için referans olması amacıyla kullanılmıştır. Öncelikle YHEG'lerden ses telleri arasındaki alanın çıkarılması amacıyla yeni bir resim işleme algoritması geliştirilmiş ve YHEG üzerine uygulanmıştır. Ardından, modellerin parametreleri, elde edilen glottal alan ve konuşma sinyalleri kullanılarak, daha önce önerilen parametre kestirim algoritmasında değişiklik yapılarak elde edilmiştir. Yapılan deneylerde konuşma sinyalinin kaynak sinyalinde olmayan bazı harmonikleri içerdiği gözlenmiştir. Bu durum konuşma üretim sisteminin doğrusal olmadığını göstermektedir. Doğrusal kaynak süzgeç modeli veya herhangi bir doğrusal sistem ile yeni harmonikler üretmek mümkün değildir. Bununla birlikte önerilen EKSM modelleri kaynak sinyalinde olmayıp konuşma sinyalinde olan harmonikleri kaynak sinyalinin harmoniklerini geliştirerek oluşturmuş ve DKSM'den daha gerçekci konuşma sinyalleri üretmiştir.

Anahtar Kelimeler: Konuşma modelleme, kaynak-süzgeç etkileşimi, doğrusal olmayan sistem modelleme, doğrusal olmayan sistem tanıma, yüksek hızlı endoskopik görüntü işleme

To my mother, Rabia
To my wife, Medine
To my lovely daughter, Nil.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gradidute to my advisor, Prof. Dr. Tolga Çiloğlu, for his guidance, patience, encouragement and contribution to my point of view during this very long Ph.D. period at METU. I greatly appreciate not only his share in every step taken in the developement of the thesis but also all support that he provided to me all these years.

I would like to thank the member of my thesis committee, Prof. Dr. Mübeccel Demirekler and Asist.Prof.Dr. Yakup Özkazanç for their support and suggestions which improved the quality of the thesis.

I am also very grateful to Eren Akdemir who is myroommate during this period for his friendship. He is always with me during my time at METU.

I also would like to thank to my fellows: Akif Durdu, Umut Tilki, Evren Ekmekçi, Sebahattin Topal, Atilla Dönük, Barış Özyer, Yücel Özbek for the ideas they share with me, their support and friendship through these years.

I would like to thank to my brother Halil Koç for all kind of support he gives me during my career and my life.

Finally, I would like to thank to my wife, Medine, for her unlimited support, patience and love at all moments of my life.

TABLE OF CONTENTS

ABSTRACT iv

ÖZ vi

ACKNOWLEDGMENTS ix

TABLE OF CONTENTS x

LIST OF TABLES xiii

LIST OF FIGURES xiv

CHAPTERS

1 INTRODUCTION 1
    1.0.1 Motivation 1
  1.1 The Outline of the Thesis 2
2 SPEECH PRODUCTION MODELS 4
  2.1 Introduction 4
  2.2 Vocal Folds 4
  2.3 Vocal Tract 8
  2.4 Acoustic models 9
    2.4.1 Wave Reflection Method 9
    2.4.2 Transmission Line Circuit Analog 10
    2.4.3 Hybrid Time - Frequency Domain Model 10
  2.5 Aerodynamic & Acoustic Interactions in Speech Production 10
  2.6 A Review of Source-Filter Coupling 13
    2.6.1 The effects of interaction on the glottal flow 14
3 MODELING 20
  3.1 Theoretical Background 20
    3.1.1 Glottal Aerodynamics 20

3.1.2 Vocal Tract Acoustics 24

3.1.3 Chain Matrix 24

3.1.4 DT-Representation of Chain Matrix for Lossless Vocal Tract 26

3.2 Interactive Source-Filter Modeling 28

3.2.1 Modification of Interactive Model 31

3.2.2 Analysis of ISFMs 33

4 PARAMETER ESTIMATION 35

4.1 Introduction 35

4.2 Parameter Estimation 35

4.3 Estimation of Vocal Tract Filter & Glottal Flow from Speech 36

4.3.1 Source-Filter Optimization 37

4.4 Nonlinear Source-Filter Model Parameter Estimation Algorithm . . . 38

4.4.1 Glottal Parameterization 40

4.4.2 Nonlinear Model Parameter Optimization 40

4.5 Optimization of the parameters of ISFMs 41

4.5.1 Gradient Based Optimization Algorithm 41

4.5.2 Optimization of the scalar parameter A_{max} 43

4.5.3 Optimization of the parameter vector $\mathbf{b_k}$ 46

4.5.4 Optimization of the parameter vector $\mathbf{q_k}$ 47

4.5.5 Equalization of linear and nonlinear models using initial parameter values 48

4.5.6 Multi-parameter optimization algorithm 48

4.6 Summary 50

5 PRELIMINARY EXPERIMENTS 51

5.1 Introduction 51

5.2 Experiments on a short speech segment 51

5.3 Summary 60

6 COMPARISON OF LSFM AND ISFM 61

6.1 Introduction 61

6.1.1 Comparison of LSFM and ISFMs 61

6.1.2 Results & Discussion 63

7 EXTRACTION OF GLOTTAL AREA FROM HIGH SPEED ENDOSCOPIC VIDEO (HSV) 69

7.1 Introduction 69

7.1.1 Automatic HSV Processing Algorithm 69

8 HSV BASED
ANALYSIS AND COMPARISON OF LSFM AND ISFMs 79

8.1 Introduction 79

8.2 Experiments 79

8.3 Summary 83

9 CONCLUSION 85

9.1 Contributions 86

REFERENCES 88

VITA 92

LIST OF TABLES

TABLES

Table 3.1 The limiting conditions of the model parameters 34

Table 6.1 The number of glottal flow cycles for each vowel used in the experiments . 63

Table 6.2 Average of MSE calculated over glottal flow cycles for each vowel. For the ISFM1 and ISFM2 the PISFM1 and PISFM2 represents the ratio over Rosenberg model 64

Table 6.3 Average Error Variances calculated over glottal flow estimates for each vowel 64

LIST OF FIGURES

FIGURES

Figure 2.1 Human Vocal System (taken from [15]) 5

Figure 2.2 The Vocal Folds (taken from [15]) 5

Figure 2.3 The movement of vocal folds (a) and corresponding glottal volume velocity (b) 6

Figure 2.4 The vocal fold models 7

Figure 2.5 The Vocal Tract Modeling 9

Figure 2.6 A global view of aerodynamic and acoustic interactions in human vocal system 11

Figure 2.7 Input - Output Representation of Vocal Fold Aerodynamics 12

Figure 2.8 Input - Output Representation of Acoustic Models 12

Figure 2.9 Vocal tract input impedance for male vowel /a/ 15

Figure 2.10 Time and amplitude based normalized glottal area (black lines) and glottal flow waveforms (red lines) obtained by the simulation of vowel /a/ with wave-reflection method for glottal area fundamental frequency, $F0$ (a) 80 Hz, (b) 120 Hz, (c) 160 Hz, (d) 200 Hz, (e) 240 Hz. 16

Figure 2.11 Vocal tract input impedance for male vowel /i/ 17

Figure 2.12 Time and amplitude based normalized glottal area (black lines) and glottal flow waveforms (red lines) obtained by the simulation of vowel /i/ with wave-reflection method for glottal area fundamental frequency, $F0$ (a) 80 Hz, (b) 120 Hz, (c) 160 Hz, (d) 200 Hz, (e) 240 Hz. 18

Figure 3.1 Glottal pulse and particle velocity 22

Figure 3.2 Linear Source-Filter Model of Speech Production 28

Figure 3.3 (a) Interactive Source - Filter Model I (ISFM1) (b) Interactive Source - Filter Model II (ISFM2) 32

Figure 4.1 Parameter estimation algorithm 39

Figure 4.2 Multi-parameter optimization algorithm for the ISFMs 49

Figure 5.1 (a) Speech Signal, (b) The spectrum of speech (red line) and estimated vocal tract (blue line), (c) Glottal flow estimate, (d) Glottal flow derivative 52

Figure 5.2 Glottal parameterization step N_p- The glottal flow peak instants, N_{GCI} - The glottal closure instants, N_{GOI} - The glottal opening Instants 53

Figure 5.3 (Top) Calculated glottal area waveform using Fant's area model (black line) versus glottal flow estimated by inverse filtering (red line) and (bottom) their derivative waveforms 54

Figure 5.4 (Top) Interactive glottal flow produced by ISFM1 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives 54

Figure 5.5 (Top) error surface and (bottom) error contours for A_{max} and P_L 56

Figure 5.6 (Top) Interactive glottal flow waveform after optimizing A_{max} (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives ($\frac{\sqrt{P_L}}{\sqrt{2}A_{\max}} \approx 165$ and $\Gamma(H)_\infty = 1$) 57

Figure 5.7 (Top) Interactive glottal flow produced by ISFM2 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line)and (bottom) their derivatives ($\Gamma(H)_\infty = 64$) 57

Figure 5.8 (Top) Calculated glottal area waveform using Rosenberg+ model (black line) versus glottal flow estimated by inverse filtering (red line) and (bottom) their derivatives 58

Figure 5.9 (Top) Interactive glottal flow produced by optimal A_{max} (red line) ($\Gamma(H)_\infty =$ 1), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives 59

Figure 5.10 (Top) Interactive glottal flow produced by ISFM1 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives ($\Gamma(H)_\infty = 99$) 59

Figure 5.11 (Top) Interactive glottal flow produced by ISFM2 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives ($\Gamma(H)_\infty = 89$) 60

Figure 6.1 Experimental setup for LSFM & ISFMs using Rosenberg model 62

Figure 6.2 (color online) MSE Distributions of the models over all glottal cycles, non-interactive Rosenberg (dash-dot line), ISFM1 (dashed line), ISFM2 (solid line) . . 65

Figure 6.3 (color online) MSE scattering of non-interactive Rosenberg model (dot) and ISFM2 (star) based on F1/F0 ratio 66

Figure 6.4 (a) MSE Distribution on utterance /a/ (b) Scattered MSEs for LSFM & ISFM2 66

Figure 6.5 (a) MSE Distribution on utterance /e/ (b) Scattered MSEs for LSFM & ISFM2 66

Figure 6.6 (a) MSE Distribution on utterance /1/ (b) Scattered MSEs for LSFM & ISFM2 67

Figure 6.7 (a) MSE Distribution on utterance /i/ (b) Scattered MSEs for LSFM & ISFM2 67

Figure 6.8 (a) MSE Distribution on utterance /u/ (b) Scattered MSEs for LSFM & ISFM2 67

Figure 6.9 (a) MSE Distribution on utterance /y/ (b) Scattered MSEs for LSFM & ISFM2 68

Figure 6.10 (a) MSE Distribution on utterance /o/ (b) Scattered MSEs for LSFM & ISFM2 68

Figure 6.11 (a) MSE Distribution on utterance /2/ (b) Scattered MSEs for LSFM & ISFM2 68

Figure 7.1 Consecutive HSV images corresponding to one glottal cycle 70

Figure 7.2 TV Measure extracted from 300 consecutive frames 71

Figure 7.3 Glottal Region Determination 72

Figure 7.4 Glottal region determinated by masking(upper figure) and corresponding vertical slice images(lower figure) 73

Figure 7.5 Histogram obtained from vertical slice image (VSI) 74

Figure 7.6 Estimated GMM model for the VSI 75

Figure 7.7 Estimated binary images for VSI 76

Figure 7.8 Estimated consecutive binary images over open phase 77
Figure 7.9 Estimated glottal area signal 78

Figure 8.1 Speech vs Glottal area (Upper) and EGG vs Glottal area (Lower) 80
Figure 8.2 Interactive glottal flow by generated NLSF Models 81
Figure 8.3 Original and synthesized speech spectrums 82
Figure 8.4 Synthesized speech by LSF (Top) and ISFM1 (Bottom) 83

CHAPTER 1

INTRODUCTION

Speech is the result of many complex neurological and physiological processes in human body in order to communicate with the environment. It conveys valuable information about the identity, the language and the emotional state of the speaker. Hence, it is the information source in signal processing applications like speech synthesis, coding, enhancement, morphing, speech/speaker recognition, speaker verification and emotion recognition. Currently, in speech processing, either data-driven or acoustics based models are used to get the desired signal or information from speech. The most widely used model is the linear source-filter model (LSFM). It is theoretically based on the acoustics of the speech production [1]. The sound source, i.e., glottal flow, is generated by the vibration of vocal folds and propagates through vocal tract and radiates from lips to produce the speech sound. Vocal tract shape determines the speech sound by spectrally weighting the source spectrum and the mouth increases the high frequency content of the generated sound.

1.0.1 Motivation

Despite the fact that the research on the physics of speech production has been continuing for a considerable amount of time, the acquired knowledge have not been utilized extensively in the development of speech signal models. The linear source-filter model, the simplest approximation of the speech production system, has been used dominantly for about half a century. LSFM assumes that the source is independent of the filter. However, acoustic simulations based on the physical speech production models show that, especially when the fundamental frequency (F0) of the source harmonics approaches to the first formant frequency (F1) of vocal tract filter, the filter has significant effects on the source due to the nonlinear coupling between them.

The aim of this study is, based on the existing knowledge of the physics of speech production, to develop an interactive source-filter model that takes the effects of the nonlinear acoustic interaction of the source and the filter into account. In particular, a way of combining the glottal nonlinearity to the linear model of vocal tract, to produce the glottal flow, is introduced. Modeling the nonlinear source-filter interaction may contribute to model based speech processing applications.

1.1 The Outline of the Thesis

In this thesis, an interactive nonlinear source-filter model for voiced speech is proposed as an alternative to linear source-filter model. The approach in the modeling is to use the physical mechanism of the speech production using the fundamental theory of aerodynamics. In this context, the interaction is modeled based on the quasi-steady Bernoulli flow assumption, a specific solution of the Navier-Stokes equation at the glottis. The simulations of the speech production based on the linear acoustics and nonlinear glottal aerodynamics are presented in Chapter 2. Typical effects of the interaction for different vowels are illustrated by using areas extracted from 3D MRI (Magnetic Resonance Images) of vocal tract and a mathematical model of vocal folds. The source-filter coupling is discussed after presenting the simulation results.

The derivation of the proposed ISFM models are presented in Chapter 3. There are two proposed ISFMs in which ISFM2 is an extension of ISFM1. An analysis of the proposed models is given. Finally, the relation between the LSFM and proposed ISFMs are established. In fact, the LSFM is an approximation of ISFMs under some conditions. These conditions are stated at the end of the chapter.

The estimation of the parameters of the suggested ISFMs only from speech is a nonlinear blind estimation problem. In Chapter 4, a robust parameter estimation algorithm is proposed in order to solve the problem. The experimental results on a speech signal are presented in this chapter. The proposed algorithm always produces a stable and better performing model compared to LSFM.

A framework for the incorporation of the source-filter coupling into classical LSFM is presented in Chapter 5. The Rosenberg source model is extended to an interactive source model. The comparison of the interactive and noninteractive source models are conducted on a large speech database collected in our laboratuary from various speakers. The results show that the proposed ISFMs consistently outperform the LSFM.

Investigating the LSFM and the proposed ISFMs on a system identification platform requires not only the output of the systems but also the input. Due to the location of the glottis, currently, there is no way to measure the glottal flow on human body. However, it is possible to monitor vocal fold vibration by high speed endoscopic cameras (HSV). In Chapter 6, a new robust algorithm is proposed to extract useful information from HSV of vocal folds. The algorithm is applied to the images from IRCAM HSV database [56] and the results are presented.

Finally, the proposed ISFM models and LSFM are investigated by using source measurements obtained from HSV Images in a system identification fashion. The results show that speech production is nonlinear and it is well modeled by the quasi - steady Bernoulli flow assumption at the glottis. The proposed nonlinear systems have shown superiour performance compared to the linear model. ISFMs, like physical simulators, enhance the source harmonics and produce more realistic speech sounds than the linear source - filter model.

CHAPTER 2

SPEECH PRODUCTION MODELS

2.1 Introduction

Speech sounds are produced by the motion of tissue and air in human vocal system. There are two functionally different sections in the human vocal system, vocal folds, vocal tract shown in Fig. 2.1. The vocal folds work as a sound source by vibrating under the lung pressure during speaking. The vocal tract is a resonator modifiying the spectral content of the sound source depending on its shape. The lips radiates the sound waves from the vocal tract to the atmosphere.

2.2 Vocal Folds

Voiced speech sounds are generated by the vocal folds. The vocal folds are soft tissue structures and located in the neck. Their ability is to abduct (move apart) during respiration and to adduct prior to and during voicing. They are the point of division between subglottal (below the vocal folds) and supraglottal (above the vocal folds) airways. The airspace between the two vocal folds is called glottis and the the airflow passing through the glottis is called glottal flow. The structure of vocal folds can be divided into two physiologically different layers, body and cover. The body layer consists of muscle fiber while the cover is pliable mucosal tissue. The cover serves as a sheath around the body layer [21]. The vibration of the vocal folds can be observed by XRAY, high speed videoendoscopy (HSV) or stroboscopy.

The ideal movement of the vocal folds in coronal plane is shown in Fig. 2.3(a). These movements can be described by eight steps;

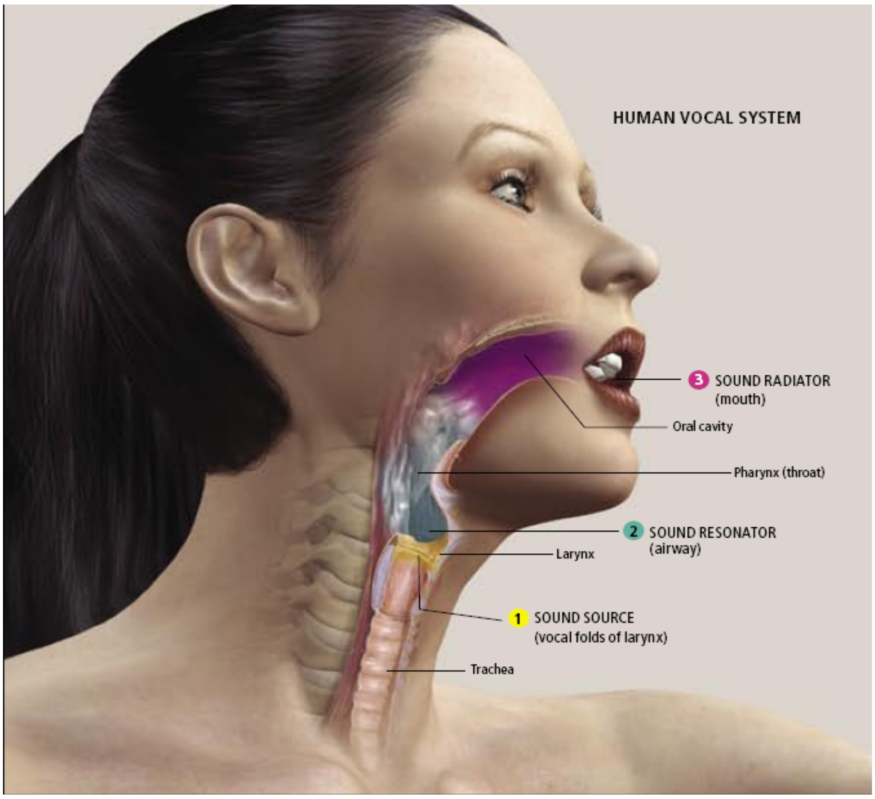


Figure 2.1: Human Vocal System (taken from [15])

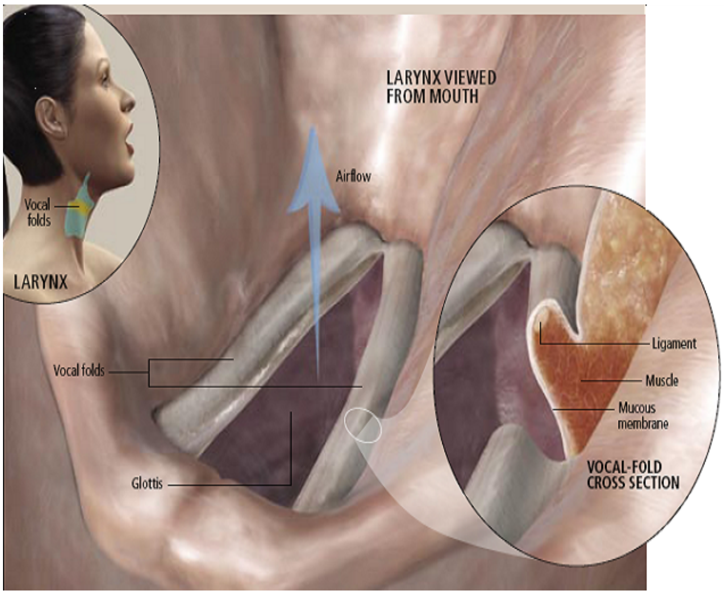
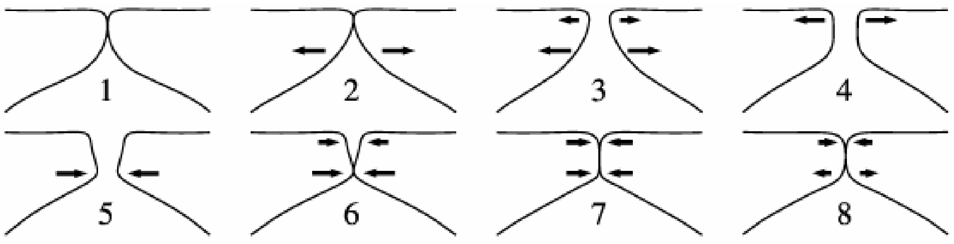
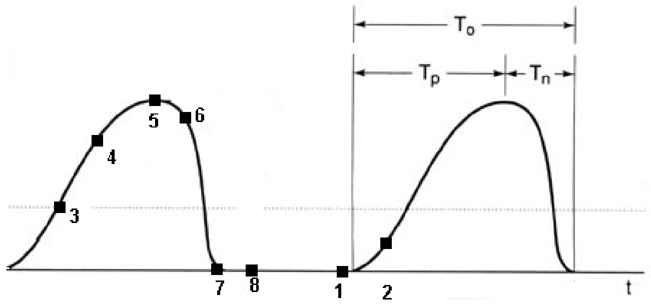


Figure 2.2: The Vocal Folds (taken from [15])



(a) The Vocal Folds Movements (Mucosal Wave)

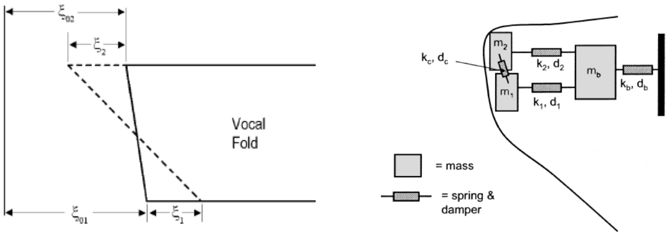


(b) The Glottal Flow

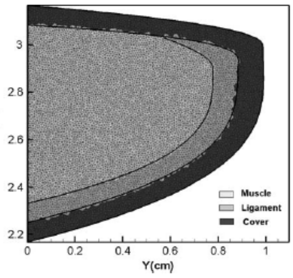
Figure 2.3: The movement of vocal folds (a) and corresponding glottal volume velocity (b)

1. Lower margins are opening while upper margins are closed.
2. Lower margins are opened while upper margins are opening.
3. Both lower and upper margins are opened. Vocal folds are apart.
4. The distance between the upper margins getting wider
5. The lower margins are getting closer
6. The lower margins are touching each-other.
7. The maximal closing of vocal folds (GCI).
8. Vocal folds are separating from lower margins towards upper margins.

It is easily seen from Fig. 2.3(a) that there is a phase difference between the movements of lower and upper portion of each vocal folds. This phase difference leads the folds to move



(a) Parametric Model (b) Mass Model



(c) Continuum Model

Figure 2.4: The vocal fold models

like a wave in the coronal (vertical) plane. This movement is called ”mucosal wave”. The entire process repeats itself quasi-periodically at the fundamental frequency of vibration (F0). The vocal fold vibration is initiated and sustained over time by the steady airflow and pressure supplied by subglottal system, lungs and trachea. As a result of this vibration, quasi-periodic glottal flow, the sound source, is generated at the vocal folds as shown in Fig. 2.3(b).

The vocal fold models used in simulation of speech production can be classified into three types as shown in Fig.2.4;

1. **Parametric area models**

   The areas between the upper and lower margins are calculated from a geometric model of vocal folds. A typical model is shown in Fig. 2.4(a) [10].

2. **Mass Models**

   Each fold is represented by single or multiple masses. The position and the velocity of each mass under pressure forces are determined by solving second order nonlinear equation.

3. **Continuum or FEM Model**

   The vocal fold geometry is approximated in 2 or 3 dimensions by millions of points shown in Fig. 2.4(c) and the Navier-Stokes equation is solved at each point by the finite element method. The continuum model is used for the investigation of fluid-structure interaction between the vocal fold tissue and the airflow in the glottis.

The vocal fold models can also be classified as mathematical (1) and biomechanical models (2-3). For articulatory speech synthesis, two mass models are usually suggested. Mathematically the movement of each mass is described by second order nonlinear differential equation. Therefore, increasing the mass approximation increases the computational complexity for synthesis applications. Due to this high computational load, despide the huge computational capacities of current computers, there is no existing real-time articulatory speech synthesizer yet.

2.3 Vocal Tract

Vocal tract determine the identity of the spoken sounds. It has two functions, generates turbulence by constricting the airflow for unvoiced or mixed speech sounds and filter the sounds produced at the vocal folds and at the turbulences. It is composed of three cavities, pharnx, oral and nasal. Each cavity has certain shape depending on the sound to be spoken. Especially, the oral cavity can create a variety of shapes by adapting the articulators, the tongue, teeth, lips, and jaw and generates new sound sources. The shapes of the cavities determine the acoustical properties of the vocal tract. Each cavity has its acoustic resonances and the vocal tract resonance frequencies, in other words the formants (F1,F2,F3,...), are determined by the acoustics of combined cavities. For any speech sound, the formants are arranged in a specific patterns depending on the age and gender of the speaker. These patterns perceptually allow us to identify the speech sounds.

The vocal tract is usually modeled by 2D or 3D geometry as shown in Fig. 2.5 [27]. The parameters of the geometric models are obtained from MRI (Magnetic Resonance Image) or CT (Computer Tomography) images as seen in left part of Fig.2.5. The geometry is approximated by the concatenation of uniform tubes in order to calculate the acoustics in the model

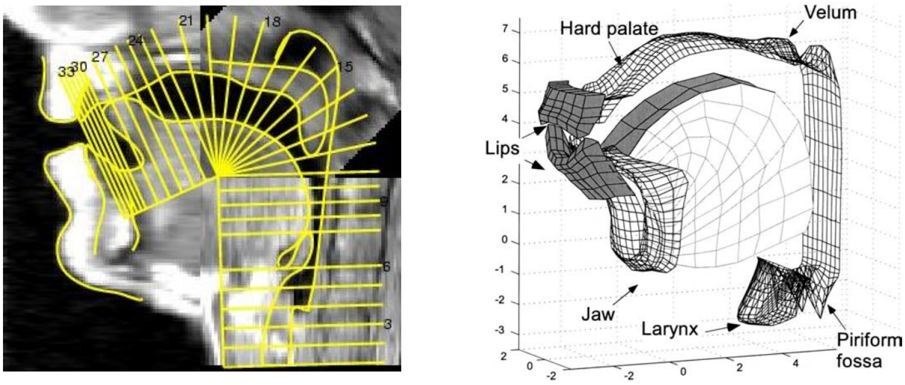


Figure 2.5: The Vocal Tract Modeling

calculated by solving wave-equation on the each tube. In general, the vocal tract acoustics is assumed to be linear except constrictions at which turbulence occurs and one dimensional which is usually valid for sound waves below 4kHz in vocal tract [34]. By assuming planar wave, the acoustics in 2D or 3D uniform geometry is calculated from the lengths, cross-sectional areas and perimeters of each tube in the spacial approximation.

2.4 Acoustic models

The task of the acoustic model is to calculate the time varying airflow and air pressure distribution within the vocal tract model and to calculate the acoustic speech signal radiated from the mouth and nose or nostril. Considering the current physics based speech synthesizers, there are three different acoustic models to calculate the sound propagation in vocal system approximated by the concatenation of uniform tubes.

2.4.1 Wave Reflection Method

Wave reflection method is based on the D'Alembert solution of the wave equation. The forward and backward traveling partial flow and pressure waves are calculated for each vocal tract tube section in time domain on the basis of scattering equations at the tube junctions.

The model parameters are either reflection coefficients or acoustic impedances. The main advantage of this method is that the nonlinear source tract interaction can be easily handled due to time domain computation. It is computationally efficient but its major drawback is that frequency dependent loses are approximate [18, 19, 7, 20, 25, 24].

2.4.2 Transmission Line Circuit Analog

Transmission line circuit analog is based on the electrical circuit equivalent of sound wave propagation in a uniform lossy tube. The vocal tract is considered as a concatenation of uniform tubes. The vocal folds are modelled by a time varying nonlinear impedance whose value depends on the glottal area. Due to the nonlinear glottal impedance, the circuit can only be simulated in time domain. Simulation of this circuit is conducted via finite difference method, solving differential equations obtained from Kirchhoff's laws and aerodynamic equations. It is computationally complex and currently cannot be simulated in real time [2, 3, 27, 29].

2.4.3 Hybrid Time - Frequency Domain Model

Hybrid method uses combination of time and frequency domain methods. Time domain model of glottis is coupled with the frequency domain model of vocal tract acoustics. The vocal tract frequency response is calculated from tube geometries by the chain matrix method and then vocal tract input impedance impulse response is obtained by inverse Fourier transform. After obtaining the impulse response, the output speech signal is calculated by the convolution of impulse response and glottal flow obtained from time domain simulation [34, 30]. Main advantage of this method is that it is possible to use frequency dependent impedance directly so that vocal tract can be represented more accurately than time domain methods. However, it is computationally demanding and cannot be simulated in real time.

2.5 Aerodynamic & Acoustic Interactions in Speech Production

Sound is produced at a source and propagates in a medium, solid, fluids or gas. Generation of sound sources in speech is a subject of biomechanics, aerodynamics and aeroacoustics, while its propagation is an acoustic process. According to the generation mechanism, there

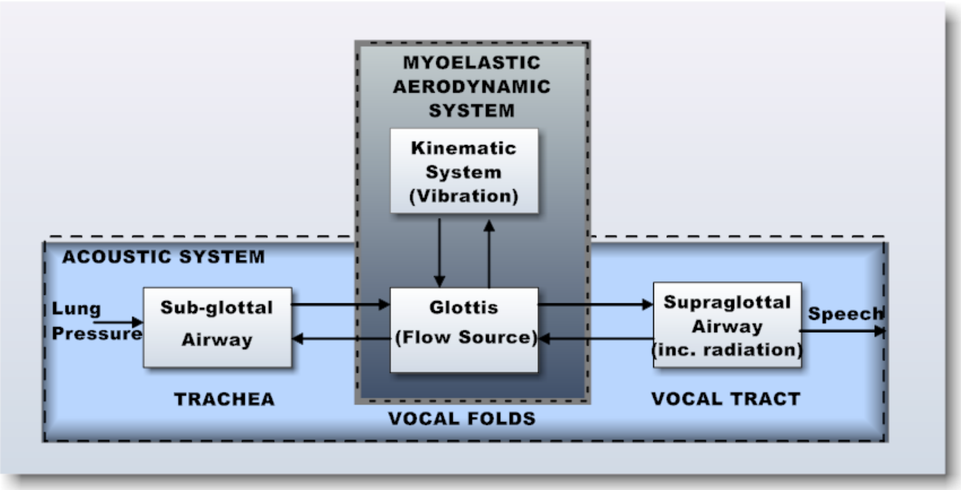


Figure 2.6: A global view of aerodynamic and acoustic interactions in human vocal system

are two types of speech sounds, voiced and unvoiced. While the voiced sounds are generated by the quasi-periodic vibration of vocal folds (aerodynamic process), the unvoiced sounds are produced at constrictions causing turbulences in vocal tract (aeroacoustic process). Speech production system can be represented by two interactive systems, aerodynamic and acoustic system.

The aerodynamic part of the system represents the vibration of vocal folds due to airflow - tissue structure interaction in the glottis. A kinematic model of vocal fold is excited by lung pressure and starts to vibrate, then produce glottal flow. In order to investigate the effect of this tissue structure interaction on the vibration of vocal folds completely, the Navier - Stokes nonlinear differential equations need to be solved on the complex soft tissue structure of the vocal folds with a set of appropriate boundary conditions. Due to the nonlinear structure of Navier Stokes equations, the numerical solution obtained by finite element based methods might not truly represent the flow - structure interaction. As a result, instead of using numerical simulation, excised canine vocal folds or kinematic solid physical models of vocal folds become more popular to investigate the aerodynamic system. In addition, the vibration frequency and the vocal fold geometry can be affected by the acoustic loading of the trachea and vocal tract. Recently, it is reported that this acoustic-aerodynamic interaction is important for self-sustained oscillations of vocal folds [23].

The aerodynamic system can be represented by a system whose inputs are lung pressure, elasticity, mass and friction of vocal folds and its output is the glottal area as shown in Fig. 2.7.

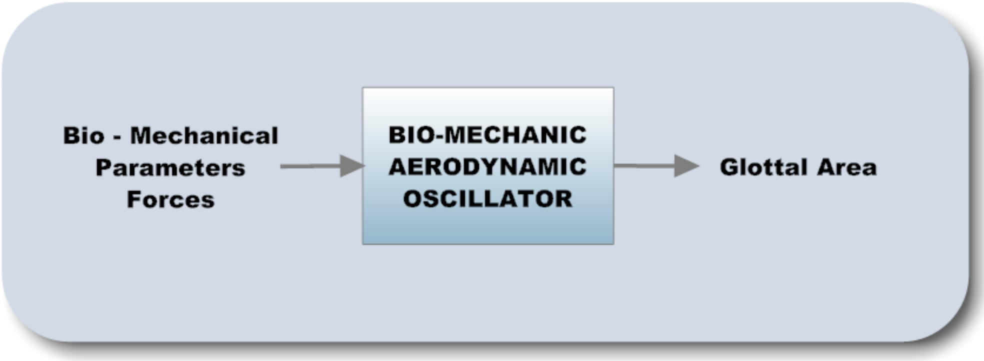


Figure 2.7: Input - Output Representation of Vocal Fold Aerodynamics

The acoustic system is composed of the supra-glottal (upper airway from glottis to lips), sub-glottal airways (lower airway from glottis to lungs) and mouth. It represents the transmission and radiation of sound generated at the vocal folds or at the turbulences in the vocal tract. Both upper and lower airway has acoustical impedance and the combined impedance affects the glottal volume velocity at the vocal folds during phonation. This is called as ”acoustic interaction” or ”source-filter interaction” in speech.

The acoustic system can be represented as an input - output system as shown in Fig. 2.8. Its inputs are pressure or flow sources, cross-sectional areas, tube lengths and its output is acoustic pressure distribution in the vocal tract and radiated speech pressure wave.

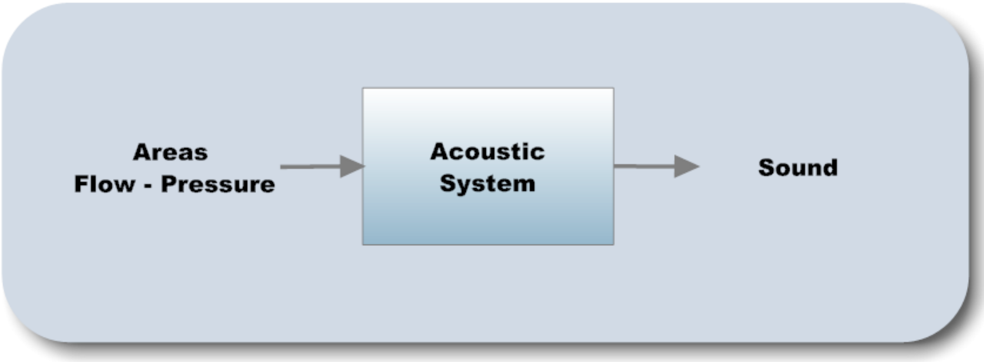


Figure 2.8: Input - Output Representation of Acoustic Models

2.6 A Review of Source-Filter Coupling

Nonlinear source-filter interaction has been studied by simulation, using acoustic and/or biomechanical models of vocal system. Due to the nonlinearity of the glottal impedance it can be simulated in time domain [20]. The one mass and the two mass models are among the first models that take into account the source-filter interaction [3]. The transmission line circuit analog of vocal system is solved by the finite difference method [1, 2, 36, 27]. The results demonstrate the acoustic interaction effects, skewness of glottal flow and superimposed ripple on it.

Rothenberg studied the acoustic interaction between the glottal source and vocal tract with a simple impedance based model of vocal tract linearly coupled to the time-varying glottal impedance [35]. The results show that the interaction effects are noticeable when the glottal impedance is comparable to the vocal tract input impedance. In particular, the vocal tract inertia is responsible for skewness of glottal flow with respect to the glottal area while the vocal tract resistance changes the amplitude of the glottal flow waveform.

Fant and Ananthapadmanabha used lumped approximation of vocal tract, trachea and glottis in order to investigate the effects of nonlinear interaction on the glottal flow for different vowels [36]. They use various configurations for the coupling of the vocal tract and trachea and solve the nonlinear system of equations by an iterative algorithm. It is concluded that the waveform of the interactive glottal flow is different for different vowels and depends more on the vocal tract resonances than subglottal resonances. They also proposed a model that approximates the nonlinear source filter coupling by varying vocal tract formants and their bandwidts in accordance with the variation of glottal area in time.

In the hybrid time-frequency articulatory speech synthesis method [30], a different approach is used. The impulse response at the input of the vocal tract is found from the frequency response obtained from vocal tract areas (subglottal part is neglected) and then it is combined with the vocal fold acoustics. However, they report that the impulse response of vocal tract is very long and it is not always possible to calculate the interactive glottal flow. Hence, they follow another approach using the input reflectance of the vocal tract (Frequency domain representation of reflection coefficients) similar to that in the wave-reflection method.

The wave-reflection method works based on particle flow and pressure variations along the subglottal, supraglottal tract and glottis [7, 8, 9, 10, 11, 13, 19, 20, 23]. For the coupling at the glottis, Titze proposed a mathematical model based on the quasi steady Bernoulli flow assumption. It is essentially the glottal flow equation used in other approaches [7, 36], however it represents the glottal flow in terms of particle velocity (instead of volume velocity).

The simulation of the interactive system model via wave-reflection method is a time-domain approach. The calculation of the parameter values is easy, and, in particular, the frequency dependent losses of the vocal tract are calculated recursively. It is considered to be a more convenient approach compared to other methods [18, 19, 20, 25, 23].

Titze and Story investigated not only the acoustic interaction but also the aerodynamic interaction of the coupling on more complicated models [7, 21, 13, 8, 9]. They show that vocal tract affects the vibration pattern of vocal folds. Zanartu investigates the acoustic loading effect of both vocal tract and trachea on a single mass model of the vocal folds under a quasi Bernoulli flow assumption at the glottis [23]. It is shown that the acoustic loading of vocal tract is dominant against that of trachea in maintaining self-sustained oscillations of vocal folds. Titze has recently introduced a theory of nonlinear source filter coupling [11]. It is based on the assumptions of linear vocal tract and a quasi-steady Bernoulli glottal flow. Two levels of interaction effects are demonstrated. The level-I interaction seen in normal speaking consists of the skewness and the ripple on the glottal flow for $F0 < F1$ and the level-II interaction involves instabilities in the vibration pattern of vocal folds when F0 crosses over the formant frequencies. The level-II effects are demonstrated in human voice by varying $F0$ [12]. In this thesis, we focus on the level-I interaction observed in normal speaking conditions.

2.6.1 The effects of interaction on the glottal flow

In this section, the dependence of glottal flow waveform on vocal tract in two different formant vowel cases, high F1 vowel /a/ and low F1 vowel /i/, for a wide F0 frequency range is demonstrated by using a parametric glottal area model. The vocal system is simulated by the wave reflection method described in [20, 23] using the areas extracted from 3D MR images provided by Story in [20, 23] and parametric glottal area model proposed by Fant [36] given by

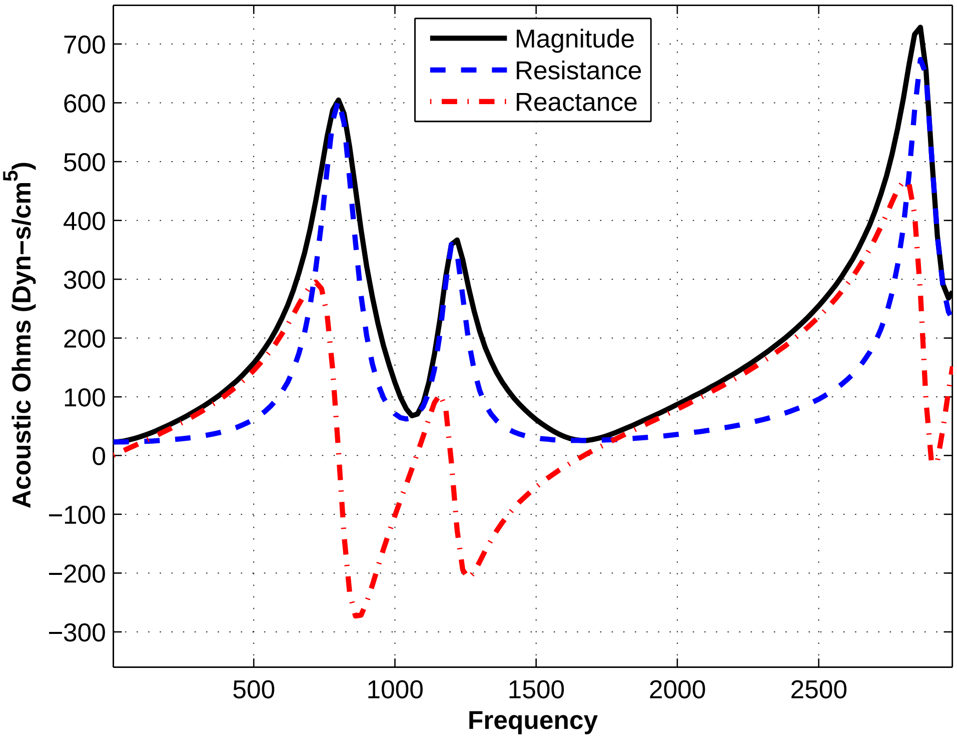


Figure 2.9: Vocal tract input impedance for male vowel /a/

$$a_g\left[n\right] = \begin{cases} A\left(0.5 - 0.5\cos\left(\frac{\pi}{2N_O}n\right)\right) & , \quad 0 < n \leq N_O \\ A\cos\left(\frac{\pi}{2N_C}n\right) & , \quad n \leq N_0 - 1 \end{cases} \tag{2.1}$$

where a_g, A, N_O and N_C are the glottal area, the peak value of the glottal area, the open phase duration and the closing phase duration, respectively. Particularly, we used the same glottal flow expression obtained by Titze [7, 11] based on the quasi-steady Bernoulli flow assumption. It is also equivalent to the interactive glottal flow equation of a single mass system in terms of volume velocity that was used by Fant for the calculation of true glottal flow [36]. The source-filter interaction depends on the vocal tract input impedance [35, 36, 11]. Considering the speech production system, there are mainly two different acoustic cavities, the vocal tract, above the glottis, and trachea, below the glottis. Since the vocal tract has much smaller crossectional area with respect to subglottal tract, its reactance is much larger than the subglottal reactance. Hence the reactance of the overall system is slightly different than the vocal tract reactance. The impedance relating the glottal flow to the transglottal pressure

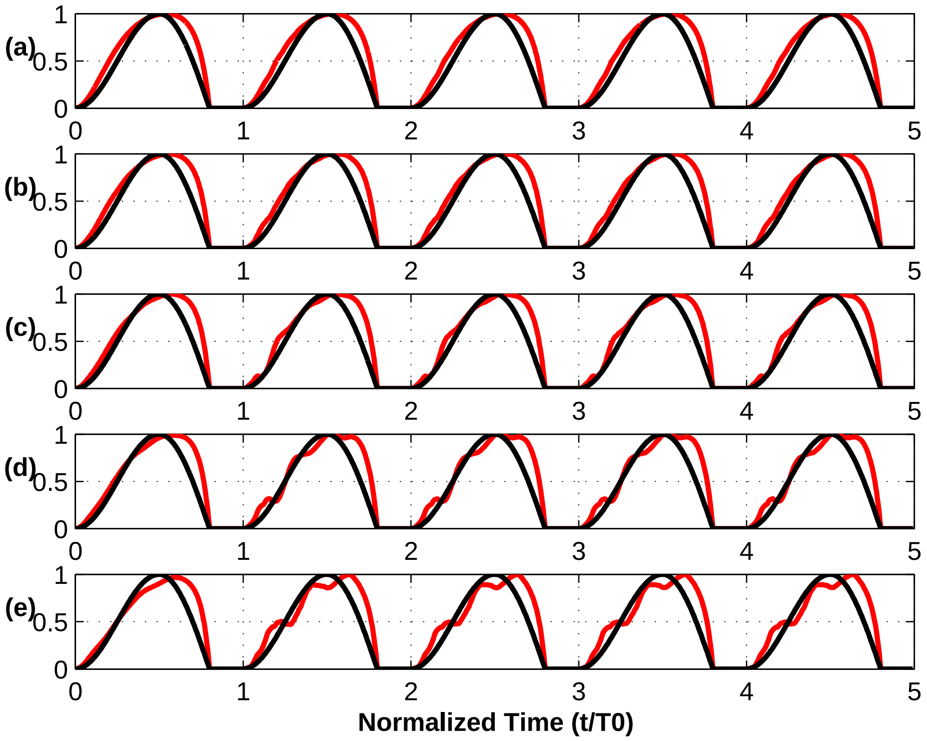


Figure 2.10: Time and amplitude based normalized glottal area (black lines) and glottal flow waveforms (red lines) obtained by the simulation of vowel /a/ with wave-reflection method for glottal area fundamental frequency, $F0$ (a) 80 Hz, (b) 120 Hz, (c) 160 Hz, (d) 200 Hz, (e) 240 Hz.

across the glottis is the sum of the subglottal and supraglottal impedances [11].

The vocal tract impedance of vowel /a/ calculated by the impulse response technique using the wave reflection method is plotted in Fig. 2.9. The magnitude and resistance curves have peaks at resonance frequencies at which the reactance curve has its zeros. It is seen in the figure that the vocal tract input impedance is inertive for frequencies below F1, approximately 0.8 kHz, and between 1.1-1.2 kHz while compliant in the frequency intervals 0.8-1.1 kHz and 1.2-1.7 kHz.

The vocal system is simulated by using constant lung pressure, $P_L = 0.8$ kPa and the maximum glottal area $A = 0.15$ cm^2. Since subglottal tract is not used in the simulation, the subglottal pressure, p_{sub}, is set to the lung pressure, P_L. The glottal area and glottal flow waveforms normalized in time and amplitude obtained for $F0$ frequencies equally spaced by 40 Hz between 80-240 Hz are plotted in Fig. 2.10. It is seen from the figure that the glottal

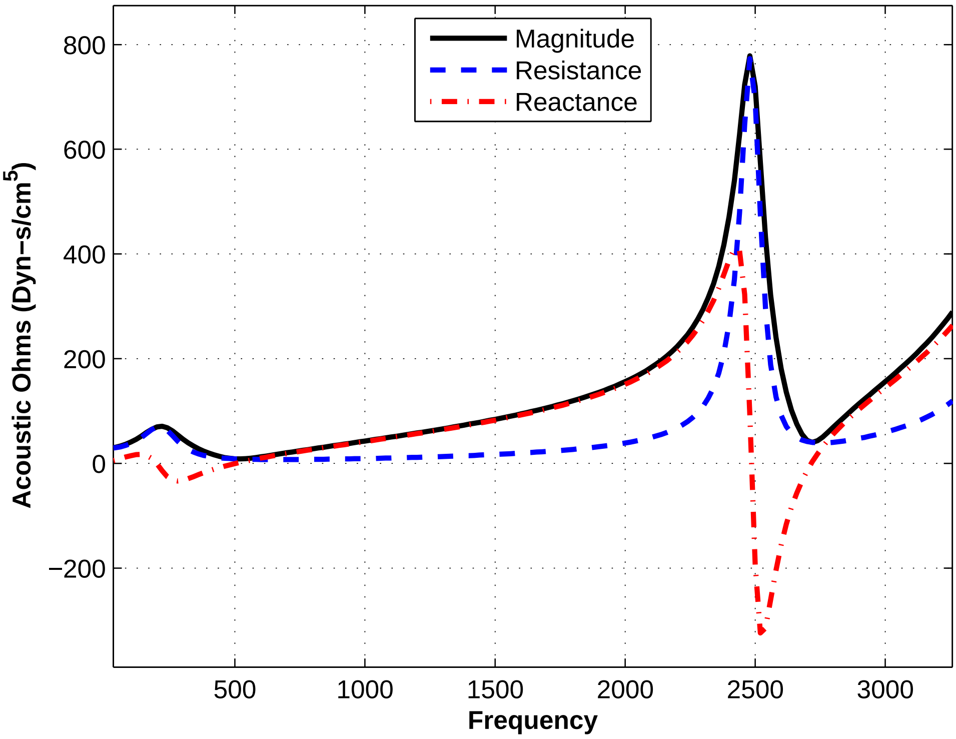


Figure 2.11: Vocal tract input impedance for male vowel /i/

flow is skewed to the right with respect to glottal area. This is the first effect of level-I interaction. This behaviour is due to the inertive vocal tract reactance. It increases the spectral slope of the source harmonic peaks. The second effect of level-I interaction is the ripple on the glottal flow waveforms. The interaction ripples can be easily seen in the glottal flow waveforms in Fig. 2.10(c)-(e). The intensity of the ripples increases as F0 is increased. On the other hand, the frequencies of the ripples are not the same in each case. They result from the variations in the magnitudes of the glottal flow harmonics. The magnitude of each harmonic in the source spectrum is mainly determined by the vocal tract input impedance. The discussion on the cause and effect of the interaction ripple requires both spectral and time domain investigation of the simulated glottal waveforms and output speech signal.

The vocal tract impedance of vowel /i/ calculated by the wave-reflection method is shown in Fig. 2.11. In this case, the first formant, ($F1$), calculated approximately as 200 Hz, is smaller and closer to $F0$ frequencies used in the simulation. For frequencies between 80-240 Hz, the simulations are conducted for /i/ and the resulting glottal flow waveforms are plotted in Fig.

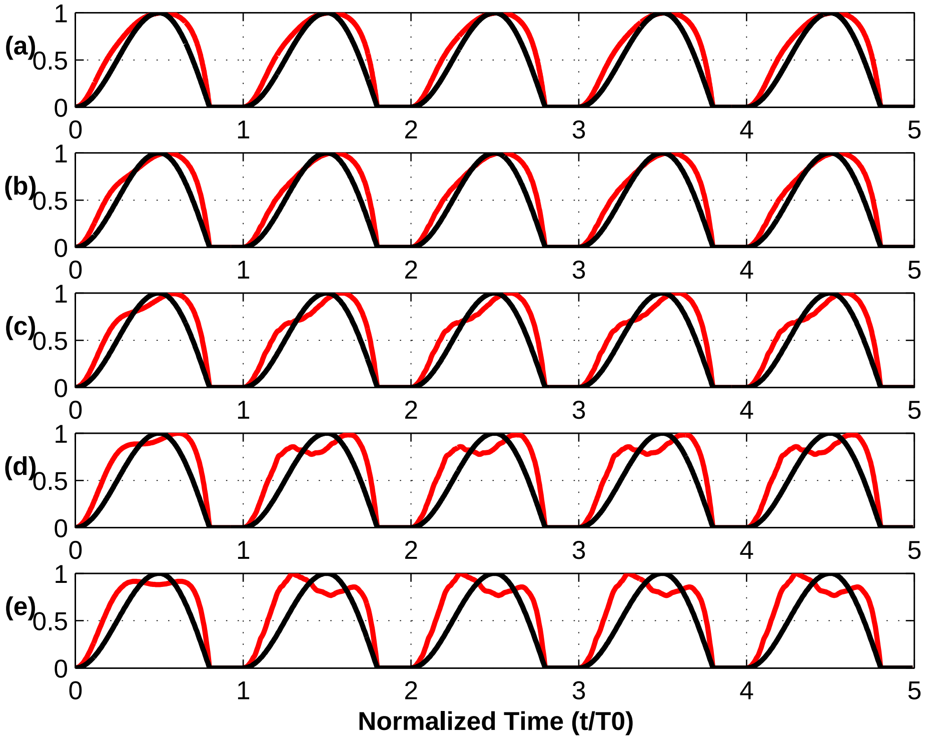


Figure 2.12: Time and amplitude based normalized glottal area (black lines) and glottal flow waveforms (red lines) obtained by the simulation of vowel /i/ with wave-reflection method for glottal area fundamental frequency, $F0$ (a) 80 Hz, (b) 120 Hz, (c) 160 Hz, (d) 200 Hz, (e) 240 Hz.

2.12. As expected, the skewness and ripple effects are observed in each case. However, an important difference between the flow waveforms of $F0 < F1$ (Fig.2.12(a)-(c)) and $F0 > F1$ (Fig.2.12(d)) is that the former waveforms are skewed to the right while the latter is skewed to the left. This is due to the first source harmonic, when $F0 = 240$ Hz is inside the negative reactance region (compliant reactance) of vocal tract input impedance of /i/. Furthermore, the size of the interaction ripple is much larger than the ripple seen in the case of /a/. The reason for the large ripple is that the vocal tract impedance has a much stronger effect on the individual harmonic magnitudes compared to that in the case of vowel /a/.

The glottal source waveform strongly depends on the vocal tract shape. It is due to the loading effect of the vocal tract on the glottis [11].

The following statements summarize the salient aspects of the source-filter interaction;

1. There are two major effects of the source-filter interaction on the source, skewness of glottal flow and ripple on the glottal flow (level - I interaction).

2. Glottal flow is affected by the vocal tract input reactance. Positive reactance increases the skewing of glottal flow to the right and increase the energy level of corresponding harmonics [35, 11].

3. The vocal tract input reactance is always positive for frequencies below F1. During speech activity, most of the time, the frequencies of source harmonics having the largest energy values are in inertive reactance interval. As a result, the glottal flow is skewed to the right [11]. The reactance usually increases up to the first vocal tract formant, F1, except some small variations around the subglottal formants. It may be concluded that for frequencies below F1, increasing F0 increases vocal tract input reactance so that skewness is increased.

4. The resistive part of vocal tract input impedance affects the peak value of the flow. Increasing resistance decreases the amplitude of the glottal flow [35, 11]. The resistance increases up to F1, then decreases. This implies that the amplitude of the glottal flow decreases as F0 increases up to F1.

5. The interaction ripple is caused by the variations in the magnitudes of the glottal flow harmonics strongly effected by the vocal tract input impedance. The size of the interaction ripple depends on the magnitudes of the corresponding source harmonics. By increasing F0, the dominant harmonics get closer to the zero reactance frequency and the ripple gets larger.

6. For different vowels, the frequency of ripple, the maximum amplitude and the skewness of the glottal flow are different due to the difference of the input impedance.

CHAPTER 3

MODELING

3.1 Theoretical Background

In this section, the aerodynamic - acoustic theory of vocal folds and vocal tract is given, then proposed nonlinear interactive system is developed based on these knowledge about the vocal system.

3.1.1 Glottal Aerodynamics

The behaviour of airflow is described based on several conservation laws, including conservation of mass, conservation of momentum and conservation of energy. These conservation laws lead to the well-known dynamical equations of aerodynamics, known as Navier Stokes equations.

Consider an air particle moving on a streamline with velocity v, according to conservation of mass law, the continuity equation is

$$\frac{\partial \rho}{\partial t} + \nabla.(\rho v) = 0 \tag{3.1}$$

where ρ and v denote the density of air and the particle velocity, respectively. If the density is constant, then flow is said to be incompressible. Hence, the equation of continuity reduces to the divergence of the particle velocity is zero on the streamline, $\nabla.v = 0$. In the aerodynamic engineering, usually when Mach number (Ma), fluid velocity over sound velocity, is below to 0.3, the flow is assumed to be incompressible [5]. In the glottis, typical particle velocity is

3000-5000 cm/s [10] and corresponding Mach number is between 0.08-0.14, hence the glottal flow can be assumed to be incompressible. Incompressible flow formulation can therefore be used to derive the velocity and pressure profiles in the glottis. Furthermore, in speech production a generally one dimensional flow is assumed, parallel streamlines. Hence, the equation of continuity turn out to be $\frac{\partial v}{\partial x} = 0$. For the glottis, the glottal volume velocity, u, is calculated by the sum of all particles passing through the glottis, $u = \int a.dv = a.v$, where a is the glottal area.

The momentum of an incompressible airflow is determined by Navier-Stokes equations,

$$\frac{\partial v}{\partial t} + v\frac{\partial v}{\partial x} = -\frac{1}{\rho}\frac{\partial P}{\partial x} + F_{gravity} + \frac{F_{viscous}}{\rho} \tag{3.2}$$

where P, F_{body}, $F_{viscous}$ are fluid pressure, force due to the gravity and viscosity, respectively. Eq. (3.2) is nonlinear due to the particle flow velocity term, v. For glottis, there is about 3-5 mm elevation change and it corresponds to a pressure change, $\rho g(z_2 - z_1) = 0.068$ Pa, which is negligle in comparison to the typical measured pressures $100-1000$ Pa in the glottis. Furthermore, if the flow is assumed to be inviscid, then the last two terms of Eq. 3.2 vanishes. For a steady incompressible inviscid flow, the momentum equation turns out to be

$$v\frac{\partial v}{\partial x} + \frac{1}{\rho}\frac{\partial P}{\partial x} = 0 \tag{3.3}$$

Eq. (3.3) is known as Euler' equation of motion. It is solved by integrating along a streamline to obtain

$$\frac{v^2}{2} + \frac{P}{\rho} = c \tag{3.4}$$

where c is a constant that may vary from one streamline to another, but remains the same for any one streamline. Hence, for two points along the same streamline, it can be written as

$$\frac{v_1^2}{2} + \frac{P_1}{\rho} = \frac{v_2^2}{2} + \frac{P_2}{\rho} \tag{3.5}$$

Eq. (3.5) is called as Bernoulli equation. Rewriting Eq. (3.5) to obtain a pressure drop in terms of velocity change between two points as

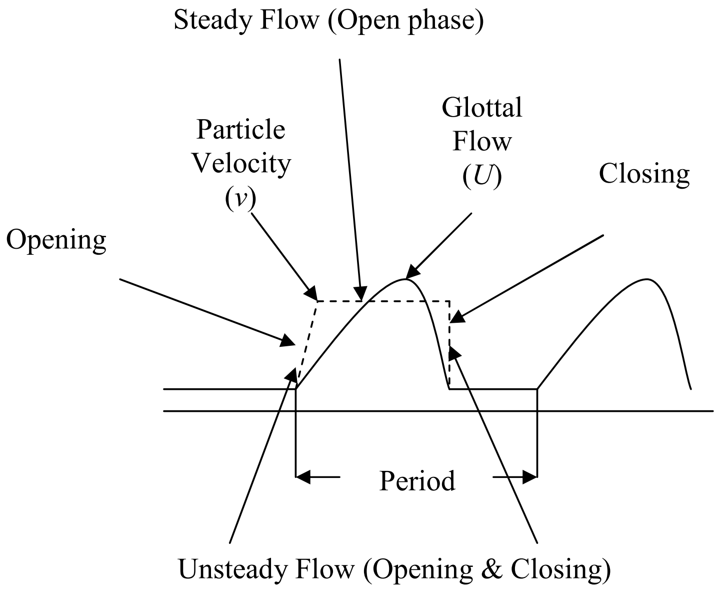


Figure 3.1: Glottal pulse and particle velocity

$$P_1 - P_2 = \frac{1}{2}\rho(v_2^2 - v_1^2) \tag{3.6}$$

In glottal aerodynamics, the particle velocity changes rapidly before glottal closing and after glottal opening, but remain more constant over most of the open phase as shown in Fig. 3.1. Therefore, the steady flow assumption can be valid over most of the open phase, despite the fact that it is violated in two short time intervals. This behaviour is called as quasi-steady flow [10]. Especially during the glottal closing, due to unsteady flow behaviour, there may be turbulent flow due to the separation of the flow from the glottal exit. As well as the flow viscosity, the turbulence can cause losing of flow energy. To account for these losses, the right side of Eq. 3.5 is multiplied by an empirical loss coefficient, k_t, varying the geometry and the Reynoulds number. k_t is determined by experimentally, measured values are usually between 1.1 - 1.3 interval [10, 11, 36].

Considering the speech production system, the cross-section of trachea is much larger compared to glottis, hence, particle velocity at the input of the glottis, v_1, is much smaller than v_2 and it can be neglected. Therefore, the pressure-flow relation in the glottis can be written in terms of glottal flow, u_g, as

$$p_{TG} = k_t \frac{1}{2} \rho \frac{{u_g}^2}{{a_g}^2} \tag{3.7}$$

where u_g is the glottal flow, P_{TG} is the transglottal pressure (the difference between the subglottal and supraglottal pressures, $P_{TG} = P_{sub} - P_{sup}$), a_g is the minimum glottal area along the vocal folds, ρ is the density of air and k_t is the transglottal pressure recovery coefficient determined experimentally [14]. If the subglottal acoustic impedance is neglected, then P_{sub} is equal to the lung pressure P_L. Substituting P_L into Eq.(3.7) and writing vocal tract input pressure, p_{IN}, instead of P_{sup}, the glottal flow can be written as

$$\begin{aligned} u_g &= a_g \sqrt{2 \frac{P_L - p_{IN}}{k_t \rho}} \qquad (3.8) \\ &= a_g \beta \sqrt{P_L - p_{IN}} \qquad (3.9) \end{aligned}$$

where $\beta = \sqrt{\frac{2}{k_t \rho}}$. The experimental results obtained by the simulation of mechanical model of vocal folds show that the unsteady Bernoulli flow assumption is valid during almost 70% of a glottal cycle. It has been found that, it is a good approximation of actual glottal flow during the open phase [16].

The linear source-filter theory [1] assumes that the glottal flow does not depend on the vocal tract acoustics. Considering Eq.(3.9), this assumption holds if

- The vocal tract input pressure or supraglottal pressure is negligible compared to the subglottal pressure ($P_{sub} >> P_{sup}$ or $P_L >> p_{IN}$).
- The acoustic impedance of the glottis is very large compared to the vocal tract input impedance ($Z_g >> Z_{IN}$). It also means that the minimum glottal area, a_g, is much smaller than the equivalent vocal tract area, A_{eq} defined by $A_{eq} = \frac{A_{sub} A_{sup}}{A_{sub} + A_{sup}}$ ($a_g << A_{eq}$) [11].

Noninteractive glottal flow is obtained by neglecting the vocal tract input pressure, p_{IN}, in Eq.(3.9). It is a linear function of glottal area.

Based on the quasi-steady Bernoulli flow assumption, the glottal flow is related to the vocal tract input pressure as given by Eq.(3.9). As we aim to develop a coupled speech production model, it is needed to have another expression, representing vocal fold dynamics, in terms of the vocal tract input pressure, p_{IN}, glottal flow and vocal tract output.

In the hybrid synthesis method, p_{IN}, is calculated from vocal tract input impedance obtained by the chain matrix method [30]. The impulse response of the vocal tract is calculated by inverse Fourier transform, and it is convolved with the glottal flow to obtain the output speech. However, it is reported that one drawback of this approach is that when the duration of the impulse response is very long, calculation of convolution gets hard. Hence, instead of using convolution directly, input reflectance of the vocal tract is used. In this work, in order to accomplish the coupling, the input pressure, p_{IN}, is estimated by a recursive filter determined using the chain matrix method.

3.1.2 Vocal Tract Acoustics

There are three types of acoustic models for vocal tract; transmission line circuit analog, wave-reflection method and chain matrix method. The first two methods work in time-domain while the last one works in frequency domain. Since it can model the frequency dependent losses due to viscosity and wall vibration, the vocal tract formants are estimated more accurately. Among these methods, the chain matrix method serves as a reference model for the others [30]. The chain matrix method is reviewed in the following section.

3.1.3 Chain Matrix

The chain or ABCD matrix representation of one uniform tube is written in the frequency domain as

$$\begin{bmatrix} U_{in} \\ P_{in} \end{bmatrix} = \begin{bmatrix} A & C \\ B & D \end{bmatrix} \begin{bmatrix} U_{out} \\ P_{out} \end{bmatrix} \tag{3.10}$$

where P_{in} and U_{in} are pressure and volume velocity at the input end of the tube. P_{out} and U_{out} are the quantities at the output end of the tube. The A, B, C, D frequency dependent elements of the matrix represents the propagation of acoustic waves through the tube in frequency domain. The elements of chain matrix for a single tube section of vocal tract is given in [30].

It depends on the solution of equation of continuity and equation of motion. The losses of viscous friction and wall vibration are approximated by modifying the equation of motion and the equation of continuity. The resulting second order linear partial differential equation has two independent solutions. For a lossy uniform tube, the matrix is given as,

$$\begin{bmatrix} A & C \\ B & D \end{bmatrix} = \begin{bmatrix} \cosh\sigma l & \frac{1}{\beta}\sinh\sigma l \\ \beta\sinh\sigma l & \cosh\sigma l \end{bmatrix} \tag{3.11}$$

where σ and β are as follows,

$$\sigma^2 = (\rho s + AR)\left(\frac{s}{\rho c^2} + \frac{Y}{A}\right) \tag{3.12}$$

$$\beta = \sqrt{\left[Y + \frac{As}{\rho c^2}\right] / \left[R + \frac{\rho s}{A}\right]} \tag{3.13}$$

where Y and R represent the wall vibration and viscous losses, l is the length of the tube, c is the speed of sound, ρ is the air density and A is the cross-sectional area of the tube. For the vocal tract, the matrix elements are calculated for each tube, then multiplied together with matrices for the other tubes, builds a matrix for the entire tract as

$$K_{TRACT} = K_N K_{N-1} K_{N-2} ... K_2 K_1 = \begin{bmatrix} A_{TRACT} & C_{TRACT} \\ B_{TRACT} & D_{TRACT} \end{bmatrix} \tag{3.14}$$

Then, the transfer matrix for vocal tract is written as

$$\begin{bmatrix} U_{IN} \\ P_{IN} \end{bmatrix} = \begin{bmatrix} A_{TRACT} & C_{TRACT} \\ B_{TRACT} & D_{TRACT} \end{bmatrix} \begin{bmatrix} U_{LIP} \\ P_{LIP} \end{bmatrix} \tag{3.15}$$

The frequency response from glottal input to the lip pressure is given by the ratio of lip pressure to the glottal volume velocity in frequency domain, $V(w) = \frac{P_{LIP}}{U_{IN}}$. In terms of lip volume velocity, the lip pressure is $P_{LIP} = Z_L U_{LIP}$ where Z_L is the radiation impedance. $V(w)$ can be written from Eq.(3.15) as

$$V(w) = \frac{P_{LIP}}{U_{IN}} = \frac{Z_L}{A_{TRACT} + C_{TRACT}Z_L} \tag{3.16}$$

If the transfer function is defined as the ratio of lip volume velocity to glottal volume velocity, it can be written as

$$H(w) = \frac{U_{LIP}}{U_{IN}} = \frac{1}{A_{TRACT} + C_{TRACT}Z_L} \tag{3.17}$$

The vocal tract input impedance can be calculated as the ratio of input pressure to input volume velocity

$$Z_{IN} = \frac{P_{IN}}{U_{IN}} = \frac{D_{TRACT}Z_L + B_{TRACT}}{A_{TRACT} + C_{TRACT}Z_L} \tag{3.18}$$

3.1.4 DT-Representation of Chain Matrix for Lossless Vocal Tract

The chain matrix method leads to the classical linear all pole filter representation of the vocal tract. Consider the vocal tract transfer function $H(s)$ given by

$$H(s) = \frac{U_{LIP}}{U_{IN}} = \frac{1}{A_{TRACT} + C_{TRACT}Z_L} \tag{3.19}$$

If the radiation impedance and losses are assumed to be zero, then

$$H(s) = \frac{1}{A_{TRACT}} \tag{3.20}$$

For the lossless case, R and Y in the $ABCD$ matrix representation are zero, so from Eq.(3.13), we can calculate that $\sigma = \frac{s}{c}$ and $\beta = \frac{A}{\rho c}$.

$$\begin{bmatrix} A & C \\ B & D \end{bmatrix} = \begin{bmatrix} \cosh \frac{s}{c}l & -\frac{A}{\rho c} \sinh \frac{s}{c}l \\ -\frac{\rho c}{A} \sinh \frac{s}{c}l & \cosh \frac{s}{c}l \end{bmatrix} \tag{3.21}$$

By dividing vocal tract into $N = l/\Delta$ equal length concatenated tube sections with length Δ, choosing sampling period $T = \frac{2\Delta}{c}$, and $z = e^{\frac{2s\Delta}{c}}$ to transform to discrete time by impulse invariance technique, for a single lossless tube with length Δ it can be written as [34]

$$\begin{aligned}
\cosh\left(\frac{\Delta s}{c}\right) &= \frac{e^{\frac{s\Delta}{c}} + e^{-\frac{s\Delta}{c}}}{2} \\
&= \frac{z^{1/2} + z^{-1/2}}{2} = z^{1/2}\left(\frac{1 + z^{-1}}{2}\right) \qquad (3.22)
\end{aligned}$$

$$\begin{aligned}
\sinh\left(\frac{\Delta s}{c}\right) &= \frac{e^{\frac{s\Delta}{c}} - e^{-\frac{s\Delta}{c}}}{2} \\
&= \frac{z^{1/2} - z^{-1/2}}{2} = z^{1/2}\left(\frac{1 - z^{-1}}{2}\right) \qquad (3.23)
\end{aligned}$$

Substituting Eq.(3.22) and Eq.(3.23) into the chain matrix yields

$$\begin{bmatrix} A & C \\ B & D \end{bmatrix} = z^{1/2}\begin{bmatrix} \frac{1}{2}\left(1 + z^{-1}\right) & -\frac{A}{2\rho c}\left(1 - z^{-1}\right) \\ -\frac{\rho c}{2A}\left(1 - z^{-1}\right) & \frac{1}{2}\left(1 + z^{-1}\right) \end{bmatrix} \qquad (3.24)$$

Since the matrix for whole tract is the product of all matrices, the elements of K_{TRACT} can be written as $z^{N/2}$ times N^{th} degree polynomial in z^{-1} as follows

$$K_{TRACT} = z^{N/2}\begin{bmatrix} \sum\limits_{n=0}^{N} a_k z^{-k} & \sum\limits_{n=0}^{N} c_k z^{-k} \\ \sum\limits_{n=0}^{N} b_k z^{-k} & \sum\limits_{n=0}^{N} d_k z^{-k} \end{bmatrix} \qquad (3.25)$$

K_{TRACT} can be used to calculate acoustic input-output transfer function of the vocal tract. For example, using Eq.(3.25) in Eq.(3.20), the discrete time transfer function from glottis to lips for the lossless vocal tract is written as

$$\frac{U_{LIPS}(z)}{U_{IN}(z)} = \frac{z^{-N/2}}{\sum\limits_{k=0}^{N} a_k z^{-k}} \qquad (3.26)$$

Here, $z^{-N/2}$ is a delay term due to time required transmission of the sound from glottis to lips [34]. If the vocal tract filter, $H(z)$, is defined as $H(z) = \frac{1}{\sum\limits_{k=0}^{N} a_k z^{-k}}$, then Eq.(3.26) is written as

$$U_{LIPS}(z) = z^{-N/2} H(z)\, U_{IN}(z) \qquad (3.27)$$

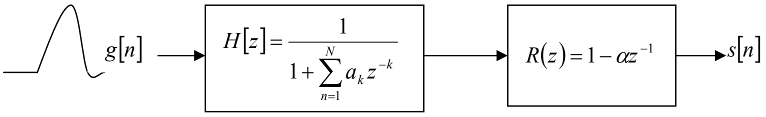


Figure 3.2: Linear Source-Filter Model of Speech Production

By defining $u_g[n] = u_{IN}\,[n - N/2]$ as the glottal flow, in time domain, Eq.(3.27) can be written as

$$\sum_{k=0}^{N} a_k u_{lips}[n-k] = u_g[n] \tag{3.28}$$

It must be noted that Eq.(3.28) is obtained under closed glottis and no lip radiation assumptions. In addition, source-filter interaction is neglected, assuming that they are independent. In general, the lip radiation is approximated by a first order high pass filter, $R(z)$, [37]. Hence, the transfer function from glottal volume velocity to lip pressure can be written as $\frac{P_{LIPS}(z)}{U_G(z)} = H(z)\,R(z)$. As a result, the linear source-filter model or linear predictive representation of voiced speech in shown in Fig. 3.2.

In order to properly couple the source and the filter, according to unsteady Bernoulli equation in Eq.(3.9), the vocal tract input pressure, p_{IN}, is needed. Now, we will obtain the input impedance of the vocal tract that relates the glottal flow with p_{IN} from chain matrix and then construct two nonlinear interactive source-filter models for voiced speech.

3.2 Interactive Source-Filter Modeling

We consider two ways to calculate p_{IN} from the chain matrix; from lip velocity and from the glottal flow. Let us first write P_{IN} in terms of U_{LIPS} using K_{TRACT} assuming that radiation impedance is zero ($P_{LIP} = 0$),

$$\frac{P_{IN}}{U_{LIPS}} = B_{TRACT} = z^{N/2}\sum_{n=0}^{N} b_k z^{-k} \tag{3.29}$$

Defining the transfer function from lip velocity to vocal tract input pressure as $B(z) = \sum_{n=0}^{N} b_k z^{-k}$, then $P_{IN}(z)$ is

$$P_{IN}(z) = z^{N/2} B(z)\, U_{LIPS}(z) \tag{3.30}$$

In time domain,

$$p_{IN}[n] = \sum_{k=0}^{N} b_k u_{LIPS}[n + N/2 - k] \tag{3.31}$$

Note that in Eq.(3.31), $p_{IN}[n]$ depends on the future values of $u_{LIPS}[n]$ and it is a noncausal representation for $p_{IN}[n]$.

Combining Eq.(3.26) and (3.30), we can write P_{IN} in terms of U_g as

$$\begin{aligned} P_{IN}(z) &= B(z)\, H(z)\, U_g(z) \\ &= \frac{\sum\limits_{n=0}^{N} b_k z^{-k}}{\sum\limits_{n=0}^{N} a_k z^{-k}} U_g(z) \end{aligned} \tag{3.32}$$

From Eq.(3.32), the vocal tract input pressure, $p_{IN}[n]$, in terms of glottal flow is

$$p_{IN}[n] = \frac{1}{a_0} \sum_{k=0}^{N} b_k u_g[n-k] - \frac{1}{a_0} \sum_{k=1}^{N} a_k p_{IN}[n-k] \tag{3.33}$$

Eq.(3.33) is a causal representation for calculating $p_{IN}[n]$. In order to calculate interactive glottal flow, $p_{IN}[n]$ is added into $u_g[n]$ expressed by Eq.(3.9) as follows

$$u_g[n] = \beta a_g[n] \sqrt{(P_L - p_{IN}[n])} \tag{3.34}$$

There are two alternatives for the inclusion of p_{IN} in Eq.(3.34)

1. p_{IN} in terms of $u_{Lips}[n]$

   Substituting equation 3.31 into 3.34, we get

   $$u_g[n] = \beta a_g[n] \sqrt{\left(P_L - \sum_{k=0}^{N} b_k u_{LIPS}[n + N/2 - k]\right)} \tag{3.35}$$

   where $u_{LIPS}[n]$, from Eq.(3.27) is

   $$u_{LIPS}[n] = \frac{1}{a_0} u_g[n - N/2] - \frac{1}{a_0} \sum_{k=1}^{N} a_k u_{LIPS}[n-k] \tag{3.36}$$

Note that Eq.(3.36) is a nonlinear noncausal recursive representation for interactive glottal flow. The current glottal flow, $u_g[n]$, depends on future values of the lips flow, $u_{LIPS}[n]$. The system representation is not realizable.

2. p_{IN} in terms of $u_g[n]$

   Normalizing the filter coefficients, a_k and b_k by a_0, from Eq.(3.33), we can write

$$p_{IN}[n] = \sum_{k=0}^{N} b_k u_g[n-k] - \sum_{k=1}^{N} a_k p_{IN}[n-k] \tag{3.37}$$

   Substituting Eq.(3.37) into (3.34), we get

$$u_g[n] = \beta a_g[n] \sqrt{P_L - \sum_{k=0}^{N} b_k u_g[n-k] + \sum_{k=1}^{N} a_k p_{IN}[n-k]} \tag{3.38}$$

   which can be written as

$$u_g{}^2[n] + b_0\beta^2 a_g{}^2[n] u_g[n] - a_g{}^2[n]\beta^2 \left(P_L - \hat{p}_{IN}[n]\right) = 0 \tag{3.39}$$

   where $\hat{p}_{IN}[n]$ is

$$\begin{aligned} \hat{p}_{IN}[n] &= \sum_{k=1}^{N} b_k u_g[n-k] - \sum_{k=1}^{N} a_k p_{IN}[n-k] && (3.40) \\ &= p_{IN}[n] - b_0 u_g[n] && (3.41) \end{aligned}$$

   Eq.(3.39), a recursive second order function of $u_g[n]$, has to be solved to get two solutions for $u_g[n]$ at any time instant n. The solution is

$$u_{g_{1,2}}[n] = \frac{\beta a_g[n]\left(-b_0\beta a_g[n] \pm \sqrt{b_0{}^2\beta^2 a_g{}^2[n] + 4\left(P_L - \hat{p}_{IN}[n]\right)}\right)}{2} \tag{3.42}$$

   In order to have a real solution, according to Eq.(3.34), it is required that $P_L \geq p_{IN}[n]$. To avoid any ambiguous solution for $u_g[n]$, Eq.(3.42) must have only one positive result. It is satisfied if

$$\left|\sqrt{b_0{}^2\beta^2 a_g{}^2[n] + 4\left(P_L - \hat{p}_{IN}[n]\right)}\right| \geq \left|b_0\beta a_g[n]\right| \tag{3.43}$$

$$b_0{}^2\beta^2 a_g{}^2[n] + 4\left(P_L - \hat{p}_{IN}[n]\right) \geq b_0{}^2\beta^2 a_g{}^2[n] \tag{3.44}$$

$$P_L \geq \hat{p}_{IN}[n] \tag{3.45}$$

Eq.(3.45) is a required condition for unique $u_g\,[n]$. By combining $P_L \geq \hat{p}_{IN}\,[n]$ and $P_L \geq p_{IN}\,[n]$, the required condition for real and unique $u_g\,[n]$ is $P_L \geq max(p_{IN}\,[n]\,, \hat{p}_{IN}\,[n])$. If b_0 is chosen to be positive ($b_0 > 0$), according to Eq.(3.41), $p_{IN}\,[n] > \hat{p}_{IN}\,[n]$ and the condition for real and unique glottal flow, $u_g\,[n]$, turns out to be $P_L \geq p_{IN}\,[n]$.

The solution of Eq.(3.42), is used for the calculation of $p_{IN}\,[n]$ in Eq.(3.33). At each time instant n, calculation of the parameters, $\hat{p}_{IN}[n], u_g[n]$, and $p_{IN}[n]$ is required.

3.2.1 Modification of Interactive Model

In the previous section, an interactive model considering only vocal tract loading on the source is obtained for voiced speech. This model is nonlinear and its stability depends on the amplitude of the input, glottal area. In order to have a control over the glottal area amplitude, the glottal area function $\beta a_g\,[n]$ in Eq.(3.38) is represented as a product of a normalized waveform, $a_{g_{UNITY}}\,[n]$, whose maximum value is unity, and an amplitude scale parameter, $A_{\max}$;

$$u_g\,[n] = A_{\max}.a_{g_{UNITY}}\,[n]\,\sqrt{P_L - p_{IN}\,[n]} \tag{3.46}$$

This modification helps to develope a robust parameter estimation algorithm.

The second modification is related with combining the subglottal tract to the model. In the previous section, we consider only supraglottal tract in calculating the input impedance, $Z_{IN} = \frac{P_{IN}}{U_g}$. It is known that the impedance relating the glottal flow to the transglottal pressure across the glottis is the sum of the subglottal and supraglottal impedances [11]. In other words, if we consider both subglottal and supraglottal impedance, the transglottal pressure can be written as

$$\begin{aligned} k_t \frac{1}{2} \rho \frac{{u_g}^2}{{a_g}^2} &= P_L - U_g * Z_{sub} - U_g * Z_{sup} && (3.47) \\ &= P_L - U_g * (Z_{sub} + Z_{sup}) && (3.48) \end{aligned}$$

By representing $U_g * (Z_{sub} + Z_{sup})$ in Eq.(3.48) as p_{IN}, we obtain the same glottal flow equation given by Eq.(3.46). However, in this case, p_{IN} represent the pressure drop due to not only the supraglottal impedance but also subglottal impedance. Therefore, if the subglottal impedance is taken into account, overall impedance is the sum of the subglottal and supraglottal impedances. It is shown in the previous section that, by using the chain matrix model, the subglottal and supraglottal impedances can be expressed by rational transfer functions. As

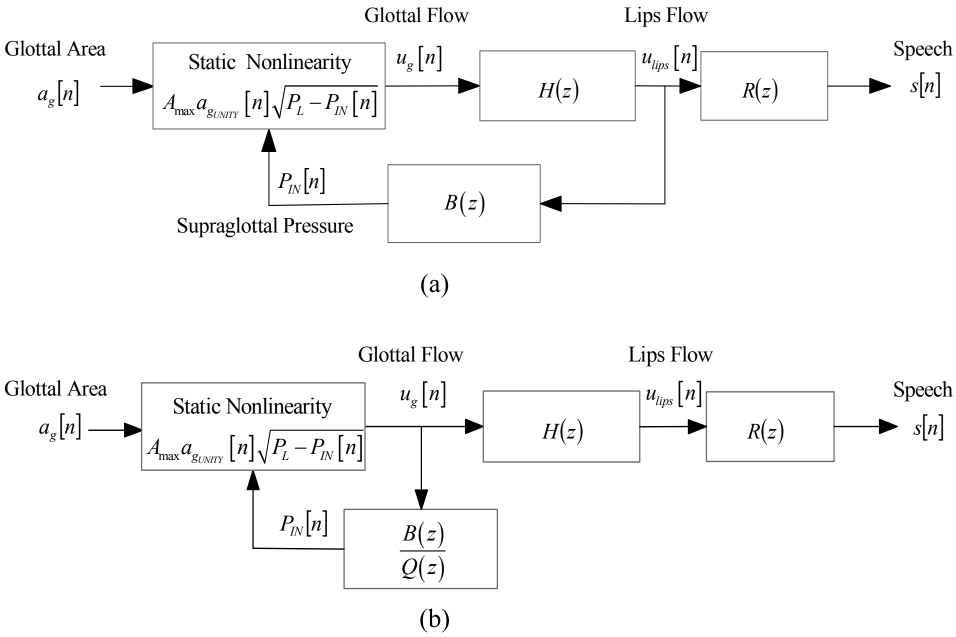


Figure 3.3: (a) Interactive Source - Filter Model I (ISFM1) (b) Interactive Source - Filter Model II (ISFM2)

a result, the overall system transfer function is the sum of the two rational functions, which is a rational function. Let q_k's denote the coefficients of the denominator of input impedance for the combined system and the order of the transfer function be N. Then,

$$p_{IN}[n] = \sum_{k=0}^{N} b_k u_g[n-k] - \sum_{k=1}^{N} q_k p_{IN}[n-k] \tag{3.49}$$

Indeed b_k's in Eq.(3.49) are different than those in Eq.(3.37), however we prefer to keep the notation.

Based on the way $p_{IN}[n]$ is calculated, we propose two interactive models. The first model using Eq.(3.37) together with Eq.(3.46) is called as interactive source-filter model I (ISFM1) and the second model using Eq.(3.49) for $p_{IN}[n]$ in Eq.(3.46) is named as interactive source-filter model II (ISFM2). The proposed interactive models, ISFM1 and ISFM2, are shown in Fig. 3.3.(a)-(b), respectively.

3.2.2 Analysis of ISFMs

The first point in the analysis of the interactive models is the stability. The interactive systems produce bounded output if the vocal tract filter, $H(z)$, is stable and the interactive glottal flow, $u_g[n]$, is bounded. Since the ISFMs are nonlinear, the stability of the system depends on the amplitude of the input, particularly the maximum amplitude of glottal area, A_{max}, and the other model parameters, P_L, b_k's and a_k's. The maximum of the glottal flow amplitude, $\left\|u_g\right\|_\infty$, is expressed by

$$\begin{aligned} \left\|u_g\right\|_\infty &= \left\|A_{\max} a_{gUNITY}\sqrt{P_L - p_{IN}}\right\|_\infty & (3.50)\\ \left\|u_g\right\|_\infty &\leq \left\|A_{\max} a_{gUNITY}\right\|_\infty \left\|\sqrt{P_L - p_{IN}}\right\|_\infty & (3.51) \end{aligned}$$

A finite gain [54], $\gamma(H)_\infty$, can be defined for the square root nonlinearity as

$$\begin{aligned} \gamma(H)_\infty &\geq \max \frac{\left\|\sqrt{P_L - p_{IN}}\right\|_\infty}{\|P_L - p_{IN}\|_\infty} & (3.52)\\ \gamma(H)_\infty &\geq \frac{\left\|\sqrt{P_L - p_{IN}}\right\|_\infty}{P_L - \|p_{IN}\|_\infty} & (3.53)\\ \gamma(H)_\infty \left(P_L - \|p_{IN}\|_\infty\right) &\geq \left\|\sqrt{P_L - p_{IN}}\right\|_\infty & (3.54) \end{aligned}$$

By substituting Eq.(3.54) into Eq.(3.51), $\left\|u_g\right\|_\infty$ is written as

$$\left\|u_g\right\|_\infty \leq A_{\max}\gamma(H_1)_\infty \left(P_L - \|p_{IN}\|_\infty\right) \qquad (3.55)$$

$\|p_{IN}\|_\infty$ can be expressed by a gain function, $\Gamma(H)_\infty$, as $\|p_{IN}\|_\infty = \Gamma(H)_\infty \left\|u_g\right\|_\infty$, which is substituted into Eq. (3.55) to get a closed form expression for $\left\|u_g\right\|_\infty$ as follows,

$$\begin{aligned} \left\|u_g\right\|_\infty &\leq A_{\max}\gamma(H)_\infty \left(P_L - \Gamma(H)_\infty \left\|u_g\right\|_\infty\right) \\ \left\|u_g\right\|_\infty &+ A_{\max}\gamma(H)_\infty\Gamma(H)_\infty \left\|u_g\right\|_\infty \leq A_{\max}P_L\gamma(H)_\infty \\ \left\|u_g\right\|_\infty &\leq \frac{A_{\max}P_L\gamma(H)_\infty}{1 + A_{\max}\gamma(H)_\infty\Gamma(H)_\infty} & (3.56) \end{aligned}$$

Since $\|p_{IN}\|_\infty = \Gamma(H)_\infty \left\|u_g\right\|_\infty$, the closed form expression for $\|p_{IN}\|_\infty$ can be obtained by multiplying Eq.(3.56) by $\Gamma(H)_\infty$, $\|p_{IN}\|_\infty$ can be calculated as

$$\|p_{IN}\|_\infty \leq \frac{A_{\max}P_L\gamma(H)_\infty\Gamma(H)_\infty}{1 + A_{\max}\gamma(H)_\infty\Gamma(H)_\infty} \qquad (3.57)$$

Some important limiting conditions obtained from Eq.(3.56) and Eq.(3.57) for $\left\|u_g\right\|_\infty$ and $\|p_{IN}\|_\infty$ are given in table 3.1,

Table 3.1: The limiting conditions of the model parameters

Limit	$\left\|u_g\right\|_\infty$	$\|p_{IN}\|_\infty$
$A_{\max} \to 0$	0	0
$A_{\max} \to \infty$	$\frac{P_L}{\Gamma(H)_\infty}$	P_L
$P_L \to 0$	0	0
$P_L \to \infty$	∞	∞
$\Gamma(H)_\infty \to 0$	$A_{\max}\sqrt{P_L}$	0
$\Gamma(H)_\infty \to \infty$	0	P_L

The limiting conditions inform about the states of the interactive system. Note that $\Gamma(H)_\infty \to 0$ is the asymptotical condition that the interactive system is in non-interactive mode and equivalent to linear source-filter model. In this mode, glottal flow is a linear function of the glottal area and vocal tract has no influence on it. Another interesting observation is that if $\Gamma(H)_\infty \to \infty$, $\left\|u_g\right\|_\infty \to 0$ and $\|p_{IN}\|_\infty$ is bounded by P_L. An unstable p_{IN} filter can produce bounded glottal flow, $u_g[n]$, and $p_{IN}[n]$ due to the nonlinear feedback in the model. However, in practice when the magnitudes of the poles of the p_{IN} filter are very large and outside the unit circle, p_{IN} can exceed P_L in a small fraction of time and $u_g[n]$ becomes complex. To avoid this problem, a sufficient condition for the p_{IN} filter is $\Gamma(H)\left\|u_g\right\|_\infty \leq P_L$. If we rewrite $\left\|u_g\right\|_\infty$ as $\left\|u_g\right\|_\infty \leq A_{max}\sqrt{P_L}\left\|\sqrt{1-\frac{p_{IN}}{P_L}}\right\|_\infty$, it is easy to show that the maximum value of $\left\|u_g\right\|_\infty$ is less than or equal to $A_{max}\sqrt{2P_L}$ when $p_{IN} \leq P_L$. Substituting it into $\Gamma(H)\left\|u_g\right\|_\infty \leq P_L$,

$$\begin{aligned} \Gamma(H)\left\|u_g\right\|_\infty &\leq P_L \\ \Gamma(H)A_{max}\sqrt{2P_L} &\leq P_L \\ \Gamma(H) &\leq \frac{\sqrt{P_L}}{\sqrt{2}A_{max}} \end{aligned} \tag{3.58}$$

Eq.(3.58) is a sufficient condition that the interactive system produces real and bounded glottal flow, $u_g[n]$.

According to limiting conditions, if $\Gamma(H)_\infty \to 0$ glottal flow is non-interactive meaning that it is a function of only glottal area and P_L, hence, source and filter are independent. Therefore, if $\Gamma(H) << \frac{\sqrt{P_L}}{\sqrt{2}A_{\max}}$ is satisfied, ($p_{IN} << P_L$), the ISFMs run in non-interactive mode and perform the same as LSFM. Using this property, it is possible to have ISFM performing always equal or better than LSFM. A_{max} and P_L are the control parameters for the mode of the ISFMs.

CHAPTER 4

PARAMETER ESTIMATION

4.1 Introduction

In speech analysis, the estimation of model parameters is a blind deconvolution problem. Since there is nothing known about either the source, glottal flow, or the filter, vocal tract. This type of problems can be solved using some a priori informations. The most widely used information for voiced speech analysis is the glottal closed phase of vocal folds. When the vocal folds are closed completely, the glottal flow is zero. Therefore, the vocal tract filter can be estimated accurately from speech signal during closed-phase. Later the estimated filter is used for glottal flow extraction by inverse filtering speech. The same approach can be used for estimating the nonlinear interactive source-filter models (ISFMs) since when the vocal folds are closed, the outputs of both LSFM and ISFMs are the same. The vocal tract is estimated by using closed-phase data, however estimation of the other parameters, P_L, A_{max}, a_k, b_k and q_k, make the problem become more complex. In this section, a robust method is developed to solve this complex nonlinear blind deconvolution problem

4.2 Parameter Estimation

Let $s[n]$ and $\hat{s}[n]$ denote speech samples of length M corresponding to a voiced utterance and its estimate reconstructed by the model respectively. Defining the error as

$$e[n|\theta] = s[n] - \hat{s}[n|\theta] \tag{4.1}$$

where θ is the parameter vector of length N. The reconstruction error is a vector in $R^M$and the size of the error can be measured by any norm, l. The most widely used norm is l_2 and it represents the mean squared error

$$E(\theta) = \frac{1}{M}\sum_{n=1}^{M}(e[n|\theta])^2 \tag{4.2}$$

The parameter estimation problem can be written as

$$\overset{\wedge}{\theta} = \underset{\theta \in R^N}{\arg\min}\,(E(\theta)) \tag{4.3}$$

4.3 Estimation of Vocal Tract Filter & Glottal Flow from Speech

The parameters of the linear source - filter model are generally estimated in a two stage process, inverse filtering and optimization of the voice source model. It is called as joint source filter optimization [43]. According to linear source - filter model, after removing the lip radiation effect, the voiced speech signal can be written as

$$\sum_{k=0}^{N} a_k s[n-k] = u_g[n],\ \text{where } a_0 = 1 \tag{4.4}$$

By inspecting Eq. (4.4), it is seen that accurate vocal tract filter estimates are obtained using the speech signal corresponding to time interval at which the vocal folds are closed. For completely closed vocal folds, the glottal flow is zero and according to Eq. (4.4), the covariance analysis of speech signal in the closed phase provides vocal tract filter coefficients, a_k [46, 47, 48]. It is written as

$$a_k = \Phi^{-1}.r \tag{4.5}$$

where Φ is the covariance matrix of $s[n]$ calculated over the closed phase interval and r is the vector containing the covariance values from 1 to N [49]. The covariance analysis may produce unstable vocal tract filter. In this case, the poles of the estimated vocal tract filter are reflected to their conjugate reciprocal locations to obtain a stable filter [49]. In addition, sometimes positive real poles that deteriorate the glottal flow waveform estimate may occur. These positive real poles are removed from the estimated vocal tract filter [46, 49]. After estimating the vocal tract filter, the glottal flow waveform is obtained by filtering the speech signal by $A(z)$,

$$u_g[n] = \sum_{k=0}^{N} a_k s[n-k] \tag{4.6}$$

4.3.1 Source-Filter Optimization

The inverse filtered speech signal provides glottal flow estimate. It is used as a reference for the source model of speech production, $u_{g_{MOD}}[n]$.

Let $\hat{s}[n]$ denote the reconstructed speech signal by the source model, $\hat{u}_{g_{MOD}}[n]$, then the reconstruction error can be written in terms of glottal flow as

$$e_u[n] = \sum_{k=0}^{N} a_k e[n-k] \quad = \sum_{k=0}^{N} a_k s[n-k] - \sum_{k=0}^{N} a_k \hat{s}[n-k] \tag{4.7}$$

$$= u_g[n] - \hat{u}_{g_{MOD}}[n] \tag{4.8}$$

The cost function in terms of glottal flow can be written as

$$E_u = \frac{1}{M}\sum_{n=1}^{M}\left(u_g[n] - \hat{u}_{g_{MOD}}[n]\right)^2 \tag{4.9}$$

The problem is to minimize the error, E_u, between the glottal flow estimate and its model. As a source model, $u_{g_{MOD}}[n]$, the most widely used models are Liljencrants-Fant (LF) , Rosenberg, Rosenberg+ and KLGLOTT [31, 43]. Among the all source models, the LF model is used as a reference model due to its physiological meaning. Estimation of the LF model parameters is a highly nonlinear nonconvex optimization problem. There are two stage approaches, first estimate a much simpler model whose global minimum is guaranteed, then use the parameter of the estimated model for the initial values of LF model. Using trust region type optimization, more accurate model parameters can be estimated [43].

4.4 Nonlinear Source-Filter Model Parameter Estimation Algorithm

We follow the steps below to find the model parameters, $A_{max}, P_L, a_k, b_k, q_k$.

1. Closed - Phase Analysis (CPA)

   Estimation of glottal flow waveform by removing the vocal tract and radiation filters from speech signal using constrained closed phase analysis [49]. In this step, both glottal flow waveform, $u_g[n]$, and vocal tract filter coefficients, a_k's, are estimated from speech.

2. Glottal parameterization

   For pitch synchronous analysis, the glottal closure and the glottal opening instants and, the location and the amplitude of glottal flow maxima are obtained from glottal flow estimate.

3. Glottal area model initialization

   Using the glottal parameters, the time sequence denoting the samples of glottal area model, $a_g\,[n]$, is generated. There are parametric glottal area models used for articulatory synthesis and acoustic analysis in the literature [36]. It is also possible to use parametric glottal flow models instead of glottal area model, such as Rosenberg, LF etc.

4. Multi-parameter optimization

   A method is developed to estimate the interactive glottal flow model parameters with respect to the error between the target glottal flow, $u_g\,[n]$, obtained by inverse filtering and reconstructed glottal flow, $\hat{u}_g\,[n;\theta]$, by the interactive system

   $$E_u\,(\theta) = \frac{1}{N_0}\sum_{n=1}^{N_0}\left(u_g\,[n] - \hat{u}_g\,[n;\theta]\right)^2 \tag{4.10}$$

   where θ denotes the model parameters. Since the interactive glottal flow is a nonlinear function of the model parameters, A_{max}, P_L, a_k and b_k, minimization of $E_u\,(\theta)$ is a nonlinear optimization problem. Due to the particular nonlinearity, there is no analytical solution for the minimization problem. However, by choosing good initial values, it is possible to solve the problem iteratively.

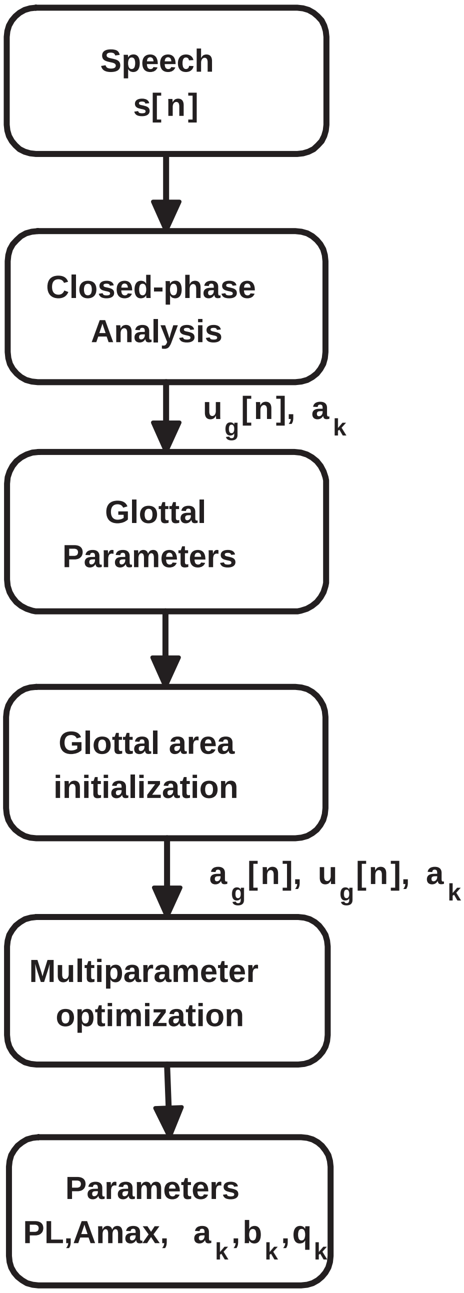


Figure 4.1: Parameter estimation algorithm

4.4.1 Glottal Parameterization

In the glottal parameterization phase, the time and amplitude instants of glottal flow obtained by CPA is marked to synthesize glottal area model waveform. Some of the parameters of the most widely used glottal source models are as follows

N_p : the time instant at which the glottal flow has a maximum.

N_{GCI} : the time instant at which the glottal flow derivative has negative sharp peak corresponding to glottal closure event of vocal folds

N_{GCO} : the time instant at which the vocal folds start to opening

4.4.2 Nonlinear Model Parameter Optimization

The parameters of the proposed nonlinear models are estimated based on a target signal. It can be glottal flow or glottal flow derivative estimated from inverse filtering or directly speech signal. The target is chosen as the glottal flow estimated by inverse filtering since the source-filter interaction effects are easily tracked on the glottal flow estimates.

The objective in the parameter estimation problem is to minimize the sum of the squared error between the glottal flow waveform obtained from inverse filtering and glottal flow waveform produced by the proposed nonlinear system. After obtaining the glottal flow by inverse filtering, the task is to approximate it by using proposed glottal flow model. This requires the optimization of the nonlinear model parameters $A_{\max}, P_L, b_k, q_k$ with respect to error E,

$$E = \frac{1}{M}\sum_{n=1}^{M}\left(u_g[n] - A_{\max}.a_{g_{UNITY}}[n]\ \sqrt{P_L - p_{IN}[n]}\right)^2 \tag{4.11}$$

Since the relation between the error, $e[n]$, and system parameters are nonlinear, it is a nonlinear least squares problem. The nonlinear least squares problems are generally solved by two types descent algorithms, Gauss-Newton or Levenberg-Marquate [53]. The main feature of these algorithms is that they do not require Hessian of the cost function by assuming that the error, $e[n]$, is small. They require the gradient of the cost function with respect to parameters [53]. Sometimes calculation of analytic gradient may be computationally very high, in that case either finite difference approximation for the gradient or non-gradient based optimization algorithms like simplex method can be used [55]. The most important problem for the nonlinear optimization is to find robust initial values for the parameters due to the non-convex structure of the error surface that may lead to iterations to stuck on a local minimum.

4.5 Optimization of the parameters of ISFMs

The coupled system equations are expressed as

$$\begin{aligned}\hat{u}_g[n] &= a_g[n]\sqrt{P_L - p_{IN}[n]}\\ &= A_{\max}.a_{g_{UNITY}}[n]\sqrt{P_L - p_{IN}[n]}\end{aligned} \tag{4.12}$$

and

$$\begin{aligned}p_{IN}[n] &= \sum_{k=1}^{N} b_k\hat{u}_g[n-k] - \sum_{k=1}^{N} q_k p_{IN}[n-k]\\ &= \sum_{k=1}^{N} b_k a_g[n-k]\sqrt{P_L - p_{IN}[n-k]} - \sum_{k=1}^{N} q_k p_{IN}[n-k]\end{aligned} \tag{4.13}$$

The following sections describe the parameter optimization routines of the ISFM models.

4.5.1 Gradient Based Optimization Algorithm

The sample error in glottal flow can be written as

$$e[n|\theta] = u_g[n] - A_{\max}.a_{g_{UNITY}}[n]\sqrt{P_L - p_{IN}[n|\theta]} \tag{4.14}$$

where θ denotes the model parameters $A_{\max}$, P_L, b_k and q_k. The objective function is the mean squared error

$$E(\theta) = \frac{1}{M}\sum_{n=1}^{M} e^2[n|\theta] = \frac{1}{M}\sum_{n=1}^{M}\left(u_g[n] - A_{\max}.a_{g_{UNITY}}[n]\sqrt{P_L - p_{IN}[n|\theta]}\right)^2 \tag{4.15}$$

Since the model parameters are nonlinearly related with the error function, this is a nonlinear least squares problem. It can be solved by gradient based iterative algorithms[53]. In vector form, Eq.(4.15) becomes

$$E(\theta) = \frac{1}{M}\underline{e}^{T}.\underline{e} \tag{4.16}$$

where $\underline{e} = \begin{bmatrix} e\,(1|\theta) & e\,(2|\theta) & . & . & . & . & e\,(M|\theta) \end{bmatrix}^{T}$. The derivative of $E(\theta)$ with respect to the elements, θ_k, of the parameter vector θ is

$$\frac{\partial E\,(\theta)}{\partial \theta_k} = \frac{2}{M}\sum_{n=1}^{M} e\,[n]\,\frac{\partial e\,[n]}{\partial \theta_k} \tag{4.17}$$

and the Jacobian is

$$\underline{J} = \begin{bmatrix} \frac{\partial e[1]}{\partial \theta_1} & \frac{\partial e[1]}{\partial \theta_2} \\ . & . \\ . & . \\ . & . \\ \frac{\partial e[M]}{\partial \theta_1} & \frac{\partial e[M]}{\partial \theta_2} \end{bmatrix} \tag{4.18}$$

Therefore, the gradient can be written in terms of Jacobian and error as follows

$$\nabla_{\theta} E\,(\theta) = \frac{2}{M} J^{T}\underline{e} \tag{4.19}$$

The entries of the Hessian matrix of the cost function is obtained by taking the derivatives of the gradient vector with respect to θ as

$$H_{k,l} = \frac{\partial^2 E\,(\theta)}{\partial \theta_k \partial \theta_l} = \frac{2}{M}\sum_{n=1}^{M}\left(\frac{\partial e\,[n]}{\partial \theta_k}\frac{\partial e\,[n]}{\partial \theta_l} + e\,[n]\,\frac{\partial^2 e\,[n]}{\partial \theta_k \partial \theta_l}\right) \tag{4.20}$$

in terms of Jacobian

$$H = \frac{2}{M}J^{T}J + \frac{2}{M}\sum_{n=1}^{M} e\,[n]\,.\frac{e^2\,[n]}{\partial \theta_k \partial \theta_l} \tag{4.21}$$

using $S = \sum_{n=1}^{M} e\,[n]\,.\frac{e^2[n]}{\partial \theta_k \partial \theta_l}$ the Hessian matrix can be written as

$$H = \frac{2}{M}J^{T}J + \frac{2}{M}S \tag{4.22}$$

Depending on the size of S, there are two approaches for the calculation of Eq.(4.22), small residue and large residue algorithms[53]. If S is large with respect to $J^{T}J$, it has to be calculated either analyticaly or by using finite difference approximation. This type of approaches are called large residue algorithms. However, if the error, $e[n]$, is small enough to neglect the term S, the Hessian can be calculated by using only the first derivative of the cost function. Algorithms that use this approximation are called small residue algorithms[53]. The Gauss-Newton and Levenberg-Marquate are the two algorithms that are based on the small error assumption. In the Gauss-Newton method, the Hessian is approximated by $H \approx \frac{2}{M} J^T J$, hence during the iterations, the parameters are updated as

$$\theta_{k+1} = \theta_k - \eta_k \left({J_k}^T J_k \right)^{-1} {J_k}^T e_k \tag{4.23}$$

where η is step size that can be chosen either as a constant or computed using line search[53].

4.5.2 Optimization of the scalar parameter A_{max}

The system error in glottal flow, $e[n]$, is defined as $e[n] = u_g[n] - \hat{u}_g[n]$. The derivative of the mean squared error, E, with respect to A_{max}, $\frac{\partial E}{\partial A_{max}}$ is

$$\begin{aligned} \frac{\partial E}{\partial A_{max}} &= \frac{2}{M} \sum_{n=1}^{M} e[n] \frac{\partial e[n]}{\partial A_{max}} \\ &= \frac{2}{M} J_{A_{max}} \underline{e} \end{aligned} \tag{4.24}$$

where $J_{A_{max}}$ and $\underline{e}$ are the gradient and error vectors of length M, respectively. Using Eq.(4.12), the derivative is expressed as

$$\frac{\partial E}{\partial A_{max}} = \frac{\partial \left(\sum_{n=1}^{M} \left(u_g[n] - a_g[n] \sqrt{P_L - p_{IN}[n]} \right)^2 \right)}{\partial A_{max}} \tag{4.25}$$

$$= \frac{2}{M} . \sum_{n=1}^{M} e[n] \frac{\partial \left(-a_g[n] \sqrt{P_L - p_{IN}[n]} \right)}{\partial A_{\max}} \tag{4.26}$$

$$= -\frac{2}{M} \sum_{n=1}^{M} e[n] . \left(\frac{\partial a_g[n]}{\partial A_{\max}} \sqrt{P_L - p_{IN}[n]} + a_g[n] \frac{\partial \left(\sqrt{P_L - p_{IN}[n]} \right)}{\partial A_{\max}} \right) \tag{4.27}$$

$$= \frac{2}{M} \sum_{n=1}^{M} e[n] . \left(\frac{\hat{u}_g[n]}{2\left(P_L - p_{IN}[n]\right)} \frac{\partial p_{IN}[n]}{\partial A_{\max}} - \frac{\hat{u}_g[n]}{A_{max}} \right) \tag{4.28}$$

The elements of $J_{A_{max}}$ used by Gauss-Newton method in Eq. (4.24) is as follows

$$J_{A_{max}}[n] = \frac{\hat{u}_g[n]}{2\left(P_L - p_{IN}[n]\right)}\frac{\partial p_{IN}[n]}{\partial A_{\max}} - \frac{\hat{u}_g[n]}{A_{max}} \tag{4.29}$$

where $\frac{\partial p_{IN}[n]}{\partial A_{max}}$ can be calculated by taking the derivative of $p_{IN}[n]$ given by Eq.(4.13) with respect to A_{max}.

$$\frac{\partial p_{IN}[n]}{\partial A_{\max}} = \frac{\partial \sum\limits_{k=0}^{N} b_k A_{\max} a_{g_{UNITY}}[n-k]\sqrt{P_L - p_{IN}[n-k]}}{\partial A_{\max}} - \frac{\partial \sum\limits_{k=1}^{N} q_k p_{IN}[n-k]}{\partial A_{\max}} \tag{4.30}$$

$$= \sum_{k=0}^{N} b_k a_{g_{UNITY}}[n-k]\left(\sqrt{P_L - p_{IN}[n-k]} + A_{\max}\frac{\partial\left(\sqrt{P_L - p_{IN}[n-k]}\right)}{\partial A_{\max}}\right) - \sum_{k=1}^{N} q_k \frac{\partial p_{IN}[n-k]}{\partial A_{\max}} \tag{4.31}$$

$$\frac{\partial p_{IN}[n]}{\partial A_{\max}} = \sum_{k=0}^{N}\left(\frac{b_k\hat{u}_g[n-k]}{A_{max}} - \frac{b_k\hat{u}_g[n-k]}{2\left(P_L - p_{IN}[n-k]\right)}\frac{\partial p_{IN}[n-k]}{\partial A_{\max}}\right) - \sum_{k=1}^{N} q_k \frac{\partial p_{IN}[n-k]}{\partial A_{\max}} \tag{4.32}$$

$$\begin{aligned}\left(1 + \frac{b_0\hat{u}_g[n]}{2\left(P_L - p_{IN}[n]\right)}\right)\frac{\partial p_{IN}[n]}{\partial A_{\max}} &= \sum_{k=0}^{N}\left(\frac{b_k\hat{u}_g[n-k]}{A_{max}}\right)\ldots \\ &- \sum_{k=1}^{N}\left(\frac{b_k\hat{u}_g[n-k]}{2\left(P_L - p_{IN}[n-k]\right)} + q_k\right)\frac{\partial p_{IN}[n-k]}{\partial A_{\max}}\end{aligned} \tag{4.33}$$

$$\frac{\partial p_{IN}[n]}{\partial A_{\max}} = \frac{1}{1 + \frac{b_0\hat{u}_g[n]}{2(P_L - p_{IN}[n])}}\left(\sum_{k=0}^{N}\frac{b_k\hat{u}_g[n-k]}{A_{max}} - \sum_{k=1}^{N}\left(\frac{b_k\hat{u}_g[n-k]}{2\left(P_L - p_{IN}[n-k]\right)} + q_k\right)\frac{\partial p_{IN}[n-k]}{\partial A_{\max}}\right) \tag{4.34}$$

When $a_k = b_k$ ($\Gamma(H)_\infty = 1$, see in the parameter estimation algorithm Step-1 in the article), which is the usual case in the optimization of A_{max}, $P_L >> p_{IN}$, hence Eq. (4.34) can be approximated by

$$\frac{\partial p_{IN}[n]}{\partial A_{\max}} = \sum_{k=0}^{N} \frac{b_k \hat{u}_g[n-k]}{A_{max}} - \sum_{k=1}^{N} q_k \frac{\partial p_{IN}[n-k]}{\partial A_{\max}} \tag{4.35}$$

The gradient vector given by Eq. (4.35) can be considered as an output of a time-invariant IIR filter. Infact, the output is equal to $p_{IN}[n]/A_{max}$. Therefore, $J_{A_{max}}$ in Eq.(4.24) can be approximated as

$$J_{A_{max}}[n] = \frac{\hat{u}_g[n]}{2(P_L - p_{IN}[n])} \frac{p_{IN}[n]}{A_{\max}} - \frac{\hat{u}_g[n]}{A_{max}} \tag{4.36}$$

Furthermore, using the same assumption used in Eq.(4.35), $J_{A_{max}}[n]$ turns out to be

$$J_{A_{max}}[n] = -\frac{\hat{u}_g[n]}{A_{max}} \tag{4.37}$$

Substituting Eq. (4.37) into Eq. (4.24) and equating to zero yields,

$$\frac{\partial E}{\partial A_{max}} = 0 \tag{4.38}$$

$$\frac{2}{M} J_{A_{max}} \underline{e} = 0 \tag{4.39}$$

$$-\sum_{n=1}^{M} \frac{\hat{u}_g[n]}{A_{max}} \left(u_g[n] - a_g[n] \sqrt{P_L - p_{IN}[n]} \right) = 0 \tag{4.40}$$

$$\sum_{n=1}^{M} \hat{u}_g[n] \, u_g[n] = \sum_{n=1}^{M} \hat{u}_g[n] \, a_g[n] \sqrt{P_L - p_{IN}[n]} \tag{4.41}$$

$$\sum_{n=1}^{M} \hat{u}_g[n] \, u_g[n] = A_{max} \sum_{n=1}^{M} \hat{u}_g[n] \, a_{g_{UNITY}}[n] \sqrt{P_L - p_{IN}[n]} \tag{4.42}$$

$$A^*_{max} = \frac{\sum_{n=1}^{M} \hat{u}_g[n] \, u_g[n]}{\sum_{n=1}^{M} \hat{u}_g[n] \, a_{g_{UNITY}}[n] \sqrt{P_L - p_{IN}[n]}} \tag{4.43}$$

The optimal A^*_{max} value is obtained by Eq.(4.43).

4.5.3 Optimization of the parameter vector $\mathbf{b_k}$

The gradient of E with respect to $\{b_l\}_{l=0}^{N}$ can be written as

$$\begin{aligned}\frac{\partial E}{\partial b_l} &= \frac{2}{M}.\sum_{n=1}^{M}\left(u_g[n]-a_g[n]\sqrt{P_L-p_{IN}[n]}\right)\frac{\partial\left(-a_g[n]\sqrt{P_L-p_{IN}[n]}\right)}{\partial b_l} &(4.44)\\ &= \frac{2}{M}.\sum_{n=1}^{M}e[n]\frac{a_g[n]}{2\sqrt{P_L-p_{IN}[n]}}\frac{\partial p_{IN}[n]}{\partial b_l} &(4.45)\end{aligned}$$

then, the elements of Jacobian is

$$J_{b_l}[n]=\frac{a_g[n]}{2\sqrt{P_L-p_{IN}[n]}}\frac{\partial p_{IN}[n]}{\partial b_l} \tag{4.46}$$

where $\frac{\partial p_{IN}}{\partial b_l}$ is calculated from Eq. (4.13) as follows

$$\frac{\partial p_{IN}[n]}{\partial b_l}=\frac{\partial}{\partial b_l}\left(\sum_{k=0}^{N}b_k a_g[n-k]\sqrt{P_L-p_{IN}[n-k]}-\sum_{k=1}^{N}q_k p_{IN}[n-k]\right) \tag{4.47}$$

$$\begin{aligned}\frac{\partial p_{IN}[n]}{\partial b_l} &= \hat{u}_g[n-l]-\sum_{k=0}^{N}\frac{b_k a_g[n-k]}{2\sqrt{P_L-p_{IN}[n-k]}}\frac{\partial p_{IN}[n-k]}{\partial b_l}...\\ &\quad-\sum_{k=1}^{N}q_k\frac{\partial p_{IN}[n-k]}{\partial b_l} &(4.48)\\ \left(1+\frac{b_0\hat{u}_g[n]}{2\left(P_L-p_{IN}[n]\right)}\right)\frac{\partial p_{IN}[n]}{\partial b_l} &= \hat{u}_g[n-l]-\sum_{k=1}^{N}\frac{b_k\hat{u}_g[n-k]}{2\left(P_L-p_{IN}[n-k]\right)}\frac{\partial p_{IN}[n-k]}{\partial b_l}...\\ &\quad-\sum_{k=1}^{N}q_k\frac{\partial p_{IN}[n-k]}{\partial b_l} &(4.49)\end{aligned}$$

Finaly form of $\frac{\partial p_{IN}}{\partial b_l}$ is

$$\frac{\partial p_{IN}[n]}{\partial b_l}=\frac{1}{\left(1+\frac{b_0\hat{u}_g[n]}{2(P_L-p_{IN}[n])}\right)}\left(\hat{u}_g[n-l]-\sum_{k=1}^{N}\left(q_k+\frac{b_k\hat{u}_g[n-k]}{2\left(P_L-p_{IN}[n-k]\right)}\right)\frac{\partial p_{IN}[n-k]}{\partial b_l}\right) \tag{4.50}$$

Closed form solution of Eq.(4.45) does not exist. Optimal parameter b_l is estimated for all $l = 0, 1, 2, .., N$ by solving Eq.(4.45) together with Eq.(4.50) iteratively.

4.5.4 Optimization of the parameter vector $\mathbf{q_k}$

The gradient of E with respect to $\{q_l\}_{l=0}^{N}$ can be written as

$$\begin{aligned}
\frac{\partial E}{\partial q_l} &= \frac{2}{M}.\sum_{n=1}^{M}\left(u_g[n] - a_g[n]\sqrt{P_L - p_{IN}[n]}\right)\frac{\partial\left(-a_g[n]\sqrt{P_L - p_{IN}[n]}\right)}{\partial q_l} && (4.51)\\
&= \frac{2}{M}.\sum_{n=1}^{M} e[n]\frac{a_g[n]}{2\sqrt{P_L - p_{IN}[n]}}\frac{\partial p_{IN}[n]}{\partial q_l} && (4.52)
\end{aligned}$$

then, the elements of Jacobian is

$$J_{q_l}[n] = \frac{a_g[n]}{2\sqrt{P_L - p_{IN}[n]}}\frac{\partial p_{IN}[n]}{\partial q_l} \tag{4.53}$$

where $\frac{\partial p_{IN}}{\partial q_l}$ is calculated from Eq. (4.13) as follows

$$\frac{\partial p_{IN}[n]}{\partial q_l} = \frac{\partial}{\partial q_l}\left(\sum_{k=0}^{N} b_k a_g[n-k]\sqrt{P_L - p_{IN}[n-k]} - \sum_{k=1}^{N} q_k p_{IN}[n-k]\right) \tag{4.54}$$

$$\begin{aligned}
\frac{\partial p_{IN}[n]}{\partial q_l} &= -\sum_{k=0}^{N}\frac{b_k a_g[n-k]}{2\sqrt{P_L - p_{IN}[n-k]}}\frac{\partial p_{IN}[n-k]}{\partial q_l} - p_{IN}[n-l]...\\
&\quad - \sum_{k=1}^{N} q_k\frac{\partial p_{IN}[n-k]}{\partial q_l} && (4.55)\\
\left(1 + \frac{b_0\hat{u}_g[n]}{2\left(P_L - p_{IN}[n]\right)}\right)\frac{\partial p_{IN}[n]}{\partial q_l} &= -p_{IN}[n-l] - \sum_{k=1}^{N}\frac{b_k\hat{u}_g[n-k]}{2\left(P_L - p_{IN}[n-k]\right)}\frac{\partial p_{IN}[n-k]}{\partial q_l}...\\
&\quad - \sum_{k=1}^{N} q_k\frac{\partial p_{IN}[n-k]}{\partial q_l} && (4.56)
\end{aligned}$$

$$\frac{\partial p_{IN}[n]}{\partial q_l} = \frac{1}{\left(1+\frac{b_0 \hat{u}_g[n]}{2(P_L - p_{IN}[n])}\right)} \left(-p_{IN}[n-l] - \sum_{k=1}^{N} \left(q_k + \frac{b_k \hat{u}_g[n-k]}{2(P_L - p_{IN}[n-k])}\right) \frac{\partial p_{IN}[n-k]}{\partial b_l}\right) \tag{4.57}$$

Closed form solution of Eq.(4.52) does not exist. Optimal parameter q_l is estimated for all $l = 1, 2, .., N$ by solving Eq.(4.52) together with Eq.(4.57) recursively.

4.5.5 Equalization of linear and nonlinear models using initial parameter values

Initial parameter values are critical for the success of the nonlinear optimization. Physical insight about the system would be useful in choosing the initial values. As far as the parameters A_{max} and P_L are concerned, a "good" choice can be obtained by setting $P_L >> p_{IN}$ and a small value for A_{max}. For large P_L, the interactive glottal flow can be written as

$$\hat{u}_g[n] \cong A_{\max} a_{g_{UNITY}}[n] \sqrt{P_L} \tag{4.58}$$

which implies that the maximum value of glottal flow is approximately $A_{\max}\sqrt{P_L}$. Hence, the corresponding initial value for A_{max} is

$$A_{\max_0} = \frac{\max\left(u_g\right)}{\sqrt{P_L}} \tag{4.59}$$

P_L gives us a flexibility to choose the operating point for ISFMs. We can operate the system either in interactive or noninteractive mode. This property allows the performance of ISFM be equal or better than the linear source-filter model. At its worst, ISFM would be performing like LSFM.

The parameter a_k estimated by CPA are the coefficients of the vocal tract filter and its poles define the vocal tract formants. It can be inferred from the spectrum of the vocal tract impedance [13] that a zero is located between any two consecutive formants. Hence, a suitable initial value for b_k can be a_k. Similarly, for ISFM2, the initial value for q_k can be taken as a_k so that ISFM1 and ISFM2 are equivalent before optimizing ISFM2.

4.5.6 Multi-parameter optimization algorithm

The parameter estimation procedure is as follows

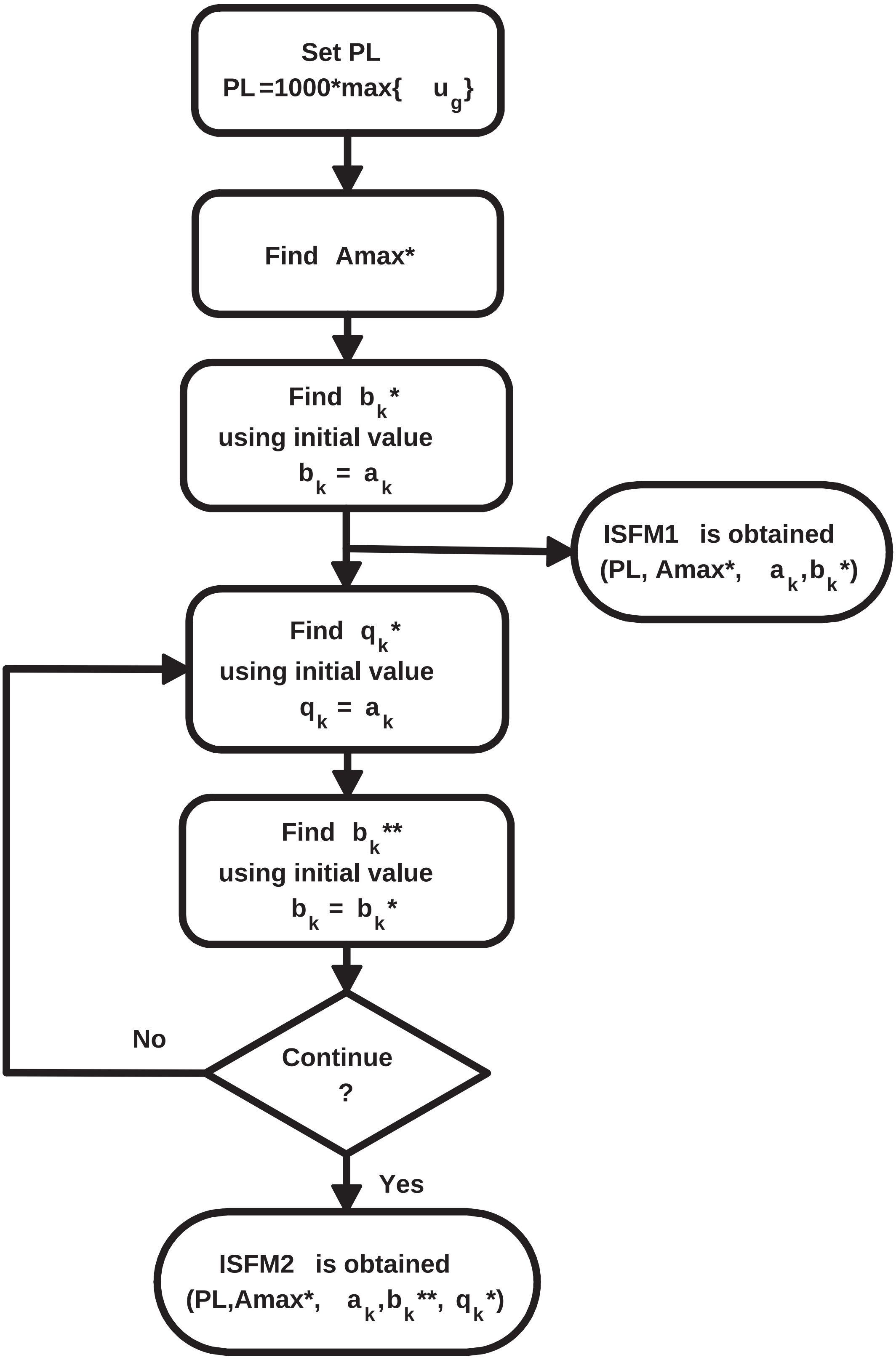


Figure 4.2: Multi-parameter optimization algorithm for the ISFMs

1. Choose P_L as a large positive number, such as $1000 * \max\left(u_g[n]\right)$ in a glottal cycle, so that initially the nonlinear system works in non-interactive mode ($\Gamma(H)_\infty = 1$ since ${b_k}^0 = a_k$).

2. Minimize $E_u(\theta)$ (Eq.(4.10)) with respect to A_{max} with initial values $A_{\mathrm{max}_0} = \frac{\max(u_g)}{\sqrt{P_L}}$, ${b_k}^0 = a_k$ and ${q_k}^0 = a_k$, obtain the glottal area maximum, A_{max}. If the error is larger than the error of LSFM increase P_L and obtain a new value for A_{max}.

3. Minimize $E_u(\theta)$ with respect to b_k with initial values ${b_k}^0 = a_k$ and ${q_k}^0 = a_k$, obtain ${b_k}^*$. The parameters of ISFM1 are now estimated!

4. Minimize $E_u(\theta)$ with respect to q_k with the initial value ${b_k}^*$ and ${q_k}^0 = a_k$, obtain ${q_k}^*$.

5. Minimize $E_u(\theta)$ with respect to b_k with initial values ${b_k}^*$ and ${q_k}^*$, obtain ${b_k}^{**}$ and then continue to iterate as it is in steps 4 and 5 until finding a minimum value. Finally, the parameters of ISFM2 are estimated!

In the first three steps, the parameters of ISFM1 are calculated. After the iterations in the fourth and the fifth steps, the parameters of ISFM2 are found. The flowchart of the algorithm is shown in Fig. 4.2.

4.6 Summary

In this section, a parameter estimation algorithm is developed for the nonlinear interactive source-filter models. Thanks to the algorithm, it is always possible to obtain better ISFMs compared to LSFM. First, it equalizes ISFM1 to LSFM, by setting the operation mode of ISFM1 to non-interactive mode by using the control parameters P_L and A_{max}, then decreases the ISFM1's error further by optimizing b_k. In a similar way, using ISFM1 as an initial model of ISFM2, the algorithm produces better ISFM2 model compared to ISFM1. In the next section, the effectiveness of the proposed algorithm is demonstrated by experiments conducted on a short speech segment.

CHAPTER 5

PRELIMINARY EXPERIMENTS

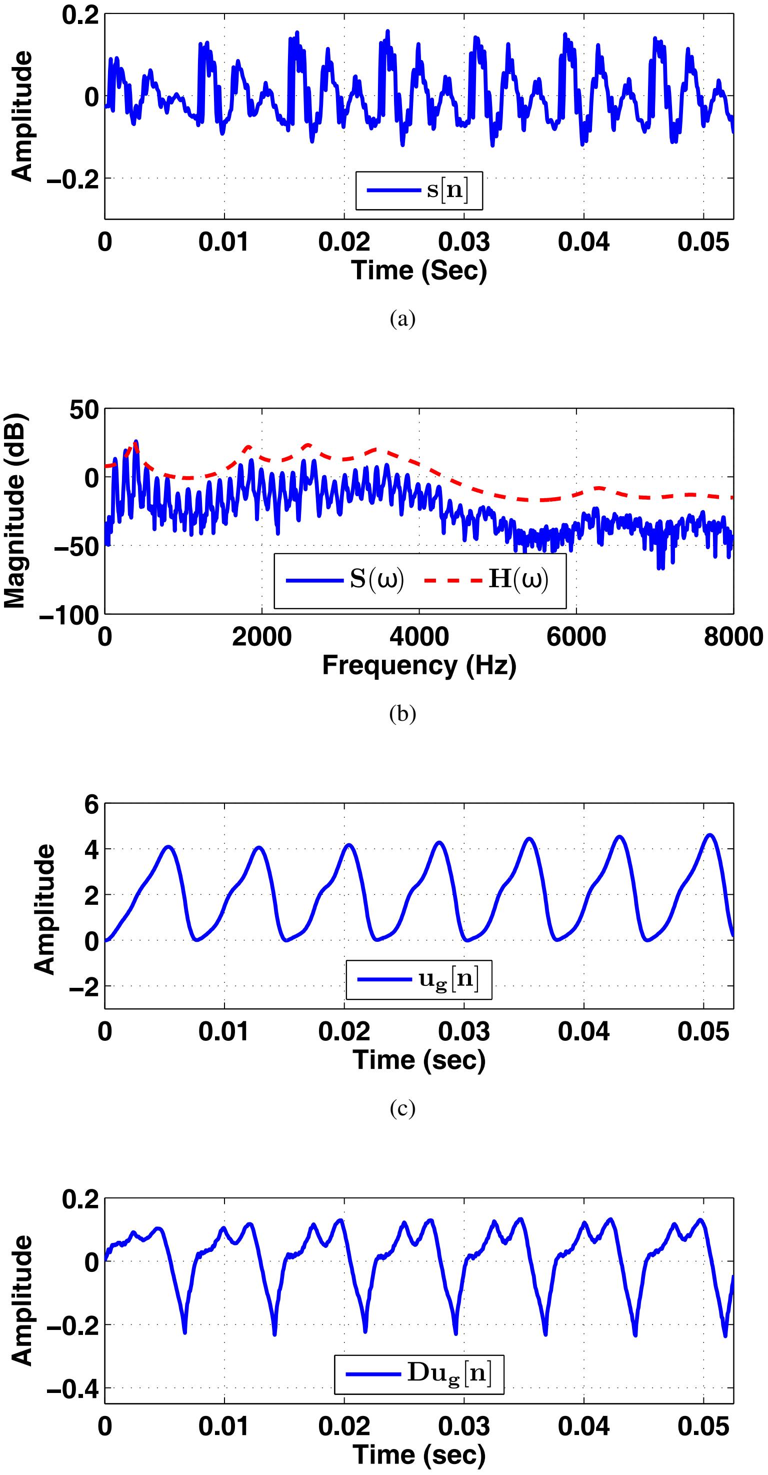
5.1 Introduction

In this section, experiments are conducted to investigate both the parameter estimation algorithm and the glottal flow waveforms produced by ISFM models. The parameter estimation algorithm described in the previous section is applied on a short speech segment. The glottal flow signal is estimated from speech by inverse filtering, then glottal area model waveform is generated. The parameters of the ISFMs are optimized, then interactive glottal flow waveforms are computed using the optimal parameters. The resulting interactive glottal flow waveforms are demonstrated.

5.2 Experiments on a short speech segment

A voiced speech segment shown in Fig. 5.1 is used for the first experiment. The first step of parameter estimation is to estimate the glottal flow waveform from speech signal using closed-phase analysis. The spectrum of speech signal and estimated vocal tract filter are plotted in Fig. 5.1(b). Glottal flow and glottal flow derivative waveforms obtained by using the vocal tract filter are plotted in Fig. 5.1(c) and Fig. 5.1(d), respectively. The glottal flow estimate is similar to the simulated glottal flows obtained by physical models in Sec. II. There is a ripple on the glottal flow and derivative estimate due to the interaction between glottal flow and vocal tract.

The next step is the glottal parameterization in which time instants and amplitude values of the glottal flow and glottal flow derivative are found for the glottal area model. A set of results



(a)

(b)

(c)

(d)

Figure 5.1: (a) Speech Signal, (b) The spectrum of speech (red line) and estimated vocal tract (blue line), (c) Glottal flow estimate, (d) Glottal flow derivative

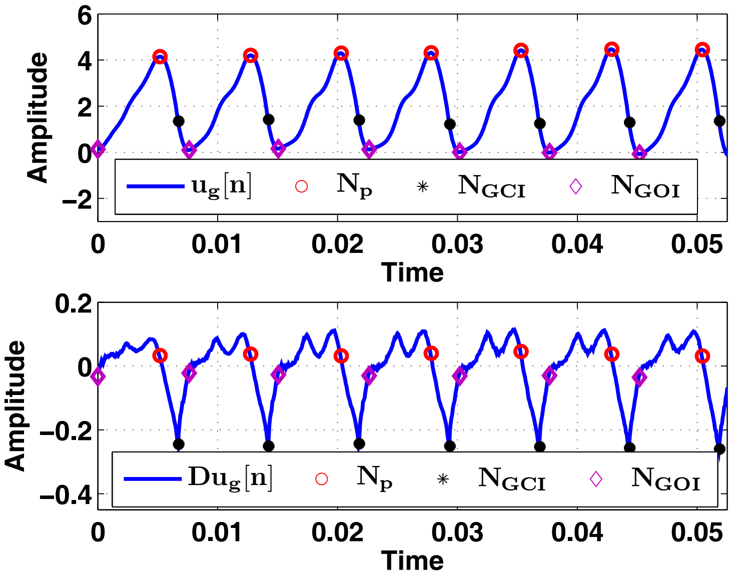


Figure 5.2: Glottal parameterization step N_p- The glottal flow peak instants, N_{GCI} - The glottal closure instants, N_{GOI} - The glottal opening Instants

are shown in Fig. 5.2.

As a glottal area input, we first use Eq. (2.1) [36]; The parameters of the model are estimated from glottal flow estimate assuming that the coarse shapes of the glottal area and glottal flow are close to each other. The glottal area waveform, $a_g[n]$, obtained according to Eq. (2.1) and its derivative are plotted in Fig. 5.3.

The first example interactive glottal flow waveform synthesized by ISFM1 using a non-optimal A_{max} and P_L is shown in Fig. 5.4 (where $A_{max} = 1$, $P_L = 1000$, $b_k = q_k = a_k$ obtained by inverse filtering). In the top figure, the blue, red and black lines show target glottal flow, $u_g[n]$, reconstructed glottal flow, $\overset{\wedge}{u_g[n]}$, and glottal area, $a_g[n]$, respectively, while their derivatives are shown in the bottom. It is seen in both figures that there are two types of error between the target and reconstructed signals. First type of error is due to the magnitude mismatch, while the second type is due to the ripples observed on the target flow. Our approach in the parameter optimization consists of two steps. In the first step, $A_{\max}$ and P_L are optimized so that the first type of error is minimized while the model works in non-interactive mode. In the second step, the parameters of the p_{IN} filter are optimized to produce the fine ripples on the interactive glottal flow waveform.

The error surface and contours obtained for optimization of $A_{\max}$ and P_L are shown in Fig.

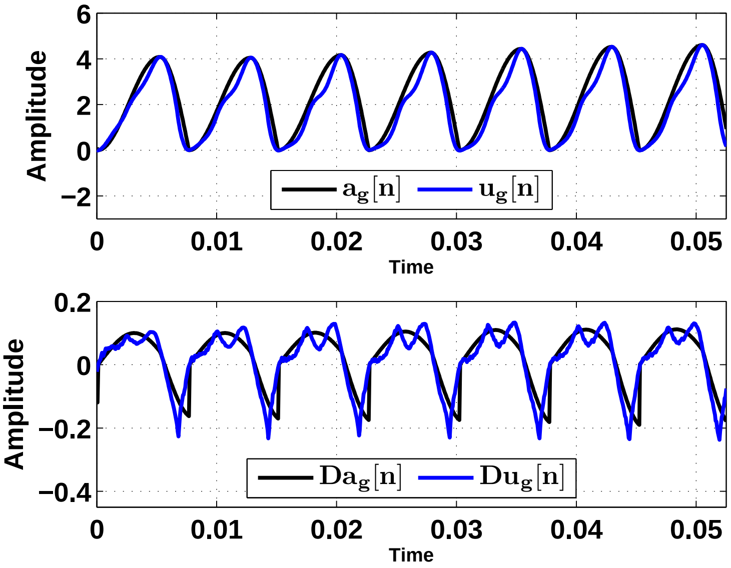


Figure 5.3: (Top) Calculated glottal area waveform using Fant's area model (black line) versus glottal flow estimated by inverse filtering (red line) and (bottom) their derivative waveforms

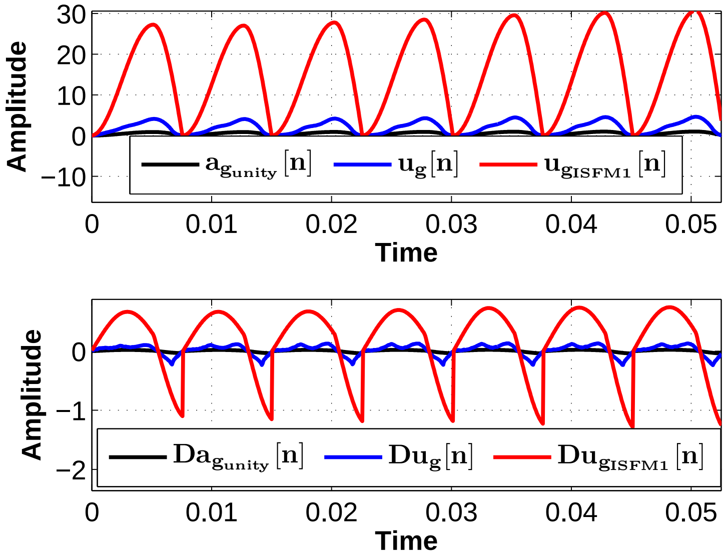


Figure 5.4: (Top) Interactive glottal flow produced by ISFM1 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives

5.5. They are obtained by sampling the cost function at equally spaced values of $A_{\max}$ and P_L. It is seen in the upper figure that the error increases if both parameters are increased. This is reasonable since both A_{max} and P_L are directly proportional to the amplitude of glottal flow. The top plot shows that at each P_L value there is a minimum that corresponds to one unique $A_{\max}$ value. These observations are seen much better on the error contour in the bottom plot. It shows that there is a contour passing through the minimum of the cost function at the middle of the contours.

The optimal value for A_{max} is found to be 0.13 as described in Sec. VI by setting $P_L = 1000$ initially. Note that it can be extracted from the error surface and contour plots in Fig. 5.5. There is a minimum at $A_{max} = 0.13$. In this configuration, b_k is equal to a_k (impulse response of the p_{IN} filter is an impulse located at zero), hence the gain is $\Gamma(H)_\infty = 1$, thus it is expected that the model is in non-interactive mode. Before optimizing the parameter, b_k, the resulting interactive glottal flow and its derivative are plotted in Fig. 5.6. It is seen that there is no ripple on the interactive glottal flow that is almost the same as the glottal area waveform in Fig. 5.3. Interactive glottal flow is seem to be an amplitude scaled version of the glottal area, meaning that glottal flow is a linear function of the glottal area.

By optimizing the parameters b_k and q_k as described in the optimization algorithm, all the parameters of the interactive models are obtained. In this case, for ISFM2, $\Gamma(H)_\infty$ is estimated as 64, less than a half of $\frac{\sqrt{P_L}}{\sqrt{2}A_{\max}} \approx 165$ and interaction ripple is expected on the resulting interactive glottal flow. The interactive glottal flow produced by ISFM2 is shown in Fig. 5.7. It is seen from the figure that ISFM2 produces the ripple on the target glottal flow from the given glottal area. However, there is a large mismatch between the closing phases of the target glottal flow and the estimated glottal flow. The mismatch is also clearly observed in Fig. 5.3 and 5.6. The glottal area model is not successful in modeling the closing-phase and it yields a large modeling error. Since the optimization algorithm attempts to reduce this error, despide its ability to produce the interaction ripple, estimated interactive glottal flow is severely affected by this modeling error. Its performance can be increased by avoiding closing phase or return phase errors in the glottal area model.

In this case, as a glottal area input, we use Rosenberg+ model [45]. It models the closing phase by an exponential function. After optimizing its time parameters, the resulting glottal area waveform and its derivative, $a_g[n]$, obtained by Rosenberg model, are shown in Fig. 5.8.

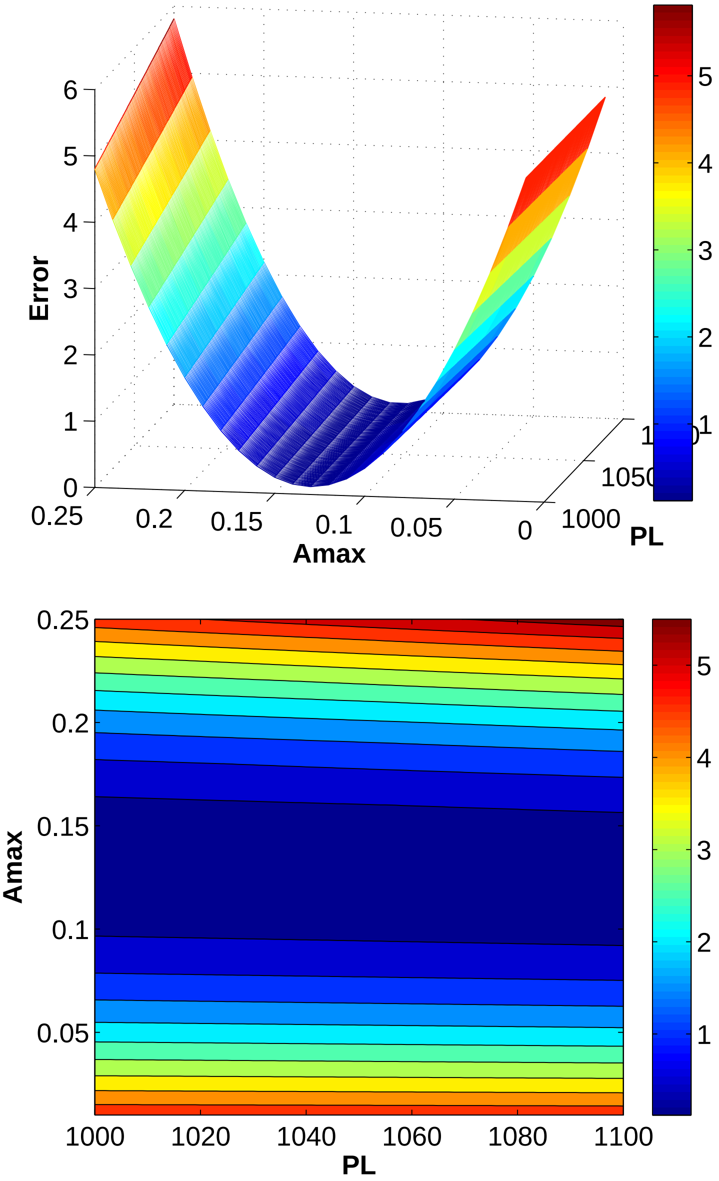


Figure 5.5: (Top) error surface and (bottom) error contours for A_{max} and P_L

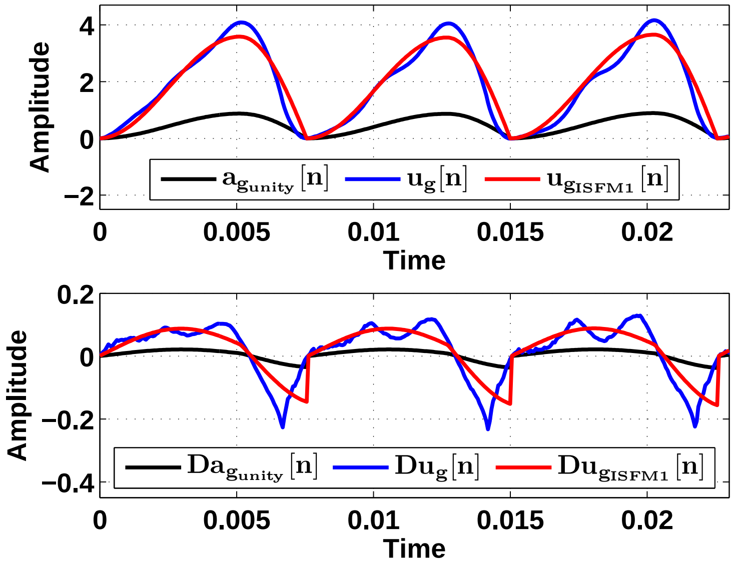


Figure 5.6: (Top) Interactive glottal flow waveform after optimizing A_{max} (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives ($\frac{\sqrt{P_L}}{\sqrt{2}A_{\max}} \approx 165$ and $\Gamma(H)_\infty = 1$)

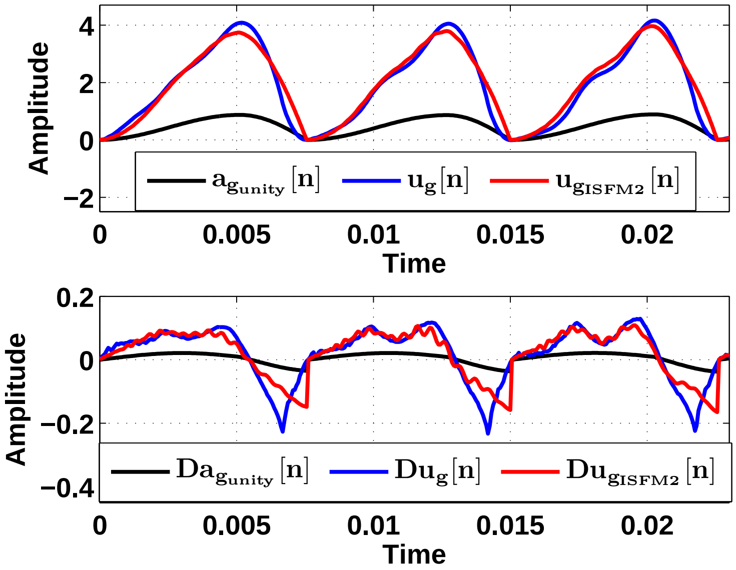


Figure 5.7: (Top) Interactive glottal flow produced by ISFM2 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line)and (bottom) their derivatives ($\Gamma(H)_\infty = 64$)

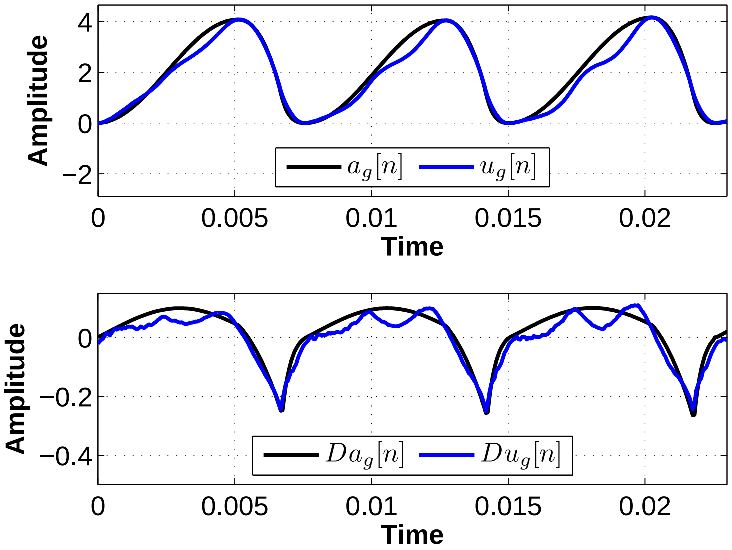


Figure 5.8: (Top) Calculated glottal area waveform using Rosenberg+ model (black line) versus glottal flow estimated by inverse filtering (red line) and (bottom) their derivatives

It is a coarse shape approximation of the glottal flow waveform.

The optimal parameter A_{max} is estimated as 0.136 when $P_L = 1000$. Since, b_k is equal to a_k, the gain is $\Gamma(H)_\infty = 1$, thus it is expected that the model is in non-interactive mode. The synthesized interactive glottal flow and its derivative are plotted in Fig. 5.9. Again no ripple is observed on the interactive glottal flow because the ISFM1 runs in non-interactive mode.

By optimizing the parameters b_k and q_k as described in the optimization algorithm, all the parameters of the interactive models are obtained. In this case, for ISFM1, $\Gamma(H)_\infty$ is estimated as 99, less than $\frac{\sqrt{P_L}}{\sqrt{2}A_{\max}} \approx 164$ and interaction ripple is expected on the resulting interactive glottal flow. The interactive glottal flow produced by ISFM1 is shown in Fig. 5.10. It is seen from the figure that ISFM1 produces the ripple on the target glottal flow from the given glottal area. It is more clearly observable on the glottal flow derivative waveform shown in the lower part of the figure.

Finally, using ISFM1 as an initial model, ISFM2 is estimated. The interactive glottal flow generated by ISFM2 is plotted in Fig. 5.11. ISFM2 refines the glottal flow produced by ISFM1 and decreases the error further. Now $\Gamma(H)_\infty$ is calculated as 89. The results show that both ISFMs can produce the ripple with fine granularity by working in the interactive mode.

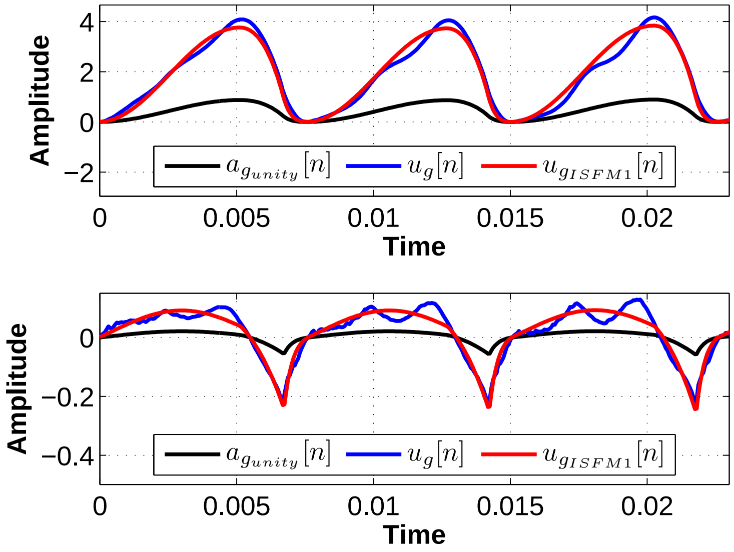


Figure 5.9: (Top) Interactive glottal flow produced by optimal A_{max} (red line) ($\Gamma(H)_\infty = 1$), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives

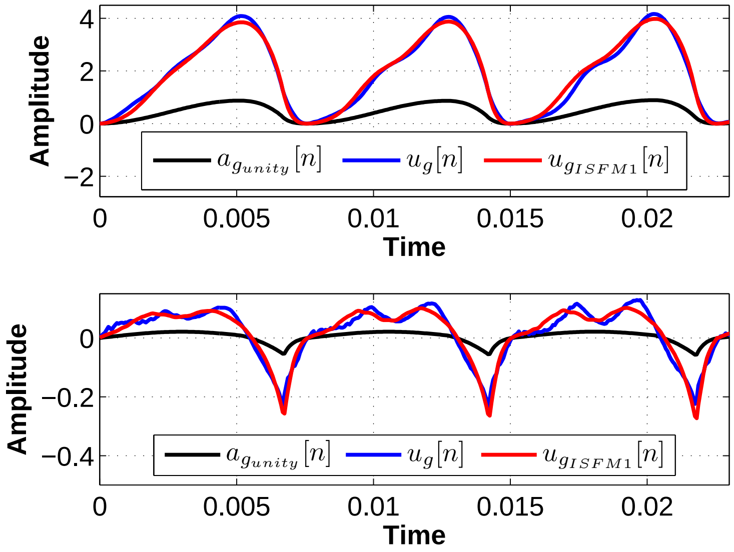


Figure 5.10: (Top) Interactive glottal flow produced by ISFM1 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives ($\Gamma(H)_\infty = 99$)

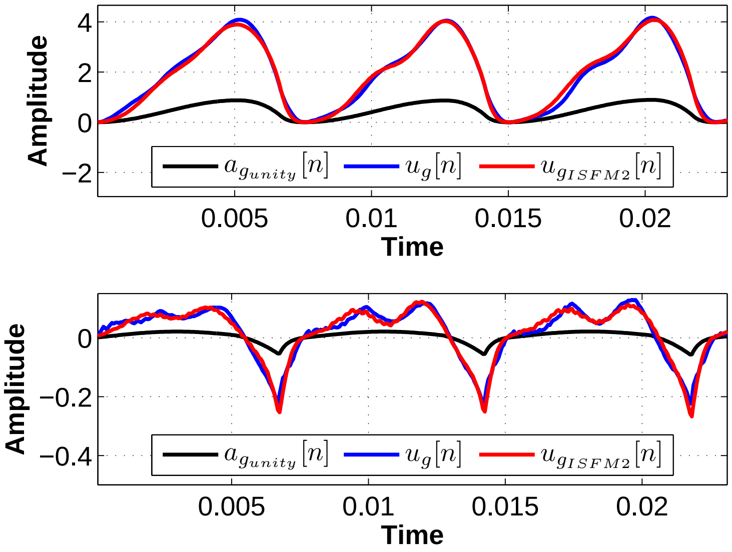


Figure 5.11: (Top) Interactive glottal flow produced by ISFM2 (red line), unity glottal area (black line) and glottal flow estimated by inverse filtering (blue line) and (bottom) their derivatives ($\Gamma(H)_\infty = 89$)

5.3 Summary

In this section, the parameters of the ISFMs are estimated from a short speech segment and interactive glottal flow waveforms synthesized by using the estimated model parameters are presented. During optimization of the model parameters, the error between the interactive glottal flow waveform produced by the proposed models and the glottal flow estimate obtained from closed-phase analysis decreases further. The glottal area model is important for obtaining a good interactive glottal flow. If it is not model the glottal flow waveform shape sufficiently, estimated model may have a large bias. For a good glottal area model, both ISFM1 and ISFM2 produce the ripple on the target glottal waveform accurately. The results show that ISFMs can produce the ripples of the glottal flow due to nonlinear source-filter interaction and the proposed parameter estimation algorithm is powerful enough to estimate a good model producing fine details of the glottal flow. In the next section, the algorithm is applied on a large speech database for a comparison of the ISFMs and the LSFM.

CHAPTER 6

COMPARISON OF LSFM AND ISFM

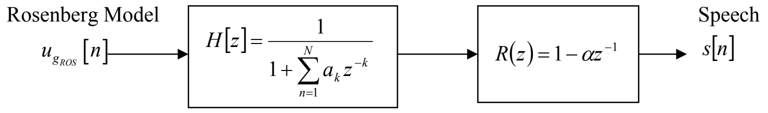
6.1 Introduction

In order to compare the linear and nonlinear models, an experiment is conducted on a large speech database. For the experiment, a speech database containing eight sustained Turkish vowels is formed by recording speech from 12 male and 8 female speakers with electroglottograph signal (EGG). The glottal closure instants (GCI) are marked by first applying the threshold method on EGG signal [51] and then manually corrected. Then, covariance analysis over closed phase of the speech signal were done for estimating the glottal flow [46, 47, 48, 49, 52].

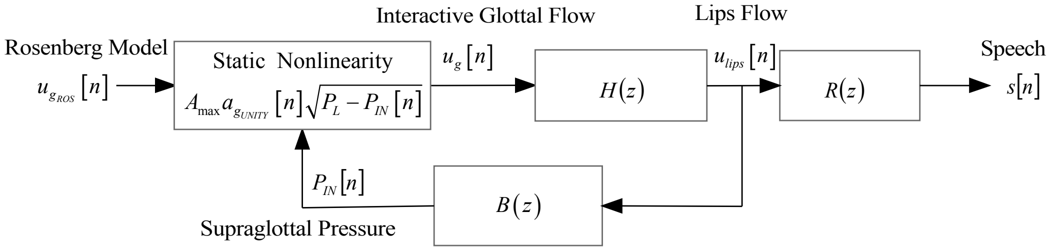
The Rosenberg model parameters are estimated and applied to both linear - nonlinear systems as shown in Fig.6.1. The primary objecive is to minimize reconstruction error for the glottal flow obtained from closed phase covariance analysis.

6.1.1 Comparison of LSFM and ISFMs

To compare the interactive and non-interactive models, we have constructed a special large speech database containing eight sustained Turkish vowels recorded from 12 male and 8 female speakers together with the electroglottograph (EGG) signals in anechoic room. Since inverse filtering requires high quality recording conditions [47], three high quality microphone , two Sennheiser M64 and Studio projects C1, with flatt frequency response down to 30 - 40 Hz, were used to record speech pressure signal with sampling rates 16 kHz and 48 kHz. Simultaneuously corresponding EGG signal were recorded using multichannels audio recording device, UA-1000.



(a) Experimental setup for LSFM



(b) Experimental setup for ISFMs

Figure 6.1: Experimental setup for LSFM & ISFMs using Rosenberg model

After constructing the database, the glottal closure instants (GCI) of each recording in the database are marked by thresholding the EGG signal [51] and then by manual correction. Then, the glottal flow is estimated by covariance analysis over closed phase (CPC) of the speech signal [46, 52, 47, 48, 49]. One of the problems in estimating the glottal flow waveform using the CPC is that when the $F0$ is high, accurate vocal tract filter may not be obtained due to the limited data in the closed phase. Arroabarren and Carlosena [50] demonstrates this type of erroneous glottal flow estimates obtained from high pitch voicing. This problem makes it difficult to use female speech in the experiments. Therefore, we only use the male speech database. For the comparison of interactive and noninteractive systems, we use Rosenberg glottal flow model given as

$$u_R[n] = \begin{cases} \frac{27AVn^2}{4(OQ^2.N_0)} - \frac{27AVn^3}{4(OQ^3.N_0{}^2)} & , \quad 0 \le n \le OQ.N_0 \\ 0 & , \quad OQ.N_0 < n \end{cases} \tag{6.1}$$

where AV, OQ and N_0 are the amplitude, the open quotient and the fundamental period a glottal cycle, respectively. The model has two free parameters, AV and OQ. Estimation of AV is a convex optimization problem, while that of OQ is non-convex [43]. The general procedure is to first obtain AV by solving a linear LS problem, then to perform a grid search, over $0 < OQ < 1$, on the error surface given by Eq.(6.2) [43].

$$E_u(\theta) = \frac{1}{N_0}\sum_{n=1}^{N_0}\left(u_g[n] - u_R[n]\right)^2 \tag{6.2}$$

The Rosenberg model parameters are estimated according to Eq. (6.2) and used in interactive systems as a glottal area waveform, $a_g[n]$. In this way, Rosenberg model is extended to an interactive glottal flow model.

6.1.2 Results & Discussion

The performance of the systems are compared based on the mean square error (MSE) between the reconstructed glottal flow and the reference glottal flow estimated from CPC analysis over each glottal cycle:

$$MSE_{Ug}(model) = \frac{1}{N_0}\sum_{n=1}^{N_0}\left(u_g[n] - \hat{u}_{g\,\mathrm{mod}\,el}[n]\right)^2 \tag{6.3}$$

Table 6.1: The number of glottal flow cycles for each vowel used in the experiments

Vowel	# of Glottal Cycles
/a/	407
/e/	306
/l/	375
/i/	490
/u/	369
/y/	293
/o/	281
/2/	247

The numbers of analyzed glottal cycles are shown in table 6.1. There are totally 2768 glottal cycles included in the experiments.

The averages of MSE and error variances calculated for each glottal cycle in the database are shown in table 6.2 and table 6.3, respectively. The PISMF1 and PISMF2 in table II represent the percentages of the average errors of ISFM1 and ISFM2, respectively, over the Rosenberg model. When the Rosenberg model is applied to ISFM1, the reconstruction error decreases at most by 45 % for /1/ and at least by 19 % for /u/. Similarly, ISFM2 decreases the error of Rosenberg model between 35 - 59 %. Among the interactive models, ISFM2 performs slightly better than ISFM1. The results show that the performances of ISFM1 and ISFM2 are much better than the performance of the Rosenberg model. Incorporating nonlinear interaction into Rosenberg model improves the performance.

Table 6.2: Average of MSE calculated over glottal flow cycles for each vowel. For the ISFM1 and ISFM2 the PISFM1 and PISFM2 represents the ratio over Rosenberg model

Vowel	Rosenberg	ISFM1	PISFM1	ISFM2	PISFM2
/a/	0.0914	0.0572	62.58	0.0531	58.10
/e/	0.4187	0.2571	61.40	0.2207	52.71
/l/	0.2768	0.1538	55.56	0.1162	41.98
/i/	0.9053	0.5214	57.59	0.3777	41.72
/u/	0.0534	0.0434	81.27	0.0348	65.17
/y/	0.2835	0.1636	57.71	0.1239	43.72
/o/	0.0278	0.0182	65.47	0.0164	58.99
/2/	0.1212	0.0831	68.56	0.0692	57.10

The cummulative error distributions of the Rosenberg and proposed models are plotted in Fig. 6.2. The vertical axis is the percentage of the number of the glottal cycles that yield a smaller MSE than that indicated by the horizontal axis. The dash-dot, dashed and solid curves represent the MSE distributions of the Rosenberg, ISFM1 and ISFM2, respectively. About 20 % of the glottal cycles by the Rosenberg model produce an MSE of greater than 0.5 while the glottal cycles with an MSE in the same range is about 7 % for ISFM2. MSE distributions show that ISFM1 and ISFM2 are far better than the non-interactive Rosenberg model.

Figure 6.3 shows the scatter plot of the MSEs of all vowels based on the ratio of first formant ($F1$) and fundamental frequency ($F0$), $F1/F0$. It is seen from the scatter plot that the MSE levels can vary. This level change may occur due to the aspiration noise and small ripples in the closed phase resulting from erroneous inverse filtering for some cases in the study. The closed phase ripples are usually reported as a bad inverse filter configuration flag in inverse

Table 6.3: Average Error Variances calculated over glottal flow estimates for each vowel

Vowel	Rosenberg Model	ISFM1	ISFM2
/a/	0.0733	0.0432	0.0394
/e/	0.2777	0.1832	0.1473
/l/	0.2219	0.1416	0.1049
/i/	0.7822	0.4754	0.3388
/u/	0.0396	0.0339	0.0258
/y/	0.2507	0.1481	0.1108
/o/	0.0233	0.0147	0.0130
/2/	0.0989	0.0696	0.0556

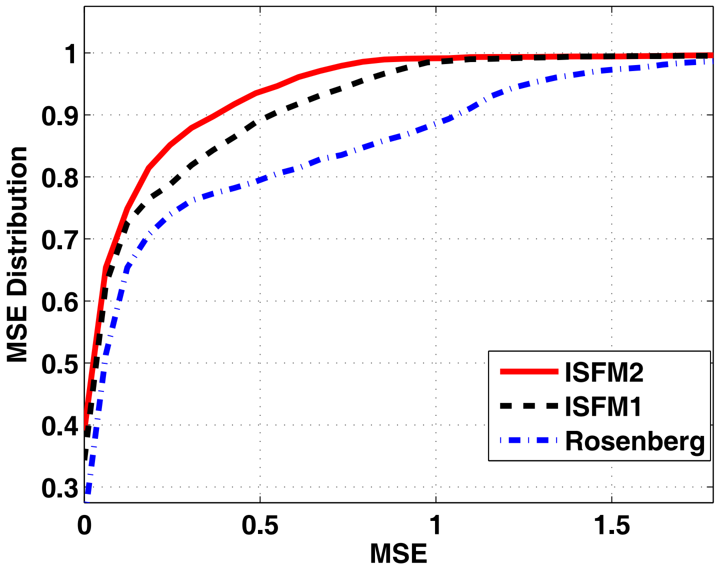


Figure 6.2: (color online) MSE Distributions of the models over all glottal cycles, non-interactive Rosenberg (dash-dot line), ISFM1 (dashed line), ISFM2 (solid line)

filtering literature and they cannot be eliminated easily [46, 49]. Since we do not use an aspiration model for closed phase, high noise in this time interval can significantly affect the total MSE calculated over a glottal cycle. It is observed that all of the vowels in the experiments are enhanced by incorporating the coupling into the non-interactive Rosenberg model. In the introduction section, based on the acoustical simulation, it is stated that increasing the $F0$ towards to $F1$ increases the nonlinear coupling effects on the glottal flow. Hence, when $F1/F0$ is large, the glottal flow can be represented by a coarse shape model more successfully. However, when it is small, the performance of a coarse shape glottal flow model decreases. These relationships can be observed on the MSE scatter plot in Fig. 6.3. The figure shows that when $F1/F0$ is large, MSE values are very close for ISFMs and Rosenberg models. However, when $F1/F0$ decreases, ISFMs reduce the MSE of Rosenberg model and provides significant improvement.

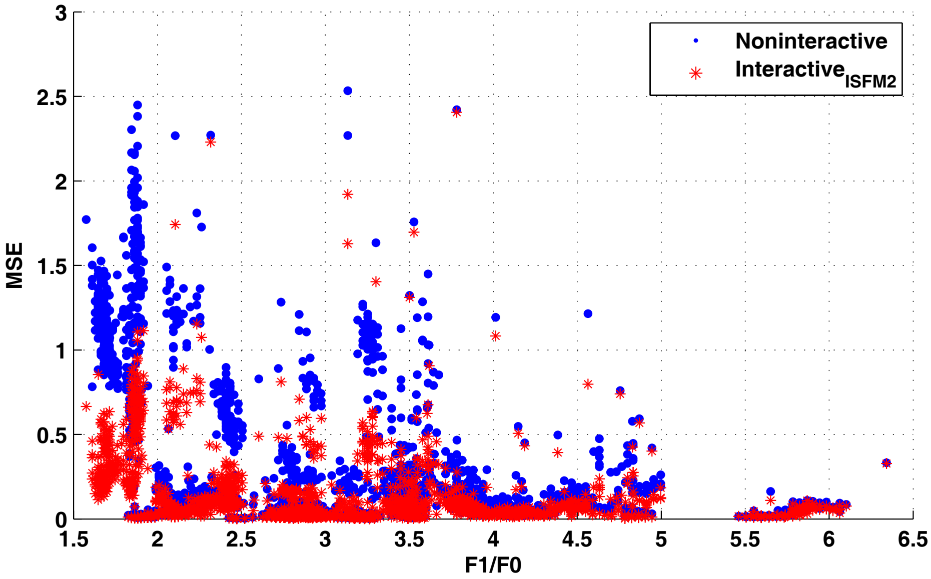
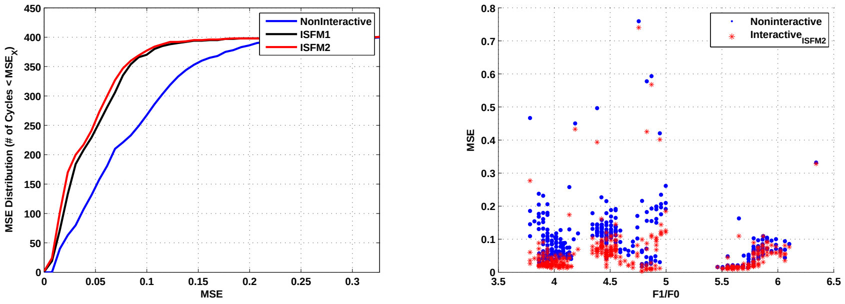


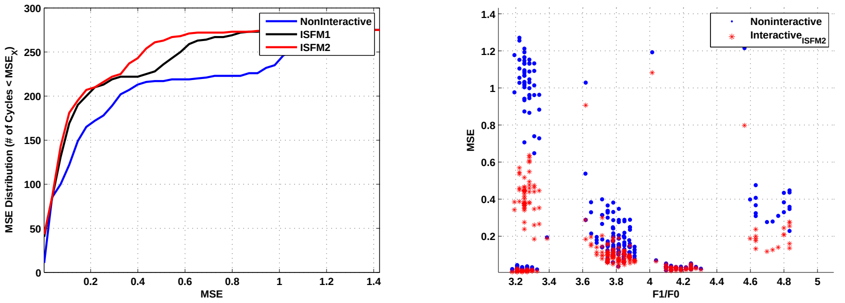
Figure 6.3: (color online) MSE scattering of non-interactive Rosenberg model (dot) and ISFM2 (star) based on F1/F0 ratio



(a)

(b)

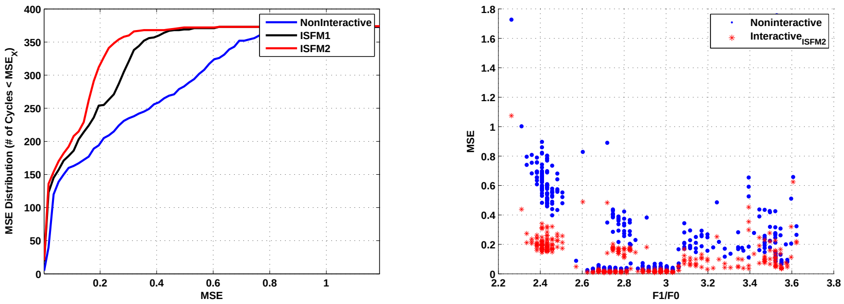
Figure 6.4: (a) MSE Distribution on utterance /a/ (b) Scattered MSEs for LSFM & ISFM2



(a)

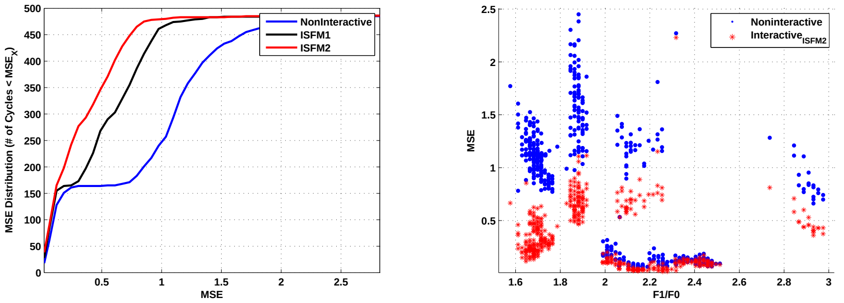
(b)

Figure 6.5: (a) MSE Distribution on utterance /e/ (b) Scattered MSEs for LSFM & ISFM2



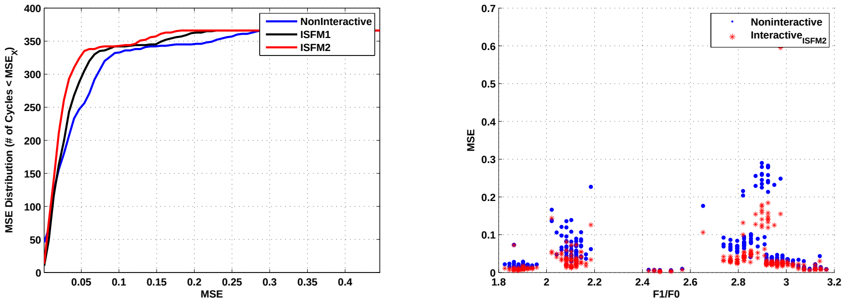
(a) (b)

Figure 6.6: (a) MSE Distribution on utterance /1/ (b) Scattered MSEs for LSFM & ISFM2



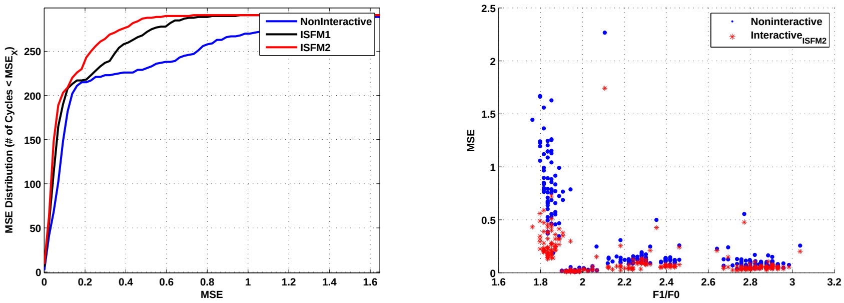
(a) (b)

Figure 6.7: (a) MSE Distribution on utterance /i/ (b) Scattered MSEs for LSFM & ISFM2



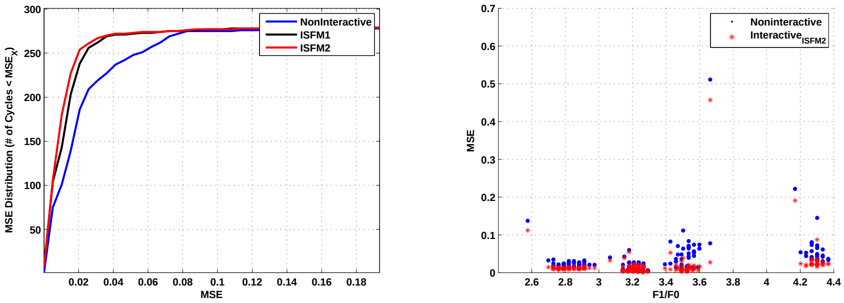
(a) (b)

Figure 6.8: (a) MSE Distribution on utterance /u/ (b) Scattered MSEs for LSFM & ISFM2



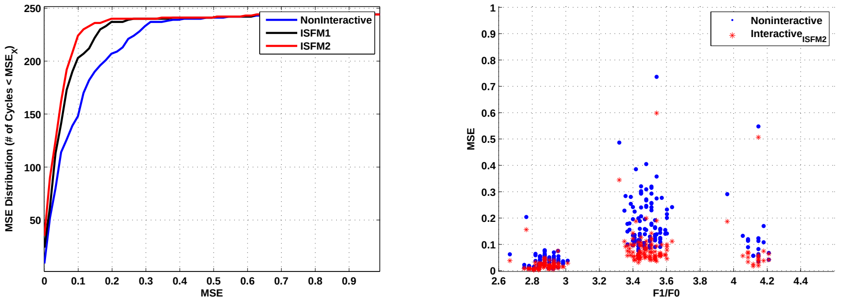
(a) (b)

Figure 6.9: (a) MSE Distribution on utterance /y/ (b) Scattered MSEs for LSFM & ISFM2



(a) (b)

Figure 6.10: (a) MSE Distribution on utterance /o/ (b) Scattered MSEs for LSFM & ISFM2



(a) (b)

Figure 6.11: (a) MSE Distribution on utterance /2/ (b) Scattered MSEs for LSFM & ISFM2

CHAPTER 7

EXTRACTION OF GLOTTAL AREA FROM HIGH SPEED ENDOSCOPIC VIDEO (HSV)

7.1 Introduction

Measuring the source is one of the most difficult tasks in speech processing. Due to the location of the glottis, measuring the glottal flow during speaking is not possible at the moment. However, it is possible to monitor the vocal folds using either stroboscopy or high speed video cameras. Among the two methods, the stroboscopy is much older while high-speed videoendoscopy (HSV) is an increasingly common new technology. A image sequence from HSV recording for one open cycle of vocal folds is shown in Fig. 7.1. It allows the estimation of the minimum glottal area. With this estimation it is possible to compare the linear and nonlinear source - filter models. The estimated glottal area samples construct the excitation signal for both systems, hence a system identification approach is used to evaluate the models. In this section, a new automatic algorithm for extracting the glottal area waveform from HSV images is introduced and its results are presented.

7.1.1 Automatic HSV Processing Algorithm

In order to extract the area between the vocal folds in HSV images, an image processing algorithm has been developed. The HSV data are obtained from IRCAM database[56]. The algorithm has the following four steps;

1. **Mask Calculation**

   In this step vocal fold locations are determined in the whole image using total varia-

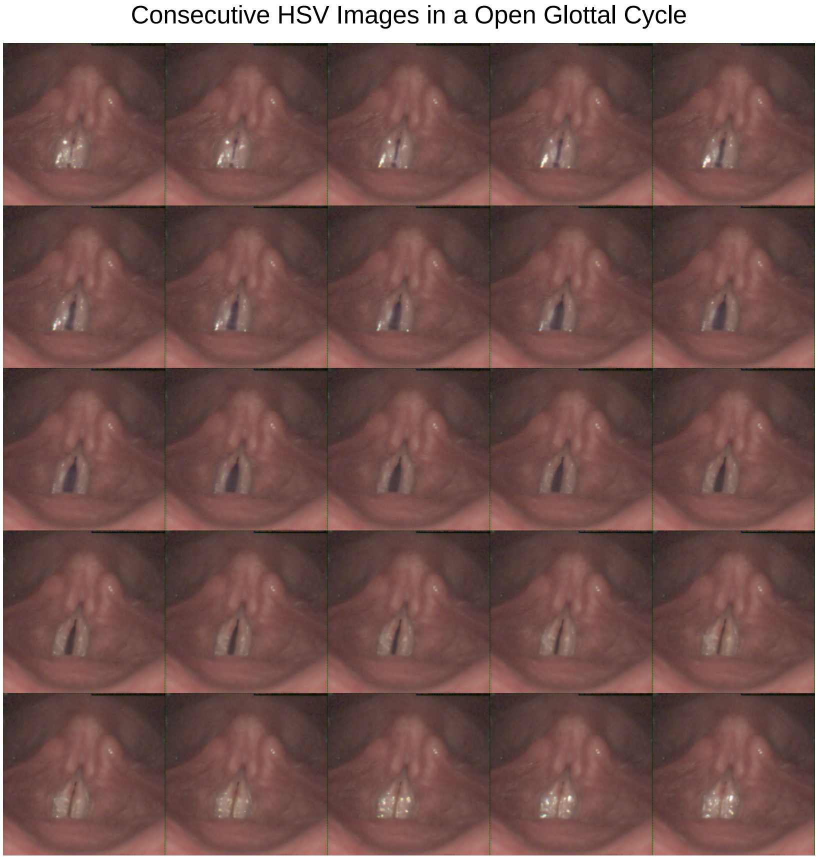


Figure 7.1: Consecutive HSV images corresponding to one glottal cycle

tional norm over the consecutive frames. Total variation norm is defined as

$$TV(x,y) = \sum_{n=1}^{N-1} |I(x,y,n+1) - I(x,y,n)| \forall \text{x,y} \in I \tag{7.1}$$

where x,y and n denotes horizontal axis,vertical axis and frame number respectively. $I(x,y,n)$ is the n^{th} gray level image frame and N is the number of frames. The measure produces large values for moving objects in the video while small for the background. Figure 7.2 illustrates the TV measure extracted over 300 consecutive HSV frames. Note that glottis has the largest TV measure while small noisy measures are seen on the background.

Using a combination of a threshold, morphologic operation and edge detection operation, the mask is located on vocal folds boundaries. The TV image is converted to

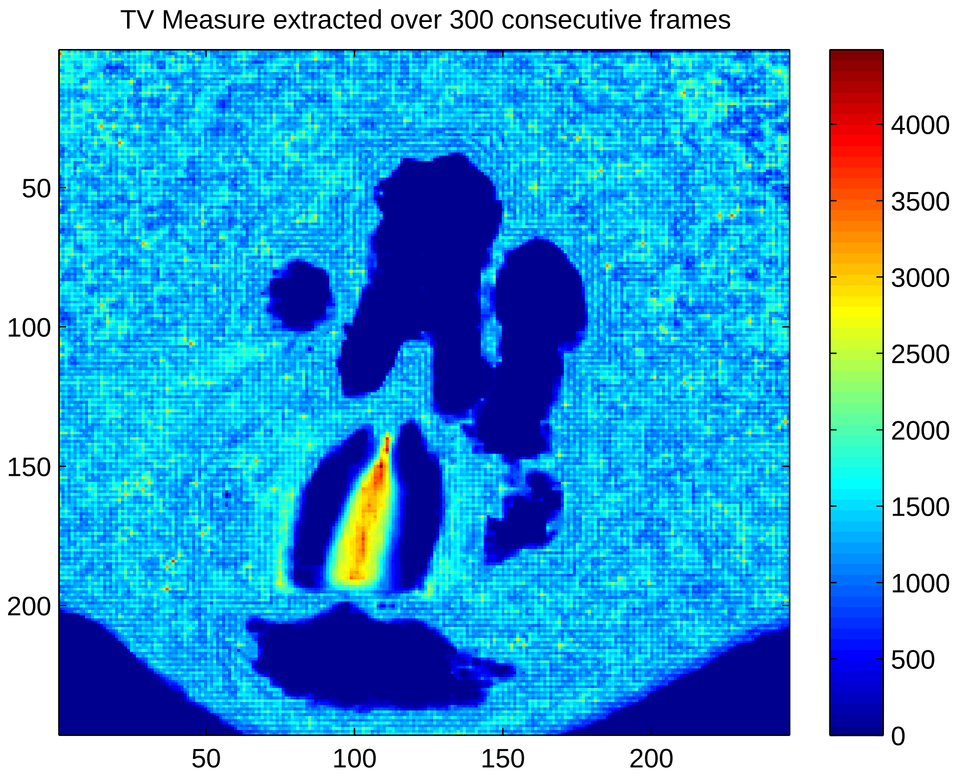


Figure 7.2: TV Measure extracted from 300 consecutive frames

binary image by a threshold with size $T = \mu_{TV} + \sigma_{TV}$, where noisy binary image is passed through morphologic opening operation in order to reduce the effect of small connected components in the image. The edge detection operation is applied on the final image and the boundary corresponding to largest TV norm region is obtained as shown in figure 7.3.

The final mask is constructed by using the maximum and the minimum location values of the boundary as follows

$$Mask\,(x, y) = \begin{cases} 1 & x_{\min} - \delta \leq x \leq x_{\max} + \delta, y_{\min} - \delta \leq y \leq y_{\max} + \delta \\ 0 & e.w \end{cases}$$

where $x_{\min}$ and $x_{\max}$ are the minimum and maximum horizontal lines passing through the boundary while $y_{\min}$ and $y_{\max}$ are vertical extremum points. The parameter δ is a margin for compensating the boundary errors, in this experiment, $\delta = 20$ is used. The masked image is calculated by product of original image and the mask

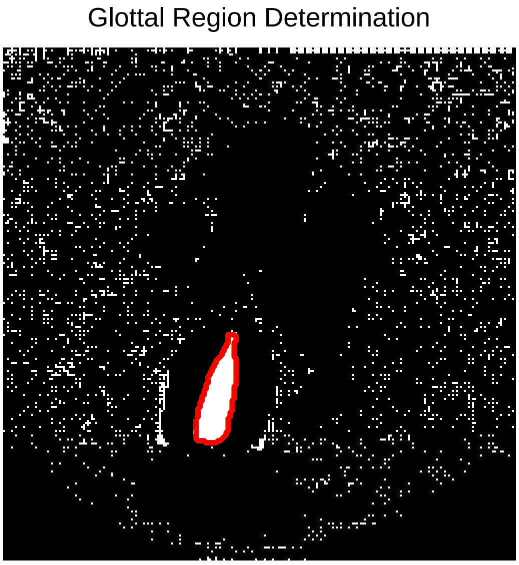


Figure 7.3: Glottal Region Determination

$$I_{MASKED}(x,y,n) = I(x,y,n).Mask(x,y) \tag{7.2}$$

The masked image is shown in the upper part of figure 7.3. It clearly locates the vocal folds in the corresponding image.

2. **GMM Based Histogram Modeling**

   After masking the image, the next process is to determine the glottal opening. In order to segment the open glottis from the background, the histogram of the masked image can be used with a simple thresholding approach. However the masked image contains intensity information of the glottis only the time at which the photo were taken. A more robust way is to compute the intensity histograms over horizontal or vertical slices. If we get a horizontal slice, the resulting image is called kymogram. However, instead of horizontal, we work with vertical slices due to their compatibility of better representing the glottis intensity. A vertical slice of the masked video frames arranged in time makes a new image as follows

$$VSI(n,y) = I_{MASKED}(x_{center},y,n) \tag{7.3}$$

where x_{center} is the horizontal centroid point of the mask. The masked image shown in Fig. 7.4 corresponds to the intensity changes in the center vertical line of the glottis during phonation in time. During the glottal opening the intensity of the vertical slice of the center of the glottis is very small, hence it is dark, while it has high intensity during the closed phase. Therefore, the histogram of the vertical slice image can provide better intensity distribution than the intensity of the original unmasked image in distinguishing the glottis and the vocal fold edges.

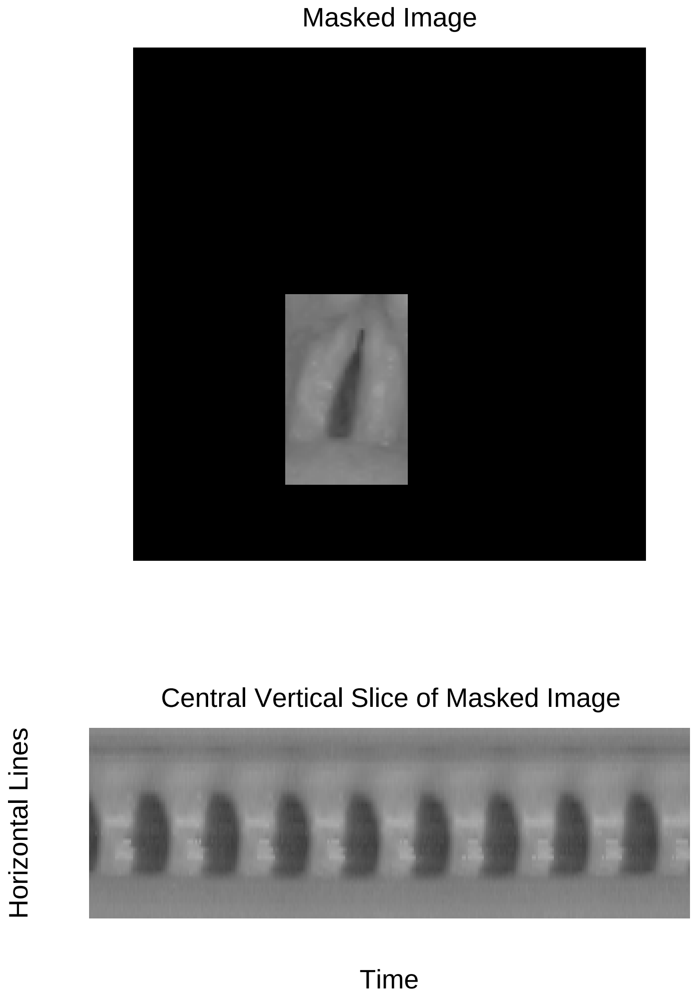


Figure 7.4: Glottal region determinated by masking(upper figure) and corresponding vertical slice images(lower figure)

The histogram of the VSI image is shown in Fig. 7.5. It can easily be considered as the sum of two weighted Gaussian densities. Hence, it can be modeled by a mixture of two Gaussians (GMM) representing glottal opening (GO) and background(B).

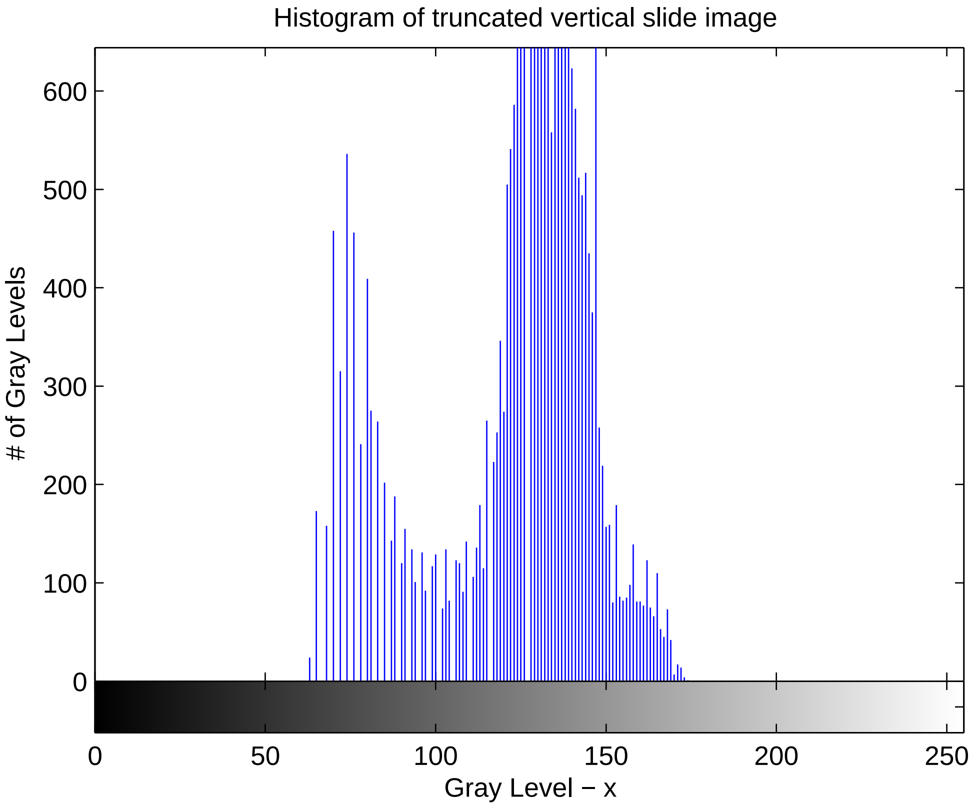


Figure 7.5: Histogram obtained from vertical slice image (VSI)

The conditional distribution of glottis, $P(I|GO)$, background $P(I|B)$ and the VSI image intensities,$P(I)$, are written as

$$P\left(I|GO\right) = N\left(I|\mu_1, \sigma_1\right) = \frac{1}{\sqrt{2\pi\sigma_1{}^2}} e^{-\frac{1}{2}\left(\frac{I-\mu_1}{\sigma_1}\right)^2} \tag{7.4}$$

$$P\left(I|B\right) = N\left(I|\mu_2, \sigma_2\right) = \frac{1}{\sqrt{2\pi\sigma_2{}^2}} e^{-\frac{1}{2}\left(\frac{I-\mu_2}{\sigma_2}\right)^2} \tag{7.5}$$

$$P\left(I\right) = \sum_{i=1}^{2} w_i N\left(I|\mu_i, \sigma_i\right) \tag{7.6}$$

where w_i's are Gaussian weights representing $P(GO)$ and $P(B)$ respectively. The parameters of the GMM, μ_i, σ_i and w_i, are estimated by EM (Expectation and Maximization) algorithm. The posterior probabilities, $P(GO|I)$ and $P(B|I)$, are written based on the Total probability theory as follows

$$P(i|I) = \frac{w_i N(I|\mu_i, \sigma_i)}{\sum\limits_{k=1}^{2} w_k N(I|\mu_k, \sigma_k)} \tag{7.7}$$

where $i = 1, 2$ denotes glottal opening (GO) and background(B) respectively. Estimated GMM for the VSI image is plotted in figure 7.6. Note that the GMM is fitted on the intensity histogram well.

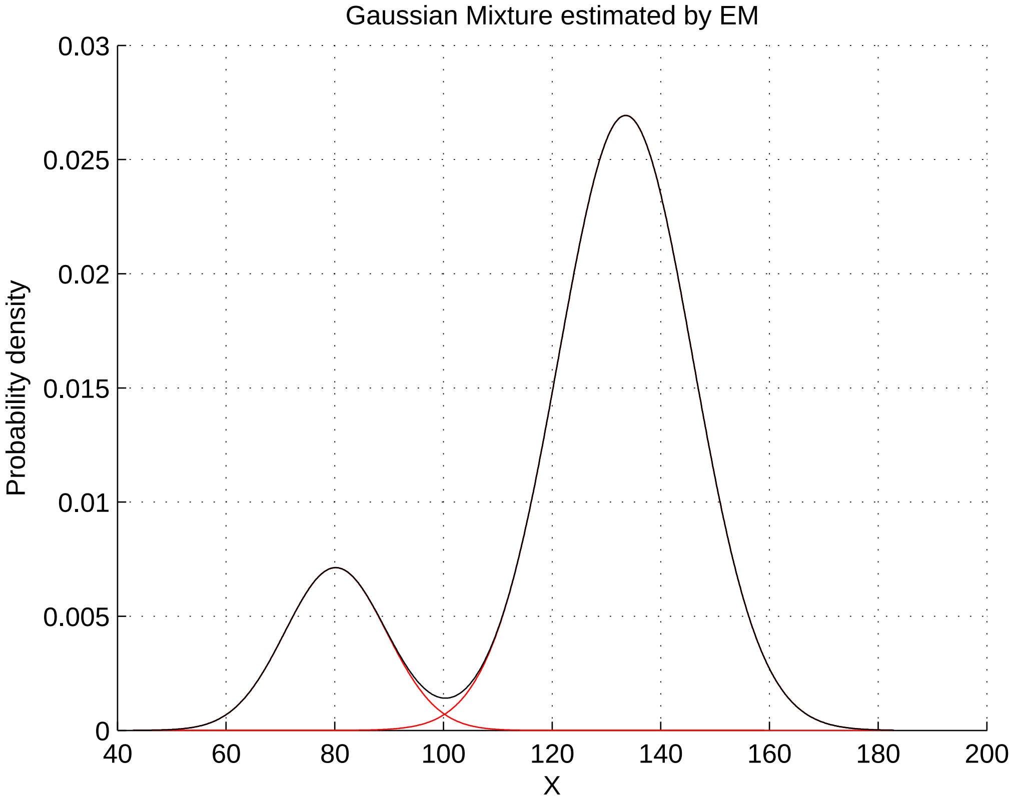


Figure 7.6: Estimated GMM model for the VSI

3. **Segmentation**

The segmentation step is based on the Bayes decision theory. The classifier for intensity of each pixel in the masked image frame is written as

$$g(I) = \underset{i}{\operatorname{argmax}} \left(P(i|I)\right) = \begin{cases} 1 & \multirow{2}{*}{for} \quad P(GO|I) > P(B|I) \\ 0 & \quad else \end{cases} \tag{7.8}$$

For the two scalar valued classes, the boundary is a scalar corresponding to the intersecting point of the Gaussian densities. This is a threshold, T_{cls}, for segmenting the

HSV images. The segmented image is written in terms of decision boundary as

$$I_S\left(x, y, n\right) = g\left(I_{Masked}\left(x, y, n\right)\right) \tag{7.9}$$

It can be seen in Fig. 7.6 that T_{cls} is around the intensity 100 for the corresponding VSI. The segmented VSI image based on the GMM thresholding is shown in Fig. 7.7. The contours of the segmented image are plotted below the same figure. The results indicate almost perfect segmentation over the glottal region.

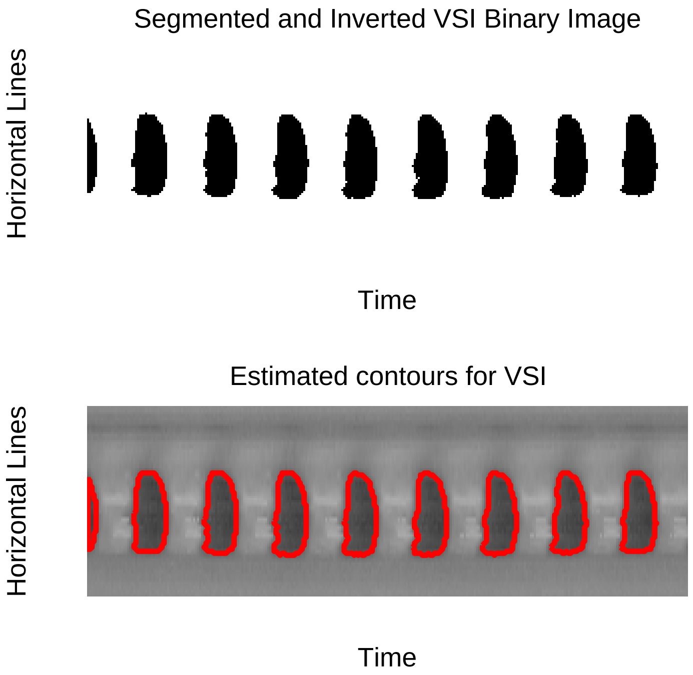

Segmented and Inverted VSI Binary Image
Horizontal Lines
Time
Estimated contours for VSI
Horizontal Lines
Time

Figure 7.7: Estimated binary images for VSI

The segmented binary image $I_S\left(x, y, n\right)$ from HSV images in Fig. 7.1 are shown in Fig. 7.8. The image shows 25 sequentially concatenated images from left to right with respect to time or frame number, n. It is seen that the resulting glottal openings are well defined and clearly extracted from the background. Furthermore, note that the rate of growth of white region is smaller than its rate of shrink. This behaviour is due to the fact that the vocal folds generally close faster than they open.

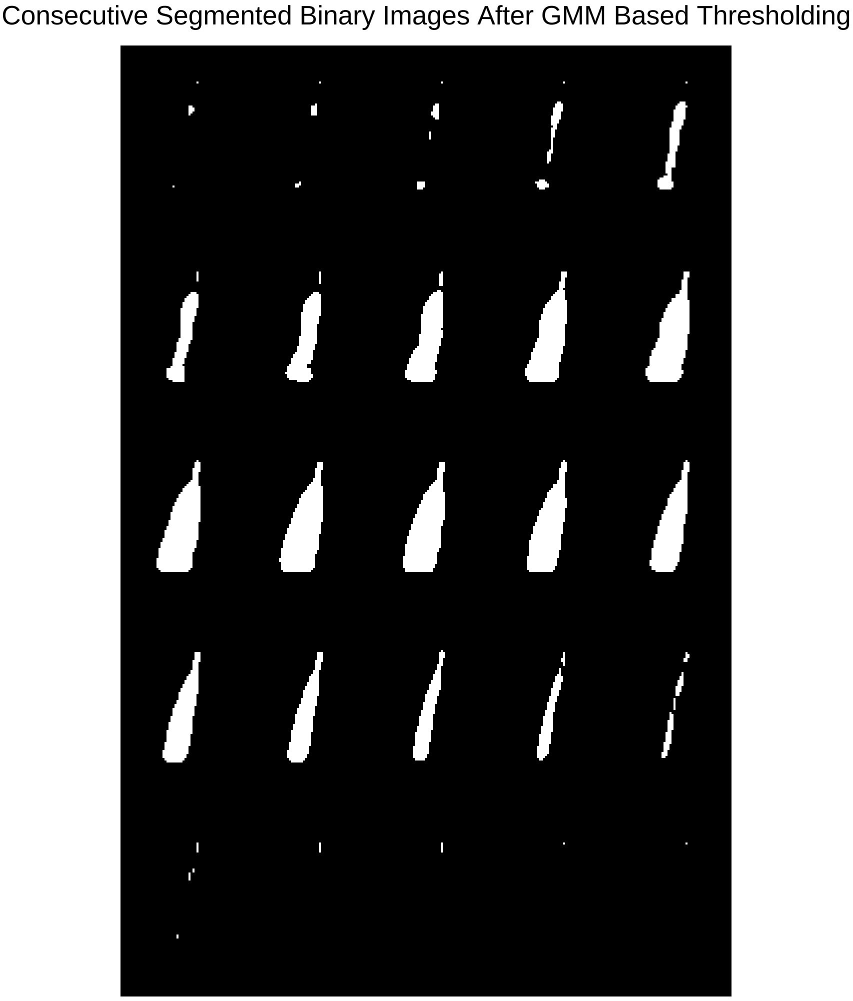

Consecutive Segmented Binary Images After GMM Based Thresholding

Figure 7.8: Estimated consecutive binary images over open phase

4. **Calculation of Glottal Area**

The glottal area is the area between the vocal folds edges. It is proportional to the number of white pixels in the segmented masked image and calculated as

$$a_g[n] = \sum_{x \in Mask} \sum_{y \in Mask} I_S(x, y, n) \tag{7.10}$$

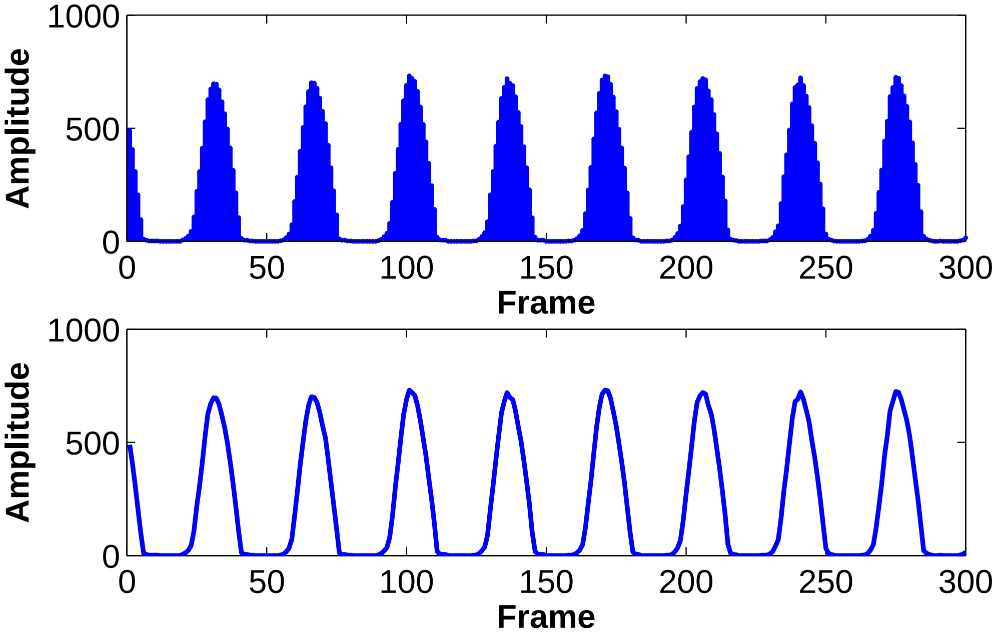


Figure 7.9: Estimated glottal area signal

The extracted glottal area samples from HSV image series are shown in figure 7.9. It is a -stationary waveform corresponding to sustained speech signal.

CHAPTER 8

HSV BASED
ANALYSIS AND COMPARISON OF LSFM AND ISFMs

8.1 Introduction

In this section, the linear and nonlinear source - filter models are compared with the help of glottal area waveform extracted from HSV images. LSFM and ISFMs are excited with the glottal area waveform and the resulting speech signals are compared.

8.2 Experiments

As a source of HSV images, we use data from IRCAM speech, HSV and EGG signal database [56]. In the database the frame rate of HSV is 4kHz, which is the sampling rate for glottal area, at a resolution of 256x256 pixels while the speech and EGG signals were recorded at a sampling frequency 44150 Hz. In order to process the signals it is necessary to synchronize them, hence the sampling rate of glottal area, speech and EGG signals are converted to 8kHz, which is a reasonable frequency for speech analysis. The resampled glottal area, electroglottograph and corresponding speech signals are shown in Fig. 8.1. It is seen from the figure that the voice quality of the speaker is modal with approximately 115 Hz fundamental frequency ($F0$).

The spectrum of the glottal area is plotted in Fig. 8.2(a). The HSV images are obtained at a sampling frequency of 4kHz, hence the frequency spectrum is displayed as 0 - 2kHz band. Since the glottal area is a quasi-periodic signal, its spectrum consists of harmonics of fundamental frequency of vocal fold oscillations or pitch which is about 115 Hz. It is seen in

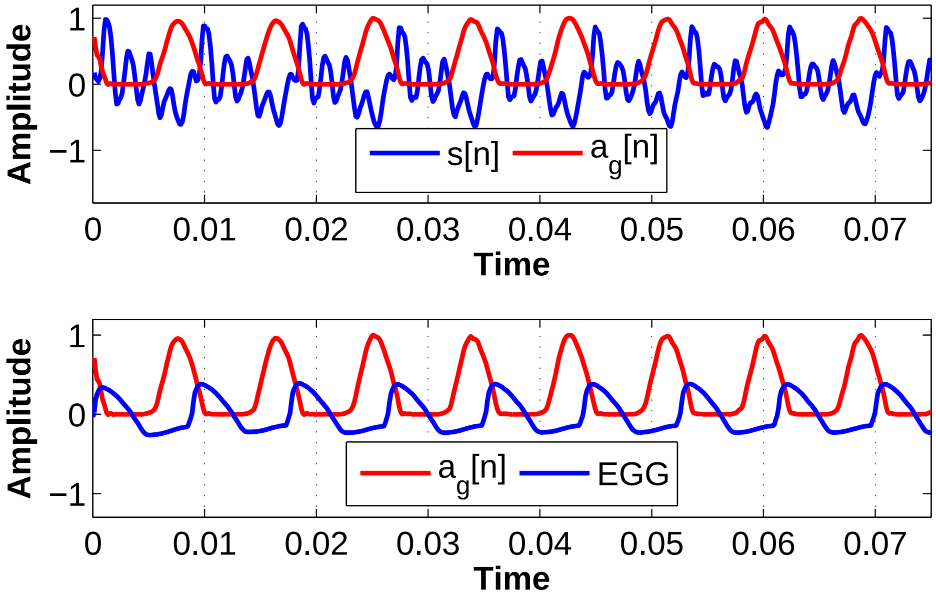
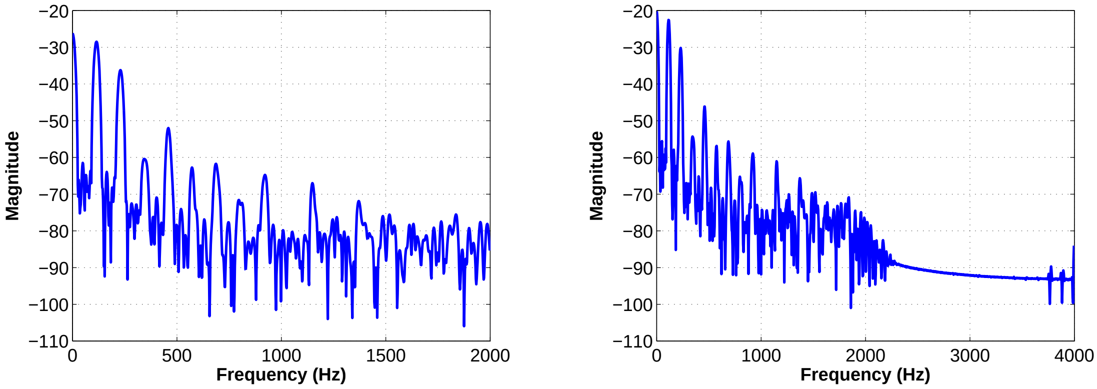


Figure 8.1: Speech vs Glottal area (Upper) and EGG vs Glottal area (Lower)

the figure that the weigth of the harmonics are not uniform and some of the higher harmonics are missing due to their low energy. However, harmonics below the 1kHz pronounced well. The upsampled glottal area spectrum is shown in figure 8.2(b). As it is expected, the spectrum is almost the same as the original area for frequencies below 2kHz and there is no energy above 2 kHz due to antialiasing filter used by the resampling process.

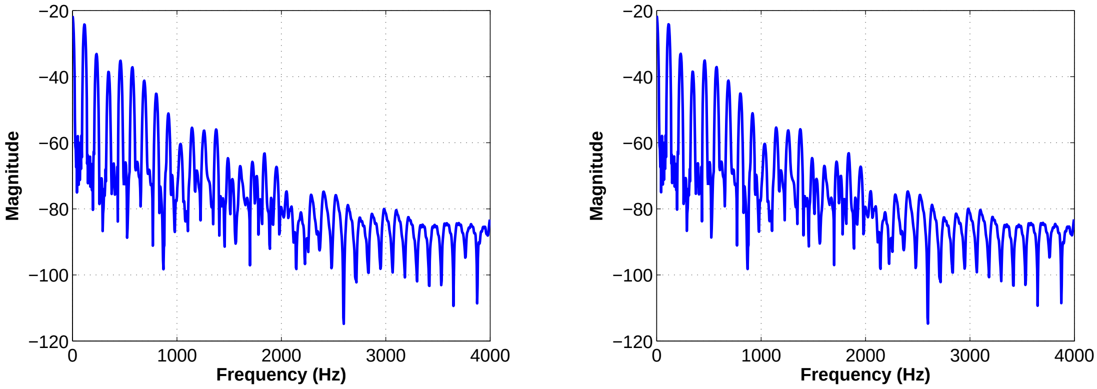
The spectrum of the interactive glottal flow for ISFM1 model is shown in Fig. 8.2(c). It is seen that the harmonic structure of the glottal area is preserved. In fact, it is enhanced by generating new harmonics of the fundamental frequency. When the harmonic structure is investigated closely, it can be seen that the 9^{th} and 11^{th} harmonics just above 1kHz missing in the glottal area is generated by the ISFM1 model. This is a favorable property of the nonlinear source filter model. It generates new source harmonics in order to produce speech signal as in the physical system models or articulatory speech synthesizers. This result is expected since the nonlinear models are based on the physical system approximation of speech production.

The spectrum of interactive glottal flow for ISFM2 model is shown in figure 8.2(d). It is almost the same with the ISFM1, there are no visible differences in the harmonic structure. Theoretically, the difference between the ISFM1 and ISFM2 is the addition of subglottal



(a) Glottal area spectrum

(b) Upsampled glottal area spectrum



(c) Interactive glottal flow generated by NLSF1

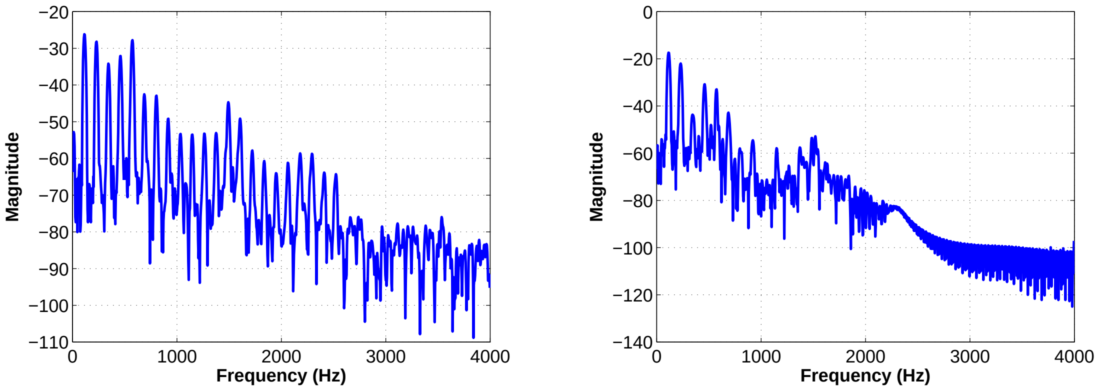
(d) Interactive glottal flow generated by NLSF2

Figure 8.2: Interactive glottal flow by generated NLSF Models

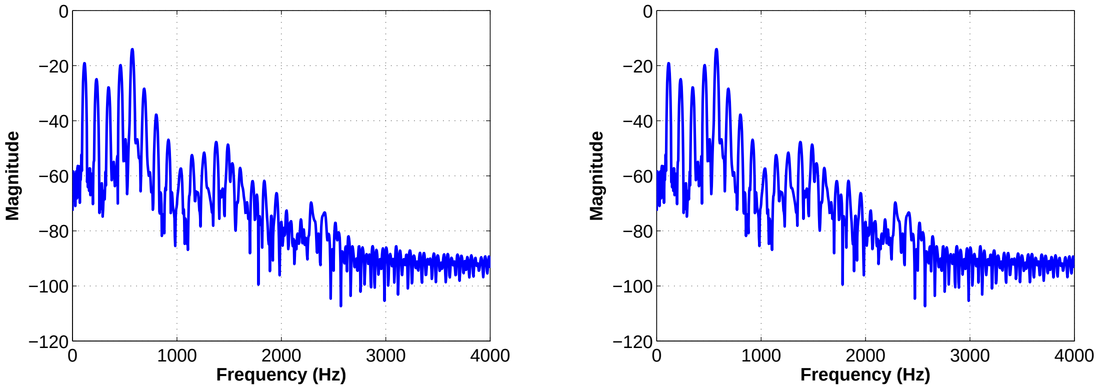
impedance, the result implies that the subglottal impedance has no or very little effect on the glottal flow spectrum for this specific case.

The spectrum of the original and synthesized speech signals are shown in figure 8.3. The original speech spectrum in part (a) of the figure contains all harmonics of the fundamental below 2.5kHz, above which the noise spectrum is dominant. The synthesized speech with the linear source - filter model in part (b) of the figure does not have all harmonics due to the missing harmonics in the glottal area source illustrated previous figure, hence it is a poor approximation of the original speech.

The spectrum of synthesized speech with nonlinear source - filter model, ISFM1, is shown in Fig. 8.3(c). The ISFM1 speech spectrum is a much better approximation of the original speech. Its spectrum has all harmonics of the original speech spectrum in part (a) except the two components located just above the frequency 2kHz around fourth formant (F4). Another



(a) Speech Spectrum (b) Spectrum of synthesized Speech by LSFM



(c) Spectrum of synthesized Speech by ISFM1 (d) Spectrum of synthesized Speech by ISFM2

Figure 8.3: Original and synthesized speech spectrums

difference is that the energy of the harmonics between the frequencies 1.5-2 kHz are small than the original case. This can be solved by increasing vocal tract gain for those frequencies. The spectrum for ISFM2 is plotted in Fig. 8.3(d). As it is for the glottal flow, the results are almost the same and there are no visible differences between the two spectra.

Finally, the synthesized speech waveforms are shown in Fig. 8.4. The LSFM speech shown in the upper is a very poor approximation of the original speech due to missing harmonic components in the source. Since the vocal tract formants are not excited enough, the resulting waveform doesn't show the all formants except the F1 and F3. The speech synthesized by ISFM1 shown in the lower part of the figure is much better approximation of the original speech. Since the ISFM1 glottal flow has almost all the harmonics in the speech, it excites the vocal tract formants much better. Therefore, resulting speech signal exhibits formants much better. Since the spectrum of ISFM2 speech is almost the same as ISFM1, its result is not

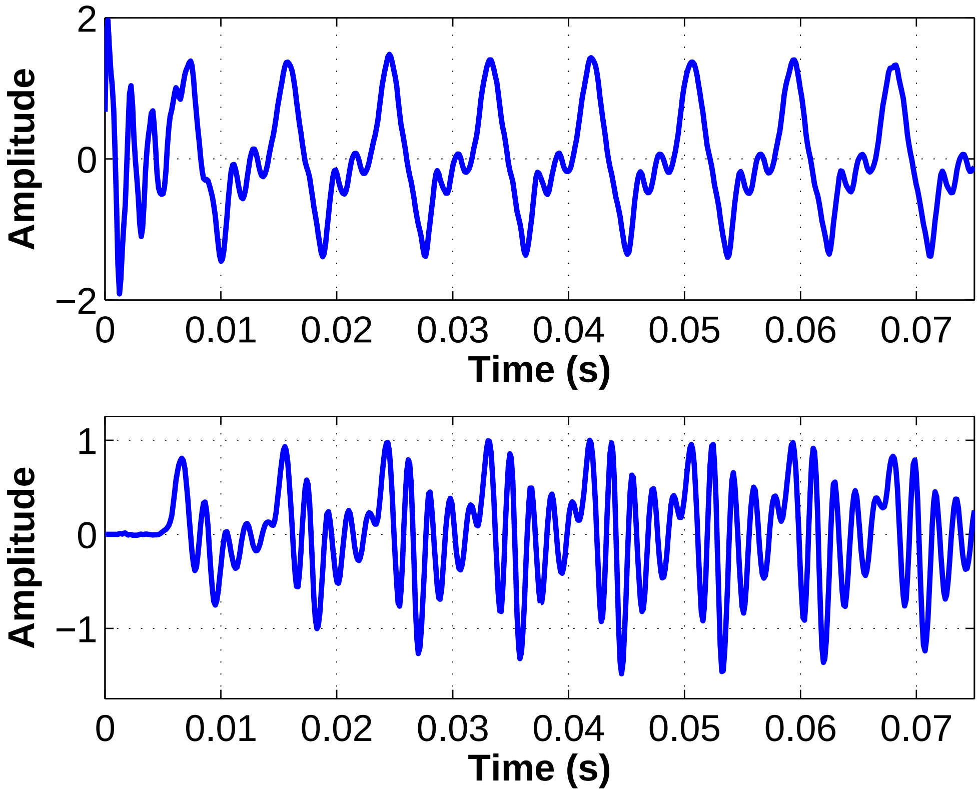


Figure 8.4: Synthesized speech by LSF (Top) and ISFM1 (Bottom)

presented.

We have compared the linear and nonlinear source filter models based on the HSV database recordings. The results show that nonlinearity in the speech production is well modeled by the quasi - steady Bernoulli flow assumption in the glottis.

8.3 Summary

In this study, we have compared the nonlinear source-filter models with the classic linear source filter model based on the high speed endoscopic video database (HSV) instead of using glottal flow estimated by inverse filtering. First, we have presented a novel image processing algorithm for extracting glottal area from HSV images. Then, the extracted area is used as an excitation source measure for the systems. The results show that nonlinear systems are superiour than the linear model. They exhibit the same behaviour observed in the physical simulators. ISMFs enhance the source harmonics in order to produce more realistic speech sounds than the linear source - filter model. As a result, it is shown that the nonlinear source filter models successfully incorporate the source - filter coupling effects seen in the physical simulators.

CHAPTER 9

CONCLUSION

In this work we present a framework for modeling nonlinear source filter coupling in voiced speech sounds. Based on this framework, two nonlinear interactive source filter models, ISFM1 and ISFM2, are developed and their relation to the linear source filter model, LSFM, and each other are exposed. It is shown that LSFM is an approximation of the ISFMs under certain conditions. The major difference between the ISFMs and LSFM is that, in the LSFM the source and the filter are assumed to be independent and do not affect each other however in the ISFMs they nonlinearly interact according to Bernoulli equation.

A parameter estimation algorithm is proposed for estimating the parameters of ISFMs and tested on speech signals. Experimental results show that proposed nonlinear interactive models exhibit the source-filter interaction effects seen in speech production. Since ISFMs are based on physical modeling, they have the capability of producing the fine ripples on the glottal flow waveform observed in the simulation of speech production. They accurately generate fine ripples on the glottal flow obtained by inverse filtering by the help of the nonlinear interaction between the glottal flow and vocal tract filter. This feature is incorporated into the Rosenberg model and used to synthesize glottal flow waveforms obtained from a large speech database. The results indicate that the interactive Rosenberg model is always better than its non-interactive counterpart.

Proposed interactive model takes the glottal area waveform as its input. A model based glottal area waveform is synthesized using the glottal flow waveform obtained by inverse filtering assuming that the glottal flow and the glottal area waveforms achieve their peaks synchronously. In this context, the optimization of the skewing quotient of the glottal area can be studied further. In generating the glottal area waveform, Rosenberg model has been used mainly due to its suitability in the optimization of its parameters. The generation of the glottal area waveform can be a branch of future studies. Alternatively, the use of a dynamic system model to generate the glottal flow waveform can be considered. Another way is to obtain the glottal area waveform by using High-Speed Endoscopic Video (HSV) of vocal folds. It is the subject of the second part of this thesis in which the glottal area is extracted from HSV images.

In the second part LSFM and ISFMs are compared by using not only the speech signal but also HSV images of the vocal folds in a system identification fashion. A new robust HSV processing algorithm is developed. It is applied on HSV images in order to extract glottal area waveform. The experimental studies using HSV and speech signal indicate that speech signal can contain some more harmonics of the fundamental frequency of the glottal area than the harmonics presented in the glottal area signal. This implies that speech production is nonlinear if the source is assumed to be a linear function of the glottal area. In the physical model of speech production, the vocal tract enhances the glottal flow harmonics. ISFMs show the same behavior and produce the source harmonics necessary for produce more realistic speech sounds.

One branch of the future studies is that ISFMs can be used in an articulatory speech synthesizer to couple the glottal flow, and sub- and supra-glottal tracts. The transfer function of the combined sub- and supra-glottal tracts has to be calculated from the time-varying areas. They can be parameterized by estimating the p_{IN} filter. Hence, instead of areas, the filter coefficients, b_k and q_k, can be stored in a codebook to represent the transfer function of the system.

Another branch of the future studies can be to investigate the inter- and intra-speaker variability of the parameters of the ISFMs since they form a ground for the definition of new features for speaker identification and speech recognition applications.

9.1 Contributions

The major contributions of this thesis can be summarized as follows

1. Nonlinear interactive source-filter models for voiced speech are developed. These models are extensions of the linear source-filter model with a capability of produc-

ing source-filter interaction effects. ISFMs can be used in speech signal processing applications as an alternative to the linear source-filter model.

2. A framework for the incorporation of the interaction into classical source models is presented. In this context, the Rosenberg model is extended to an interactive model having capability of fine details of glottal flow waveform.

3. A parameter estimation algorithm for ISFMs is developed. The algorithm can produce always stable and better ISFMs compared to the LSFM.

4. A new HSV analysis algorithm is developed to extract the glottal area from HSV images. It is necessary to obtain the glottal area waveform for the evaluation of speech production models. By the help of the developed algorithm, it is possible to extract glottal area form HSV images accurately.

5. A system identification platform for HSV based analysis of speech is constructed. This platform can be used for the investigation of speech models.

6. Multi-channel Turkish speech & EGG databases are constructed. The databases contain wide variety of Turkish words spoken by both male and female speakers, therefore it can be used for any speech or speaker based research on Turkish.

REFERENCES

[1] Fant, G.,The Acoustic Theory of Speech Production Moulton, The Hague. 1960

[2] Flanagan, Speech analysis synthesis and perception, Springer-Verlag, 1965

[3] Ishizaka, K., and Flanagan, J. L., Synthesis of voiced source sounds from a two-mass model of the vocal cords, Bell Syst. Tech. J. 51, 1233- 1268, 1972

[4] P.M. Morse and K.U. Ingard, Theoretical Acoustics, McGraw-Hill, New York, 1968

[5] Anderson, J.D.: ”Fundamentals of Aerodynamics”, Mcgraw Hill Series in Aeronautical and Aerospace Engineering,2010.

[6] Van den Berg, J., Myoelastic-aerodynamic theory of voice production, Journal of Speech and Hearing Research 3(1): 227-244, 1958

[7] Titze, I. R.,Parameterization of the glottal area, glottal flow, and vocal fold contact area,J. Acoust.Soc. Am. 75, 570-580,1984

[8] Titze, I. R.,Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model,J. Acoust. Soc. Am. 111, 367-376. 35,2002

[9] Titze, I. R., A theoretical study of F0-F1 interaction with application to resonant speaking and singing voice, J. Voice 18, 292-298,2004

[10] Titze, I. R., Theoretical analysis of maximum flow declination rate versus maximum area declination rate in phonation, J. Speech Lang.Hear. Res. 49, 439-447,2006

[11] Titze, I. R ,Nonlinear source-filter coupling in phonation: Theory,J. Acoust. Soc. Am. 123 5, May 2008

[12] Titze, I. R., Riede, T., and Popolo, P., Nonlinear source-filter coupling in phonation: Vocal exercises, J. Acoust. Soc. Am. 123, 1902-1915,2008

[13] Titze, I. R., and Story, B. H.,Acoustic interactions of the voice source with the lower vocal tract, J.Acoust. Soc. Am. 101, 2234-2243,1997

[14] Titze, I. R., The Myoelatic Aerodynamic Theory of Phonation, Iowa City:National Center for Voice and Speech, 2006.

[15] Titze, I. R.,The Human Instrument, Scientific American, Jan. 2008.

[16] Jong Beom Park,Luc Mongeau,Instantaneous orifice discharge coefficients of driven physical models of the human larynx, Journal of the Acoustical Society of America 121(1),January 2007

[17] M. R. Portnoff, A quasi - one dimensional digital simulation for the time varying vocal tract, M.S. Thesis, Dept. of Elect. Engr., MIT, Cambridge, Mass., June 1973.

[18] J. L. Kelly, C. Lochbaum, Speech Synthesis, Proc. of the Speech Communication Seminar,vol. II, 1963

[19] Liljencrants, J., Speech synthesis with a reflection-type line analog, DS Dissertation, Dept. of Speech Comm. and Music Acoust., Royal Inst. of Tech., Stockholm, Sweden, 1985

[20] Story, B. H., Speech Simulation with an Enhanced Wave-Reflection Model of the Vocal Tract, Ph. D. Dissertation, University of Iowa, 1995

[21] Story B.H., Titze IR. Voice simulation with a body-cover model of the vocal folds. J. Acoust. Soc. Am., 97(2), 1249-1260, 1995

[22] Story, B. H., Simulation of sentence-level speech with kinematic models of the vocal tract shape and vocal folds,Proc. Of Third Intl. Sym. Biomech., Human Func., and Inf. Science, Kanazawa, Japan, Feb. 20-22, Vol. II, pp. 55-61,2009

[23] Zañartu, M., Mongeau, L. and Wodicka, G. R., Influence of acoustic loading on an effective single mass model of the vocal folds. J. Acoust. Soc. Am., 121(2), 1119-1129, 2007

[24] M. Zañartu, "Influence of Acoustic Loading on the Flow-Induced Oscillations of Single Mass Models of the Human Larynx," M.S. Thesis, School of Electrical and Computer Engineering, Purdue University, May 2006

[25] S. Mathur, B. H. Story, and J. J. Rodriguez, Vocal-tract modeling: Fractional elongation of segment lengths in a waveguide model with half-sample delays, IEEE Trans. on Audio, Speech and Language Processing, 2006

[26] Meyer, P., Wilhelms, R., Strube, H.W. A quasiarticulatory speech synthesizer for German language running in real time, J. Acoust. Soc. Am., 86, 523-540, 1989

[27] Birkholz, P., Jackèl, D., Kröger, B.J.: Simulation of losses due to turbulence in the time-varying vocal system. IEEE Transactions on Audio, Speech, and Language Processing 15, 1218-1225, 2007

[28] Allen, D.R., Strong, W.J.: A model for synthesis of natural sounding vowels. Journal of the Acoustical Society of America 78, 58-69, 1985

[29] P.Mokhtari, H. Takemoto, and T. Kitamura, Single-matrix formulation of a time domain acoustic model of the vocal tract with side branches, Speech Communication, vol. 50, pp. 179-190, 2008.

[30] Sondhi, M.M., Schroeter, J.: A hybrid time-frequency domain articulatory speech synthesizer. IEEE Transactions on Acoustics, Speech, and Signal Processing 35, 955-967, 1987

[31] Fant, G., Liljencrants, J., Lin, Q.: A four-parameter model of glottal flow. Speech Transmission Laboratory - Quarterly Progress and Status Report 4/1985. Royal Institute of Technology, Stockholm, pp. 1-13, 1985

[32] Alipour, F., Berry, D.A., Titze, I.R.: A finite-element model of vocal-fold vibration. Journal of the Acoustical Society of America 108, 3003-3012, 2000

[33] Sondhi, MM , Model for the wave propagation in a lossy vocal tract, J. Acoust. Soc. Am.,55 (5), 1070-1075, May,1974

[34] Sondhi, MM. , Schroeter, J., ”Speech Production Models and Their Digital Implementations” Digital Signal Processing Handbook Ed. Vijay K. Madisetti and Douglas B.Williams Boca Raton: CRC Press LLC, 1999

[35] Rothenberg M. (1981)., Acoustic interaction between the glottal source and the vocal tract, in Vocal Fold Physiology, edited by Stevens K. N. and Hinano M. (University of Tokyo Press, Tokyo), 305-328

[36] Ananthapadmanabha, T.V. and Fant, G. (1982). Calculation of true glottal flow and its components. STL-QPSR 1/198, 1-30; also in Speech Communication 1, 167-184, 1982

[37] Thomas F. Quatieri ”Discrete-Time Speech Signal Processing: Principles and Practice” Prentice Hall PTR — 2001-11-08 — ISBN: 013242942X

[38] Vojnovic’ M., Mijic’ M., An improved model for the acoustic radiation impedance of the mouth based on an equivalent electrical network, Applied Acoustics, 66, pp. 481-499, 2005

[39] Peter Birkholz, Dietmar Jackèl, Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. In Interspeech 2004-ICSLP, pp. 1125-1128, Jeju, Korea, 2004

[40] Peter Birkholz, Dietmar Jackèl, Simulation of flow and acoustics in the vocal tract. In 30th Deutsche Jahrestagung für Akustik (CFA/DAGA ’04), pp. 895-896, Strasbourg, France, 2004

[41] Peter Birkholz, Dietmar Jackèl, Noise sources and area functions for the synthesis of fricative consonants. Rostocker Informatik Berichte, 30, pp. 17-23, 2006

[42] Alku P., Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering., Speech Communication, 1992

[43] Qiang Fu, Peter Murphy, Robust glottal source estimation based on joint source - filter model optimization, IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No.2, March 2006

[44] Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels. Journal of the Acoustical Society of America, 49, 583-590.

[45] Veldhuisa, R. (1998). A computationally efficient alternative for the liljencrants-fant model and its perceptual evaluation, J. Acoust.Soc. Am. 103, 566-571.

[46] Wong D., Markel J., and Gray A. Jr., Least squares glottal inverse filtering from the acoustic speech waveform, IEEE Trans. Acoust. Speech Signal Process. 27, 350-355,1979

[47] Krishnamurthy, A., and Childers, D. Two-channel speech analysis, IEEE Trans. Acoust., Speech, Signal Process. 34, 730-743, 1986

[48] Childers, D., and Wong, C.-F., Measuring and modeling vocal source-tract interaction, IEEE Trans. Biomed. Eng. 41, 663-671,1994

[49] Alku P, Magi C, Yrttiaho S, Bäckström T, Story B., Closed phase covariance analysis based on constrained linear prediction for glottal inverse filtering, J Acoust Soc Am.,125(5):3289-305,May 2009

[50] Arroabarren, I. and Carlosena, A. (2006). Inverse filtering in singing voice: a critical analysis,IEEE Trans. on Audio, Speech and Lang. Process. 14, 1422-1431.

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

[52] Veeneman, D., and BeMent, S., Automatic glottal inverse filtering from speech and electroglottographic signals, IEEE Trans. Acoust., Speech, Signal Process. 33, 369-377,1985

[53] Nelles Oliver, Nonlinear system identification from classical approaches to neural networks and fuzzy models,ISBN 3-540-67369-5,Springer-Verlag Berlin Heidelberg 2001,

[54] Marquez, H. J. (2003). Nonlinear Control Systems Analysis and Design (John Wiley & Sons, Inc.),155-182.

[55] Lagarias, J.C., J. A. Reeds, M. H. Wright, and P. E. Wright, "Convergence Properties of the Nelder-Mead Simplex Method in Low Dimensions," SIAM Journal of Optimization, Vol. 9 Number 1, pp. 112-147, 1998.

[56] Gilles Degottex, Erkki Bianco, IRCAM Databases of High Speed Videoendoscopy,UPMC-Ircam, France, 2010

[57] http://www.claymath.org/millennium/

VITA

PERSONAL INFORMATION

Surname, Name: Turgay, Koç

Nationality: Turkish (TC)

Date and Place of Birth: 03 Feb 1976 , Kayseri

Marital Status: Married

Phone: +90 505 230 08 79

email: tkoc@metu.edu.tr and turgaykoc2011@gmail.com

EDUCATION

Degree	Institution	Year of Graduation
BS	Niğde University Electrical and Electronics Engineering	2001
Technician	Erciyes University Industrial Electronics	1996
High School	Kayseri Lisesi (1991-92), Selçuklu Lisesi (1993)	1993

WORK EXPERIENCE

Year	Institution
2002-2012	METU Electrical and Electronics Engineering, Research and Teaching Assistant
2007-2008	ECESS - European Center of Excellence in Speech Synthesis, METU Pitch determination and epoch marking (PDA/PMA evaluation campaign) and METU VAD+ algorithm developer

Publications:

- Koç, T., Çiloğlu, T., ”An Interactive Source-Filter Model For Voiced Speech,” Signal Processing, Communication and Applications Conference, 2012. SIU 2012. IEEE 20th , pp.1-4, 18-20 April 2012.
- Koç, T., Çiloğlu, T., ”Implementation Of A VAD+ Algorithm And Its Evaluation,” Signal Processing, Communication and Applications Conference, 2008. SIU 2008. IEEE 17th , pp.1-4, 9-11 April 2009.
- B.Kotnik,P.Sendorek,S.Astrov,T.Koc,T.Ciloglu,L.D.Fernández,E.R.Banga, H.Höge,Z.K, ”Evaluation of Voice Activity and Voicing Detection.”, INTERSPEECH 2008, 2008

Submitted:

- Koç T., Çiloğlu T., Nonlinear modeling of source-filter coupling in voiced speech, Submitted to: Journal of Acoustical Society of America,September 2012

To be submitted:

- Koç T., Çiloğlu T., High Speed Endoscopic Video (HSV) Based Analysis of Speech, Submitted to: Journal of Acoustical Society of America
- Koc T., Çiloğlu T., Robust Extraction of Glottal Area Waveform from HSV Images, IEEE Transactions on Biomedical Engineering