Departament de Teoria del Senyal i Comunicacions

UNIVERSITAT POLITÈCNICA DE CATALUNYA

Ph.D. Dissertation

Voice Source Characterization for Prosodic and Spectral Manipulation

Javier Pérez Mayos

Thesis supervisor:
Dr. Antonio Bonafonte Cávez

TALP Research Centre, Speech Processing Group
Department of Signal Theory and Communications
Universitat Politècnica de Catalunya

Barcelona, 2012

A la meva dona que m’estima molt

Resum

L'objectiu d'aquesta tesi és estudiar la descomposició del senyal de veu en els seus components principals (tracte vocal i font de veu) mitjançant tècniques de filtratge invers, amb la major part dels esforços centrats a la parametrització del pols glotal. Volem explorar la seva utilitat en diverses tecnologies de la parla, com ara síntesi de veu, conversió de veu o detecció d'emocions. Per tant, estudiarem diverses tècniques per la manipulació prosòdica i espectral del senyal de veu. Un altre requeriment important és que els mètodes han de ser prou robustos com per poder treballar amb grans bases de dades típiques de la síntesi de veu. En el nostre treball, hem adoptat un model de producció de veu on les cordes vocals vibren per generar el pols glotal, que travessa el tracte vocal i és radiat pels llavis. Eliminar l'efecte del tracte vocal de la senyal de veu per obtenir el pols glotals es coneix com a filtratge invers. Els mètodes tradicionals són basats en predicció lineal durant la fase tancada de la glotis i tenen problemes quan aquesta és molt curta. Com a conseqüència, les estimacions acostumen a ser sorolles i difícils de parametritzar després. Nosaltres proposem de solucionar el problema fent servir un model paramètric del pols glotal directament a la fase d'estimació del tracte vocal i descomposició del senyal de veu. Com a resultat, obtenim millors estimacions del pols glotal (amb menys soroll) i una primera parametrització fent servir el model KLGLOTT88, que fem servir després per a estimar el model LF, més apropiat. També incloem un model per al residu (que inclou part del soroll d'aspiració) estimat dins l'algorisme principal.

Per tal de validar el funcionament del mètode de parametrització, hem construït un corpus sintètic fent servir paràmetres del model LF obtinguts de la literatura, complementats amb resultats propis. Així, podem generar senyals sintètiques amb paràmetres coneguts que podem fer servir posteriorment com a referència per calcular l'error de parametrització. Els resultats proven que el nostre mètode funciona bé per a una àmplia gama de paràmetres amb diferents nivell de soroll. També hem fet un test MOS

d'opinió comparant la qualitat de dos mètodes proposats per nosaltres amb la del vocoder STRAIGHT, una referència d'alta qualitat. El nostre mètode fent servir residus blanquejats resulta guanyador, amb una puntuació molt alta. El segon mètode fent servir parametrització total pel residu queda en tercer lloc, però encara amb puntuacions superiors al llindar d'acceptació.

A continuació hem proposat dos algorismes per a fer modificacions prosòdiques, segons quin dels dos mètodes pel residu fem servir: un fa servir la parametrització complerta i interpolació de trames per als canvis prosòdics, i l'altre una tècnica de remostreig aplicada a les formes d'ona. Les dos opcions han estat avaluades en un test MOS d'opinió, comparant-les amb STRAIGHT i el mètode PSOLA fet servir al nostre sintetitzador Ogmios. Tots dos mètodes reben puntuacions semblants, per sobre el llindar d'acceptació però no de prou qualitat, així que el tractament del residu és un tema que cal seguir investigant.

Hem inclòs el nostre model de producció de veu dins d'un sistema de conversió de veu, per tal d'avaluar l'impacte de la nostra parametrització en la conversió. El sistema de conversió de veu va ser desenvolupat al nostre grup com a part del projecte europeu TC-STAR i ha participat en diverses campanyes d'avaluació. Hem fet servir el model paramètric i les mateixes condicions per l'entrenament i el test, per tal de poder comparar els resultats directament amb els de l'última avaluació del projecte. El nostre mètode ha estat un èxit en aquesta tasca, ja que ha estat puntuat amb millor nota pels participants a l'avaluació.

Com a part d'aquesta tesi també hem treballat en el camp de les qualitats de veu (de l'anglès *voice quality*). Hem enregistrat una petita base de dades que consisteix en una locutora professional generant vocals sostingudes en espanyol, amb diferents qualitats de veu (modal, aspre, trencada i falset). Llavors les hem analitzat totes fent servir el nostre algorisme de descomposició i parametrització mitjançant figures estadístiques dels paràmetres LF. Comparant els resultats amb d'altres publicats anteriorment, hem trobat que les nostres conclusions coincidien en major part amb les d'altres investigadors. Les diferències es podrien atribuir a la mida de la base de dades i a les dificultats en comparar qualitats de veu produïdes per persones diferents.

Per a finalitzar la nostra feina en aquesta tesi, hem treballat també en el camp del reconeixement automàtic d'emocions fent servir mètodes estadístics basats en GMM. Per cada emoció, hem entrenat un model específic fent servir diferents característiques, comparant la nostra parametrització amb un sistema de referència que fa servir mesures espectrals (MFCC) i prosòdiques (F_0 i $\log F_0$). Hem fet servir una base de dades d'emocions d'aproximadament 5000 frases de dos locutors diferents, que conté exemples de sis emocions i l'estat neutre. Els resultats són molt satisfactoris, ja que la noves característiques disminueixen un 20% l'error respecte el mètode de referència. L'encert en la detecció de les emocions també supera els publicats anteriorment amb la mateixa base de dades

fent servir només característiques prosòdiques. La conclusió que en traiem és que els paràmetres del pols glotal obtinguts amb el nostre mètode tenen un impacte positiu en el camp del reconeixement d'emocions.

Abstract

The objective of this dissertation is to study and develop techniques to decompose the speech signal into its two main components: voice source and vocal tract. Our main efforts are on the glottal pulse analysis and characterization. We want to explore the utility of this model in different areas of speech processing: speech synthesis, voice conversion or emotion detection among others. Thus, we will study different techniques for prosodic and spectral manipulation. One of our requirements is that the methods should be robust enough to work with the large databases typical of speech synthesis. We use a speech production model in which the glottal flow produced by the vibrating vocal folds goes through the vocal (and nasal) tract cavities and its radiated by the lips. Traditional inverse filtering approaches based on closed-phase linear prediction have the problem of having to work only with samples corresponding to the glottal closed phase, which in some cases can be quite short. This results in a large amount of noise present in the estimated inverse-filtered waveforms, which poses further problems when parameterizing them. We overcome this problem by using a parametric model for the glottal waveform and thus including the glottal open phase in the estimation. As a result of the source-filter decomposition, we not only obtain a better (i.e., less noisy) inverse-filtered glottal estimation, but also a first parametrization using the simpler KLGLOTT88 model used to estimate the more appropriate LF model. A parametric model for the residual comprising the aspiration noise was also proposed as part of the parametrization.

In order to validate the accuracy of the parametrization algorithm, we designed a synthetic corpus using LF glottal parameters reported in the literature, complemented with our own results from the vowel database. Since the parameters used in synthesis were known a priori, they were used as reference to compute the parametrization error with the estimated parameters. The results show that our method gives satisfactory results in a wide range of glottal configurations and at different levels of SNR. We also conducted

an on-line evaluation in which the quality of two proposed methods was compared to that of the well established vocoder STRAIGHT. Our method using the whitened residual compared favorably to this reference, achieving high quality ratings (Good-Excellent). Our full parametrized system scored lower than the other two ranking in third place, but still higher than the acceptance threshold (Fair-Good).

Next we proposed two methods for prosody modification, one for each of the residual representations explained above. The first method used our full parametrization system and frame interpolation to perform the desired changes in pitch and duration. The second method used resampling on the residual waveform and a frame selection technique to generate a new sequence of frames to be synthesized. Both options were again evaluated, using two standard algorithms as reference (STRAIGHT as in the resynthesis test, and the PSOLA-like algorithm as implemented in our speech synthesizer Ogmios). The results showed that both methods are rated similarly (Fair-Good) and that more work is needed in order to achieve quality levels similar to the reference methods.

Our speech production model was incorporated to an existing voice conversion (VC) system to evaluate the impact of the parametrization on the conversion performance. The system used for VC was developed in our group as part of the TC-STAR project, and it had participated in several evaluation campaigns. The same testing conditions were replicated and used with our parametrization model, using the full parametrization for the residual. The waveforms were generated and compared with those obtained with the original VC system, again using an on-line MOS evaluation. The results showed that the evaluators preferred our method over the original one, rating it with a higher score in the MOS scale.

As part of this dissertation, we conducted a study in the field of voice quality. We recorded a small database consisting of isolated, sustained Spanish vowels in four different phonations (modal, rough, creaky and falsetto) and were later also used in our study of voice quality. Each of them was analyzed using our decomposition and parametrization algorithm, and boxplot of the glottal and residual parameters were produced. The LF parameters were compared with those reported in the literature, and we found them to generally agree with previous findings. Some differences existed, but they could be attributed to the difficulties in comparing voice qualities produced by different speakers.

We have also evaluated the performance of an automatic emotion classifier using glottal measures. The classification was performed by statistical GMM classifiers trained for each emotion using different features. We have compared our parametrization to a baseline system using spectral (MFCC) and prosody (F_0 and $\log F_0$) characteristics. The results of the test using an emotional database of almost 5000 utterances and two speakers were very satisfactory, showing a relative error reduction of more than 20% with respect to the baseline system. The accuracy of the different emotions detection was also high, improving the results of previously reported works using the same database. Overall, we can

conclude that the glottal source parameters extracted using our algorithm have a positive impact in the field of automatic emotion classification.

Acknowledgments

Quan finalment arriba el moment de donar les gràcies per tota l'ajuda i el suport moral rebut durant aquest temps (més l'econòmic, el físic, el tècnic, el logístic, el...) un s'asseu davant l'Emacs i realment no sap ni per on començar... però per acabar d'una vegada s'ha de començar per algun lloc, així que primer de tot, gràcies al meu director de tesi, el Toni Bonafonte, per haver-me donat l'oportunitat de treballar al grup de síntesi de veu i d'haver-me aguantat durant tant de temps, amb les idees esbojarrades, depressions, eufòries, etc. Sense ell aquest dia no hagués arribat.

I ja que he començat pel grup de veu, deixeu-me que aprofiti per mostrar el meu agraïment a tota la gent que forma i ha format part d'aquest increïble grup humà: l'Asunción, l'Helenca, el Dani, el Jordi, el Pablo, l'Ignasi, la Tatyana... Han estat molts bon moments compartits, reunions, projectes, cafès de les 5 al bar de la FIB, pizzes al vespre preparant avaluacions...

Moltes gràcies també a tota aquella gent que durant aquests anys ha dedicat part del seu temps a ajudar-me amb les meves avaluacions online! I com vaig prometre, menció especial per a aquells que van participar a l'últim temps, gràcies a vosaltres vaig creure'm per fi que ja acabava la tesi: Jaume Padrell, Frank Diehl, Jordi Adell, Javi Ruiz, Daniel Erro, Huc Castells, Guillermo Vila, Sergio Oller, David Escudero, Ramon Morros, Alberto Irurueta, Rosa Morros, Anderson Fraiha Machado, Adrià de Gispert, Daniel Delgado, Ignatz Links, Federico Flego i Clara Panozzo.

La tornada a la universitat pel doctorat m'ha permet de conèixer i fer amistat directament i indirecta amb gent molt maca a qui espero de seguir veient durant molt de temps: Javi, Ramon, Jaume, Marta, Adrià, Aleyda, Frank, Alberto, Josep Ma, Ali... Ah! I menció especial per la meva fillola l'Alba, que va néixer just a temps d'esperonar-me per l'últim esforç!

I com moltes coses importants de la meva vida, sense Estocolm no es pot entendre la meva història. Allà vaig descobrir una família que va començant sent de Lappis, una part va marxar a Barcelona on va seguir creixent, una altra part va seguir a Estocolm, una altra va seguir voltant pel món... A tots, gràcies de tot cor! Laieta, Laia, Eva, Xavi, Jordi, Cynthia, Ernesto, Kristina, Paco, Paquito, Naoko, Ken, Anna, Victor, Sandra, Merche, Veera... I a tots aquests a qui durant aquests anys he dit que "trabajo en ordenadores y cosas de esas" (Pepe, Bea, David, Lidia...) des d'aquí una abraçada molt forta!

Com oblidar tota la gent d'Alemanya que encara que no entenen gaire què he estat fent allà tot aquell temps sempre ens han fet costat? Guille, Dani, Bettina, Loisl, Marcos, Julia, Stephan, Miriam, Benjamin, Lena, Hedda... vielen Dank!

I com no, a la meva família vull agrair la paciència que durant tot aquest temps han tingut i el suport de tot tipus que m'han donat i m'ha ajudat a acabar la tesi: papa, mama, Laura, Marta, se acabó! I per descomptat a les àvies (Ana i Pepita), tiets i tietes (Marta, Marimar, Ma José, Arturo, Juan Andrés, Rafa, Ma Luisa), cosinets i cosinetes (Guille, Irene, Lluc, Fer) i resta de la família també va dedicada aquesta tesi. Vull dedicar aquesta tesi, com no, també a la branca Lifante, tant a Barcelona com a Suïssa (Pere, Conchita, Susana, Adi, dolent i dolenta).

Segurament em deixo molta gent que es mereix sortir als agraïments com qui més, així que des d'aquí les meves sinceres disculpes per l'omissió en aquesta impressió!

Menció especial per les meves gates, la Vispgrädde i la Chokladkoppen, que sobretot a partir de l'etapa alemanya han compartit moltes hores *de feina* amb mi a casa (és a dir, d'estirar-se sobre el teclat de l'ordinador mentre jo treballava per a que els rasqués la panxa). Malauradament, moltes de les contribucions no eren gaire profitoses, sort del Ctrl+_, o això estaria ple de zyzcccccccccccccccxxyyyyyyasadassssssssss :)

I acabar, com no, amb la persona que m'ha hagut de patir de més a prop, que durant les èpoques dolentes m'ha fet sempre costat i durant les bones ho ha gaudit i celebrat de ben a prop: a tu, Conxita, la meva dona, gràcies de tot cor... ja la tenim aquí!!! T'estimo molt.

Barcelona, 2012

Contents

List of Figures

List of Tables

Glossary

ABX	Preference test
AM	Amplitude Modulation
APE	Average Percentile Error
AQ	Amplitude Quotient
AR	Auto-regressive model
ARMA	Auto-regressive moving average model
ARMAX	Auto-regressive moving average model with exogenous input
ARX	Auto-regressive model with exogenous input
AV	Amplitude of voicing
AWGN	Additive white Gaussian Noise
CART	Classification and Regression Trees
CCD	Complex Cepstrum Decomposition
ClQ	Closing Quotient
CP	Closed-phase
CPLP	Closed-phase Linear Prediction
DAP	Digital All-Pole modeling
dEGG	Differentiated laryngograph signal
EGG	Laryngograph signal
EM	Expectation-Maximization
FFT	Fast Fourier Transform

FIR	Finite Impulse Response
GA	Genetic Algorithm
GCI	Glottal Closure Instant
GDF	Group Delay Function
GFM	Glottal Flow Model
GMM	Gaussian Mixture Model
GOI	Glottal Opening Instant
GQM	Glottal Quality Measure
HMM	Hidden Markov Model
IAIF	Iterative Adaptive Inverse Filtering
IF	Inverse Filtering
IHMD	Inverse Harmonic Mean Distance
KLGLOTT88	Ronsenberg-Klatt glottal model with tilt
LF	Liljencrants-Fant model
LP	Linear Prediction
LPC	Linear Prediction Coding
LSF	Line Spectral Frequencies
MA	Moving Average
MFCC	Mel-Frequency Cepstral Coefficients
MOS	Mean Opinion Score
NAQ	Normalized Amplitude Quotient
OLA	OverLap and Add
OP	Open Phase
OQ	Open Quotient
PCA	Principal Component Analysis
PDF	Probability Density Function
PE	Percentile Error
PSOLA	Pitch-Synchronous OverLap and Add

PSP	Parabolic Spectral Parameter
QP	Quadratic Programming
SNR	Signa-to-Noise Ratio
SQ	Speed Quotient
STRAIGHT	Speech Transformation and Representation using Adaptive Interpolation of weiGHTed spectrum
TTS	Text-To-Speech
VC	Voice Conversion
VQ	Voice Quality
VT	Vocal Tract
ZCR	Zero-Crossing Rate
ZZT	Zeros of the Zeta Transform

CHAPTER 1

Introduction

In a world where images and visual information appear in all the aspects of everyday life, the human voice continues to play a key role in the communication between people. Speech contains the message a person intends to transmit to their audience, but also conveys extra information about the speaker's mood, intention, gender, etc. Native speakers can easily recognize where the person they are speaking with comes from using only acoustic cues. Speech contains also information intimately tied to the speaker, so that he or she can be uniquely identified by other people. In short, the information speech conveys is not only language-related, but also speaker-related. In relation to the particular language we can agglutinate the speech properties determined by the selection of words or by the grammar governing the sentence structure. Among the speaker-related features we can find pitch (fundamental frequency), loudness (magnitude of the auditory sensation produced), voice quality (characteristics tied to the transmission of affect) and rhythm (particular melody).

Speech being the most common way of communication between people, it seems just natural that it should find its way into man-machine interfaces. The widespread of computers and the increasing necessity of finding better ways of communicating with them, have resulted in an increased popularity of speech-related technologies. The two main components of a voice-driven user interface are speech recognition and speech synthesis. In speech recognition, the main task is to translate what the speaker says into text that can be processed and acted upon by a computer. But as in regular human interaction, not only the *what* is important, but also the *how* and *why*. And this information is contained in speech properties not yet recognizable by current speech recognition algorithms. In speech synthesis, during many years the trend has been to work towards concatenative speech synthesis. Earlier approaches like formant synthesis, in which knowledge of the way human produce speech was at the core of the methods, were slowly abandoned for mainstream applications and were relegated to specific applications (Ding and Campbell, 1997; Klatt, 1980; Liljencrants, 1967). Although very flexible and able to imitate many of the underlying processes in speech production, they required laborious tuning and computer power to achieve a minimal degree of naturalness, still insufficient for satisfactory communication. The idea with modern concatenative systems is to use parts of prerecorded speech to form the synthetic utterances, thus being able to achieve high quality and naturalness, assuming, of course, that the concatenation is properly done (Charpentier and Moulines, 1988; Moulines and Charpentier, 1990). The main problem here is that we are limited to the quality and style of the recordings, since they only aim at perfectly reproducing the original recorded voice. In order to achieve high quality synthesis we need large databases, which has a high cost in terms of storage space, processing power and human time (high-quality recording is very time-demanding for both the professional speakers and the operators). And there are new applications, like voice conversion of emotional speech synthesis, requiring large degrees of prosody modification that concatenation systems are unable to fulfill (Dutoit, 1997). Several alternatives making use of the periodic/aperiodic components of the speech signal have been proposed in order to overcome some of these difficulties (Dutoit and Leich, 1992, 1993). McAulay and Quatieri (1986) introduced a speech sinusoidal representation with a higher resilience to distortion when performing prosodic and spectral modifications. In Laroche et al. (1993) the authors presented an speech modification algorithm based on the decomposition of the speech into harmonic and noise, which was later applied to concatenative speech synthesis by Stylianou (2001).

However, although these algorithms generally succeed in permitting greater prosody modifications without the penalty of degrading the resulting quality to unacceptable levels, they still fail to incorporate the characteristics responsible for emotion and affect. Is in this context where glottal waveform analysis has the ability to play an important role. Fant (1970) proposed a speech production model in which the glottal flow produced by the vibrating vocal folds goes through the vocal (and nasal) tract cavities and its radiated by the lips. According to this model, a source-filter decomposition should be possible, in which the two main contributions (from the glottis and the vocal tract) could be independently analyzed and modified. The focus of this thesis is on the extraction of the glottal source information. This is a fundamental problem in speech processing, having applications in a wide range of speech related technologies: voice conversion (Childers and Ahn, 1995; Gutiérrez-Arriola et al., 1998; Mori and Kasuya, 2003), speech synthesis (Cabral et al., 2007; Childers and Hu, 1994; Pinto et al., 1989), voice pathology detection (Drugman et al., 2009b; Gómez-Vilda et al., 2009), emotion analysis/synthesis (Burkhardt and Sendlmeier, 2000; Gobl and Chasaide, 2003b), etc.

The objective of this dissertation is to study and develop techniques to decompose the speech signal into its two main components: voice source and vocal tract. Our main efforts are on the glottal pulse analysis and characterization. We want to explore the utility of this model in different areas of speech processing: speech synthesis, voice conversion or emotion detection among others. Thus, we will study different techniques for prosodic and spectral manipulation. One of our requirements is that the methods should be robust enough to work with the large databases typical of speech synthesis. We use a speech production model in which the glottal flow produced by the vibrating vocal folds goes through the vocal (and nasal) tract cavities and its radiated by the lips. Traditional inverse filtering approaches based on closed-phase linear prediction have the problem of having to work only with samples corresponding to the glottal closed phase, which in some cases can be quite short. This results in a large amount of noise present in the estimated inverse-filtered waveforms, which poses further problems when parameterizing them. We overcome this problem by using a parametric model for the glottal waveform and thus including the glottal open phase in the estimation. As a result of the source-filter decomposition, we not only obtain a better (i.e., less noisy) inverse-filtered glottal estimation, but also a first parametrization using the simpler KLGLOTT88 model used to estimate the more appropriate LF model. A parametric model for the residual comprising the aspiration noise was also proposed as part of the parametrization. The dissertation is organized as follows:

Chapter 2 explains the human voice production system used in this work and introduces the different parametric models for the glottal model that have been presented in the literature, with special focus on those used in our method. Furthermore, a thorough review of the different methods for source-filter decomposition is presented, from the early attempts to the current state-of-the-art systems. Measures for the automatic rating of the glottal estimations are also explained. The main characteristics of the aspiration noise generation are detailed here.

Chapter 3 describes in detail our proposed algorithm for source-filter decomposition. The chapter starts with an outline of the whole method and the details of the convex decomposition algorithm used thorough the chapter. The extraction of the glottal timing information from the laryngograph signal is presented, and an algorithm to optimize this initial estimation follows. Then the source-filter estimation technique is described, and the algorithms to parametrize the glottal flow and the aspiration are described next. The chapter ends with the schema for synthetic speech generation and the conclusions.

Chapter 4 presents the methodology used to validate and evaluate our source-filter analysis system, using both a synthetic corpus designed using real glottal parameters reported in the literature, and real corpus specifically recorded for this purpose, containing sustained vowels in Spanish uttered with different voice qualities. The resynthesis capabilities of the method are evaluated using an online listening test.

Chapter 5 describes our two proposed algorithms for prosody modification. The required changes in the parametrization are explained and the two methods to change pitch and duration are described next. We follow by presenting and discussing the results of an online listening test comparing them to two well-known and established algorithms.

Chapter 8 describes our work in the field of CART-based voice conversion using GMM. The underlying theory is presented and the current algorithms are described. Next the baseline system used is introduced, with the necessary changes to introduce our parametrization detailed. There follows the results of an online evaluation test and correspondent analysis and discussion.

Chapter 7 presents our work in the field of voice quality analysis. The literature on this topic is reviewed and an analysis of the corpus presented in Chapter 4 is conducted. There follows a listening test focused on the characterization and identification of voice quality using our analysis/synthesis algorithm. Our work in automatic emotion recognition is also presented here.

Chapter 8 ends this dissertation by presenting the conclusions and future lines of research that could follow the work presented here.

CHAPTER 2

Background

2.1 The human voice production system

Voicing sounds are produced when the airflow expelled by the lungs arises from the trachea and reaches the glottis, causing the vocal folds to vibrate in a quasi-periodic manner. This vibration generates the glottal volume-velocity waveform, or glottal flow, which then travels through the cavities of the vocal (and nasal) tract, and it is radiated by the lips resulting in the speech pressure waveform. Fant (1970) proposed a model of the human speech production system using a source-filter approach, adding detailed knowledge of the glottal flow signal. In this model, the glottal source excites the vocal tract, and it is radiated by the lips (as illustrated in figure 2.1). Both the vocal tract filter and the lip-radiation effect are often modeled using linear filters: the vocal tract is usually assumed to be an all-pole filter, a simplification that excludes the possibility of source-filter interaction, but that works generally well in the majority of occasions; the lip-radiation effect is often modeled by a first-order derivative filter. Since these two filters are linear and invariant (at least over short periods of time), they can be linearly exchanged and the derivative filter can be combined with the source, resulting in the differentiated glottal volume-velocity waveform.

One method to obtain this representation is to use inverse filtering techniques to acquire an estimation of the glottal volume-velocity waveform by canceling the effect of the vocal tract. The goal is to design a filter that cancels the poles introduced by the vocal tract in order to eliminate its contribution from the speech signal. This can be used to estimate either the glottal velocity waveform or its first derivative, depending on whether the lip-radiation and the vocal tract filters have been commuted. Thus, the resulting (simplified) source-filter consists of the derivative glottal waveform, together with the differentiated

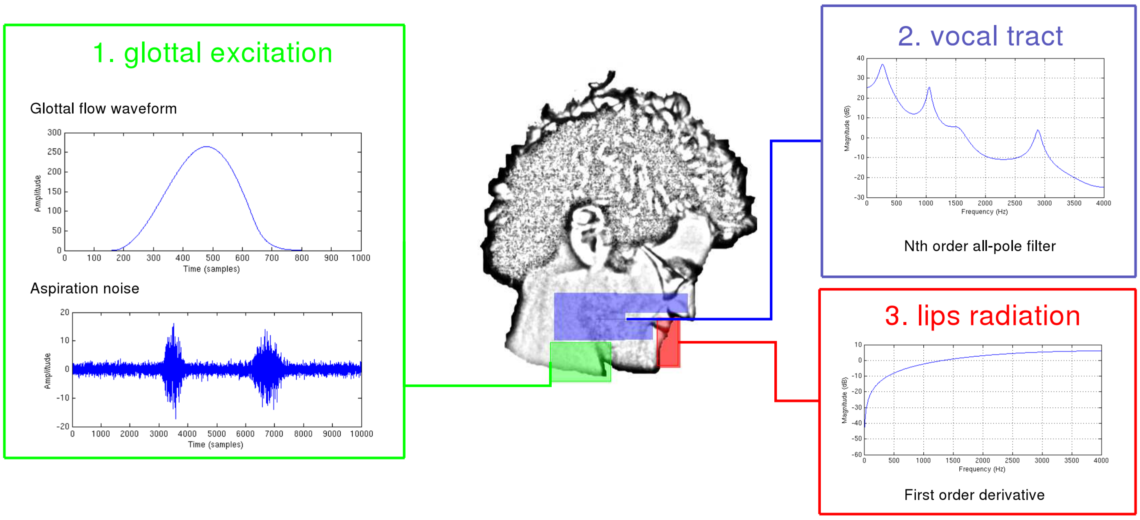


Figure 2.1: *Speech production system using the source-filter approach.*

aspiration noise, passing trough an all-pole filter modeling the vocal tract effects. This process is illustrated in figure 2.2 and described next. According to this model, the speech waveform $S(z)$ is obtained as:

$$S(z) = U_g(z) \cdot V(z) \cdot L(z), \tag{2.1}$$

where $V(z)$ represents the vocal tract, $L(z)$ the lip radiation effect, and $U_g(z)$ is the glottal volume-velocity waveform. Since we are modeling the vocal tract using an all-pole filter, we can write:

$$V(z) = \frac{1}{A(z)} = \frac{1}{1 - \sum_{k=1}^{N} a_k \, z^{-k}}, \tag{2.2}$$

where $A(z)$ is the polynomial in the z-domain, whose roots are the poles of the vocal tract filter $V(z)$. The lip radiation effect can be effectively modeled using a differentiator filter:

$$L(z) = \frac{1}{1 - z^{-1}}. \tag{2.3}$$

Furthermore, we can exchange $V(z)$ and $L(z)$ since they are both linear, and we can incorporate the differentiator into the glottal source, effectively working with the differentiated glottal volume-velocity waveform $G(z)$ as the glottal source:

$$G(z) = U_g(z) \cdot L(z). \tag{2.4}$$

Taking into account equations 2.2 and 2.4, we can rewrite (2.1) as:

$$S(z) = G(z) \cdot \frac{1}{1 - \sum_{k=1}^{N} a_k \, z^{-k}}, \tag{2.5}$$

which in the time-domain can be written as:

$$s(n) = g(n) + \sum_{k=1}^{N} a_k s(n-k). \tag{2.6}$$

Thus, the value of the speech signal at sample n can be computed using the current value of the glottal source $g(n)$ and a linear combination of the previous N speech samples using the filter coefficients a_k.

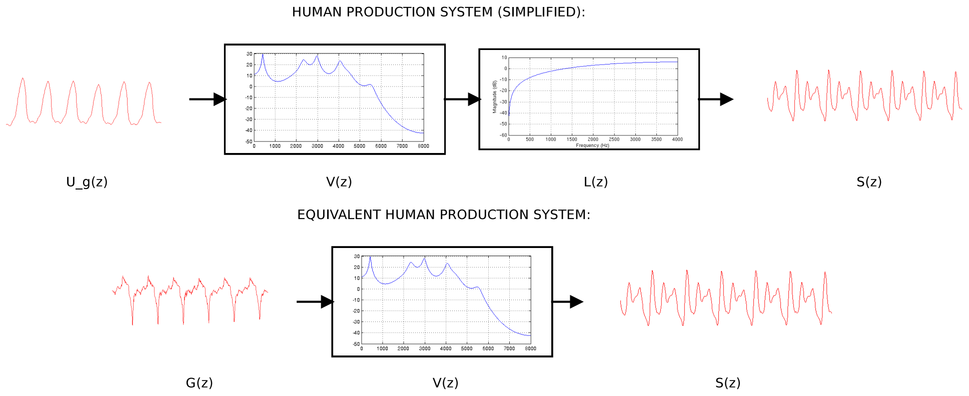

HUMAN PRODUCTION SYSTEM (SIMPLIFIED):
U_g(z)
V(z)
L(z)
S(z)
EQUIVALENT HUMAN PRODUCTION SYSTEM:
G(z)
V(z)
S(z)

Figure 2.2: *Block-diagram of the main source-filter components.*

The dynamics of the glottis can be separated into two phases. During the first phase, called the *open phase*, the airflow coming from the trachea increases the sub-glottal pressure, causing the vocal folds to open in a progressively manner. When the vocal folds reach the elastic displacement limit, they abruptly return to the initial state (*return phase*), and remain closed until the sub-glottal pressure increases again (*closed phase*) include it in the closed phase we . Figure 2.3 serves as an illustration of the glottal phases, showing a cycle of the glottal flow and its time derivative using the LF model (Fant et al., 1985) that will be explained in detail in section 2.2.2. The open phase consists of two sub-phases: the *opening phase*, from the moment the glottis starts to open at time 0, until the glottal flow (upper figure) reaches its maximum at T_p, and the *closing phase*, from this moment until the derivative of the glottal flow (lower figure) reaches its minimum at T_e. The closed phase begins at T_e with the *return phase* [1] , that lasts until T_c, where the vocal folds are completely closed for the rest of the cycle until T_0.

[1] Whether the return phase is considered part of the open or closed phases is a matter of opinion and varies from author to author. In this thesis we consider the closed phase to begin at T_e in figure 2.3, thus comprising

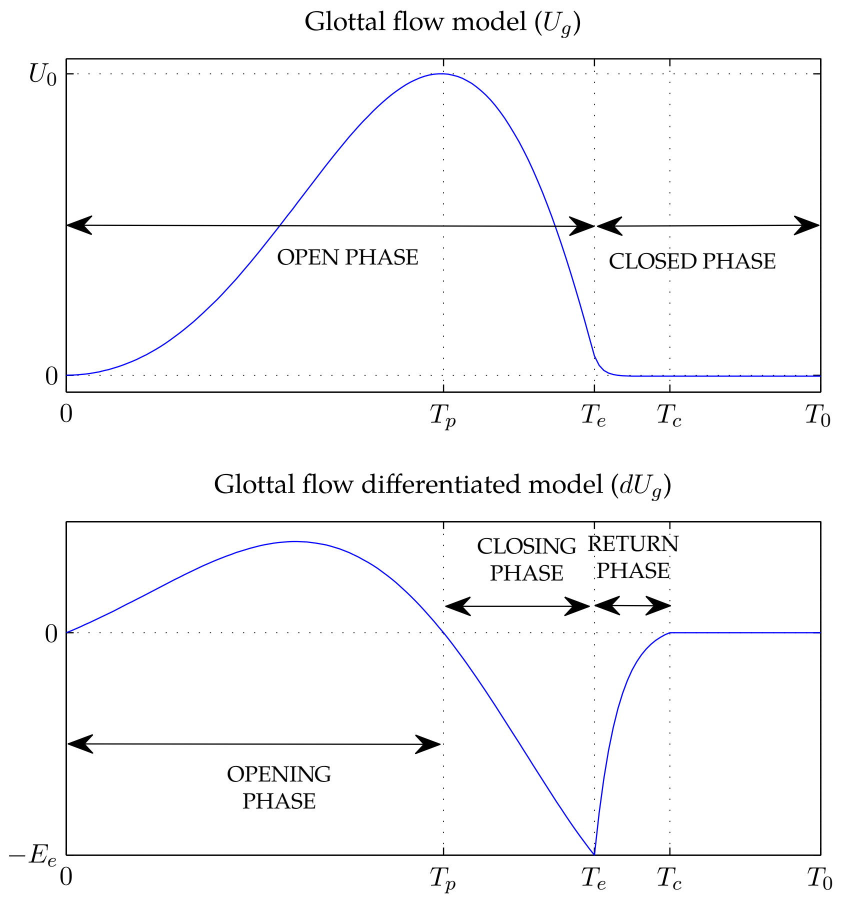


Figure 2.3: *Phases of the glottis in the glottal flow and differentiated glottal flow (illustrated using the LF model nomenclature from Section 2.2.2)*

2.2 Glottal models

While some applications require directly working with the high definition inverse filtered waveform (e.g., pathological speech analysis or the study of voice disorders), there are many others that would immensely benefit if these waveforms could be reduced to a smaller set of parameters. Several models have been proposed in the literature for parametrization of the glottal flow pulses. One of the earlier models was the Rosenberg (1971) C, a trigonometric model defined by four parameters (amplitude, fundamental period, maximum of the glottal flow waveform and the time interval between this maximum, and the glottal closure instant). Based on this model, several other parametric models were presented: the Liljencrants-Fant model (LF) was introduced by Fant et al. in 1985, the KLGLOTT88 model (Klatt and Klatt, 1990) and the R++ model (Veldhuis, 1998). Both the LF and the KLGLOTT88 model have been widely adopted and extensively reported

the return phase.

in the literature, thus they will be reviewed below. The LF model is more flexible than the KLGLOTT88 model in its ability to model different voice types. On the other hand, it has a more complex mathematical formulation, which makes it difficult to use in some algorithms.

Although several parametric models have been proposed in the literature, all of them share some common properties as reported in Doval and d'Alessandro (1997, 1999, 2006). We will outline here the general guidelines for characterizing the (derivative) glottal volume-velocity waveforms, and then we will present the details of the two models used in this work: the KLGLOTT88 and the LF models. A glottal flow model (GFM) is a continuous, differentiable function of time (with some exceptions at the instant of glottal closing), always positive or null. The GFM is quasi-periodic and bell-shaped during a fundamental period: first increasing, then decreasing, and then becoming null. This is shown in the upper part of figure 2.3. A differentiated glottal flow model (dGFM) is a quasi-periodic function, and during a fundamental period it is positive (increasing glottal flow), then null (glottal flow maximum), then negative (decreasing glottal flow) and finally null (when the glottal flow is null). This is shown in the lower part of figure 2.3. Arroabarren and Carlosena (2003c) also presents a unified spectral analysis of the KLGLOTT88 and LF models. Analyzing the general properties of open quotient, asymmetry coefficient and spectral tilt, the authors conclude both models are equivalent. The main difference reported there is the glottal asymmetry not being independent of the open quotient and the tilt in the KLGLOTT88 model.

2.2.1 KLGLOTT88 model

The KLGLOTT88 model (Klatt and Klatt, 1990) is a time-domain model, defined using a Rosenberg-Klatt waveform (Klatt, 1980) followed by a first-order, low-pass filter $TL(z)$ controlling the glottal closure abruptness. The Rosenberg-Klatt glottal flow is described by a third degree polynomial:

$$u_{rk}(t) = \begin{cases} at^2 - bt^3 & , 0 \leq t < O_q \cdot T_0 \\ 0 & , O_q \cdot T_0 \leq t < T_0, \end{cases} \tag{2.7}$$

and, accordingly, the Rosenberg-Klatt glottal flow derivative is expressed as:

$$g_{rk}(t) = \begin{cases} 2at - 3bt^2 & , 0 \leq t < O_q \cdot T_0 \\ 0 & , O_q \cdot T_0 \leq t < T_0, \end{cases} \tag{2.8}$$

where

$$a = \frac{27\,AV}{4\,O_q^2\,T_0}, \tag{2.9}$$

$$b = \frac{27\,AV}{4\,O_q^3\,T_0^2}, \tag{2.10}$$

and T_0 is the pitch period of the voice, O_q is the open quotient (ratio of the open phase of the glottal cycle to the duration of the cycle) and AV is the amplitude parameter. Note that the parameters a and b are related as:

$$\frac{a}{b} = O_q\,T_0. \tag{2.11}$$

The complete KLGLOTT88 model is then obtained by filtering the Rosenberg-Klatt waveform with the aforementioned low-pass filter:

$$TL(z) = \frac{1}{1-\mu z^{-1}}, \tag{2.12}$$

resulting in:

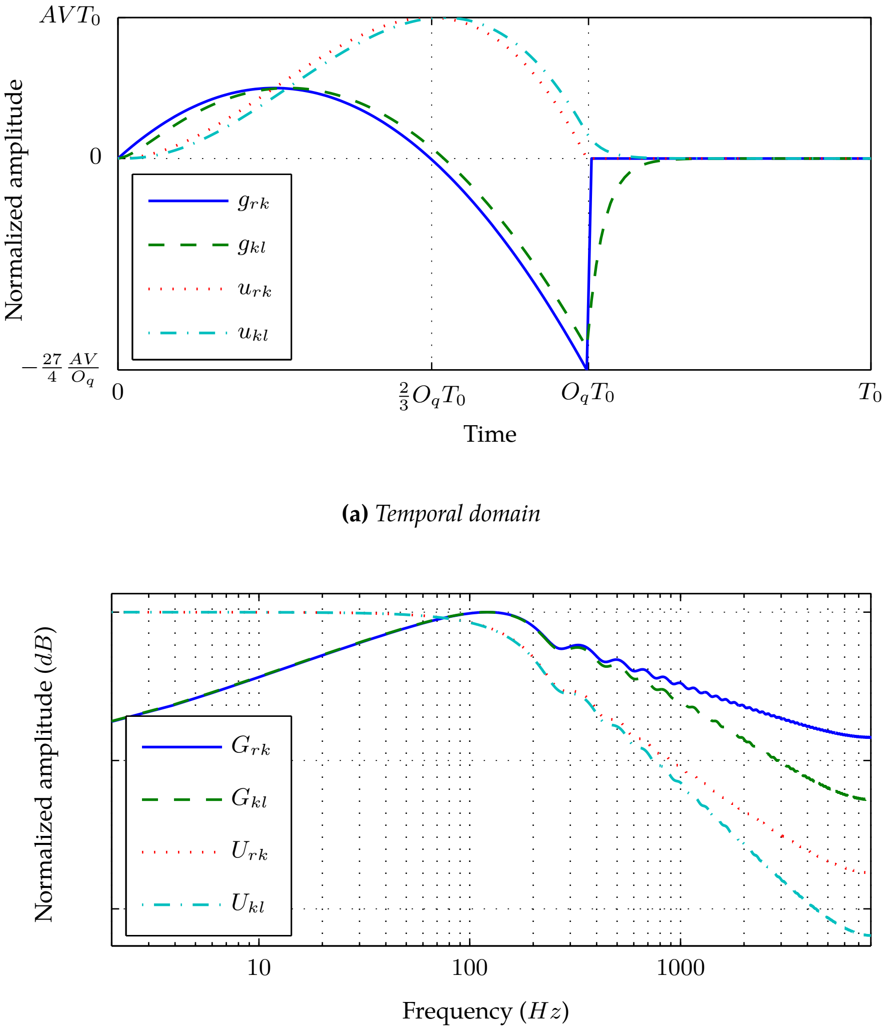
$$G_{kl}(z) = G_{rk}(z)\,TL(z) = G_{rk}(z)\,\frac{1}{1-\mu z^{-1}}. \tag{2.13}$$

Thus, the KLGLOTT88 model contains four parameters: AV, O_q, T_0 and μ. Figure 2.4 shows the time-domain glottal cycles (2.4a) and the corresponding spectra (2.4b), respectively, for both the glottal waveform flow and its derivative.

Although in this thesis we work almost exclusively in the temporal domain, we will explain some of the properties of the spectrum of the glottal models. The analytical spectrum of the Rosenberg-Klatt flow was computed in Doval and d'Alessandro (1997) and is given by:

$$U_{rk}(f) = \frac{27j\,AV}{2\,O_q\,(2\pi f)^2}\left(\frac{je^{-j2\pi f O_q T_0}}{2} + \frac{1+2e^{-j2\pi f O_q T_0}}{2\pi f O_q T_0} + 3j\frac{1-e^{-j2\pi f O_q T_0}}{(2\pi f O_q T_0)^2}\right). \tag{2.14}$$

As we can see in Figure 2.4b, this spectrum is flat in the lower range of the spectrum and has a slope of $-12\,\text{dB}/oct$ for higher frequencies. This behavior can be approximated by a second order low-pass filter with cutoff frequency $f_k = \frac{\sqrt{3}}{\pi}\,\frac{1}{O_q T_0}$, which depends only on O_q and T_0 (Doval and d'Alessandro, 1997). Differentiating the glottal flow to obtain the differentiated glottal waveform adds $6\,\text{dB}/oct$, as we can see in the figure. The tilt filter $TL(z)$ from eq. (2.12) adds $-6\,\text{dB}/oct$ attenuation at its cutoff frequency f_t. Thus, the spectrum of the KLGLOTT88 model can be divided in three regions: flat between 0 and f_k, with a spectral slope of $-12\,\text{dB}/oct$ between f_k and f_t, and with a spectral slope of $-18\,\text{dB}/oct$ from this point on.



(a) *Temporal domain*

(b) *Spectral domain*

Figure 2.4: *Temporal and spectral plot of the Rosenberg-Klatt and KLGLOTT88 glottal flow (u_{rk} and u_{kl}) and differentiated (g_{rk} and g_{kl})*

The KLGLOTT88 model has been used in the study of glottal characteristics for female and male speakers (Hanson, 1997; Hanson and Chuang, 1999; Klatt and Klatt, 1990). It is also of common use to use this model in the initialization stage of several source-filter decomposition algorithms working with more complex models (e.g., Ding and Campbell, 1997; Lu, 2002).

2.2.2 Liljencrants-Fant (LF) model

The LF model is a well established and powerful model, capable of characterizing the shape of the derivative glottal wave for a wide range of voices, both in the open and closed phases. Its parameters have been derived from and correlated to physiological and acoustic features (Childers and Ahn, 1995; Gobl, 1989), and researchers from several disciplines have adopted it for tasks such as speaker identification (Plumpe et al., 1999), voice conversion (del Pozo and Young, 2008), singing speech (Kim, 2003; Lu, 2002) among others. The LF model is more flexible than the KLGLOTT88 model from previous section, although this flexibility comes at a price: its formulation is more complicated, which makes its inclusion in optimization algorithms more difficult to accomplish. The mathematical description of the LF model is:

$$g_{lf}(t) = \begin{cases} E_0 e^{\alpha t} \sin(w_g t) & , 0 \leq t \leq T_e, \\ -\frac{E_e}{\epsilon T_a}[e^{-\epsilon(t-T_e)} - e^{-\epsilon(tc-te)} & , T_e < t \leq T_c, \\ 0 & , T_c < t \leq T_0, \end{cases} \tag{2.15}$$

In figure 2.5 a glottal LF cycle is presented, ranging from 0 to the fundamental period T_0. The other time marks are:

- T_p, representing the maximum of the glottal flow (and thus a value of 0 for the derivative),
- T_e, the time instant of the minimum in the derivative,
- T_a, defined as the point where the tangent to the exponential return phase crosses 0,
- T_c the moment when the return phase reaches 0,
- Ee as the absolute value of the minimum of the derivative.

The rest of the parameters in eq. (2.15) (α, w_g, E_0 and ϵ) are computed from the temporal ones by fulfilling some requirements of area balance and continuity (Fant, 1995; Gobl, 2003; Lin, 1990):

$$\int_0^{T_0} g_{lf}(t) = 0 \tag{2.16}$$

$$w_g = \frac{\pi}{T_p} \tag{2.17}$$

$$\epsilon T_a = 1 - e^{-\epsilon(tc-te)} \tag{2.18}$$

$$E_0 = -\frac{E_e}{e^{\alpha T_e}\sin w_g T_e}. \tag{2.19}$$

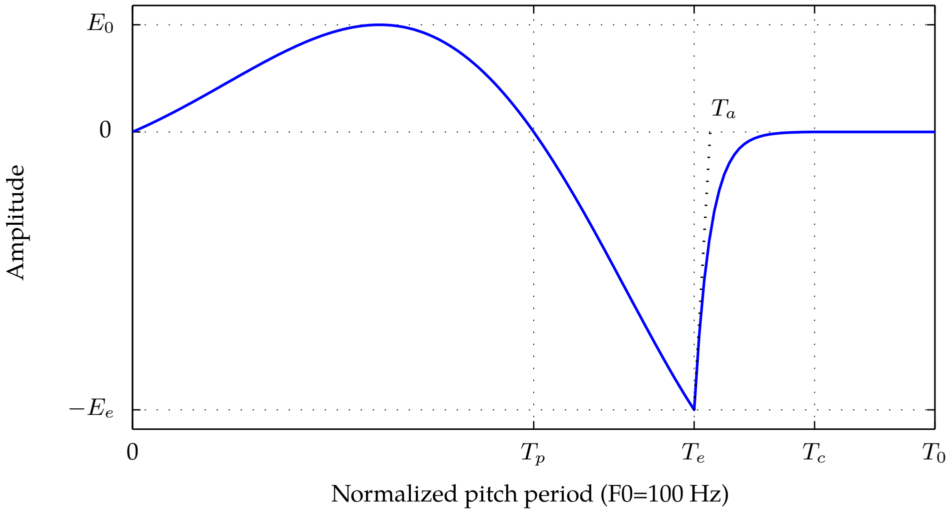


Figure 2.5: *Glottal cycle of the LF model with parameters* $T_p = 5.0, T_e = 7.0, T_c = 8.5, T_a = 0.2$ *and* $T_0 = 10.0$ *in milliseconds, and* $E_e = 2$.*)*

In this thesis we follow the common approach of setting T_c to T_0, effectively working with only three temporal parameters (T_p, T_e and T_a) (Fant, 1995). These time measures are not suitable for direct interpolation, since they are absolute time instants inside cycles of different duration. Furthermore, it is not possible to directly compare pulse shapes of different glottal cycle duration. Fant proposed the use of *extended parameter set*, an alternative representation solving these problems:

$$R_a = \frac{T_a}{T_0}, \tag{2.20}$$

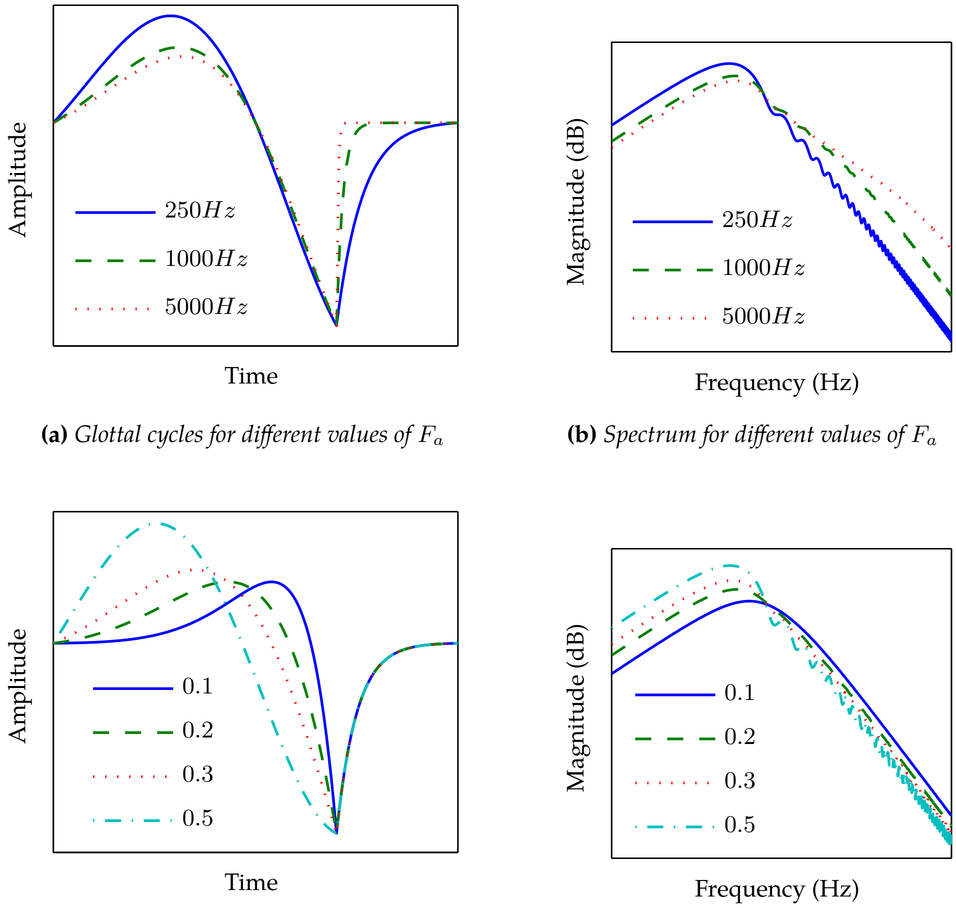
$$R_g = \frac{T_0}{2T_p}, \tag{2.21}$$

$$R_k = \frac{T_e - T_p}{T_p}. \tag{2.22}$$

R_a (or T_a) defines the abruptness of the glottal closure, a measure related to the degree of spectral tilt. This is often made explicit by writing:

$$F_a = \frac{1}{2\pi T_a} = \frac{T_0}{2\pi R_a} \tag{2.23}$$

R_k is the relative duration of the LF model's falling branch (from T_p to T_e), and R_g is inversely proportional to the duration of the opening phase (from 0 to T_p). One of the main advantages of the LF model over the KLGLOTT88 model is its ability to model a wider range of glottal pulse shapes, thanks to the more flexible formulation. As Figure 2.6a shows, it makes possible the use of a variable asymmetry coefficient (ratio of the opening phase over the closing phase) as opposed to the KLGLOTT88 fixed value of $\frac{2}{3}$. Figure 2.6c illustrates the control of the abruptness of the glottal closure using the R_a or equivalent F_a parameter.



(a) *Glottal cycles for different values of* F_a

(b) *Spectrum for different values of* F_a

(c) *Glottal cycles for different values of* R_k

(d) *Spectrum for different values of* R_k

Figure 2.6: *Effect of* R_k *and* F_a *on the glottal cycle and spectrum*

The analytical expression for the spectrum of the LF model was computed in Doval and d'Alessandro (1997):

$$
\begin{aligned}
G_{lf}(f) = E_0 \frac{1}{(\alpha - j2\pi f)^2 + w_g^2} \cdot \Big(& e^{(\alpha - j2\pi f)T_e}((\alpha - j2\pi f)\sin(w_g Te) \\
& - w_g \cos(w_g Te))\Big) + E_e \frac{e^{-j2\pi f Te}}{\epsilon T_a j2\pi f(\epsilon + j2\pi f)} \cdot \\
& \Big(\epsilon(1 - \epsilon Ta)(1 - e^{-j2\pi f(T_0 - T_e)}) - \epsilon Taj2\pi f\Big). \qquad (2.24)
\end{aligned}
$$

Figures 2.6b and 2.6d shows the spectrum of the LF model for both the glottal flow and its derivative for several values of F_a and R_k. As in the KLGLOTT88 case, the spectral behavior of the model can be approximated by a second order low-pass filter of cutoff frequency F_a (Fant, 1995, 1997). Doval and d'Alessandro (1997) computed its exact value using the analytical expression of the spectrum, and demonstrated that the cutoff frequency also depended on R_k and R_g, although for normal values of these parameters the approximation using only F_a worked well.

Up to this point we have introduced the generic characteristics of the glottal flow and its derivative, and we have presented in detail the parametric models of the glottal waveforms that we used in this thesis. Next we will review the state-of-the-art in glottal extraction and parametrization algorithms.

2.3 Source-filter decomposition algorithms

Glottal inverse filtering (IF) encompasses the techniques whose goal is to obtain an estimation of the glottal volume velocity waveform, the source of voicing in speech, using acoustical measures. Although our main interest here lies on the acoustical domain, it is worth mentioning other voice production analysis techniques using different approaches. In glottography, for instance, the dynamics of the vocal folds during phonation are directly recorded using electric (e.g., Henrich et al., 2004; Lecluse et al., 1975) or electromagnetic sensors (e.g., Holzrichter et al., 1998; Titze et al., 2000). Although its application is usually restricted to clinical applications, there are several techniques that perform visual analysis of the vibrating vocal folds: video stroboscopy (e.g., Hirano, 1981), digital high-speed stroboscopy (e.g., Eysholdt et al., 1996) and kymography (Švec and Schutte, 1996), among others. Multi-channel approaches combining some of the previous techniques are also used for specific studies (e.g., Granqvist et al., 2003; Larsson et al., 2000). We will proceed now to review the main methodologies in IF.

2.3.1 Early approaches

Early attempts at inverse vocal tract filters were analog networks, constructed using discrete components, and required arduous manual efforts for fine-tuning. For instance, in Miller (1959) only the first two formants were canceled, resulting in recognizable glottal waveforms, but which contained a considerable amount of ripple during the closed phase. Computers started being incorporated into the task, but still required large amounts of manual intervention. Mathews et al. (1961) identified individual pitch periods using an oscilloscope, and computed the Fourier transform to obtain the speech spectrum using a computer. The spectrum was then analyzed using pole-zero techniques, and source-filter decomposition was achieved by assigning the zeros to the glottal waveform and the poles to the vocal tract. Rosenberg (1971) presented one of the first attempts at parameterizing the glottal waveform. Using the same analysis technique developed by Mathews et al. (1961), Rosenberg incorporated pitch-synchronous resynthesis with synthetic simulations, using different shapes for the glottal waveform (triangular, polynomial, trigonometric and trapezoidal). The resulting speech was then judged in terms of naturalness by means of a listening test. It was concluded that the shapes with a spectral decay of 12 dB/octave produced the most natural sounding speech, consistent with earlier findings (Mathews et al., 1961).

In order to overcome some of the difficulties associated with inverse filtering in those early works (e.g., amplitude calibration, low frequency distortion due to ambient noise, incorrect DC offset level), Rothenberg (1973) introduced the Rothenberg mask, a pneumotachograph mask permitting the direct measurement of the oral volume velocity (as opposed to the speech pressure waveform). The mask allowed for better identification of the glottal closure phases, but was somewhat limited in its frequency response, which reached to only 1 kHz. Despite the limitations, the mask was successfully used in several studies (Holmberg et al., 1988; Karlsson, 1985; Sundberg and Gauffin, 1978). Sondhi (1975) introduced a new inverse-filtering approach using a reflectionless tube with rigid walls to cancel out the vocal contribution, thus allowing a microphone, embedded into the wall of the tube, to directly record the glottal waveform. Monsen and Engebretson (1977) used this equipment to study the variations of the glottal waveform in males and females, across several voice qualities. Sondhi and Resnik (1983) performed vocal tract area function estimations in real time (18 frames per second) using the tube, and were successful in synthesizing intelligible speech from the estimated area functions.

2.3.2 Closed-phase linear prediction

Despite the possible limitations, inverse filtering using closed-phased linear prediction (CPLP) continues to be one of the most popular techniques for glottal waveform estimation. One of the main problems is properly identifying the regions where the glottis is closed. Wong et al. (1979) proposed a pitch-synchronous algorithm using the linear prediction error of the covariance method to detect the closed phase. A similar approach was taken by Childers and Lee (1991), where a two-pass method was used to first identify the regions of closed phase, using a fixed frame LP analysis, and then a pitch synchronous covariance CPLP analysis was performed to estimate an improved filter. Plumpe et al. (1999) presented a technique for automatic estimation and parametrization of the glottal flow, using the lack of modulation of the formant frequencies during the closed phase to identify the regions of glottal closure, applied to the task of speaker identification. Akande and Murphy (2005) presented a new iterative method to improve the estimation of the vocal tract transfer function. Traditional fixed, single-pole pre-emphasis was substituted by a multi-pole (high-pass) filter. Then, the covariance analysis was performed in an adaptive loop, using the phase information of the filter candidates to select the optimal frame position and filter order. Cabral et al. (2007) used a similar CPLP analysis method to integrate the glottal waveform, modeled using the LF model, into a HMM synthesizer. The results of the evaluation tests showed that the glottal source resulted in more natural speech than the previously used pulse train source.

2.3.3 Joint decomposition

To avoid some of the problems associated to closed-phase linear prediction (e.g., being dependent on the correct location —or existence— of the glottal closed phase, analysis of high pitched voices) Alku (1992) proposed an iterative method (IAIF) to obtain the glottal waveform. A gross initial estimation of the glottal waveform was initially obtained and removed from the speech signal, using a first order LP analysis (like an adaptive pre-emphasis filter). Next, the vocal tract was estimated using a higher order LP analysis on the pre-emphasized speech, and a new estimation of the glottal waveform was obtained by inverse-filtering the original speech with this filter. This process was iterated using different orders for the LP analysis to obtain the final, refined estimation of the glottal waveform. In a following work (Alku et al., 2004), the discrete all-pole modeling (DAP) algorithm (El-Jaroudi and Makhoul, 1991) was adopted, instead of LP analysis. The DAP method uses the Itakura-Saito distortion measure to model the discrete spectral envelope of voiced speech. El-Jaroudi and Makhoul showed that DAP modeling results in better all-pole spectral modeling, and also provided an extended version of the algorithm using frequency weighted error functions. In the modified IAIF, the speech was high-pass filtered to eliminate low-pass distortions, and then inverse filtered using a first-order all-pole filter, estimated using DAP. A DAP analysis was performed on the output of this stage, the input speech was inverse filtered using the resulting filter, and the result was integrated to obtained an estimation of the true glottal flow. This process was again repeated twice with a different order DAP analysis stage, to further refine the estimated glottal flow. The results obtained showed an encouraging small error between the synthetic (real) and estimated results. The "simultaneous inverse filtering and model matching" method proposed by Frölich et al. (2001), also based on the DAP algorithm introduced above, included the spectrum of the parametric LF model into the iterative DAP algorithm. All glottal measures were then computed from the LF parameters, avoiding the use of the noisy inverse filtered signal. The authors concluded that the algorithm performed with a high accuracy when analyzing synthetic data, although natural data presented more difficulty. Lu (2002) investigates the parametrization of singing speech, using the KLGLOTT88 model to perform an initial joint estimation using convex optimization methods. This is further refined using the LF model and a non-linear optimization stage. Kim (2003) uses a modified version of this approach introducing a warping factor in the filter coefficients in order to improve the accuracy of the matching filter. A similar approach was used in del Pozo (2008) for the purposes of voice source and duration modelling, applied to the tasks of voice conversion and speech repair.

2.3.4 Pole-zero modeling

There are several authors using more advanced production models for speech, where both zeros and poles are used to model the vocal tract. Ding et al. (1997) used and autoregressive model with an exogenous input (ARX) for the speech production process:

$$s(n) + \sum_{i=1}^{p} a_i(n)s(n-i) = g(n) + \sum_{j=1}^{q} b_j(n)g(n-j) + e(n). \tag{2.25}$$

The (observed) speech signal and the (unknown) glottal waveform at time n are represented by $s(n)$ and $g(n)$ respectively. $e(n)$ agglutinates both the input noise and the model estimation error. The joint estimation is performed by means of simulated annealing (SA) and Kalman filtering, using the mean-square estimation error criterion. The iterative algorithm uses SA to select a new set of KLGLOTT88 parameters, and Kalman filtering to compute the corresponding ARX filter coefficients. This continues until the system converges to a (global) minimum. They include a final step for correcting the order of the model (i.e., the number of poles) using a formant tracking algorithm to discard any ghost formants. The method is validated by analyzing a two-channel speech database, and achieving reliable estimates of the different parameters (voice source and vocal tract). The same decomposition algorithm was also used to improve the unit selection module in the CHATR speech synthesizer (Ding and Campbell, 1997).

According to Ohtsuka and Kasuya (2000), the previous method presented by Ding et al. suffers from two main drawbacks. First, the Kalman filtering provides a new set of coefficients on a point-by-point basis, and they average all the valid formant values to obtain a single set per pitch period. Furthermore, analyzing female voices with a high pitch presents difficulties, as well as weak voiced consonants. Ohtsuka and Kasuya presented an improved algorithm based on the previously reviewed one. To solve the first problem, they propose using a least-square method to reduce the Kalman coefficients to a single filter per period. They claim that the second problem owes to the formant tracking algorithm introducing spectral distortion by simply excluding the roots of the filters not associated with valid formants. Instead, they include an adaptive prefilter to compensate for the effects of the disregarding roots on the spectral tilt. Dynamic programming is also introduced to improve the formant tracking. According to their perceptual experiments performing analysis-synthesis of a natural speech sentence, their algorithm outperforms both the method presented in Ding and Campbell (1997) and the mel cepstral method (Tokuda et al., 1994).

Funaki et al. (1997) propose a combination of a time varying ARMA exogenous model with glottal excitation together with white Gaussian noise. The MIS method is used to estimate the ARMAX filter coefficients, and an hybrid approach using genetic algorithms (GA) and simulated annealing (SA) is employed for estimating the parameters of the KLGLOTT88 model (Klatt and Klatt, 1990). Funaki et al. (1997) report that both natural and synthetic speech can be accurately analyzed with the proposed algorithm. One of the main drawbacks of this method is the high computational load. For this reason, they further introduce the use of Haar or QMF filter-banks for sub-band processing (Funaki et al., 1998). This not only reduces the computation time, but also achieve a more accurate estimation of the glottal source model parameters due to the improved frequency resolution.

All the methods so far reviewed use the KLGLOTT88 parametric model although it does not model certain types of voice as well as, say, the LF model. The main reason to choose this model, is that it presents a simpler formulation, more suitable for certain optimization procedures. Fu and Murphy (2004) presented a joint source-filter estimation method based on the LF model for the glottal source. They separated the joint estimation procedure in two parts, stating that if the parametric glottal flow derivative were known, the ARX could be uniquely obtained using the Kalman algorithm. Thus, they proposed an iterative procedure where at each step, the LF parameters are refined using a descent algorithm based on the interior trust region method (Coleman and Li, 1996), and where the filter coefficients were updated applying the Kalman algorithm. Since the problem is not convex, they needed a good starting point in order to avoid getting stuck into a local minimum. They used the procedure explained in (Strik, 1998) to obtain an initial KLGLOTT88 model. The parameters were then directly mapped into the corresponding LF ones. Fu and Murphy reported the method to be accurate and robust both for synthetic and natural speech. In their previous, more detailed report, Fu and Murphy (2003) compared their proposed method with the closed-phase linear prediction (CPLP) and the iterative adaptive inverse filtering (IAIF) methods. They reported a similar performance to that of CPLP, and a better one than the obtained with IAIF. One of the main advantages of their method is that no closed-phase information (quite often difficult to obtain) is needed. Furthermore, separation of non-parametric inverse filtered glottal signal and glottal noise is inherent to the algorithm.

2.3.5 Spectral methods

So far, we have only dealt with methods using a time-domain parametrization of the voice source. However, several authors have reported evidence that for the estimation of some parameters it would improve the accuracy to do it in the frequency domain (Childers and Ahn, 1995; Childers and Lee, 1991; Fant and Lin, 1998). Arroabarren and Carlosena (2003b) present an inverse filtering technique using the analytical expression of the spectrum of the KLGLOTT88 model. They analyze the speech signal using a hamming window to frame three or four pitch periods. They compute the several glottal spectrum candidates using different values of the open quotient (OQ) parameter, and subtract each of them from the short time spectrum of the speech signal. DAP modeling is then applied to obtain the corresponding set of coefficients of the all-pole vocal tract filter. From the different candidates for the OQ, they choose the one resulting in minimum formant ripple when inverse filtering the speech signal.

2.3.6 Mixed-phase algorithms

A different sort of method are those relying on the mixed-phase model of speech, in which speech can be divided in both minimum-phase, or causal, and maximum-phase, or anticausal, components (Bozkurt and Dutoit, 2003). Whereas the vocal tract response and the return phase of the glottis can be considered as minimum-phase signals, the glottal open phase has been proved to be a maximum-phase signal (Doval et al., 2003). In these algorithms, both the window used to extract speech frames for the analysis (due to the phase-based nature of the algorithm) and the GCI location play an crucial role in the success of the decomposition (Bozkurt et al., 2005; Drugman et al., 2009a).

One of the main techniques to achieve this mixed-phase decomposition uses the zeros of the Z-transform (ZZT) (Drugman et al., 2012). The Z-transform $X(z)$ of a series of n samples $x(n)$ of a signal ($n = 1 \dots N$):

$$X(z) = \sum_{n=1}^{N} x(n)\, z^{-n}, \tag{2.26}$$

Assuming the signal samples are obtained using a carefully designed and located window, to avoid phase distortion, the roots (zeros) of $X(z)$ are computed: those falling outside the unit circle are assigned to the glottal open phase (anticausal component), and the rest to the vocal tract (causal component) (Bozkurt et al., 2005). The ZZT method has been shown to outperform other approaches of glottal flow estimation (Sturmel et al., 2007).

The other main technique is based on the complex cepstrum decomposition (CCD) (Drugman et al., 2009a). The complex cepstrum $c_x(n)$ of a signal $x(n)$ can be formulated as (Proakis and Manolakis, 1996):

$$c_x(n) = \frac{1}{2\,\pi} \int\limits_{-\pi}^{\pi} \log X(\omega)\, e^{j*\omega*n} d\omega, \tag{2.27}$$

where $X(\omega)$ is the Fourier transform of $x(n)$. The source-filter decomposition is used by using the following propriety: when $x(n)$ is a causal signal, $c_x(n) = 0$ for $n < 0$; conversely, if $x(n)$ is anticausal, $c_x(n) = 0$ for $n > 0$. If only the negative indexes of the complex cepstrum are kept, then it is possible to obtain the glottal contribution. (Drugman et al., 2012) performed a comparative study of glottal source estimation techniques, obtaining similar results for CCD and CPLP using clean speech (the performance degraded when analyzing noisy speech).

2.4 Glottal Quality Measures

Evaluating the performance of an IF algorithm is very difficult, since there are no real (correct) values that could be used as *reference*. For certain applications it is possible to assess it using manual inspection, but for others that is unfeasible. Furthermore, only defining what the quality of an estimated glottal waveform should be is not an easy task: even among trained scientists there is often a lack of consensus regarding this matter. Many of the proposed algorithms evaluate the performance using synthetic data generated using known glottal and vocal tract data, according to the speech production model adopted in the study. This way, those values can be used as reference and a numerical evaluation of the decomposition is possible. Nevertheless, the validity of the results obtained with this approach could be questioned, since the same model is normally used both in the analysis and synthesis stages. To overcome this limitations, Alku et al. (2004) proposed the use of a physical modeling of voice production (vocal folds vibration and acoustic wave propagation) to generate the synthetic vowels used as testing references (Story and Titze, 1995; Titze, 2002; Titze and Story, 2002).

There are, however, some glottal quality measures (GQM) based on the characteristics of a theoretical glottal waveform that could be of help here. Although its practical application in automatic glottal waveform estimation techniques has not yet been successful (Moore and Torres, 2008), it is worth reviewing them here.

2.4.1 Group delay phase

In (Alku et al., 2005) it was proposed to use the group delay function (GDF), the negative derivative of the spectrum phase, to assess the quality of the inverse filtering. When computed over a single glottal cycle, the GDF is almost a linear function of frequency over the largest part of the frequency range. Thus, if the estimation of the vocal tract is accurate and its poles are correctly canceled, the inverse filtered signal (estimated glottal waveform) should also present this behavior. Figure 2.7 shows the GDF computed for the estimated glottal waveform obtained with two different candidates for the glottal epochs.

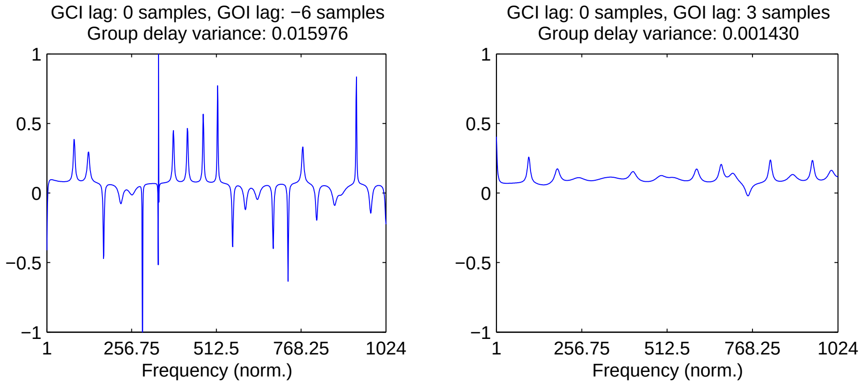


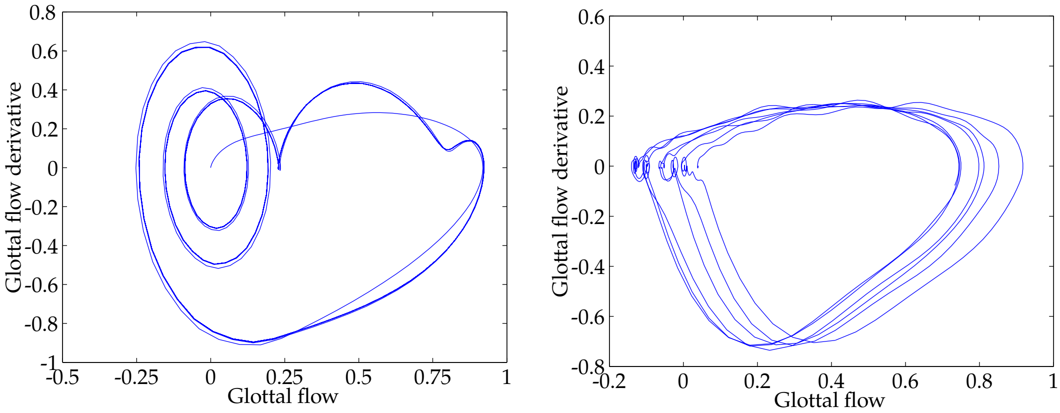
Figure 2.7: *Group delay computed for two candidate pairs (gci, goi). The left side presents a higher variance (i.e., a worst inverse filtering).*

2.4.2 Phase-plots

Visual (subjective) inspections of phase-plots were proposed in Edwards and Angus (1996) as a tool to evaluate the quality of estimated glottal waveforms. Bäckström et al. (2005) defined two objective measures based on phase-plane analysis. Since the glottal flow waveform can be modeled using a second-order polynomial, its plot on the phase-plane $(g(t), dg/dt)$ consists of one (closed) loop per glottal cycle. Errors in the estimation of the VT filter resulting in non-completely removed resonances, have the effect of secondary loops in the phase-plane, as shown in Fig. 2.8. They proposed two measures that quantify this effect:

- pp_{cper}, number of cycles per fundamental period. This number should ideally be 1, with results higher than 2 indicating the presence of formant ripple in the estimation. It is often necessary to threshold this ratio in the range $[1, 2]$ to increase the robustness.

- pp_{cyc}, mean sub-cycle length. The size of the sub-cycles is directly proportional to the amount of formant ripple still present in the estimation. The lower the length of the sub-cycles, the better the estimation is.



(a) **(b)**

Figure 2.8: *Phase-plots of two glottal candidates: the candidate in (b) (*$pp_{cper} = 1.733$*,* $pp_{cyc} = 0.0054$*) is of higher quality than that of (a) (*$pp_{cper} = 3.481$*,* $pp_{cyc} = 0.231$*). This comes from an extreme test case scenario where two of the vocal tract resonances where not removed.*

2.4.3 Kurtosis

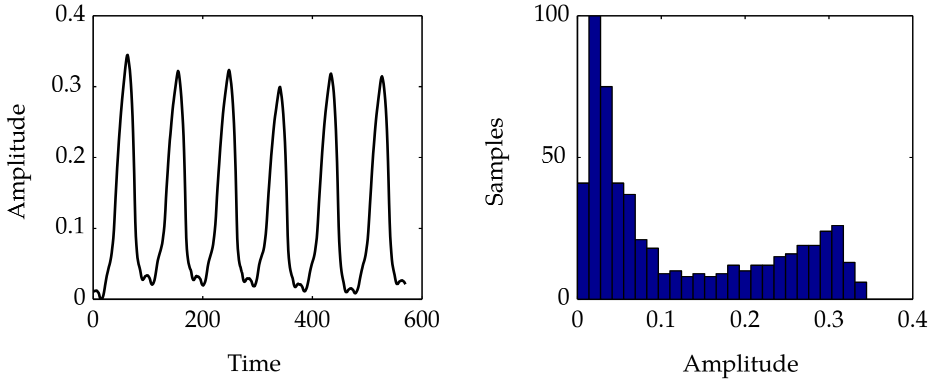
The kurtosis is a measure of similarity of a given distribution to the Gaussian distribution. The range of the kurtosis measures is $[-3, \inf]$, where positive and negative values correspond to narrower and wider peaks of the distribution, respectively (the kurtosis of a Gaussian distribution is zero). Bäckström et al. (2005) proposed it as a glottal quality measure, arguing that since the glottal glow is a subgaussian waveform by nature (it has two distinct peaks as shown in Fig. 2.9), a lower kurtosis should indicate a better estimation.

2.4.4 Harmonics ratio

Moore and Torres (2006, 2008) proposed a GQM measure based on the mean ratio of the first harmonic peak to other peaks, computed over two different frequency ranges. The motivation is that the glottal waveform should present a strictly negative spectral slope, since any resonant structure is removed by the inverse filtering process. Three measures based on this property of the glottal waveform were proposed:

- hr_{mn}, mean ratio of the first harmonic peak to other peaks,

- hr_{mx}, ratio of the first harmonic peak to the maximum pea,



(a) *Inverse-filtered glottal flow waveform* **(b)** *Histogram of the glottal flow waveform*

Figure 2.9: *Glottal flow waveform and corresponding histogram for the kurtosis measure*

- linear regression R^2 statistic of the log-spectral peaks, computed as

  $$R^2_{(X)} = \frac{\sum_{i:0<f_i<X} (\hat{v}_i - \bar{v})^2}{\sum_{i:0<f_i<X} (v_i - \bar{v})^2},$$

  where v represents the logarithmic magnitude of the spectral peaks, f contains their frequencies (Hz), and $\hat{v}$ is the best linear fit to v. The magnitudes are normalized using the mean value $\bar{v}$.

All these measures are computed over two different frequency ranges X, one comprising only the lower part of the spectrum (0–1000 Hz), and a second one extending from 0–3500 Hz).

2.5 Aspiration noise

So far, we have only dealt with the excitation of the vocal tract due the vibration of the vocal folds. During the production of unvoiced sounds, and in a lesser measure also during voicing (aspiration noise), a turbulent flow acts as the excitation source. Aspiration noise, although not so extensively research, has been proved to contribute to the naturalness of synthetic speech. From turbulent flow theory Cook (1991), the sound pressure of the turbulent noise is proportional to the square of the volume-velocity of airflow, and inversely proportional the cross-sectional area of the glottal constriction. Cook has calculated the likelihood of the existence of aspiration (turbulence) noise, and reported that the likelihood of turbulence exists during the whole open phase. Maximum sound radiation power is achieved just when the closing-phase of the glottis begins. He also concluded that another high power burst of noise is likely to occur at glottal opening instant, due to highly pressurized air passing through a small aperture. Hence, two pulses of noise per glottal cycle are to be expected, occurring at the beginning of the opening and closing phases.

Klatt (1980); Klatt and Klatt (1990) used two different noise sources in the software implementation of a formant synthesizer: aspiration source and frication source. Both were defined in the same way, although they represented noise generated at different locations: a random Gaussian-noise generator, modulated in amplitude by the fundamental frequency $f0$, and a low-pass filter (in practice canceled by the lip-radiation filter). Klatt used a square waveform of period equal to the fundamental period, with a fixed degree of amplitude-modulation (50%). The amplitude of the noise waveform was determined with an independent parameter.

Hermes (1991) simulated breathy vowels by combining low-pass filtered pulse trains and de-emphasized, high-pass filtered noise bursts. The frequency of the noise burst was equal to the pulse train frequency ($125\,\mathrm{Hz}$ in the experiments). They varied the phase difference between the pulses and the noise bursts (between 0 and 2π) to study whether the two signals were integrated or perceived as segregated. They concluded that in order to contribute to adding breathiness to the synthetic vowels, the noise bursts should be synchronized with the pulse trains, should not have excessive peaks and their energy should be about equal in each pitch period. They chose a cut-off frequency of the high-pass filter between $1200\,\mathrm{Hz}$ and $2000\,\mathrm{Hz}$, with lower values resulting in a greater degree of breathiness. The de-emphasis filter was a first order low-pass filter, with a pole at 0.9. Childers and Ahn (1995); Childers and Lee (1991) used a similar approach, without de-emphasis.

Lu (2002) neglected the existence of noise bursts at the opening points of the glottis, and modeled the aspiration noise using two components: and additive white Gaussian noise to represent the noise floor of the aspiration noise (constant leakage of air through the glottis), and a pitch synchronous amplitude modulated Gaussian noise. Lu used a Hanning window for each glottal period, centered around the glottal closing point.

Mehta and Quatieri (2005) studied the use of a modulated noise signal in their speech production model. They extracted the noise component from the speech signal using a decomposition method based on the sinusoidal/harmonic model. A linear prediction step was then performed to estimate the vocal tract filter. After inverse filtering the noise signal (whitening process), they applied a Hilbert transform followed by a low-pass filtering in order to extract the envelope. Pitch modification was then performed on the estimated envelope by means of resampling, and the result was then filtered using the vocal tract coefficients to obtain the pitch-modified noise signal.

Matthews et al. (2006) studied the synthesis of breathiness in natural speech by means of sinusoidal modeling and modulated noise. The harmonics in the sinusoidal model were used to modulate low-pass filtered white noise, and both amplitude and phase modulation results were presented. They used three parameters: the amplitude of the modulated noise, the bandwidth of the low-pass filter, and the lower cutoff frequency (i.e. minimum harmonic) that was modified. They reported a moderate success in adding breathiness to the synthetic speech, by measuring with a MOS test the number of sentences the listeners would rate as *breathy*.

Kim (2003) took a stochastic codebook approach and used the glottal derivative residuals to train a codebook using Principal Component Analysis (PCA) (Duda et al., 2000). PCA uses the eigenvectors and eigenvalues of the covariance matrix of the data. Since PCA requires all the input vectors to be of equal size, the compute the Fast Fourier Transform for each glottal residual prior to the PCA analysis. They were able to reduce the dimensions of the codebook by ignoring the largest eigenvectors, those capturing the lowest amount of variance among the input data.

2.6 Proposed approach

For this work we have decided to follow a time-based approach in which the source-filter decomposition is done using the temporal expression of the glottal models. We need a robust algorithm that could readily be applied to large amounts of data, since our main goal is the characterization of the voice source for the purpose of performing prosodic and spectral manipulations. This is necessary for including the parametrization algorithm in speech synthesis and voice conversion tasks. In our method, we use the two glottal models explained before: the KLGLOTT88 model, and the LF model. The relatively simple formulation of the former allows us to include it in a convex optimization step to obtain an optimal estimation of the vocal tract parameters. The estimated vocal tract is then used to obtain the glottal waveform by inverse-filtering the speech signal. The more powerful LF model is then fitted to this glottal waveform by means of a minimization algorithm, that due to the LF formulation is non-linear by nature and thus require a good initial estimation to guarantee convergence to the optimal solution. This initial estimate is obtained by using the previously obtained KLGLOTT88 model parameters. In next chapter we will explain in detail both the analysis and synthesis methods.

CHAPTER 3 Speech production system parametrization

In this chapter we present our complete algorithm for the parametrization of natural speech using the speech production system proposed by Fant, as seen in Sec. 2.1. First we introduce the main blocks of the algorithm, and then we proceed to explain in detail each of them.

3.1 Algorithm outline

Our algorithm is based on the idea that, knowing the opening and closing instants of a glottal cycle, we can decompose a single cycle of speech using the source-filter approach to obtain the coefficients of the all-pole vocal tract filter and the KLGLOTT88 glottal model using a convex minimization algorithm. This is the idea around which the different parts of our algorithm revolve. Thus, we start by giving the details of the convex decomposition in Section 3.2. The algorithm we have designed around this decomposition consists of the four main blocks shown in Figure 3.1 and explained below.

In the first block, explained in Section 3.3, we deal with the extraction and optimization of the glottal epochs (instants of glottal opening and closing). We follow a dual-channel approach using both the speech signal and a simultaneously recorded laryngograph signal, since it is available in all of our databases, to obtain an initial estimation of the glottal epochs. This is explained in detail in Section 3.3.1. Next, we proceed to optimize these initial locations (Section 3.3.2) by first performing a global synchronization step, which accounts for the time lag due to the different nature of the two signals being recorded (Section 3.3.2), and then a separate optimization of the individual GCI and GOI points as explained in Sections 3.3.2 and 3.3.2.

Once we have the optimal set of glottal epochs, we proceed to perform the source-

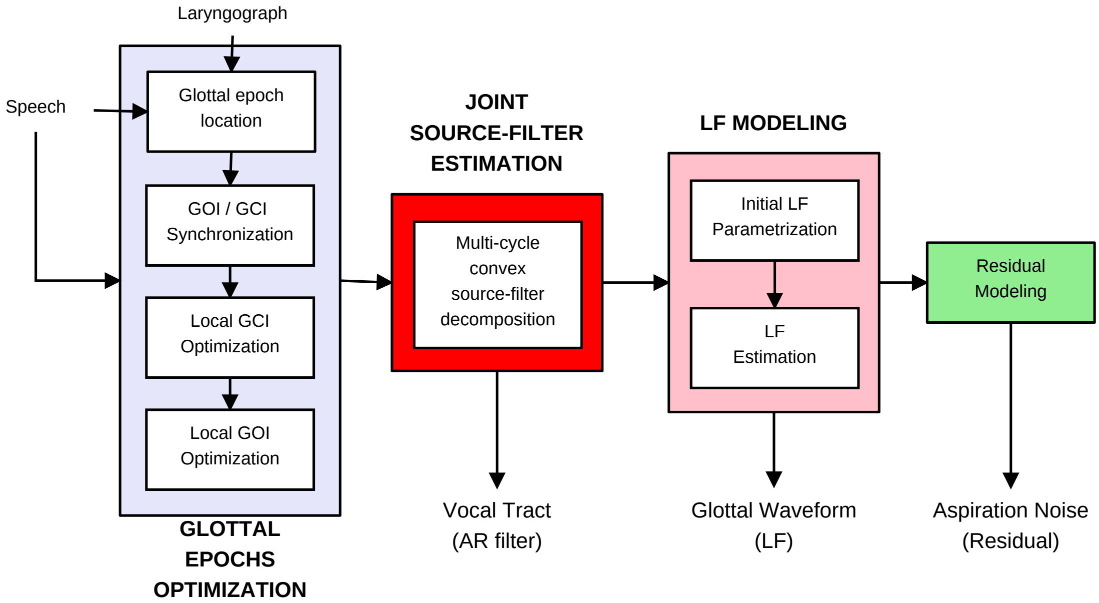

Laryngograph
Speech
Glottal epoch location
GOI / GCI Synchronization
Local GCI Optimization
Local GOI Optimization
GLOTTAL EPOCHS OPTIMIZATION
JOINT SOURCE-FILTER ESTIMATION
Multi-cycle convex source-filter decomposition
Vocal Tract (AR filter)
LF MODELING
Initial LF Parametrization
LF Estimation
Glottal Waveform (LF)
Residual Modeling
Aspiration Noise (Residual)

Figure 3.1: *Block diagram of our proposed algorithm for source-filter decomposition.*

filter estimation. This is done by extending the single-cycle decomposition algorithm of Section 3.2 to work with multiple, adjacent cycles. We explain this in detail in Section 3.4.

The next block deals with the use of the more advanced LF glottal model to obtain a better representation of the glottal source. As will be explained in Section 3.5, this is involves a non-linear least squares minimization.

In the last step of our algorithm, we extract the aspiration noise present in the residual resulting from the previous LF parametrization. Section 3.6 first outlines the aspiration noise characteristics, and then explains the details of the extraction and parametrization process.

3.2 Convex source-filter decomposition

As we have seen in Chapter 2, Section 2.1, in our speech production model the speech signal $s(n)$ is the result of filtering the true glottal waveform $g(n)$, using an all-pole filter $\frac{1}{A(z)}$ with coefficients a_k to model the vocal tract effect:

$$S(z) = \frac{1}{A(z)} G(z) = \frac{1}{1 - \sum_{k=1}^{N} a_k z^{-1}} G(z) \quad \text{frequency domain,} \tag{3.1}$$

$$s(n) = g(n) + \sum_{k=1}^{N} a_k s(n-k) \quad \text{time domain.} \tag{3.2}$$

Let us suppose that we are able to estimate the filter coefficients (a_k). We can then estimate the inverse-filtered glottal waveform g_{if} by removing the effect of the vocal tract from the speech signal:

$$G_{if}(z) = A(z)\,S(z) = \left(1 - \sum_{k=1}^{N} a_k z^{-1}\right) S(z), \tag{3.3}$$

$$g_{if}(n) = s(n) - \sum_{k=1}^{N} a_k s(n-k). \tag{3.4}$$

For the convex decomposition we will use the KLGLOTT88 model from Section 2.2.1 to model g_{if} above. Recall that the return phase (tilt) of the glottis is controlled using a low-pass filter:

$$TL(z) = \frac{1}{1 - \mu z^{-1}}. \tag{3.5}$$

We can separate the inverse-filtered glottal waveform G_{if} from eq. (3.3) into its *untilted* component $\tilde{G}_{if}$ and the tilt filter:

$$G_{if}(z) = \tilde{G}_{if}(z) \cdot \frac{1}{1 - \mu z^{-1}}, \tag{3.6}$$

$$g_{if}(n) = \tilde{g}_{if}(n) + \mu\, g_{if}(n-1) \tag{3.7}$$

$$\tag{3.8}$$

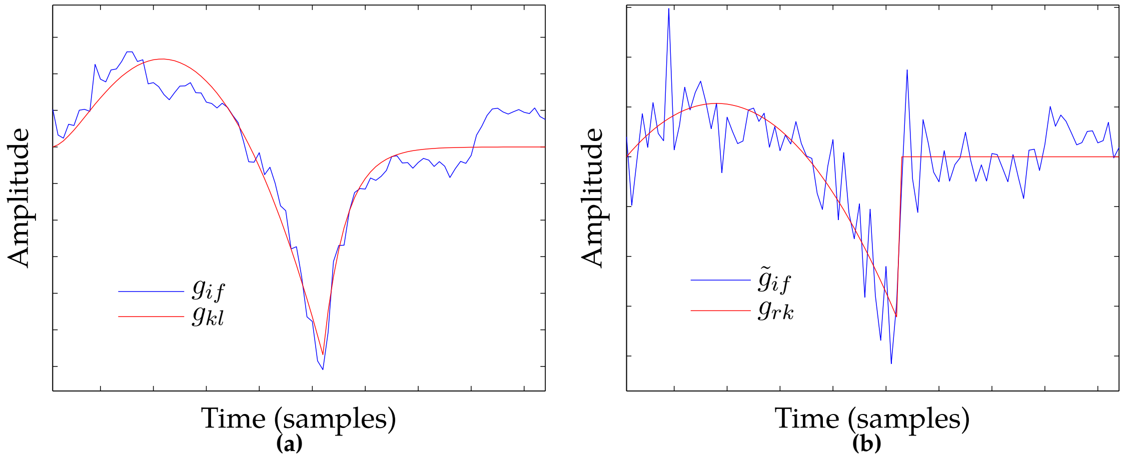
Since both the tilt $TL(z)$ and the vocal tract $\frac{1}{A(z)}$ filters are linear, we will combine them into an *extended* vocal tract filter and compute μ as part of the vocal tract estimation. This approach is similar to that in Lu (2002) or Kim (2003), as opposed to the a priori tilt elimination using pre-emphasis used in Childers and Lee (1991) or del Pozo (2008) for instance. Combining eqs. (3.3) and (3.6), we then have:

$$\begin{aligned}\tilde{G}_{if}(z) &= \left(1 - \mu z^{-1}\right)\, G_{if}(z) \\ &= \left(1 - \mu z^{-1}\right) \left(1 - \sum_{k=1}^{N} a_k z^{-1}\right) S(z) \\ &= \left(1 - \sum_{k=1}^{N+1} \tilde{a}_k z^{-1}\right) S(z), \end{aligned} \tag{3.9}$$

where the filter order has been increased from N to $N+1$ to accommodate for the extra parameter μ it now includes. In its time domain form, eq. (3.9) is written as:

$$\tilde{g}_{if}(n) = s(n) - \sum_{k=1}^{N+1} \tilde{a}_k s(n-k). \tag{3.10}$$

Figure 3.2 shows the differences between g_{if} from eq. (3.4) and $\tilde{g}_{if}(n)$ from the equation above, and their approximations using the KLGLOTT88 and Rosenberg-Klatt models respectively (with and without tilt effect).



(a) **(b)**

Figure 3.2: Real *inverse-filtered* g_{if} *and parametric KLGLOTT88* g_{kl} *glottal waveforms (a), vs* Untilted *inverse-filtered and estimated Rosenberg-Klatt* g_{rk} *glottal waveforms (b).*

The main idea behind our decomposition algorithm is that using an all-pole filter to model the vocal tract, and the Rosenberg-Klatt parametric model to represent the glottal waveform, results in the estimation of the filter coefficients and the glottal waveform parameter (amplitude) being a convex optimization problem (Lu, 2002). Since we want to model the glottal waveform $\tilde{g}_{if}$ using the Rosenberg-Klatt model from Section 2.2.1, we can rewrite (2.8) to emphasize its linear dependence on the amplitude parameter b. During the glottis' open phase we then have (in terms of $t = n/f_s$, where f_s is the sampling frequency):

$$\begin{aligned} g_{rk}(n) &= 2\,a\,\frac{n}{f_s} - 3\,b\left(\frac{n}{f_s}\right)^2 \\ &= b\,\frac{n}{f_s^2}\left(2\frac{a}{b}f_s - 3n\right) \\ &= b\,\frac{n}{f_s^2}\,(2N_{op} - 3n) \\ &= b\,C(n) \qquad 1 \le n \le N_{op}, \end{aligned} \tag{3.11}$$

where we have used (2.11) to set the duration N_{op} (in samples) of the open phase as:

$$N_{op} = \frac{a}{b} f_s = O_q\,T_0\,f_s, \tag{3.12}$$

As we can see, with this formulation, the Rosenberg-Klatt model requires only one amplitude-related parameter to be estimated (b) since

$$C(n) = \frac{n\,(2N_{op} - 3n)}{f_s^2} \tag{3.13}$$

depends only on known parameters (given that the glottal instants GOI/GCI are obtained outside this optimization step): the sample index n, the sampling frequency f_s, and the duration of the open phase N_{op}.

We can calculate the glottal source parametrization error using our Rosenberg-Klatt model by means of (3.10) and (3.11) as:

$$\begin{aligned} e(n) &= g_{rk}(n) - \tilde{g}_{if}(n) \\ &= \begin{cases} b \cdot C(n) + \sum_{k=1}^{N} \tilde{a}_k s(n-k) - s(n) & \text{open phase,} \\ \sum_{k=1}^{N} \tilde{a}_k s(n-k) - s(n) & \text{closed phase.} \end{cases} \end{aligned} \tag{3.14}$$

This is the error that needs to be minimized by a proper selection of both the filter coefficients $\tilde{a}_k$ and the KLGLOTT88 amplitude b. We will show below that when we minimize the L_2 norm of the error, this minimization is a *convex optimization* problem, guaranteed to have only one minimum (i.e., the optimal solution) (Boyd and Vandenberghe, 2004). In particular, it pertains to the family of *Quadratic Programming* (QP) mathematical optimization problems, where a quadratic function $f(\mathbf{x})$ of several variables is optimized (minimized or maximized) w.r.t. $\mathbf{x}$ subject to linear constraints on these variables. We could use L_1 or L_∞ and solve the problem using Linear Prediction methods, but choosing an L_2 norm results in AR filters less prone to instabilities (Lu, 2002). The QP problem can be formulated as:

$$\min_{\mathbf{x}} f(\mathbf{x}) = \min_{\mathbf{x}} \frac{1}{2}\mathbf{x}^T\mathbf{Q}\mathbf{x} + \mathbf{f}^T\mathbf{x}, \tag{3.15}$$

subject to one or more linear constraints:

$$\mathbf{A}\mathbf{x} \leq \mathbf{c} \qquad \text{inequality constraint,} \tag{3.16}$$

$$\mathbf{E}\mathbf{x} = \mathbf{d} \qquad \text{equality constraint,} \tag{3.17}$$

where $\mathbf{x} \in \Re^n$, $\mathbf{Q}$ is a $n \times n$ symmetric matrix, and $\mathbf{f}$ is a $n \times 1$ vector. If $\mathbf{Q}$ is also *positive semi-definite*, then $f(\mathbf{x})$ is a *convex function*, and the quadratic program has a global unique minimum. The solution of this type of problems is well known[1] and it exists in closed from due to the convex properties. We will see next that this is the case for our problem, so the solution will be a global optimum point. For this work we have chosen a trust region method implemented in Matlab (Boyd and Vandenberghe, 2004), since they have

[1] There are many implementations of these algorithms already available and listed in the Neos Optimization Guide, available on-line at `http://www.mcs.anl.gov/otc/Guide/`

good convergence properties. It is based on the concept of approximating the function to be minimized with a simpler one on a region around the current local solution and use a gradient algorithm to iteratively improve the local solution. The trust region is updated or corrected depending on whether or not the new solution results in an improvement. The process is repeated until convergence is achieved.

We will use matrix notation to write down the error (3.14) for the whole glottal cycle:

$$\begin{aligned}\mathbf{e} = \begin{pmatrix} e(1) \\ e(2) \\ \\ \vdots \\ \\ e(P) \end{pmatrix} &= \begin{pmatrix} C(1) & s(0) & \cdots & s(-N) \\ C(2) & s(1) & & s(-N+1) \\ \vdots & \vdots & & \vdots \\ C(N_{op}) & s(N_{op}-1) & & s(N_{op}-1-N) \\ 0 & s(N_{op}) & & s(N_{op}-N) \\ \vdots & \vdots & & \vdots \\ 0 & s(P-1) & \cdots & s(P-N-1) \end{pmatrix} \begin{pmatrix} b \\ \tilde{a}_1 \\ \\ \vdots \\ \tilde{a}_{N+1} \end{pmatrix} - \begin{pmatrix} s(1) \\ s(2) \\ \\ \vdots \\ \\ s(P) \end{pmatrix} \\ &\equiv \mathbf{F}\,\mathbf{x} - \mathbf{y} \end{aligned} \tag{3.18}$$

where P the cycle length and

$$\mathbf{x} = [b \;\; \tilde{a}_1 \; \cdots \; \tilde{a}_{N+1}]^T \tag{3.19}$$

is the vector containing the parameters to be estimated. We can now formulate the minimization of the L_2 norm of the error (3.18) in matrix notation:

$$\begin{aligned}\min_{\mathbf{x}} ||\mathbf{e}||^2 &= \min_{\mathbf{x}} \;\; ||\mathbf{F}\mathbf{x} - \mathbf{y}||^2 \\ &= \min_{\mathbf{x}} \;\; \mathbf{x}^T\mathbf{F}^T\mathbf{F}\mathbf{x} + \mathbf{y}^T\mathbf{y} - 2\mathbf{y}^T\mathbf{F}\mathbf{x} \\ &= \min_{\mathbf{x}} \;\; \mathbf{x}^T\mathbf{F}^T\mathbf{F}\mathbf{x} - 2\mathbf{y}^T\mathbf{F}\mathbf{x}, \end{aligned} \tag{3.20}$$

since $\mathbf{y}^T\mathbf{y}$ is independent of $\mathbf{x}$. Comparing equations (3.20) and (3.15), we can rewrite our problem to follow the QP formulation using the following identities:

$$\mathbf{Q} = 2\,\mathbf{F}^T\,\mathbf{F} \tag{3.21}$$

$$\mathbf{f} = -2\,\mathbf{F}^T\,\mathbf{y}. \tag{3.22}$$

The solution of (3.20) needs to be constrained in order to be physically meaningful. For example, the amplitude b of the glottal pulses needs to be positive and the resulting filters should be stable.

3.2.1 Constraints for the convex formulation

The solution to the convex optimization problem must fulfill the following restrictions: the resulting vocal tract filter must have a resonator characteristic, the filter must be stable, and the spectral tilt must have the characteristics of a low-pass filter. Since it is not possible to directly enforce these restrictions without breaking the convexity of the problem, we will make some compromises. The filter response of the vocal tract is the result of concatenating a series of digital resonators, second-order filters with a transfer function given by

$$R(z) = \frac{d_1}{1 - c_1\, z^{-1} - c_2\, z^{-2}}. \tag{3.23}$$

The frequency response of a single resonator is shown in figure 3.3. The coefficients d_1, c_1 and c_2 are related to the *formant frequency* F_f and *bandwidth* B_W:

$$c_2 = -e^{-2\pi B_W/f_s} \tag{3.24}$$

$$c_1 = 2e^{-\pi B_W/f_s}\cos(2\pi F_f/f_s) \tag{3.25}$$

$$d_1 = 1 - c_1 - c_2, \tag{3.26}$$

where f_s is the sampling frequency. In this work we consider that $d_1 = 1$, implicitly incorporating its effect into the amplitude of the glottal waveforms. Thus we can see that each resonator has a pair of complex conjugate poles:

$$R(z) = \frac{1}{(1 - pz^{-1})(1 - p^{*}z^{-1})}. \tag{3.27}$$

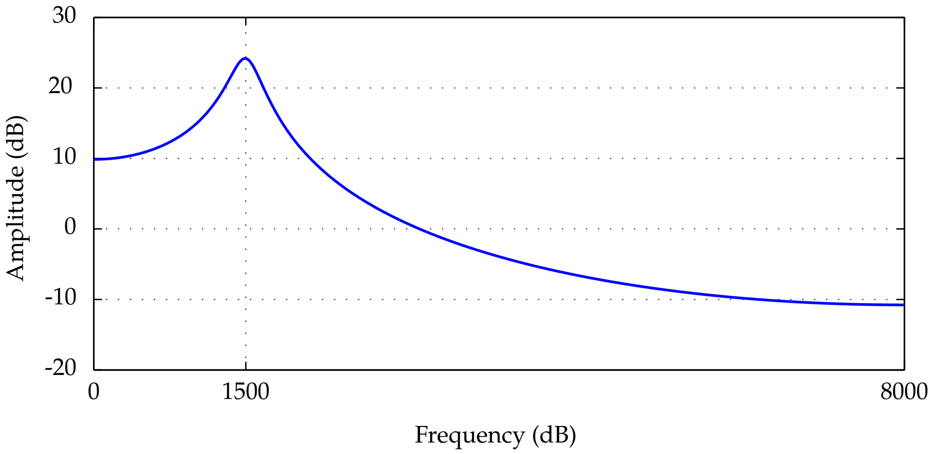


Figure 3.3: *Frequency response of a resonator with center frequency 1500 Hz and bandwidth 300 Hz (sampling frequency of 16 kHz)*

Then, for vocal tract filter length of N coefficients (i.e., $N/2$ complex conjugate pairs) we have:

$$\begin{aligned} A(z) &= 1 - \sum_{k=1}^{N} a_k\, z^{-k} \\ &= (1 - p_1 z^{-1})(1 - p_1^* z^{-1}) \dots (1 - p_{N/2} z^{-N/2})(1 - p_{N/2}^* z^{-N/2}). \end{aligned} \tag{3.28}$$

Figure 3.4 shows the transfer function of a female vocal tract, with center frequencies $640, 1100, 2850, 3750$ and formant bandwidths $270, 170, 200, 200$ (data from Karlsson, 1988). Thus, in order for the vocal tract to have a resonator characteristic, its poles must occur in complex conjugate pairs as we have seen. This can not be directly ensured given the problem formulation, and although some times extraneous roots may occur, the resulting filter often behave as expected. Some authors working on formant tracking place restrictions on the roots and prune those considered as not related to the resonators (Childers and Lee, 1991), although for this work we keep all of them. If the vocal tract were to be modified, the estimated glottal waveform would need to be recomputed using eq. (3.10) with the updated filter, and the KLGLOTT88 model would need to be reestimated using the procedure explained in Section 3.2.2.

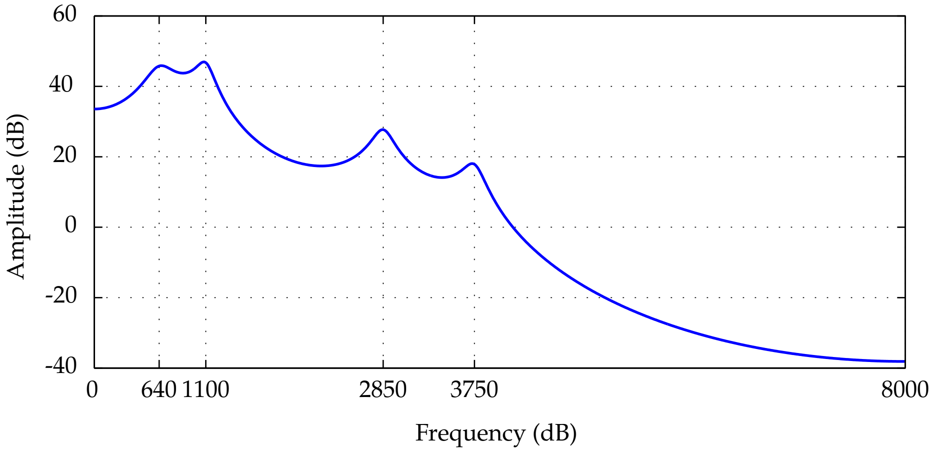


Figure 3.4: *Vocal tract transfer function, corresponding to a female subject with center frequencies (bandwidths) of: 640(270), 1110(170), 2850(200), 3750(200) Hz.*

The low-pass spectral tilt can be achieved if the filter coefficient μ is positive and smaller than 1 for stability. If we combine $A(z)$ from (3.28) with $(1 - \mu z^{-1})$, we can form the extended polynomial:

$$\begin{aligned}\tilde{A}(z) &= A(z)\cdot(1-\mu z^{-1})\\ &= (1-p_1 z^{-1})(1-p_1^* z^{-1})\dots(1-p_{N/2}z^{-N/2})(1-p_{N/2}^* z^{-N/2})(1-\mu z^{-1})\\ &= 1-\sum_{k=1}^{N+1}\tilde{a}_k\, z^{-k}. \end{aligned} \tag{3.29}$$

We see that the last coefficient of the resulting filter, $\tilde{a}_{N+1}$, is the product of all the other poles and μ:

$$\begin{aligned}\tilde{a}_{N+1} &= p_1\cdot p_1^*\cdot\dots\cdot p_{N/2}\cdot p_{N/2}^*\cdot\mu\\ &= \|p_1\|^2\cdot\dots\cdot\|p_{N/2}\|^2\cdot\mu. \end{aligned} \tag{3.30}$$

Thus, constraining the coefficient $\tilde{a}_{N+1}$ to be positive guarantees the low-pass characteristic of the spectral tilt.

The filter will be stable if all the roots lay within the unit circle in the z-plane. Stability can not be guaranteed within the formulation of the convex problem but we can place an upper bound on $\tilde{a}_{N+1}$ in order to obtain filters as stable as possible. The upper bound is computed by placing maximum values to the poles and spectral tilt. We use the values 0.9 for the glottal spectral tilt and 0.985 for the N vocal tract poles (Kim, 2003; Lu, 2002). This approximation does not always result in stable filters, so after each convex decomposition step, those roots lying outside the unit circle are inverted. This is a common approach to guarantee stable filters that does not affect the spectral response of the filter (Proakis and Manolakis, 1996).

We also need a final constraint to force the amplitude of the KLGLOTT88 glottal waveform (b) to be positive. We can now formulate the constraints for the variables $\tilde{a}_{N+1}$ and b in terms of (3.16):

$$\begin{aligned} \tilde{a}_{N+1} > 0 &\quad\rightarrow\quad (0\cdots 0\ -1)\cdot\mathbf{x}\le 0\\ \tilde{a}_{N+1}\le 0.9\cdot 0.985^N &\quad\rightarrow\quad (0\cdots 0\ \ 1)\cdot\mathbf{x}\le 0.9\cdot 0.985^N\\ b > 0 &\quad\rightarrow\quad (-1\ \ 0\cdots 0)\cdot\mathbf{x}\le 0. \end{aligned}$$

As we can see, our problem has no equality constraints, only inequality ones. From (3.16), matrix $\mathbf{A}$ and vector $\mathbf{c}$ are:

$$\mathbf{A}=\begin{pmatrix}0 & 0 & \cdots & -1\\ 0 & 0 & \cdots & 1\\ -1 & 0 & \cdots & 0\end{pmatrix}\qquad \mathbf{c}=\begin{pmatrix}0\\ 0.9\cdot 0.985^N\\ 0\end{pmatrix}$$

3.2.2 KLGLOTT88 reestimation from inverse-filtered glottal waveform

The reestimation of the KLGLOTT88 model is not needed unless the vocal tract obtained during the convex decomposition step were modified outside the algorithm, for instance by placing restrictions on the maximum allowed bandwidth or minimum formant frequencies like in Childers and Lee (1991). Although in this work we do not perform any modifications of the vocal tract, we include the KLGLOTT88 reestimation because it will be used in next chapter to fit the KLGLOTT88 model to the LF model for testing purposes.

To re-estimate the KLGLOTT88 model using g_{if}, we will apply the same method as in the full convex decomposition algorithm, but restricting the filter order to 1 since only the μ parameter needs to be estimated. From eq. (3.7) we see, that given the estimated glottal waveform g_{if}, the tilt effect can be removed as:

$$\tilde{g}_{if}(n) = g_{if}(n) - \mu\, g_{if}(n-1). \tag{3.31}$$

In this case, the error from eq. (3.14) is rewritten as:

$$\begin{aligned} e(n) &= g_{rk}(n) - \tilde{g}_{if}(n) \\ &= \begin{cases} b \cdot C(n) + \mu\, g_{if}(n-1) - g_{if}(n) & \text{open phase,} \\ \mu\, g_{if}(n-1) - g_{if}(n) & \text{closed phase.} \end{cases} \end{aligned} \tag{3.32}$$

Using matrix notation, we can write the error for the whole cycle length:

$$\begin{aligned} \mathbf{e} = \begin{pmatrix} e(1) \\ e(2) \\ \\ \vdots \\ \\ e(P) \end{pmatrix} &= \begin{pmatrix} C(1) & g_{if}(0) \\ C(2) & g_{if}(1) \\ \vdots & \vdots \\ C(N_{op}) & g_{if}(N_{op}-1) \\ 0 & g_{if}(N_{op}) \\ \vdots & \vdots \\ 0 & g_{if}(P-1) \end{pmatrix} \begin{pmatrix} b \\ \mu \end{pmatrix} - \begin{pmatrix} g_{if}(1) \\ g_{if}(2) \\ \\ \vdots \\ \\ g_{if}(P) \end{pmatrix} \\ &\equiv \mathbf{F}\,\mathbf{x} - \mathbf{y} \end{aligned} \tag{3.33}$$

where P the cycle length and $\mathbf{x} = [b \;\; \mu]'$ is the vector containing the two parameters to be estimated. We can now proceed as in the regular convex decomposition algorithm, using the inequality constraints:

$$\begin{aligned} \mu &> 0 && \rightarrow && (0\ \ -1) \cdot \mathbf{x} \leq 0 \\ \mu &\leq 0.9 && \rightarrow && (0\ \ 1) \cdot \mathbf{x} \leq 0.9 \\ b &> 0 && \rightarrow && (-1\ \ 0) \cdot \mathbf{x} \leq 0. \end{aligned}$$

As a result, a new tilt filter coefficient μ and a new KLGLOTT88 amplitude b are obtained. Figure 3.5 serves as an illustration of this procedure.

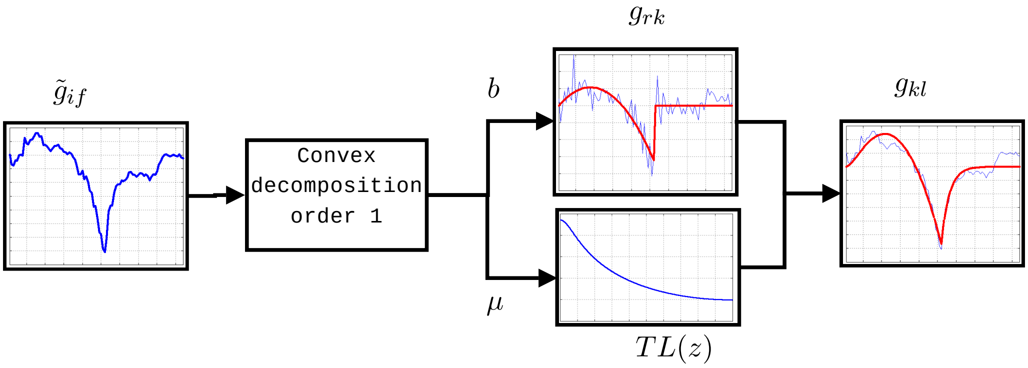


Figure 3.5: *Block diagram of the convex g_{if} re-parametrization.*

In this section we have seen how to analyze a single cycle of speech to obtain the coefficients of the all-pole vocal tract filter and the KLGLOTT88 glottal model using a convex minimization algorithm. Now we will explain how this is integrated into the complete analysis/synthesis algorithm. As we have seen, the location of the glottal epochs needs to be known a priori. The GOI are required since they delimit the speech segment corresponding to a glottal cycle as required to build the error matrix in eq. (3.18). The GCI indicates the instant when the glottis closes, so it corresponds to the duration in samples of the open phase (the Rosenberg-Klatt model parameter N_{op} in eq. (3.12)). We will start by describing the extraction and optimization of the glottal epochs (instants of glottal opening and closing), and will proceed with the glottal modeling using the LF model, and the residual parametrization.

3.3 Estimation of glottal epochs

Correct and accurate extraction of the glottal opening (GOI) and closing (GCI) instants is a key issue for the success of the glottal parametrization algorithm. In this section we will explain how an initial estimation is obtained using a dual channel approach that includes a laryngograph (Section 3.3.1, and then will see how to refine them into the final set of GOI/GCI (Section 3.3.2).

3.3.1 Initial estimation of glottal epochs

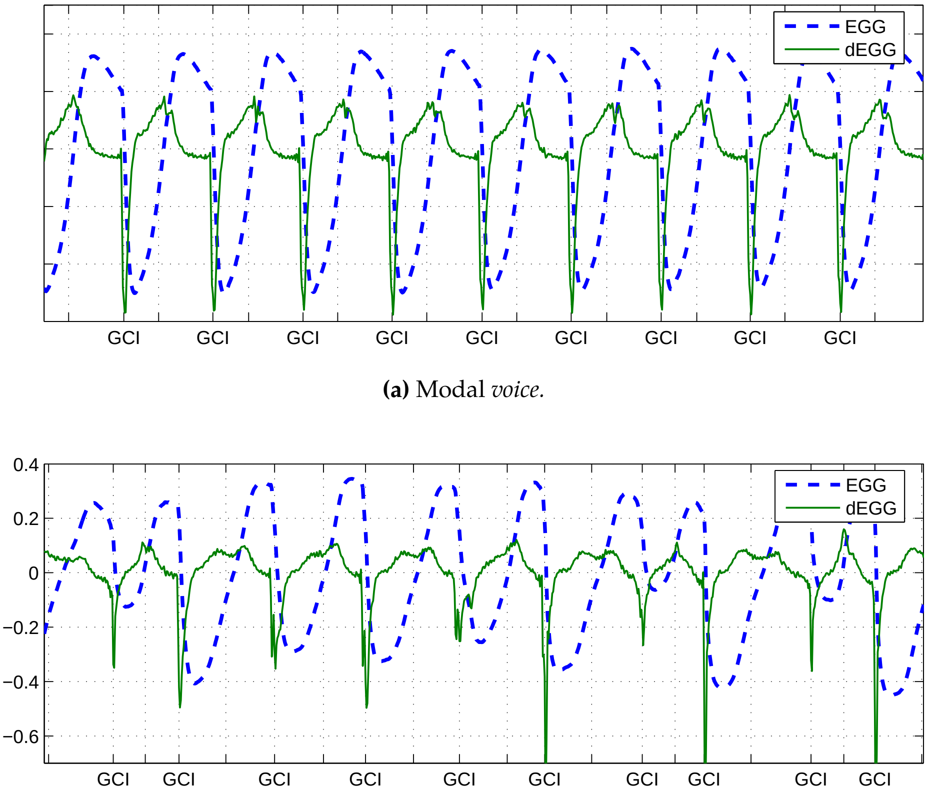
Direct observation of the glottal movements is not feasible when recording large databases, since it requires intrusive methods like X-ray imaging or high-speed cameras among others (Fourcin, 2000; Granqvist et al., 2003). We obtain the information of the glottal epochs using the simultaneously recorded laryngograph signal (EGG), which has been regarded, with some caution, as an accurate source of information of the glottal timing behavior (Henrich et al., 2004; Krishnamurthy and Childers, 1986). The laryngograph[2] monitors the variations in the conductance of a high frequency signal transmitted between two electrodes, placed to the neck on each side of the larynx. The opening and closing of the vocal folds modify the conductivity of the path between the electrodes, and thus modulate in amplitude the signal transmitted.

The GCIs can be obtained by locating the local minima of the differentiated EGG signal (dEGG), as shown in Fig. 3.7. The dEGG waveform presents sharps negative peaks[3], which are associated to the glottal closure instants (Henrich et al., 2004; Krishnamurthy and Childers, 1986; Marasek, 1997). This measurement has been shown to be a very accurate estimation of the GCIs.

Obtaining the opening instants is not such a straightforward task, since it is often difficult to assess when the glottis has completely opened. Some authors make use of dEGG as before, locating the local maxima of the signal between consecutive GCI marks (Krishnamurthy and Childers, 1986). Although this often results in accurate predictions (see Fig. 3.7a), when dealing with pathological voices or some voice types other than modal, one often founds that this estimation is inaccurate (Henrich et al., 2004). For instance, in case of a non-gradual opening of the glottis, the differentiated EGG often presents two peaks, as can be observed in Fig. 3.7b. A number of methods have been proposed to overcome this problem (Bouzid and Ellouze, 2009; Orlikoff, 1991), we have adopted a ro-

[2] Some authors refer to this instrument as *Glottograph*, since it is used to monitor the activity of the glottis. However, we prefer the term *Laryngograph* since what we actually measure is the conductivity of the larynx, from which indirect knowledge about the glottis phases can be derived.

[3] We assume a polarization where the increasing impedance measured with the laryngograph coincides with the signal raising. Speech polarity can be automatically determined and corrected if necessary (Ding and Campbell, 1998).



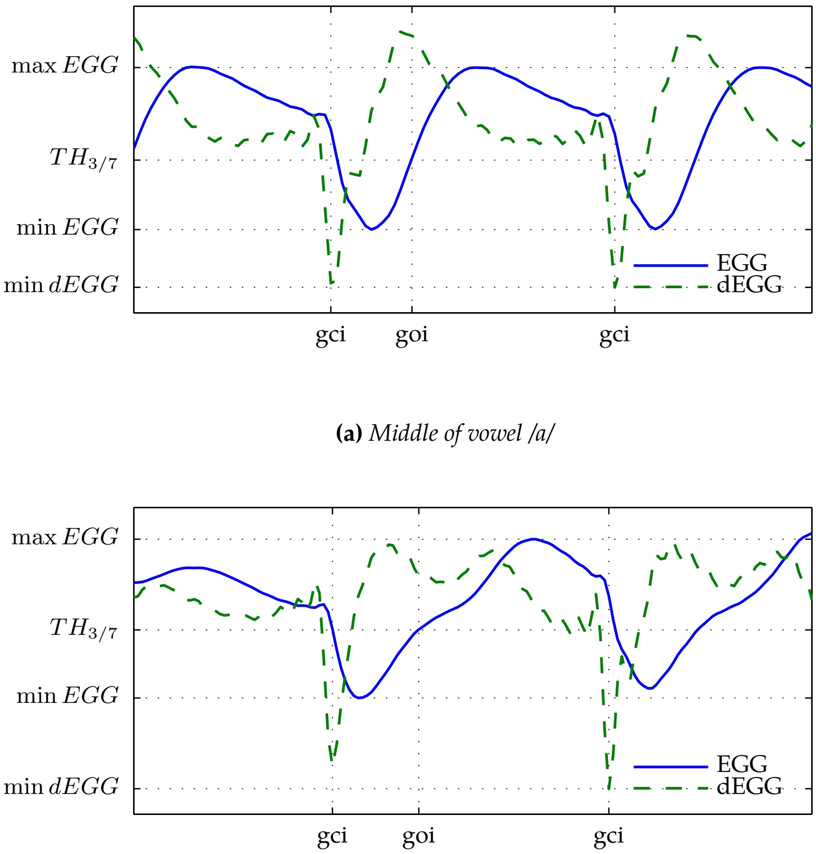
(a) Modal *voice.*

(b) Creaky *voice.*

Figure 3.6: *Laryngograph signal (EGG) and its first derivative (dEGG), corresponding to the vowel* u *uttered in two phonations: modal (3.6a) and creaky (3.6b). The vertical lines indicate the glottal closing instants (labeled GCI) and the glottal opening instants (without label).*

bust approach using threshold levels: the GOI corresponds to the time instant where the ratio of the peak value of the EGG (in that cycle) to the value at the opening instant is $7/3$ (Marasek, 1997). Figure 3.7 illustrates this procedure.

The EGG signal needs to be bandpass filtered prior to any processing, due to the adjustments in the larynx position every speaker perform while speaking. The movements of the larynx often are sufficient to alter the impedance between the electrodes, and thus add a slow variation to the dc level of the signal, as can be seen in Fig. 3.8a. This filter needs to be designed so that it does not introduce distortion in the EGG signal, otherwise the temporal measures would be inaccurate. We use in this study a 199-point bandpass linear FIR filter, with cutoff frequencies at 80 Hz and 5000 Hz (assuming a sampling frequency of $f_s = 16\,\text{kHz}$). To filter the laryngograph signal, we then use a forward-backward step, by first filtering the EGG signal as usual, reversing the resulting waveform, filtering again and then reversing the waveform one last time. Thus the lags introduced by each filtering stage cancel each other out and there is no phase distortion. This procedure works

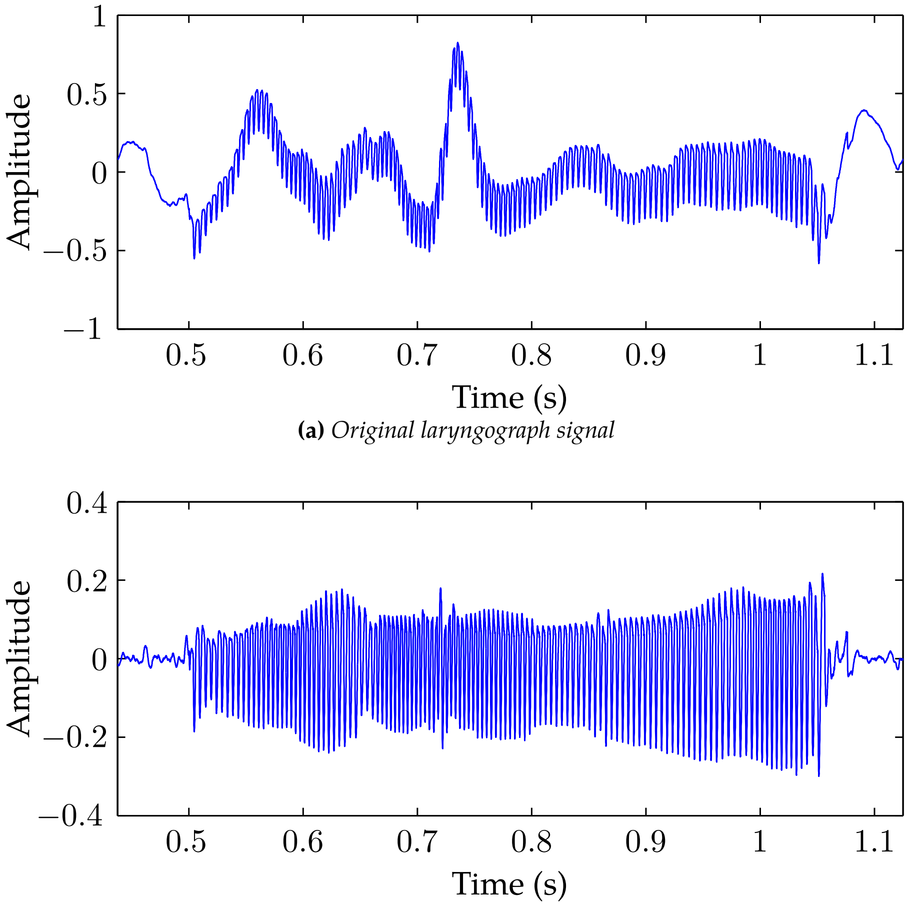


(a) *Middle of vowel /a/*

(b) *Onset of vowel /u/*

Figure 3.7: *Glottal timing detection using the laryngograph. In (a) both the* $3/7$*-ratio and maximum EGG methods give the same results (middle of a sustained vowel), but in (b) the* $3/7$ *is more robust (vowel onset).*

with any generic, nonlinear phase response filter (Proakis and Manolakis, 1996). Fig. 3.8b shows the resulting signal after removing the artifacts and noise.



(a) *Original laryngograph signal*

(b) *Band-pass filtered laryngograph signal*

Figure 3.8: *Pre-processing of the laryngograph signal to remove low frequency artifacts and noise*

Prior to the GCI/GOI location, we perform a coarse estimation of the voiced portions of the signal using the speech signal. This is necessary since the laryngograph signal often contains non-negligible bursts of energy in the unvoiced regions that would affect the performance of the algorithms. For this, we first compute a Moving Average (MA) smoothed version of the speech signal:

$$s_s(k) = \sum_{i=k}^{k+N_s-1} s(i)/N_s, \tag{3.34}$$

where s is the original speech signal, s_s is the smoothed result, k is the sample index and N_s is the length of the MA window (20 samples here at a sampling frequency $f_s = 16\,\text{kHz}$).

We now proceed to calculate the instantaneous energy of the signal as:

$$en_s(k) = \sum_{i=k-N_e/2}^{k+N_e/2} s_s(i)^2, \tag{3.35}$$

using frames of length $N_e = 160$ samples (10 ms at $f_s = 16$ kHz). The course voiced region detection is done by selecting those frames k with energy such that $en_s(k) > th \cdot e\bar{n}_s$, where $e\bar{n}_s$ is the mean energy computed over all the frames, and th is a threshold set experimentally to 0.02. Figure 3.9 shows the band-pass filtered laryngograph signal (top), the original speech signal (middle), and the smoothed speech and instantaneous energy signals (bottom) in solid lines. The voiced regions resulting from the energy thresholding are shown with a dash line. As we can see in the top figure, a noisy laryngograph region occurring during the initial silence has been successfully excluded from the analysis.

In a dual channel (speech and laryngograph signals) approach like this, the EGG signal needs to be delayed in order to account for the non-negligible period of time required for the speech flow to pass through the vocal tract. Since our main algorithm individually corrects the pitch marks (Sec. 3.3.2), at this moment we just apply a global delay of 1ms to synchronize the voiced regions in both the laryngograph and speech channels (roughly corresponding to an average vocal tract length of 25cm plus 10cm of space between the microphone and the speaker).

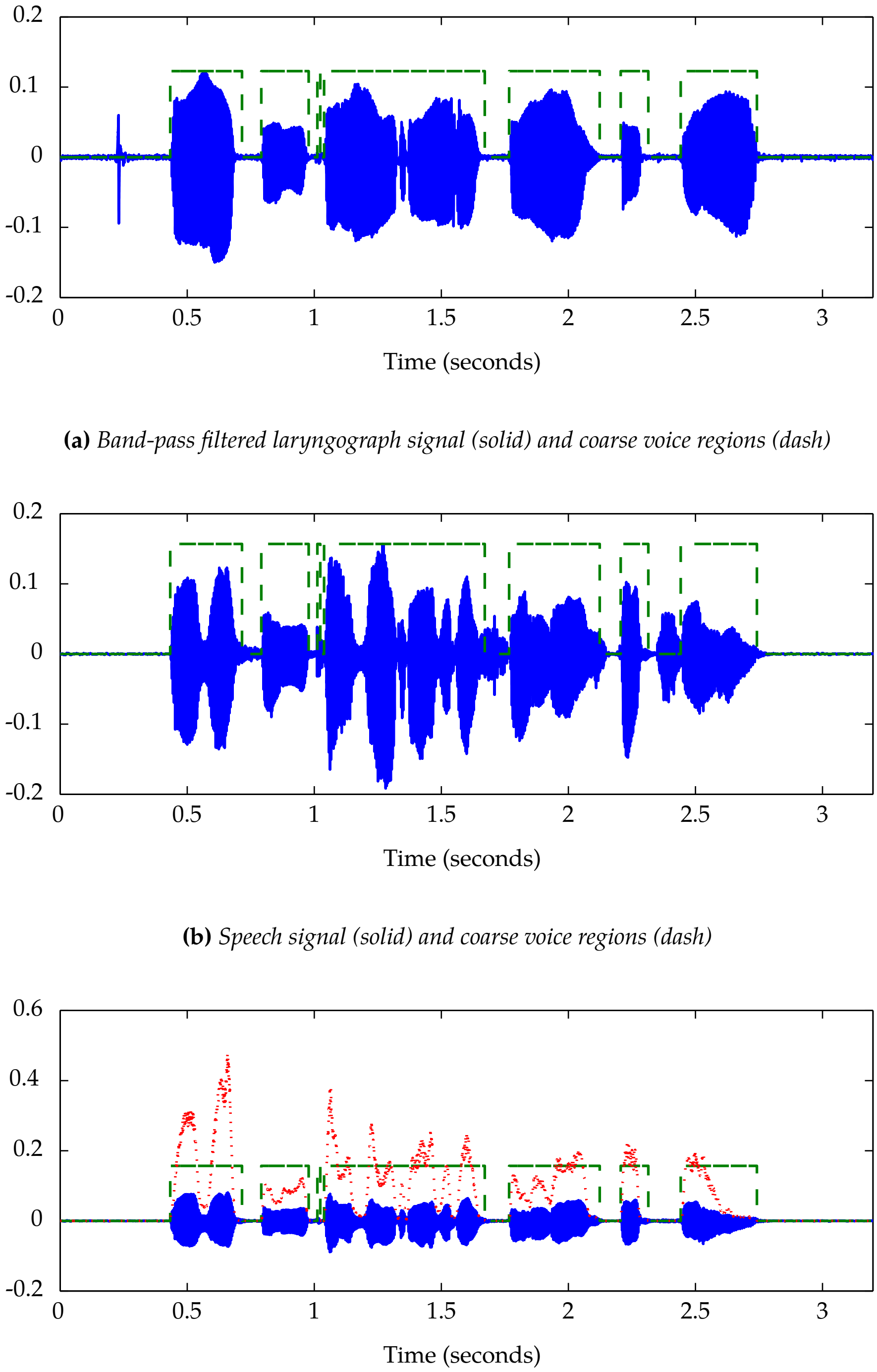
All the voiced frames are then analyzed in order to extract the GCIs marks using the minima of the differentiated EGG, as stated above. In practice, it is necessary to post-process these results in order to eliminate isolated marks or small voiced-regions (we used 5 as the minimum number of pulses in a *voiced* frame). We also set minimum and maximum values for the $F0$ in the post-processing stage (e.g., close marks due to double negative peaks in the dEGG). Once this initial set of glottal epochs is obtained, we can proceed with the optimization of the individual marks, as next section explains.

3.3.2 Optimization of glottal epochs

Using the initial set of epochs extracted from the laryngograph and the estimated voiced regions, we proceed now to optimize the glottal marks. The optimization is performed on a voiced region-by-region basis, and for each one, we proceed as follows:

- a better global delay is applied to the region's GOI/GCI marks to obtain a better synchronization with the speech signal, improving the generic 1ms delay previously applied,

- then, the location of each individual GCI in the voiced region is optimized, operation on a glottal cycle-by-cycle bases,



(a) *Band-pass filtered laryngograph signal (solid) and coarse voice regions (dash)*

(b) *Speech signal (solid) and coarse voice regions (dash)*

(c) *MA smoothed speech signal (solid), instaneous energy (dotted) and coarse voice regions (dash)*

Figure 3.9: *Coarse voiced regions detection for glottal epochs location using the instantaneous energy of the MA smoothed speech signal.*

- lastly, each single GOI is individually optimized following a similar procedure.

Each of the three steps of our optimization technique is described in detail in the following sections.

Global synchronization

In this first step, we search for a better synchronization of the glottal marks and the speech signal, since the 1 millisecond synchronization used in the previous section is good for voiced regions alignment, but it is not optimal for exact glottal period alignment. The algorithm is listed in pseudo-code in 1 and it goes as follows. We conduct an exhaustive search for the optimal global delay in the range -15 to 5 samples (approximately $-1\,\mathrm{ms}$ to $0.3\,\mathrm{ms}$ at $f_s = 16\,\mathrm{kHz}$), which we found to be sufficient by analyzing a subset of the corpus. For each lag in this range, the whole GOI/GCI set is delayed, and we perform a decomposition on a glottal cycle-by-cycle basis: for each speech period k (delimited by the k and $k{+}1$ delayed GOIs) in the voiced region, we perform the convex decomposition from Section 3.2. We obtain an estimation of the KLGLOTT88 model and the vocal tract filter, which we use to inverse-filter the speech period to obtain the estimated glottal waveform. We then evaluate the period estimation by computing the mean-squared error between the estimated (i.e., inverse-filtered) and parametrized (KLGLOTT88) glottal waveforms. The global error associated to each individual lag is the mean value of each period's glottal error. Since we are not yet considering the noise present in the source, we apply a low-pass filter to both waveforms prior to computing the error. Once we have the averaged mean-squared error for each set of delayed glottal marks, we chose as optimal global delay the one resulting in minimum distortion.

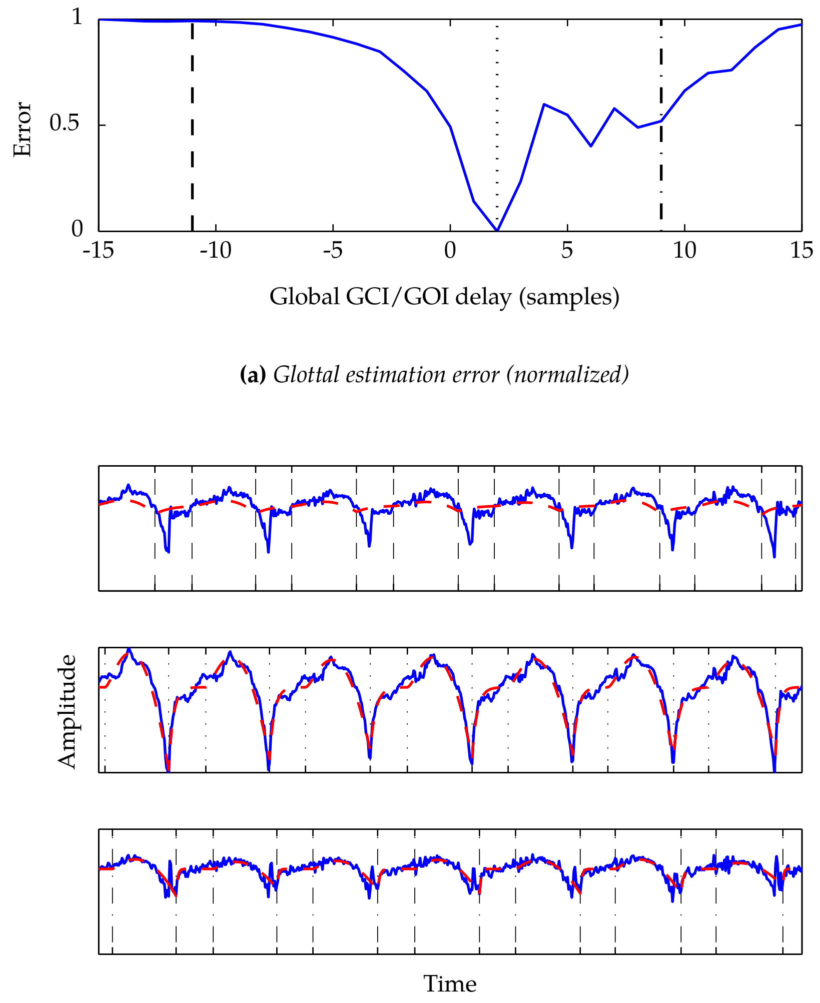
Algorithm 1: GOI/GCI synchronization

```
input  : Initial GOIs and GCIs
output: Synchronized GOIs and GCIs

foreach VoicedRegion $v$ do
    foreach $lag \in -15$ to $5$ do
        $GCI_{lag} \leftarrow (GCI \in v) + lag$;
        $GOI_{lag} \leftarrow (GOI \in v) + lag$;
        foreach GlottalCycle $k \in v$ do
            speech ←SpeechFrame($GOI_{lag}[k]$ to $GOI_{lag}[k+1]$);
            [$g_{if}$, $g_{kl}$] ←CvxDecomposition($GCI_{lag}[k]$,speech);
            $error_{cycle}[k]$ ←Norm2($g_{if}$, $g_{kl}$);
        $error[lag]$ ←Mean($error_{cycle}$);
    opt ← $\arg\min_l(error)$;
    $(GCI \in v) \leftarrow (GCI \in v) + lags$[opt ];
    $(GOI \in v) \leftarrow (GOI \in v) + lags$[opt ];
```

The top figure in 3.10 shows the resulting error function over the whole search range. As we can see, it presents a clear, sharp minimum corresponding to the optimal synchronization delay (2 samples in our case, denoted with a vertical dotted line). The bottom figure shows the inverse-filtered (solid) and KLGLOTT88 matched (dash) waveforms corresponding to each of the three synchronization lags marked in the error function plot. The quality of the inverse-filtered waveforms is clearly better for the optimal case (dotted line in 3.10a, middle figure in 3.10b), than for the other two cases (top and bottom subfigures in 3.10b, corresponding to the lags denoted with a dash and dot-dash lines in 3.10a). After finding the optimal synchronization lag, the whole set of GCI/GOI marks is delayed accordingly, and we proceed to improve each mark individually, first the GCI, and then the GOI. The procedure is detailed in the next two sections.

Local GCI optimization

Once we have globally adjusted the glottal epochs, we proceed to their individual optimization. We start by finding the optimal location of each GCI inside the corresponding glottal period. The idea is to search for the optimal GCI in the neighboring area around the initial GCI. We found that a search area of ± 10 samples ($\pm 0.6\,\mathrm{ms}$ at $f_s = 16\,\mathrm{kHz}$) around the initial GCI was sufficient to find the optimal GCI. Our optimality criterion here is the minimum parametrization error between g_{if} and g_{kl}. The algorithm is listed in pseudo code in 2. For each GCI candidate in the search area, a convex decomposition is performed to obtain the vocal tract filter and the KLGLOTT88 waveform. The inverse



(a) *Glottal estimation error (normalized)*

(b) *Glottal waveforms for three different global synchronization delays (solid: inverse-filtered, dash: KLGLOTT88)*

Figure 3.10: *Global synchronization. Top figure: error function for the synchronization lags in the range [-15:15] (chosen for illustrative purposes only). Bottom figure: inverse-filtered and KLGLOTT88 glottal waveforms corresponding to the three synchronization lags indicated in the top figure (optimal: dotted vertical line, candidates: dash and dot-dash vertical lines).*

of the filter is used to obtain an estimation of the glottal waveform, and the mean-square error is computed. After all the candidates have been tested, the optimal GCI is chosen as the one resulting in the lower error. This process is repeated for each period in the voiced region.

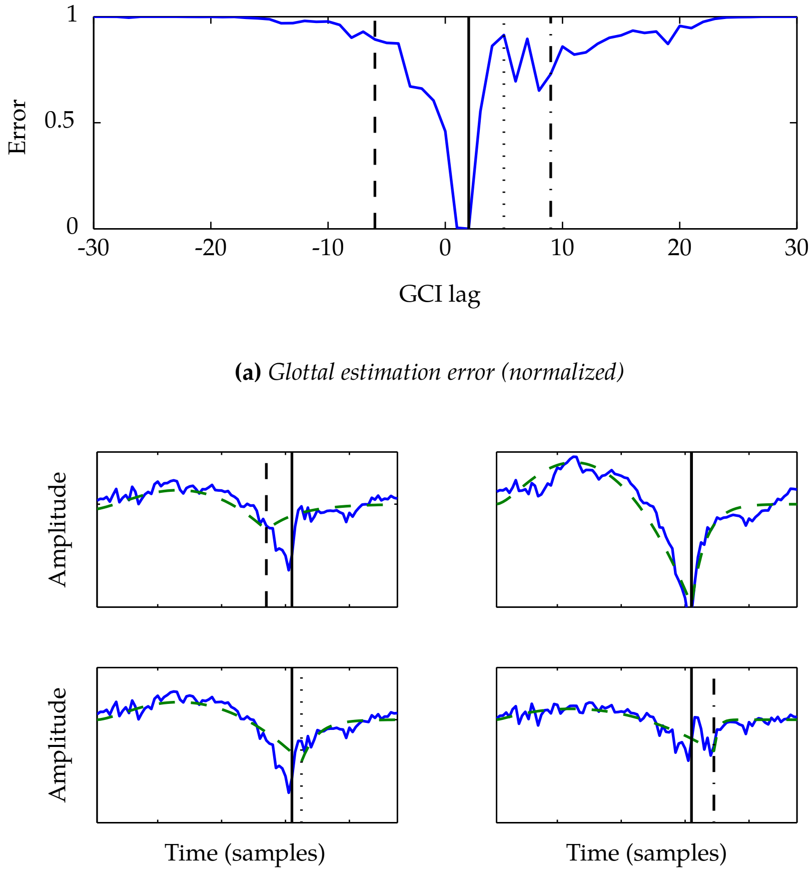
Algorithm 2: gci optimization

```
input  : Synchronized GOIs and GCIs
output: Optimized GCIs
foreach VoicedRegion $v$ do
    foreach GlottalCycle $k \in v$ do
        speech ←SpeechFrame($GOI[k]$ to $GOI[k+1]$);
        foreach $lag \in -10$ to $10$ do
            cand ← $GCI[k] + lag$;
            [$g_{if}$, $g_{kl}$] ←CvxDecomposition(cand,speech);
            $error[lag]$ ←Norm2($g_{if}$, $g_{kl}$);
        opt ← $\arg\min_l (error)$;
        $GCI[k] \leftarrow GCI[k] + lags[opt]$;
```

The top figure in 3.11 shows the glottal error function (normalized to the $[0:1]$ range) for a search area around the initial candidate of ±30 samples at $f_s = 16\,\mathrm{kHz}$ (chosen for illustrative purposes only, the actual range used in the algorithm is $[-10:10]$). The glottal waveforms (inverse-filter and KLGLOTT88) corresponding to four lags in the range (optimal: solid line, candidates: dotted, dashed and dot-dashed lines) are shown in the bottom figure 3.11b. As we can see, the error function presents a sharp minimum indicating the optimal delay, and the quality of the estimation is clearly better, both in terms of estimated glottal waveform shape and quality of the KLGLOTT88 matching.

Local GOI optimization

In the last step, for each glottal period, the optimal GOI is searched in the neighborhood of the initial GOI resulting from the synchronization step. The method, depicted in 3, is identical to the GCI optimization algorithm, with two differences. First, the search area around initial GOIs (±20 samples, $\pm 1.25\,\mathrm{ms}$ at $f_s = 16\,\mathrm{kHz}$) is increased with respect to the GCI case (±10). This is necessary since extracting GOIs using a laryngograph signal is usually less precise and more prone to errors. And second, the glottal cycles are now delimited using GCIs, so they start during the closed phase and end with the open phase. The convex decomposition algorithm needs to be reformulated accordingly. In this case,



(a) *Glottal estimation error (normalized)*

(b) *Glottal waveforms for the different GCI candidates (solid: inverse-filtered, dash: KLGLOTT88)*

Figure 3.11: *Local GCI optimization. Top figure: error function for the GCI lags in the range [-30:30] (chosen for illustrative purposes only). Bottom figure: inverse-filtered and KLGLOTT88 glottal waveforms corresponding to the four lags indicated in the top figure (optimal: solid line, candidates: dotted, dashed and dot-dashed lines).*

from (2.8) and (3.11) we can write:

$$g'_{rk}(n) = \begin{cases} 0 & 1 \leq n < N - N_{op} \\ b \cdot C(n - (N - N_{op})) & N - N_{op} \leq n \leq N. \end{cases} \tag{3.36}$$

Were, as in the GOI delimited case, N_{op} is the duration of the open phase and $C(n)$ is calculated as before using (3.13). Using (3.36), we can rewrite (3.14) as:

$$\begin{aligned} e(n) &= g_{rk}(n) - \tilde{g}_{if}(n) \\ &= \begin{cases} \sum_{k=1}^{N} \tilde{a}_k s(n-k) - s(n) & \text{CP} \\ b \cdot\, C(n - (N - N_{op})) + \sum_{k=1}^{N} \tilde{a}_k s(n-k) - s(n) & \text{OP} \end{cases} \end{aligned} \tag{3.37}$$

to emphasize the use of the GCIs as starting points of the glottal cycle (starts with the closed phase CP, and then the open phase OP). The error (3.18) for the whole glottal cycle (length P) is:

$$\begin{aligned} \mathbf{e} = \begin{pmatrix} e(1) \\ e(2) \\ \\ \vdots \\ \\ e(P) \end{pmatrix} &= \begin{pmatrix} 0 & s(0) & \cdots & s(-N) \\ \vdots & s(1) & & s(-N+1) \\ 0 & & & \\ C(1) & \vdots & & \vdots \\ C(2) & & & \\ \vdots & & & \\ C(N_{op}) & s(P-1) & \cdots & s(P-N-1) \end{pmatrix} \begin{pmatrix} b \\ \tilde{a}_1 \\ \\ \vdots \\ \tilde{a}_{N+1} \end{pmatrix} - \begin{pmatrix} s(1) \\ s(2) \\ \\ \vdots \\ \\ s(P) \end{pmatrix} \\ &= \mathbf{F}\,\mathbf{x} - \mathbf{y}. \end{aligned} \tag{3.38}$$

The problem is solved as before.

Algorithm 3: GOI optimization

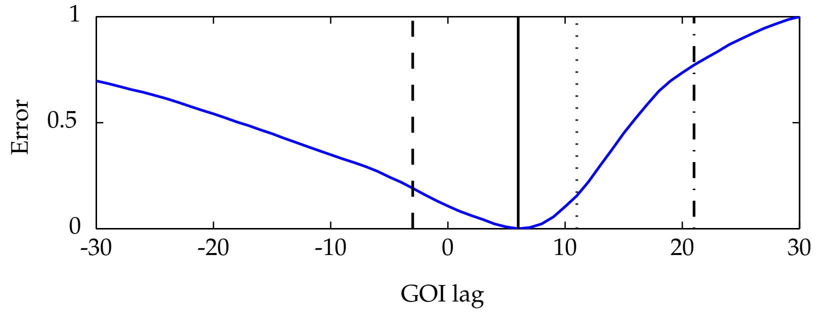
```
input  : Optimized GCIs, synchronized GOIs
output: Optimized GOIs

foreach VoicedRegion $v$ do
    foreach GlottalCycle $k \in v$ do
        speech ←SpeechFrame($GCI[k]$ to $GCI[k+1]$);
        foreach $lag \in -20$ to $20$ do
            cand ← $GOI[k] + lag$;
            [$g_{if}$, $g_{kl}$] ←CvxDecomposition(cand,speech);
            $error[lag]$ ←Norm2($g_{if}$, $g_{kl}$);
        opt ← $\arg\min_l(error)$;
        $GOI[k] \leftarrow GOI[k] + lags[opt]$;
```

The top figure in 3.12 shows the glottal error function (normalized to the $[0:1]$ range) for a search area around the initial GOI candidate of ± 30 samples at $f_s = 16\,\mathrm{kHz}$ (chosen for illustrative purposes only, the actual range used in the algorithm is $[-20 : 20]$). The glottal waveforms (inverse-filter and KLGLOTT88) corresponding to four lags in the range (optimal: solid line, candidates: dotted, dashed and dot-dashed lines) are shown in the bottom figure 3.12b. As we can see, the error function presents a clear minimum indicating the optimal delay. Compared with the GCI error function shown in 3.11a, one can clearly see that the GOI error function is much smoother. This is due to the opening of the glottis being more gradual than its closing, which is generally much sudden. As in the GCI case, the quality of the estimation is clearly better for the optimal lag, both in terms of estimated glottal waveform shape and quality of the KLGLOTT88 matching.

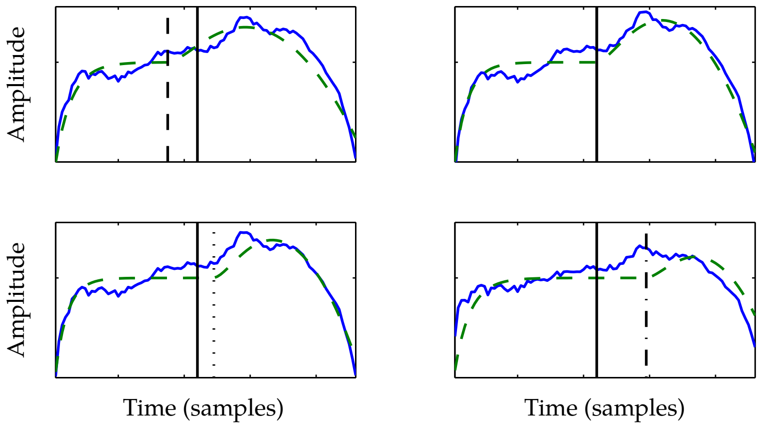
With the optimized set of GOI/GCI, we can now proceed to the main block of our algorithm in which the optimal parameters of the voice source and vocal tract filter are estimated.

3.4 Source-filter estimation

In this section we detail our method for obtaining the definite estimation of the vocal tract filter coefficients and the voice source parameters, in which several glottal cycles are simultaneously analyzed. This is convenient in order to increase the robustness of the analysis (e.g., in case of very short lengths of the glottal periods, such as occur in high-pitched female voices). Furthermore, tying some of the parameters in the multi-frame analysis can lead to better continuity properties of the estimation. So far we have used single-cycle source-filter decompositions to optimize the glottal epochs, and thus the



(a) *Glottal estimation error (normalized)*



(b) *Glottal waveforms for the different GOI candidates (solid: inverse-filtered, dash: KLGLOTT88)*

Figure 3.12: *Local GOI optimization. Top figure: error function for the GOI lags in the range [-30:30] (chosen for illustrative purposes only). Bottom figure: inverse-filtered and KLGLOTT88 glottal waveforms corresponding to the four lags indicated in the top figure (optimal: solid line, candidates: dotted, dashed and dot-dashed lines).*

convex optimization problem has been formulated on a period-by-period basis, where each glottal cycle is individually analyzed. We will first formulate the problem for the multi-cycle case, and then the constraints will be modified accordingly.

3.4.1 Multi-cycle formulation

For this work, we force the modified vocal tract filter (i.e., including the low-pass filter coefficient μ associated to the KLGLOTT88 model) to be the same for all the glottal cycles in the analyzed frame, while we allow independent KLGLOTT88 amplitudes (the parameter b) for each cycle. This approach has the benefit of providing smoother estimates while preserving the dynamic characteristics of the speech signal. We want to obtain the parameters that minimize the parametrization error for each cycle in the analysis frame:

$$\begin{aligned} e(n) &= g_{rk}(n) - \hat{g}_{if}(n) \\ &= \begin{cases} b_1\, C_1(n) + \sum_{k=1}^{N+1} \tilde{a}_k s(n-k) - s(n) & n \in OP_1 \\ \sum_{k=1}^{N+1} \tilde{a}_k s(n-k) - s(n) & n \in CP_1 \\ & \\ b_2\, C_2(n) + \sum_{k=1}^{N+1} \tilde{a}_k s(n-k) - s(n) & n \in OP_2 \\ \sum_{k=1}^{N+1} \tilde{a}_k s(n-k) - s(n) & n \in CP_2 \\ & \\ \vdots & \vdots \\ & \\ b_M\, C_M(n) + \sum_{k=1}^{N+1} \tilde{a}_k s(n-k) - s(n) & n \in OP_M \\ \sum_{k=1}^{N+1} \tilde{a}_k s(n-k) - s(n) & n \in CP_M \end{cases} \end{aligned} \tag{3.39}$$

where we have simplified the notation using $C_i(n) = n(2N_{op}^i - 3n)$, and OP_i and CP_i are the open and closed phases for glottal cycle i inside the analysis frame. Since the error is linear w.r.t. our unknown variables, we can rewrite eq. 3.39 in matrix form as:

$$\begin{aligned}\mathbf{e} = \mathbf{F}\begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_M \\ \tilde{a}_1 \\ \\ \vdots \\ \tilde{a}_{N+1} \end{pmatrix} - \begin{pmatrix} s(1) \\ s(2) \\ \\ \vdots \\ \\ s(P) \end{pmatrix} \\ \equiv \mathbf{F}\,\mathbf{x} - \mathbf{y},\end{aligned} \tag{3.40}$$

where the matrix $\mathbf{F}$ is written as:

$$\mathbf{F} = \begin{pmatrix} C_1(1) & 0 & \cdots & 0 & s(0) & \cdots & s(-N) \\ \vdots & \vdots & \ddots & \vdots & s(1) & & s(-N+1) \\ C_1(N_{op}^1) & 0 & \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \vdots & & \vdots \\ 0 & C_2(1) & \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & C_2(N_{op}^2) & \cdots & 0 & \vdots & & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 0 & \vdots & & \vdots \\ \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ \\ 0 & 0 & \cdots & C_M(1) & \vdots & & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & C_M(N_{op}^M) & \vdots & & \vdots \\ \vdots & \vdots & \ddots & \vdots & s(P-2) & & s(P-N-2) \\ 0 & 0 & \cdots & 0 & s(P-1) & \cdots & s(P-N-1) \end{pmatrix}, \tag{3.41}$$

where N_{op}^m is the GCI of the mth cycle inside the analysis frame.

3.4.2 Multi-cycle constraints

When working with multi-cycle analysis with independent glottal amplitudes, we have to include as many constraints as necessary. The vector of variables in this case is $\mathbf{x} = [b_1\, b_2 \cdots b_M\ \tilde{a}_1\ \cdots\ \tilde{a}_{N+1}]$, and the constraints are :

$$\begin{aligned}
\tilde{a}_{N+1} &> 0 && \rightarrow\quad (0 \cdots 0\ {-1}) \cdot \mathbf{x} \leq 0 \\
\tilde{a}_{N+1} &\leq 0.9 \cdot 0.985^N && \rightarrow\quad (0 \cdots 0\ \ 1) \cdot \mathbf{x} \leq 0.9 \cdot 0.985^N \\
b_1 &> 0 && \rightarrow\quad ({-1}\, 0 \cdots 0) \cdot \mathbf{x} \leq 0 \\
b_2 &> 0 && \rightarrow\quad (0\ \ {-1}\, 0 \cdots 0) \cdot \mathbf{x} \leq 0 \\
b_M &> 0 && \rightarrow\quad (0 \cdots {-1} \cdots 0) \cdot \mathbf{x} \leq 0.
\end{aligned}$$

In this case, matrix $\mathbf{A}$ and vector $\mathbf{c}$ are:

$$\mathbf{A} = \begin{pmatrix} 0 & \cdots & \cdots & 0 & 0 & \cdots & -1 \\ 0 & \cdots & \cdots & 0 & 0 & \cdots & 1 \\ -1 & 0 & \cdots & 0 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & 0 & \cdots & 0 \\ 0 & 0 & \cdots & -1 & 0 & \cdots & 0 \end{pmatrix} \qquad \mathbf{c} = \begin{pmatrix} 0 \\ 0.9 \cdot 0.985^N \\ 0 \\ 0 \\ \cdots \\ 0 \end{pmatrix}$$

3.4.3 Inverse filtering

Each analysis frame was analyzed with the convex decomposition algorithm enforcing the same vocal tract filter for all of them, but allowing for independent glottal amplitudes. The resulting vocal tract filter and corresponding glottal amplitude was used as the parameter set for the cycle. The procedure is illustrated in Figure 3.13.

3.5 LF parametrization

As a result of the previous steps, we obtain an initial estimation of the noisy glottal waveform by inverse-filtering the speech with the vocal tract filter, and an optimal approximation using the KLGLOTT88 model 3.14. As stated before, the KLGLOTT88 model is a simple model, useful for the mathematical formulation of the problem; there are other models better suited for a wider range of voice types. For this thesis we have chosen the LF model, since it is widely accepted and has been used in several research projects. The next step is then to reparametrize the differentiated glottal inverse waveform with the LF model.

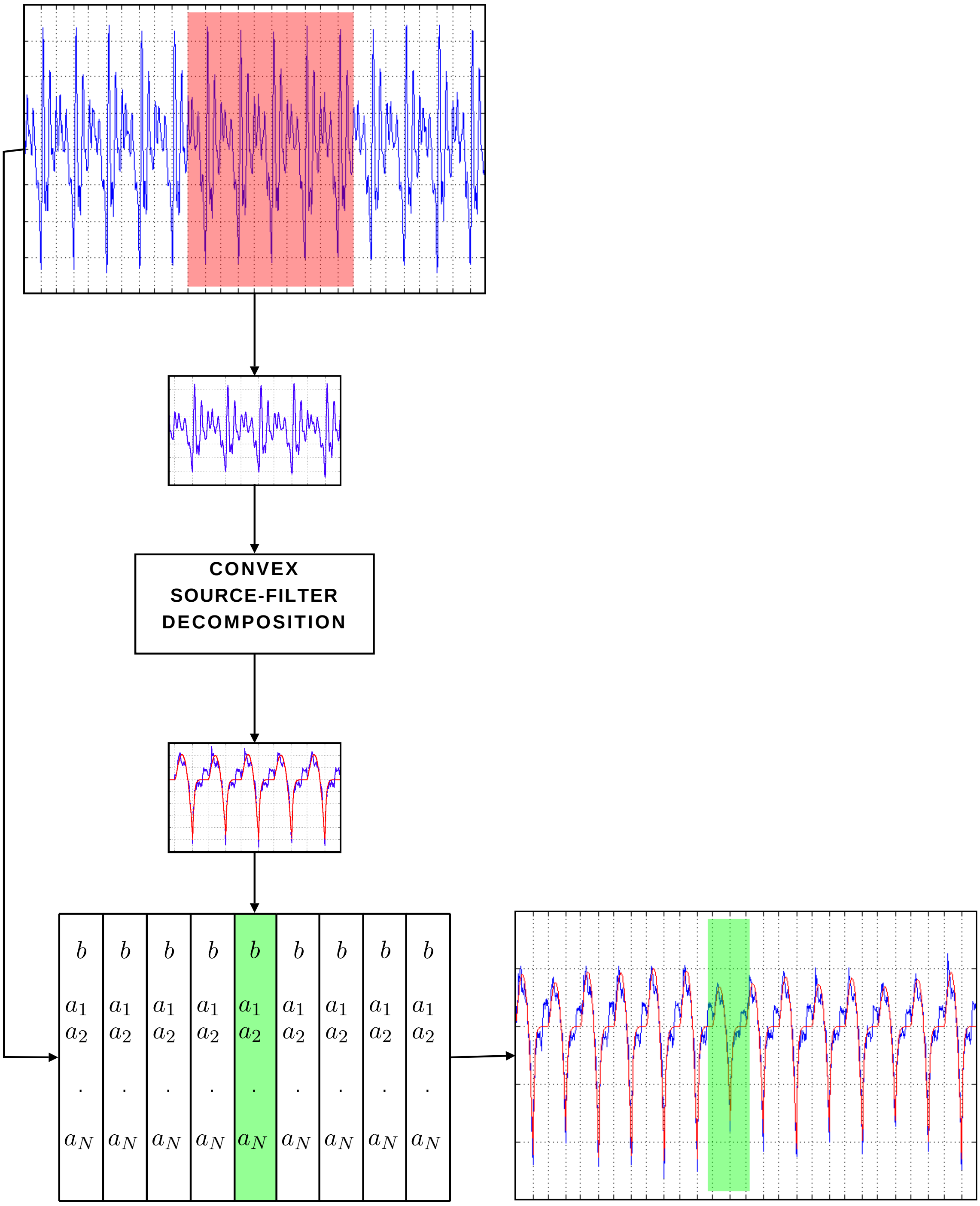


Figure 3.13: *Block diagram of the multi-cycle source-filter decomposition algorithm.*

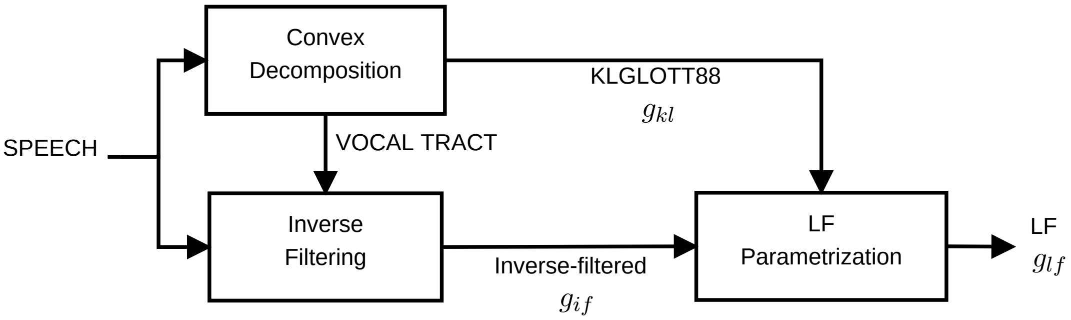


Figure 3.14: *LF parametrization scheme*

3.5.1 Optimization

With this initial estimate, the parameter set is optimized via constrained non-linear optimization methods (Lu, 2002; Strik et al., 1993). The optimization was done using a non-linear least squares algorithm, which solves problems of the form:

$$\min_{\theta} \|f(\theta)\|_2^2 = \min_{\theta}(f_1(\theta)^2 + f_2(\theta)^2 + \cdots + f_k(\theta)^2). \tag{3.42}$$

In our case, we want to minimize the L_2-norm of the error between the inverse-filtered glottal waveform $g_{if}(n)$ and the LF parametrization $g_{lf}(n)$. Prior to any processing, we will minimize the effect of the noise present in the inverse-filtered waveform by low-pass filtering both g_{if} and g_{lf}. To this effect, we will using a Blackman window of length 7(Strik et al., 1993). Figure 3.15 presents the impulse (upper part) and frequency (lower part) response of this filter. As we can see, the cut-off frequency f_c of the filter can be adjusted by changing the length of the Blackman window. For this work, after analyzing a subset of the residual waveforms, we have found a cut-off frequency of 2 kHz to be sufficient. Thus, the selected Blackman window's length is 7 samples.

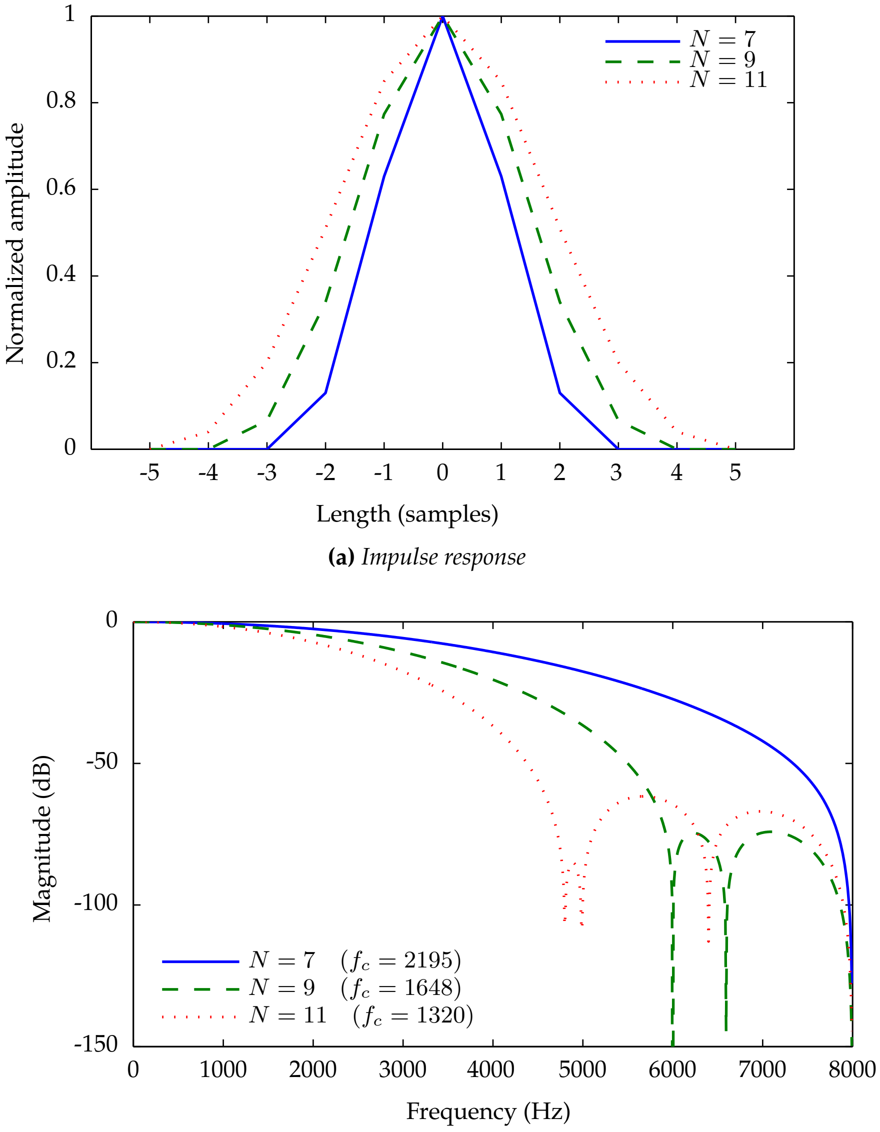
Thus, we can set the error function to:

$$\mathbf{f}(\theta) = \mathbf{g_{if}} - \mathbf{g_{lf}}|_{\theta}, \tag{3.43}$$

where $\theta = (t_0, t_p, t_e, t_a, E_e)$. And then, we construct the minimization problem (3.42) using:

$$f_k(\theta)^2 = (g_{if}(k) - g_{lf}(k)|_{\theta})^2. \tag{3.44}$$

When discretizing the LF model equation, it is important to keep the temporal values to its original value, allowing them to fall between sample values. This is to avoid the staircase effect when computing the error, which could result in the minimization algorithm getting stuck in a local, non-optimum minimum, as illustrated in Figure 3.16 for the case of the t_e value.



(a) *Impulse response*

(b) *Frequency response*

Figure 3.15: *Impulse and frequency response of the Blackman window based low-pass filter for different window lengths (7, 9 and 11 samples).*

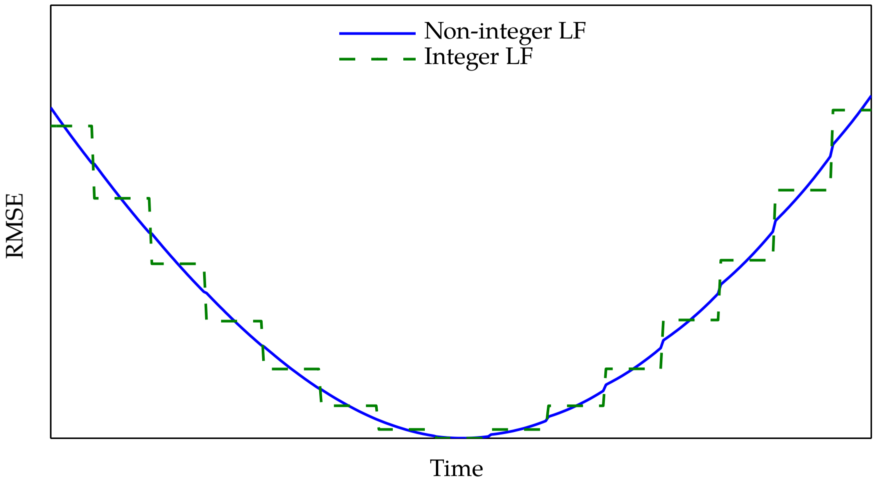


Figure 3.16: *Illustration of the staircase effect during the estimation of the t_e parameter in the LF model in the integer vs. non-integer case.*

To solve this constrained, non-linear data fitting problem, we use an optimization algorithm based on the interior trust region approach (Coleman and Li, 1996), implemented in the Matlab Optimization Toolbox. These sort of algorithms have good convergence properties and they are both reliable and robust, making them appropriate to ill-conditioned problems (Yuan, 2000). Trust region algorithms solve the minimization problem iteratively. At each iteration, a model that approximates the function to be minimized is built near the current estimate of solution. The solution of this simplified model is then taken as the next iterative point. The key aspect is that this solution is only *trusted* in a region near the current iterate (the trust region). This region is adjusted from iteration to iteration, depending on how well the approximated model fits the original model: when the fitting is satisfactory, the trust region can be kept or enlarged, otherwise, it is reduced and the iteration repeated. The model constraints are incorporated into the problem by making the trust region be a subset of the feasible region (i.e., the region in which a solution complies with the constraints and is thus valid) (Coleman and Li, 1996).

3.5.2 Initialization

The fitting of an LF model to the inverse filtered speech signal is a well-understood and documented process (Strik, 1998; Strik and Boves, 1994; Strik et al., 1993). One of the problems with these algorithms is that the inverse filtered signal is very noisy, and it requires the use of a low-pass filter to achieve proper fitting. The same low-pass filter should be used when matching the LF waveform. Since the fitting is performed by means of non-linear optimization algorithms, the starting point should be robust enough to avoid ending in a local (non-optimal) minimum.

One approach to obtain the initial estimations of the model parameters is identifying the minimum value and the zero-crossings of the estimated inverse-filtered glottal waveform g_{if}, as shown in Fig. 3.17: t_e is set to the time at which the minimum of the derivative glottal waveform occurs (t_{min}); E_e is set to the value at this point (g_{min}); t_c is set to the time at which the next zero-crossing to the right of t_{min} occurs (ZC_{end}); t_p is set to the time at which the next zero-crossing to the left of t_{min} occurs (ZC_{mid}); t_o is set to the time at which the first zero-crossing occurs ZC_{ini}, starting from ZC_{mid}. The parameter controlling the effective duration of the return phase (t_a) is generally obtained by fitting the return branch of the LF model to the inverse-filtered glottal waveform.

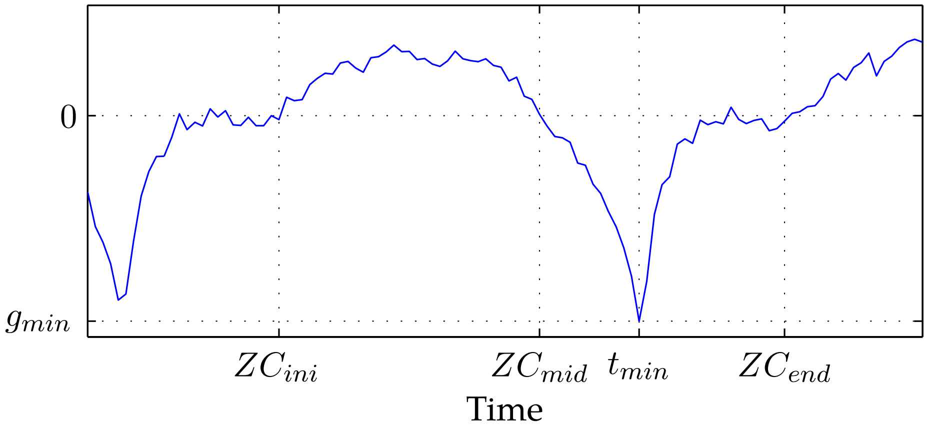


Figure 3.17: *Standard initialization of the LF parameters using amplitudes, zero-crossings and maxima/minima of the estimated glottal waveform g_{if} obtained by inverse filtering.*

One of the main problems with this approach, is that the quality of the matching is highly influenced by the level of noise present in the inverse-filtered speech signal (Strik et al., 1993). In our work, we have decided to take advantage of the similarities between the parametric KLGLOTT88 model obtained in the optimization step and the final LF model. Ee is set to the minimum value of g_{KL}, t_e to the time position of this minimum, and t_p as the zero-crossing to the left of t_e. Since the glottal cycle length are already available, t_c and t_0 are set to the end of the glottal cycle. The t_a parameter of the LF model controlling the abruptness of the glottal closure has no direct equivalent in the KLGLOTT88 model. We could obtain it by fitting the return branches of both the LF and KLGLOTT88 models, but we will use a more elegant approach. As we have seen in eq. 2.23 Section 2.2.2, t_a is proportional to a frequency F_a where an extra $-6\,\mathrm{dB/octave}$ are added to the source spectrum. Thus, an elegant way to obtain a first approximation for F_a (and then t_a) is to set it to the cut-off frequency f_c of the KLGLOTT88 tilt filter $TL(z)$[4]. Now, we can obtain

[4] The 1st-order FIR filter and its power spectrum equation are $H(\omega) = \frac{1}{1-\mu\, e^{-j\omega}}$ and $\|H(\omega)\|^2 =$

the initial estimation for t_a by setting $F_a \approx f_c$:

$$t_a = \frac{1}{2\pi F_a} \approx \frac{1}{2\pi f_c} = \frac{1}{f_s \text{ , arccos} \left(2 - \frac{1+\mu^2}{2\mu}\right)}. \tag{3.45}$$

We have found this method to be faster and more accurate than the traditional approach of obtaining t_a by fitting the return branch (Strik et al., 1993).

After the LF modeling presented in this section, we can generate synthetic speech applying the synthesis equation (3.2), using the LF train of glottal pulses and the estimated vocal tract filters. However, since we are ignoring the effects of the aspiration noise, the quality of the resulting speech will still be quite poor. We performed and informal listening test resynthesizing the original sentences, and the resulting speech was described as lacking richness and sounding as muffled, typical of a vocoder-like scheme. This was to be expected, since we are not including any of the aspiration noise present in the original speech, which greatly improves the naturalness.

3.6 Glottal residual parametrization

We will extract the noise component from the speech signal by computing the glottal residual (i.e., LF parametrization error). Then we will model the aspiration noise using three components, corresponding to the main characteristics identified by turbulence noise theory (Section 2.5):

- a constant leakage during the whole glottal cycle,
- a primary burst of noise occurring at the beginning of the closing-phase,
- and a secondary one after the glottis opens.

We will start by computing the glottal residual (i.e., LF parametrization error) as:

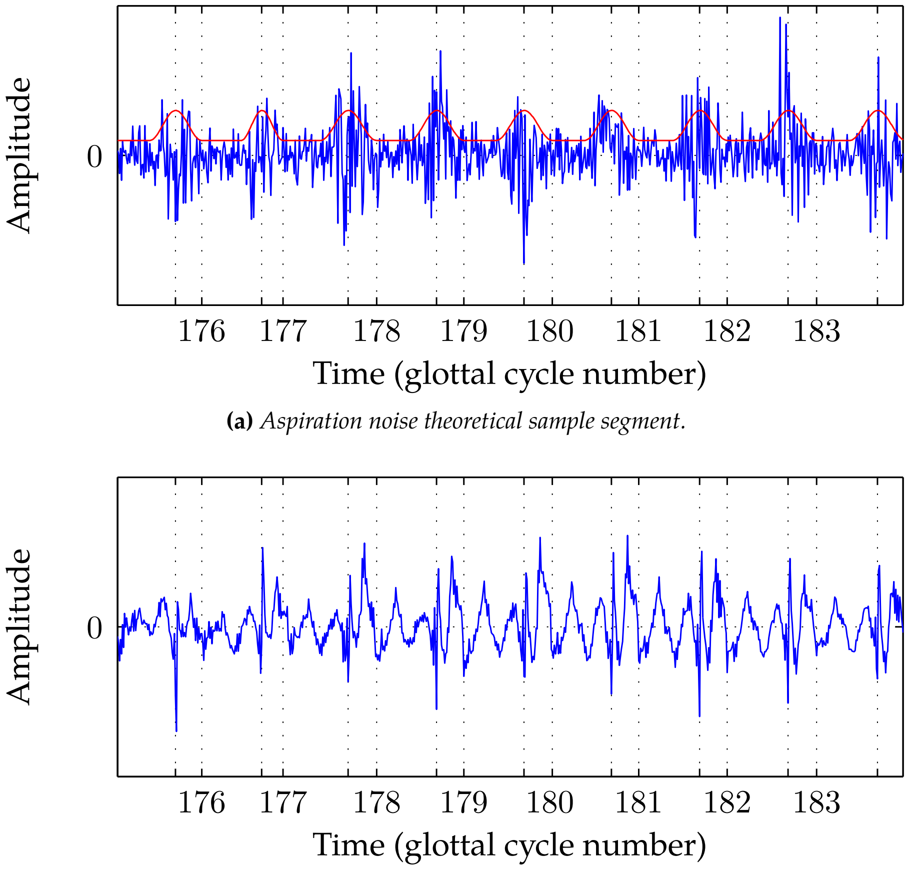
$$g_{res}(n) = g_{if}(n) - g_{lf}(n), \tag{3.46}$$

where g_{if} is the real glottal waveform obtained by inverse-filtering the speech with the vocal tract filter, and g_{lf} is the optimal parametrization using the LF model. Ideally, this residual should contain only the aspiration noise present in the original speech, but, as

$\frac{1}{1-2\mu\cos\omega+\mu^2}$. Since at the cutoff frequency f_c at 3 dB it holds $\|H(\omega = 2\pi\frac{f_c}{f_s})\|^2 = \frac{\|H(0)\|^2}{2}$, we can then compute $f_c = \frac{f_s}{2\pi}\arccos\left(2 - \frac{1+\mu^2}{2\mu}\right)$, with f_s being the sampling frequency.

expected, this is not the case due to several factors: real aspiration noise not considered in the convex formulation; modeling errors (LF glottal parametrization errors, vocal tract filter mismatch, etc.); intrinsic errors of the simplified model for the human speech production system (nasalization, coupling between vocal tract and sub-glottal cavities, etc.). As a result, the residual waveform does not correspond to the shapes one would expect from turbulent noise theory (Section 2.5, page 24).

In theory, the aspiration noise should consist of additive, white Gaussian noise present during the whole glottal cycle, and a burst of noise at the beginning of the closing phase. This can be clearly seen in Fig. 3.18: in the upper part (3.18a) we see an artificially generated segment of aspiration noise conforming to the theory; in the lower part (3.18b) we have a segment of the real glottal residual.



(a) *Aspiration noise theoretical sample segment.*

(b) *LF parametrization error real segment.*

Figure 3.18: *Comparison of a theoretical segment of aspiration noise (a) and a real segment of glottal parametrization residual (b)*

The objective here is to condition the glottal residual from (3.46), so it resembles the predicted aspiration noise waveforms, and to parametrize the resulting waveform with as few parameters as possible while retaining good resynthesis quality. The method we propose to parametrize the residual is divided in three steps:

- Residual whitening, to eliminate extraneous components in the glottal error (mainly occurring in the lower bands of the spectrum),
- Noise envelope extraction,
- Envelope parametrization.

The process is illustrated in Figure 3.19.

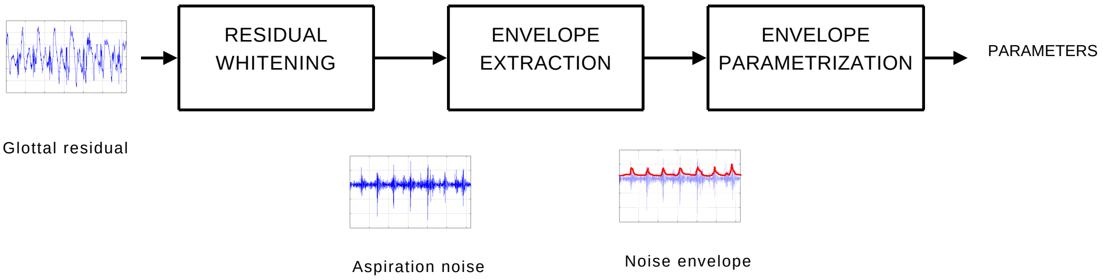


Figure 3.19: *Proposed scheme to parametrize the glottal residual.*

3.6.1 Residual whitening

The first step consist of whitening the residual by eliminating the extraneous components in the glottal error. Since they are mainly mainly present in the lower band of the spectrum, we have implemented the following three strategies:

- High-pass filtering (500, 1000, 2000, 3000 and 4000 Hz)
- Linear prediction (LP) filtering (orders 2, 4, 8 and 16),
- Combined LP and high-pass filtering.

The whole scheme is depicted in Figure 3.20. The procedure to select the best whitening technique involve visual inspection to assess that the residual waveforms have been well conditioned, and informal listening tests to evaluate whether the whitening process introduces unacceptable degradation. This was tested by adding the residual and LF waveforms, and resynthesizing them using the vocal tract filter estimated before.

The first obvious approach we tried was to high-pass filter the residual to eliminate the low-frequency fluctuations present in the signal. Figure 3.21 shows a segment of the original residual, and the result of high-pass filtering at different cut-off frequencies (500, 1000, 2000 and 4000 Hz). As we can see in the figure, the method is successful in eliminating the non-desired characteristics of the residual when we reach a cut-off frequency of

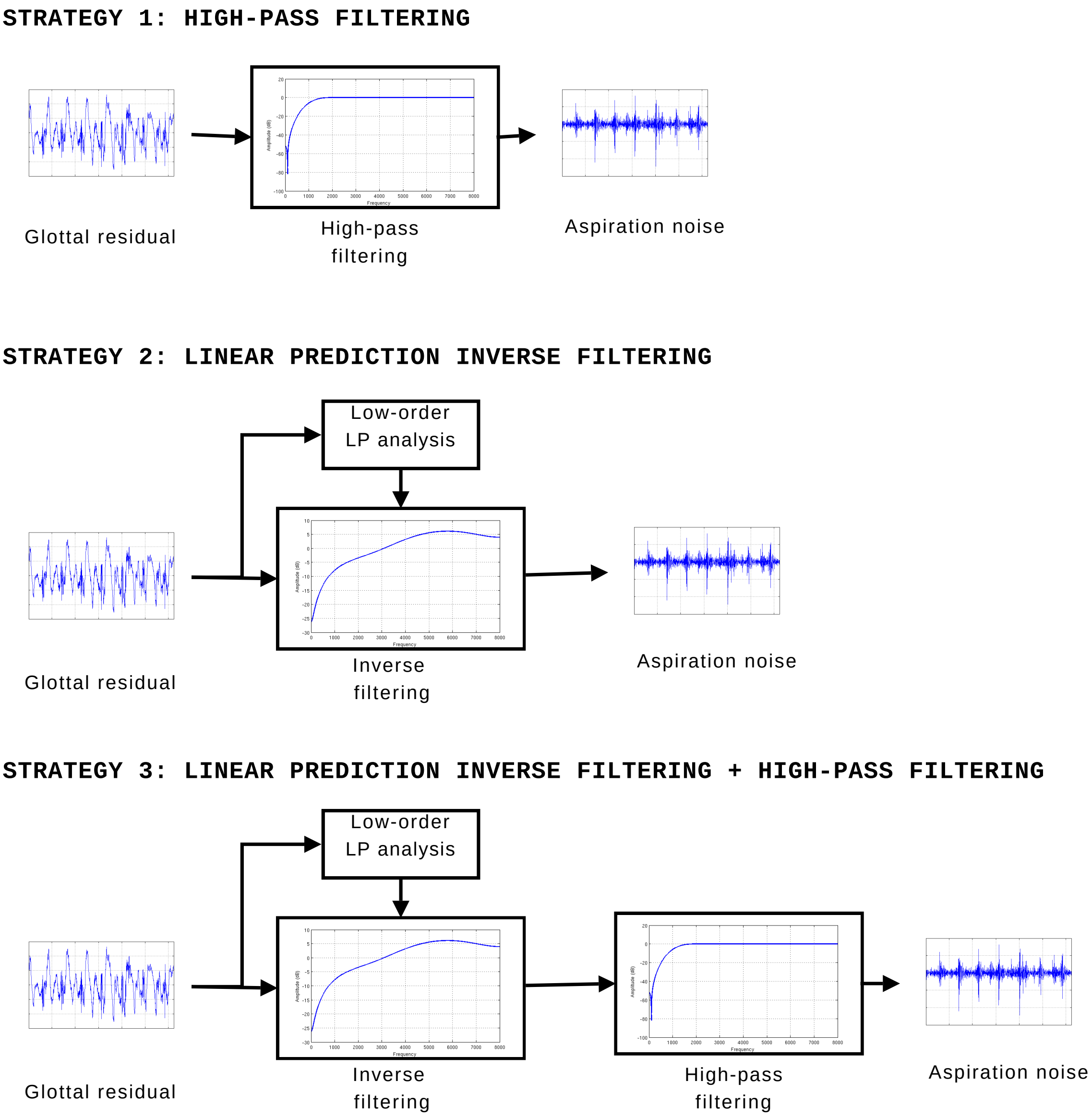


Figure 3.20: *Proposed scheme to whiten the glottal residual prior to parametrization.*

2000 Hz. In order to evaluate the quality, we performed an informal listening test where the high-pass filtered residual was added to the LF glottal waveform, and the resulting (noisy) glottal waveforms (one for each cut-off frequency) were synthesized using the vocal tract filters. The test showed that with a cutoff frequency higher than 1000 Hz the synthetic voice was noticeably degraded and it lacked naturalness.

Since with the high-pass filter at 1000 Hz the residual waveforms still contained extraneous components, we tried using linear prediction (LP) analysis to estimate a low-order filter, and then use it to whiten the residual. In Figure 3.22 we can see a segment of the

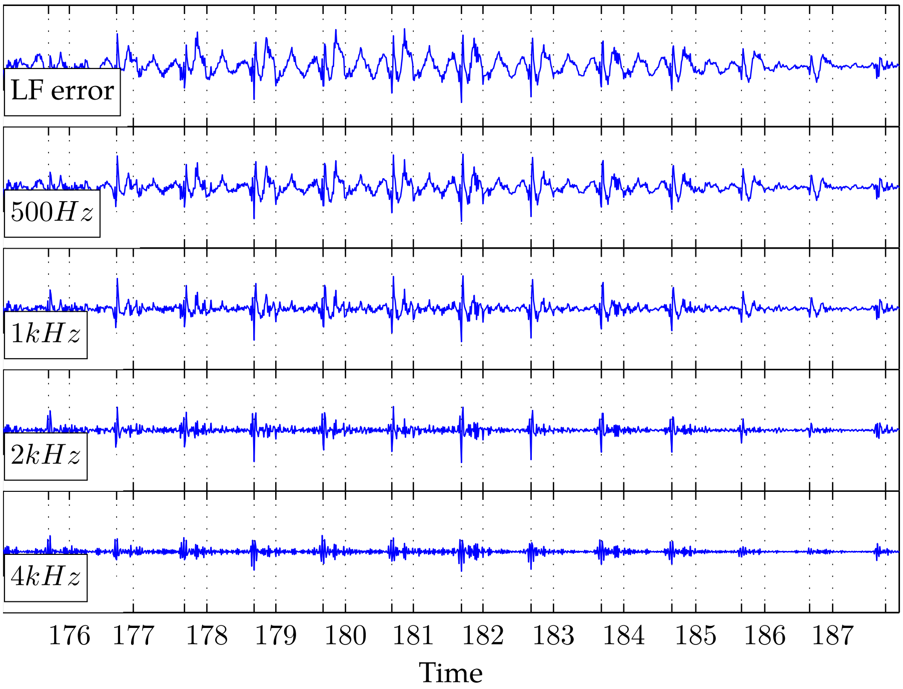


Figure 3.21: *Original LF parametrization error and residual after whitening using a high-pass filter with cut-off frequencies of* 500, 1000, 2000 *and* 4000 *Hz. The numbered* ticks *mark the gois; unnumbered* ticks *mark the gcis.*

original residual, and the result of applying a whitening filter using LP analysis of different orders. Visual assessment of the conditioned residual waveforms using this method did not show any advantage with filter orders higher than 2. However, we also ran a listening test where the whitened residual was added to the LF glottal waveform and then synthesized (one waveform for each of the LP orders), the results showing that the quality improves slightly when increasing from 2 to 4 coefficients the order of the filter. Further increases did not show any improvement in terms of quality.

However, in terms of waveform shape, the LP whitening step can be further improved by applying a high-pass filter to the LP residual. We tried again the same filters as in the high-pass filtering approach, with cut-off frequencies of 500, 1000, 2000 and 4000 Hz. Visual inspection of the results in Figure 3.23 show that from 1000 Hz onwards, the residual waveforms are more adequate for envelope extraction and parametrization.

The final whitening filter is then a combination of a 4th order LP filter (H_{lp}^{4}), followed by a high-pass filter with cut-off frequency $f_c = 1000\,\mathrm{Hz}$ ($H_{hp}^{1\,\mathrm{kHz}}$). In the z-domain, the whitened residual r_{white} is computed as:

$$R_{white}(z) = \frac{H_{hp}^{1\,\mathrm{kHz}}(z)}{H_{lp}^{4}(z)} \cdot G_{res}(z). \tag{3.47}$$

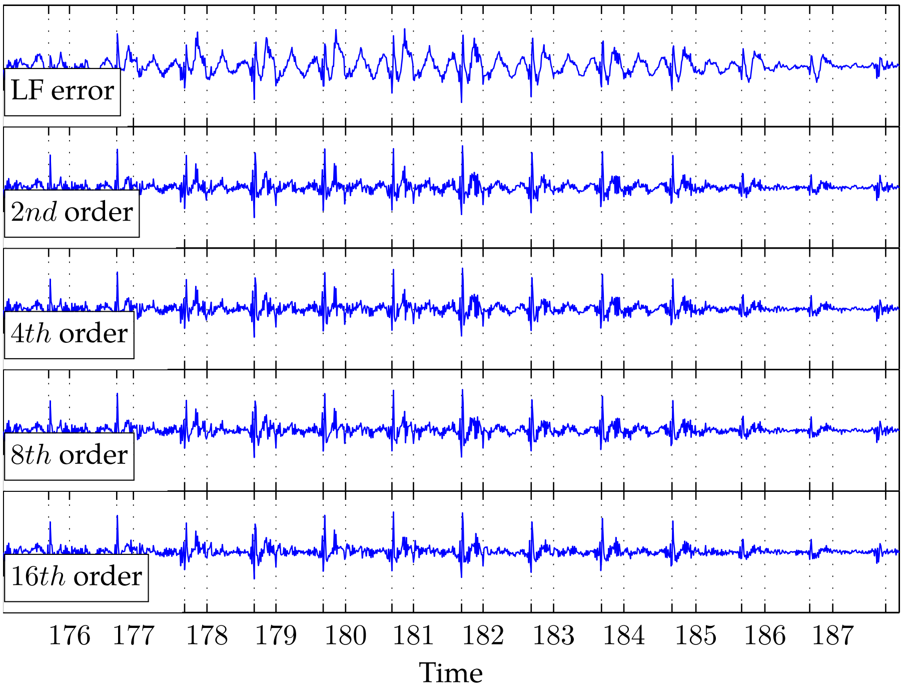


Figure 3.22: *Original LF parametrization error and residual after whitening using linear prediction (LP) filters of orders* $2, 4, 8$ *and* 16*. The numbered* ticks *mark the gois; unnumbered* ticks *mark the gcis.*

where G_{res} is the glottal residual from (3.46). Once the residual has been properly conditioned, we can proceed with the extraction and parametrization of the noise envelope.

3.6.2 Envelope extraction and parametrization

The envelope is extracted using the Hilbert transform, and then applying a low-pass filter to eliminate spurious components. Let r_{white} be the residual from (3.47) after the whitening process explained in the previous section. The envelope is computed as:

$$r_{env} = \sqrt{\check{r}_{white} \cdot \check{r}^*_{white}}, \tag{3.48}$$

where $\check{\cdot}$ denotes the Hilbert transform, and * denotes the conjugate. Prior to the parametrization step, we apply a low-pass filter $h_{lp}^{f_c}$ to r_{env} (cut-off frequency $f_c = 2\,f_0$), to reduce the effect of the noise on the fitting procedure explained below:

$$\tilde{r}_{env}(n) = h_{lp}^{f_c} \star r_{env}(n), \tag{3.49}$$

where $\star$ denotes convolution. In Fig. 3.24 we can see two glottal cycles of the whitened residual, the Hilbert envelope, and the low-pass filtered envelope used for parametriza-

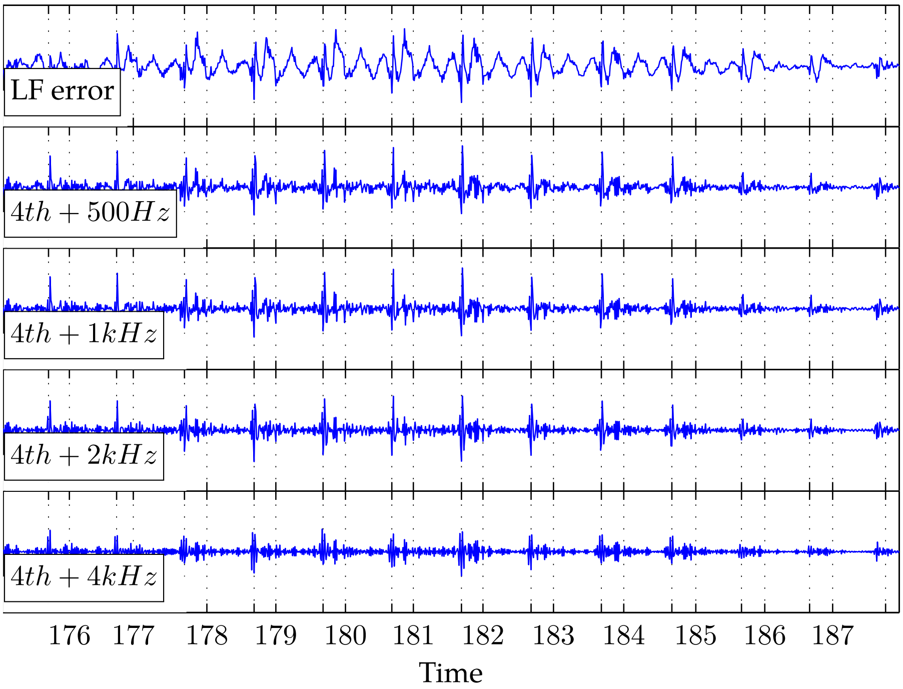


Figure 3.23: *Original LF parametrization error and residual after whitening using a combination of a 4th order LP filter, and a high-pass filter of cut-off frequencies of* 500*,* 1000*,* 2000 *and* 4000 *Hz. The numbered* ticks *mark the gois; unnumbered* ticks *mark the gcis.*

tion.

We will use amplitude modulated high-pass filtered white Gaussian noise to represent the aspiration noise. As we can see in Fig.3.25, the modulating envelope consists of three components: a constant, base-floor level (b_{lvl}), accounting for the constant leakage present during the whole glottal cycle, a window centered around the GCI, which accounts for the noise burst predicted at the beginning of the closing phase, and a second window account for the secondary burst occurring after the glottis opens. After inspecting the residuals, we experimented with three different windows for the modulation part: two bell-shaped, smooth windows (a Gaussian window W_{gau} and a Hanning window W_{han}),

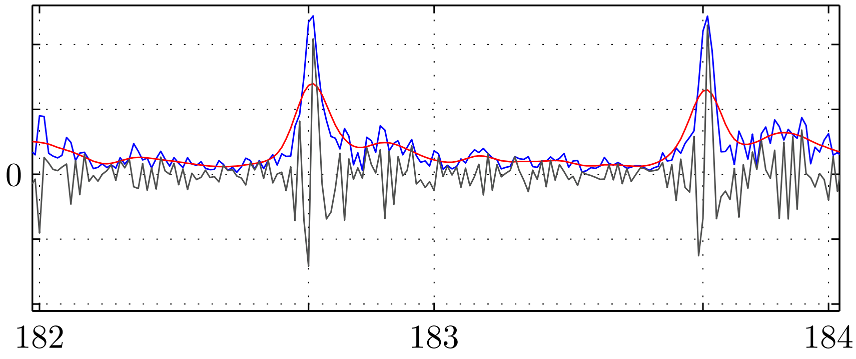


Figure 3.24: *Noise envelope extraction. Whitened glottal residual (grey), original Hilbert envelope (blue) and low-pass filtered final envelope (red) for two complete glottal cycles.*

and a sharper window (a exponential window W_{exp}). The windows are formulated as:

$$W_{gau}(t) = \begin{cases} e^{-\frac{1}{2}\left(\frac{t-w_c}{w_l}\right)^2} & 0 \leq t \leq w_c \\ e^{-\frac{1}{2}\left(\frac{t-w_c}{w_r}\right)^2} & w_c < t \leq 1. \end{cases}$$

$$W_{han}(t) = \begin{cases} 0 & 0 \leq t < w_c - w_l \\ \frac{1}{2} - \frac{1}{2}\cos\left(\pi \frac{t-w_c+w_l}{w_l}\right) & w_c - w_l \leq t \leq w_c \\ \frac{1}{2} - \frac{1}{2}\cos\left(\pi \frac{t-w_c+w_r}{w_r}\right) & w_c < t \leq w_c + w_r \\ 0 & 0 \leq t < 1. \end{cases} \tag{3.50}$$

$$W_{exp}(t) = \begin{cases} \left(e^{t\,w_l} - 1\right)/e^{w_c\,w_l} & 0 \leq t \leq w_c \\ \left(e^{w_r\,(1-t)} - 1\right)/e^{w_r\,(1-w_c)} & w_c < t \leq 1. \end{cases}$$

In all three cases, t is constrained to the normalized glottal cycle's duration: $t \in [0,1]$. w_c represents the window's center inside the glottal cycle, and w_l and w_r are the widths of the left and right parts respectively. These are not required to be equal in this generic *asymmetric* formulation.

In order to select the optimal type of window (Gaussian, Hanning or exponential), the number of windows (1 or 2) and their symmetry, for each of the three window types, we generated four synthetic envelopes using the following configurations:

$$\begin{aligned}\hat{r}^{a2}_{env}(n) &= b_{lvl} + w^1_{lvl} \cdot W^1_x(n/N; w^1_c, w^1_l, w^1_r) \\ &\quad + w^2_{lvl} \cdot W^2_x(n/N; w^2_c, w^2_l, w^2_r) \qquad (3.51)\\ \hat{r}^{s2}_{env}(n) &= b_{lvl} + w^1_{lvl} \cdot W^1_{x,sym}(n/N; w^1_c, w^1_l) \\ &\quad + w^2_{lvl} \cdot W^2_{x,sym}(n/N; w^2_c, w^2_l) \qquad (3.52)\\ \hat{r}^{a1}_{env}(n) &= b_{lvl} + w^1_{lvl} \cdot W^1_x(n/N; w^1_c, w^1_l, w^1_r) \qquad (3.53)\\ \hat{r}^{s1}_{env}(n) &= b_{lvl} + w^1_{lvl} \cdot W^1_{x,sym}(n/N; w^1_c, w^1_l) \qquad (3.54)\end{aligned}$$

where the superscripts ^{1} and ^{2} denote the window number, W_x is the window (Gaussian, Hanning or exponential), and the subscript $_{sym}$ indicates that the window is symmetric (i.e., $w_l = w_c$). The $\hat{r}_{env}$ superscripts $a2$, $s2$, $a1$ and $s1$ denote the symmetry of the window (a for asymmetric, s for symmetric), and the number of windows (1 or 2). Figure 3.25 shows a sample synthetic envelope $\hat{r}^{a2}_{env}$ using Hanning windows, and the corresponding synthetic aspiration noise.

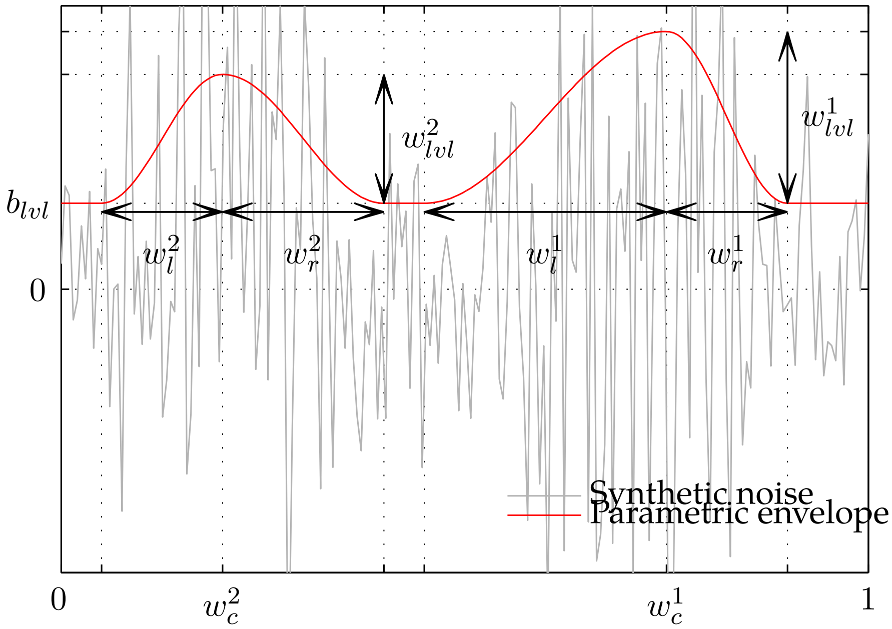


Figure 3.25: *Proposed method for residual envelope parametrization: a constant level (n_{lvl}) and two modulating Hanning windows: the primary window centered at w^1_c, with amplitude w^1_{lvl} and left and right amplitudes w^1_l and w^1_r. A secondary, optional window is situated at w^2_c, with amplitude w^2_{lvl} and left and right amplitudes w^2_l and w^2_r respectively.*

The synthetic envelope parameters from equations (3.51)–(3.54) are obtained by fitting each of them to the extracted, low-pass filtered envelope from equation (3.49). The fitting is done on a period-by-period basis, where for each period the following error norm is minimized using a non-linear least squares step:

$$\min_{\theta^{xy}} \|\hat{\mathbf{r}}^{\mathbf{xy}}_{\mathbf{env}}(\theta^{xy}) - \tilde{\mathbf{r}}_{\mathbf{env}}\|_2^2, \tag{3.55}$$

where $x \in [a, s]$ and $y \in [1, 2]$ as above, and θ^{xy} is the vector of parameters:

$$\theta^{a2} = (b_{lvl}, w^1_{lvl}, w^1_c, w^1_l, w^1_r, w^2_{lvl}, w^2_c, w^2_l, w^2_r)', \tag{3.56}$$

$$\theta^{s2} = (b_{lvl}, w^1_{lvl}, w^1_c, w^1_l, w^2_{lvl}, w^2_c, w^2_l)', \tag{3.57}$$

$$\theta^{a1} = (b_{lvl}, w^1_{lvl}, w^1_c, w^1_l, w^1_r)', \tag{3.58}$$

$$\theta^{s1} = (b_{lvl}, w^1_{lvl}, w^1_c, w^1_l)'. \tag{3.59}$$

To select the optimal parametrization (window type; symmetric or asymmetric windows; 1 or 2 windows), we compute the SNR between the real and the 12 parametrized noise envelopes as:

$$SNR_r = 10\ \log_{10}\left(\frac{\tilde{\mathbf{r}}_{\mathbf{env}}}{\hat{\mathbf{r}}^{\mathbf{xy}}_{\mathbf{env}}(\theta^{xy}) - \tilde{\mathbf{r}}_{\mathbf{env}}}\right)^2 \tag{3.60}$$

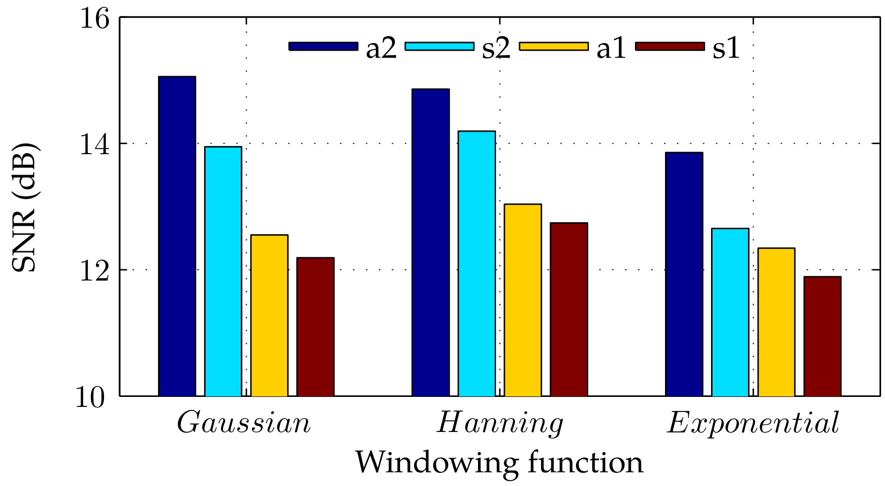


Figure 3.26: *Noise envelope parametrization results, using three different windowing functions (Hanning, Exponential, Gaussian), symmetric (s) and asymmetric (s), with 1 or 2 windows per cycle.*

Figure 3.26 shows the results in terms of averaged SNR between the real and the parametrized noise envelopes. As the figure shows, the two bell-shaped windows work better than the sharp window. Using the complete configuration (2 asymmetric windows per cycle), the Gaussian window gives slightly better results than the Hanning window. However, once we start reducing the parameter set dimensionality, the later performs better. By informal listening tests it was determined that there was no noticeable degradation when using 1 symmetric window, and since we are interested in reducing the dimensionality of our overall parameter set, we will choose the configuration with 1 symmetric Hanning windows for the rest of this work. Thus, the equation to generate the synthetic noise envelope using one symmetric Hanning window is:

$$\hat{r}^{s1}_{env}(n) = b_{lvl} + w_{lvl} \cdot W_{hann,sym}(n/N; w_c, w_l) \tag{3.61}$$

where, as explained before, $W_{hann,sym}$ is a symmetric Hanning window, b_{lvl} is the level of the noise base, w_{lvl} the Hanning window amplitude, w_c its position inside the glottal cycle, and w_l the window width.

To generate the synthetic residual, the envelope $\hat{r}^{s1}_{env}$ from eq. (3.61) is used to modulate in amplitude high-pass filtered (cutoff frequency of 1 kHz, as used when whitening the residual) white Gaussian noise for the duration of the glottal period. Mathematically, the synthetic glottal residual $\hat{g}_{res}$ can be written as:

$$\hat{g}_{res} = \hat{r}^{s1}_{env} \cdot (\mathcal{N}(0,1) \star h_{hp}), \tag{3.62}$$

where $\mathcal{N}(0,1)$ is additive white Gaussian noise (zero mean and unitary variance), $\star$ indicates convolution and h_{hp} is the high-pass filter.

3.7 Synthetic speech generation

Once the analysis of the voiced segments is completed, we have have a set of parameters for each glottal period. Each voiced period is synthesized by:

- generating the LF glottal waveform using eq. (2.15),
- generating the artificial aspiration noise using eq. (3.62),
- adding the LF and noise waveforms,
- and filtering them with the vocal tract filter using eq. (3.2).

This can be formulated as:

$$\hat{s}(n) = g^l_{lf}(n) + \hat{g}^l_{res}(n) + \sum_{k=1}^{N} \hat{a}^l_k \, \hat{s}(n-k), \quad n^l_{beg} \leq n \leq n^l_{end}. \tag{3.63}$$

where ^{l} denotes the glottal period l, g_{lf} is the LF glottal waveform, $\hat{n}_{asp}$ is the artificial aspiration noise, and n^l_{beg} and n^l_{end} indicate the beginning and end of each glottal period. The LF waveform is generated by directly using the model equation (2.15).

Before generating the synthetic speech, we need to complete the analysis of the unvoiced speech segments. We use a standard LPC approach, where each unvoiced segment is divided into 10ms frames, and each of them is analyzed by Linear Prediction (LP). As a result, for each unvoiced frame, we have a vector of LSFs coefficients representing the vocal tract filter, and a scalar noise variance representing the error. Thus, each unvoiced frame is synthesized by:

- generating zero-mean, white Gaussian noise with the appropriate variance,
- and filtering it with the vocal tract filter.

The complete synthesis scheme is depicted in Fig.3.27.

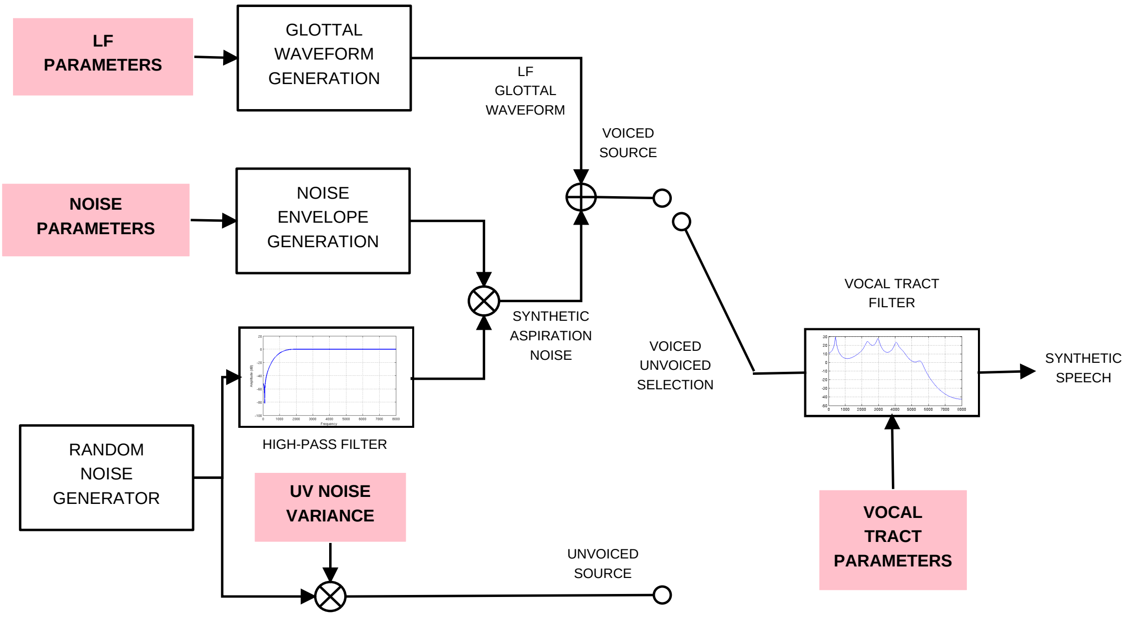


Figure 3.27: *Block diagram of the complete speech generation algorithm.*

3.8 Conclusions

The idea of exploiting the convexity of the problem resulting from minimizing the error between the inverse-filtered and parametric (Rosenberg-Klatt) glottal waveforms has been used in a number of studies. Lu (2002) presented a method for singing synthesis in which the singing speech was analyzed on a cycle-by-cycle basis, and an extra optimization step was introduced to obtain the open quotient, necessary for the location of the gci/goi. The aspiration noise was obtained by performing waveform denoising using wavelet decomposition, and then it was modeled using a similar envelope parametrization approach to the one we have presented here. Kim (2003) presented a similar procedure using warped linear-prediction integrated into a HMM-based analysis/synthesis algorithm for singing speech. The possibility of several cycles being simultaneously analyzed was included to deal with the extremely short periods corresponding to the higher frequencies common in singing speech. The aspiration noise of the model was tackled by means of a stochastic codebook obtained using principal component analysis (PCA). Our source-filter decomposition algorithm (Pérez and Bonafonte, 2005, 2009) is based on the method proposed in Lu (2002), differing in several aspects in order to increase the robustness of the algorithm. We locate the goi/gci using the laryngograph signal and then proceed to the optimization of this initial estimation using convex optimization. The multi-cycle analysis we propose to increase the robustness of the estimations differs from that in Kim (2003) in that we tie the amplitudes of the glottal cycles in the analysis frame to obtain smoother estimates of the source-filter parameters. A similar approach to that in Lu (2002) or Pérez and Bonafonte (2005) was later proposed in del Pozo (2008); del Pozo and Young (2008) to estimate the source-filter components in a voice conversion system, with some differences. The glottal tilt parameter μ was estimated outside of the convex optimization step by means of adaptive pre-emphasis. The glottal aspiration noise is also obtained by means of wavelet denoising as in Lu (2002), but it is modeled differently: additive white Gaussian noise is modulated using the LF waveform, and its energy is adjusted so to match that of the original noise estimate. In our algorithm, the parametrization of the aspiration noise is done by means of modulated white Gaussian noise, but our method differs from other approaches in that we directly use the glottal parametrization error between the inverse-filtered and fitted LF glottal waveforms to obtain the parameters of the modulating envelope. During the estimation of the LF model, we have proposed the use of the KLGLOTT88 parameters obtained in the convex decomposition step to initialize the non-linear LF estimation, thus increasing its robustness.

In this section we have presented in detail the source-filter decomposition and parametrization algorithm, together with the synthesis schema used to generate synthetic speech using the model parameters (Pérez and Bonafonte, 2005, 2009). The convex decomposition step has been detailed first since it is being used in different blocks of the algorithm, both in its single and multi-cycle formulation. We have seen how to obtain the vocal tract filter and the estimated glottal waveform by inverse-filtering the speech signal. We then have explained how to reparametrize the glottal waveform using the LF model, and how the glottal residual is analyzed and parametrized. To end the chapter, the final synthesis schema including all the human speech production system components has been depicted. In the next chapter we will present the results of the different steps and tests we carried on to validate and evaluate the performance of the proposed analysis/synthesis algorithm.

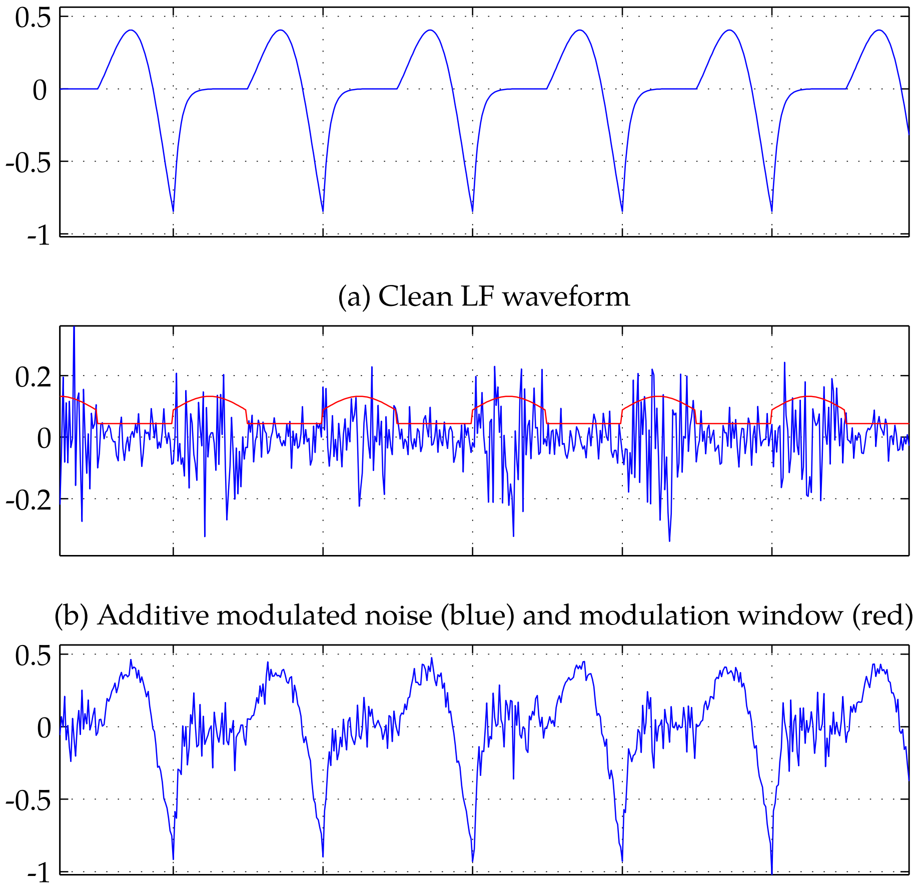
CHAPTER 4
Algorithm evaluation and resynthesis quality

In the previous chapter we have described in detail the algorithm to decompose the speech signal into the voice source (glottal and residual waveform) and the vocal tract components of the human speech production system. Now we will proceed to explain the steps we have performed to evaluate the algorithm performance. We will start by presenting the results obtained using a corpus of synthetic data created using glottal values reported in the literature and from our own work (Section 4.1). Next we will proceed to analyze the performance of the parametrization algorithm using a corpus of real vowels recorded specifically for this thesis (Section 4.2) and will present the results of an online evaluation test pairing two proposed resynthesis methods against a reference vocoder (Section 4.3).

4.1 Evaluation using synthetic data

The purpose of this work is to assess the quality of the source-filter deconvolution process, in terms of glottal and vocal tract parameter matching. For this reason, we have created a synthetic corpus combining realistic glottal source obtained during our own research with glottal and vocal tract parameters obtained in several studies and reported in the literature. Since the source parameters are known *a priori*, we use them as reference and compare them with the estimated parameters using our convex optimization method. Thus, we can compute objective measure of the performance of our algorithm. Table 4.2 in page 80 contains the 33 LF configurations used to generate the reference corpus and the phonation mode or voice quality associated to it. Configurations 1–6 are obtained from our own research and the rest have been previously reported in the literature: 7–11 are taken from van Dinther et al. (2005), 12–21 from Childers and Lee (1991) and 22–33 from Karlsson and Liljencrants (1996). The R parameters are expressed in % and the fundamental frequency F_0 in Hz.

The reference corpus of synthetic material is generated using the aforementioned configurations as follows. Each configuration is used to create a reference glottal source signal by concatenating the same pulse 30 times. For this experiment we work with signals sampled at 8 kHz. We then add white Gaussian noise to each of the glottal signals at different SNR levels (from 5 dB to 20 dB, in 5 dB increments), amplitude-modulated using a Hanning window placed at the GCIs (this is motivated by turbulence noise theory from Section 2.5). The levels of SNR were empirically determined to represent the different degrees of breathiness present in real speech data by means of an informal perceptual test. Figure 4.1 illustrates this procedure for several glottal cycles.



(a) Clean LF waveform

(b) Additive modulated noise (blue) and modulation window (red)

(c) Resulting source waveform

Figure 4.1: *Example of the noise modulation used to generate the synthetic corpus set*

These source signals are then filtered with a vocal tract filter constructed using 4 formants (8 coefficients) obtained from van Dinther et al. (2005), corresponding to a vowel /a/ in modal phonation, located at 790 Hz, 1320 Hz, 2340 Hz and 3600 Hz, with bandwidths of 90 Hz, 90 Hz, 142 Hz and 210 Hz respectively. The frequency response of the resulting vocal tract filter is shown in Figure 4.2 up to half the sampling frequency.

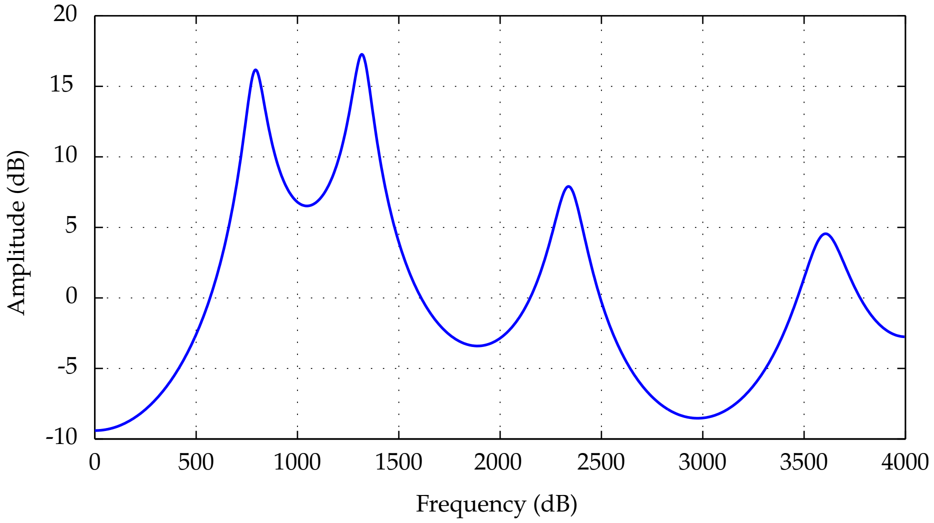


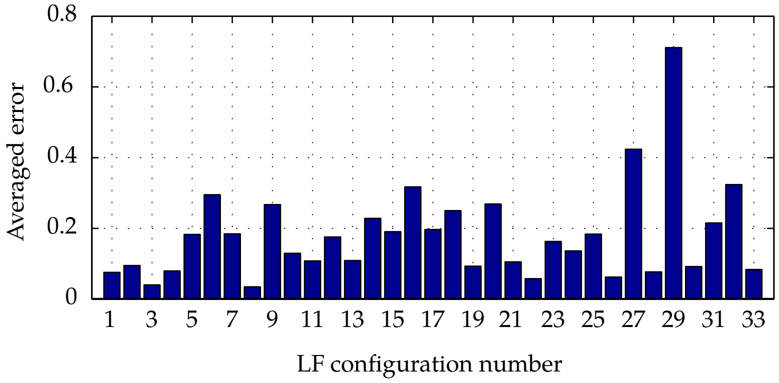
Figure 4.2: *Frequency response of the vocal tract filter used to generate the synthetic corpus. The formant frequencies are 790* Hz*, 1320* Hz*, 2340* Hz *and 3600* Hz*; the respective bandwidths are 90* Hz*, 90* Hz*, 142* Hz *and 210* Hz

To evaluate the quality of the estimation of each of the LF parameters R_a, R_k, R_g, E_e or F_0, we compute the estimation error using Percentile Error (PE) (Strik, 1998):

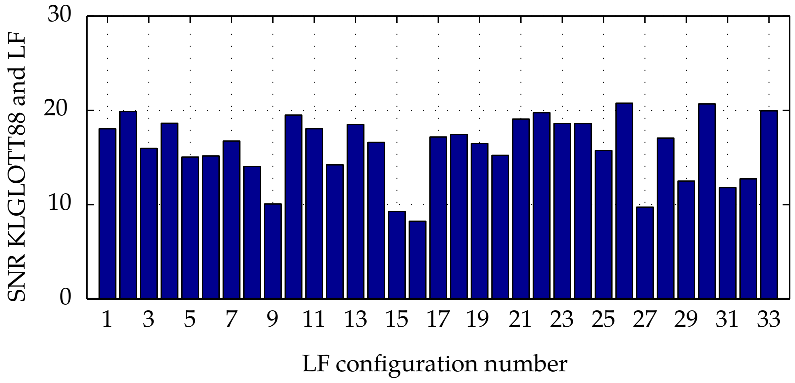
$$PE = 100 \cdot \frac{|\hat{P} - P|}{P}, \tag{4.1}$$

where P is the reference LF parameter from Table 4.2 page 80 used to generate the synthetic utterance, and $\hat{P}$ is the estimation using our algorithm. Table4.1 shows the averaged estimation error for each of the LF parameters and each of the SNR levels of additive noise. The value in parenthesis is the standard deviation of the error.

Since we are using the KLGLOTT88 model for the initial parametrization, and the test corpus is being generated using the LF model, the final performance of the parametrization algorithm may depend on the ability of the KLGLOTT88 model to match each of the LF configurations presented in Table 4.2. To check whether this holds true, we have analyzed the input LF glottal waveforms using the KLGLOTT88 reparametrization explained in Section 3.2.2. By using g_{lf} instead of g_{if} in eq. (3.32), we can obtained optimal KLGLOTT88 approximations of their LF counterparts. To evaluate the matching quality of the KLGLOTT88 parametrization we have used the standard Signal-to-Noise Ratio (SNR)



(a) *Overall error of the different LF configurations*



(b) *SNR between original (LF) and matched (KLGLOTT88) glottal waveforms*

Figure 4.3: *Relation between signal-to-noise ratio (SNR) between original LF and matched KLGLOTT88 glottal waveforms and overall error of the different LF configurations*

Table 4.1: *Mean estimation error (in %) of the LF parameters for each SNR level (in parenthesis the standard deviation)*

		SNR (dB)			
		5	10	15	20
LF param	Ra	12.1 (35.2)	8.5 (25.4)	7.5 (30.2)	7.5 (31.3)
	Rg	2.4 (7.4)	2.4 (7.4)	2.4 (7.4)	2.4 (7.1)
	Rk	4.6 (13.0)	3.7 (10.5)	3.5 (10.2)	3.4 (10.2)
	Ee	3.0 (8.1)	2.5 (6.9)	2.3 (6.5)	2.6 (7.1)
	F0	1.2 (8.5)	1.1 (7.9)	1.0 (6.1)	1.0 (6.0)

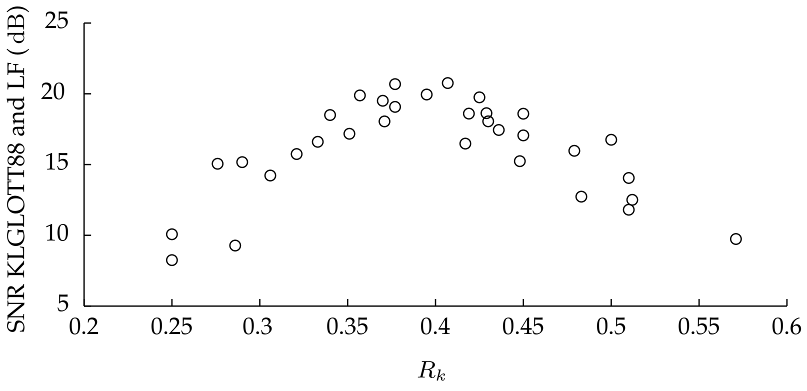
measure commonly used in speech coding:

$$SNR(x,\hat{x}) = 10 \cdot \log \frac{\frac{1}{N}\sum_n^N x(n)^2}{\frac{1}{N}\sum_n^N (x(n) - \hat{x}(n))^2}, \tag{4.2}$$

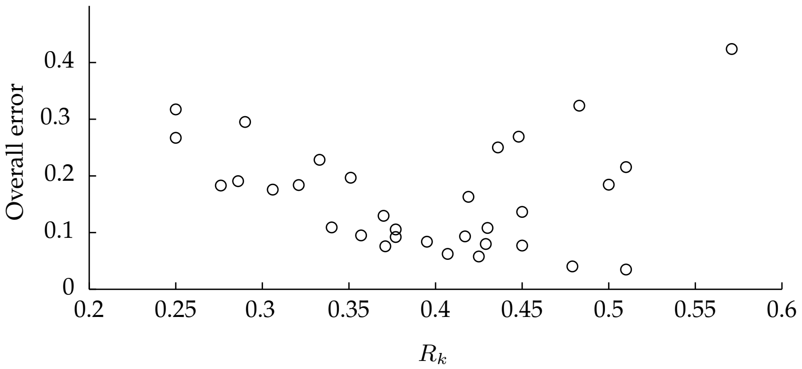
where x is the original waveform (LF in our case), $\hat{x}$ is the estimated waveform (KLGLOTT88 in our case), and N is the frame length being considered. We compute the SNR on a cycle-by-cycle basis using eq. (4.2), and then average the results to obtain an overall SNR for the whole glottal waveform. Figure 4.3 shows the results of this evaluation. The top figure (4.3a) shows the overall error for each of the LF configurations used in this experiment. The bottom figure (4.3b) shows the SNR between the original LF and matched KLGLOTT88 glottal waveforms for each of the configurations. Although it does not hold true for all the cases, we can observe that the overall performance of the algorithm tends to be lower the worse the KLGLOTT88 matching is. So it appears that when dealing with synthetic test material the matching ability of the KLGLOTT88 model has a clear impact on the overall performance, although the performance is in all cases satisfactory.

As we saw in Section 2.2.2, one of the main differences between the two models is the ability of the LF model to use a variable asymmetry coefficient (R_k), whereas in the KLGLOTT88 case it is fixed. This difference has a clear impact on the performance, as Figure 4.4 shows. In the top figure 4.4a we have the SNR between the LF and KLGLOTT88 models plotted as a function of R_k. The bottom figure 4.4a shows a similar plot for the overall performance case. As we can see, in the $[0.30.5]$ range the performance is very good, with high values of SNR and low overall error, but outside this range it starts to decrease, although it is still in the acceptable range as proven by the good overall LF parameter estimation results that Table 4.1 shows.

Once that we have evaluated the algorithm using synthetic data to obtain objective measures of its performance, we can proceed with the tests using real utterances. In this



(a) *Signal-to-noise ratio (SNR) between original LF and matched KLGLOTT88 glottal waveforms versus R_k LF parameter*



(b) *Overall error of the different LF configurations versus R_k LF parameter*

Figure 4.4: *Impact of the R_k parameter on the overall error and SNR between LF and KLGLOTT88 waveforms*

case, the evaluation needs to be more subjective, since no reference parameters exist to be compared with. Next section explains the details of the real data corpus and its evaluation.

	R_a	R_k	R_g	F_0	Voice Quality
1	4.1	37.1	134.1	170	Modal
2	7.6	35.7	83.9	170	Modal
3	1.3	47.9	126.0	132	Low F0
4	3.5	42.9	88.1	144	Low F0
5	13.1	27.6	98.3	281	High F0
6	11.7	29.0	75.8	340	High F0
7	0.6	50.0	108.7	110	Modal
8	2.0	51.0	92.1	110	Lax
9	1.1	25.0	152.4	110	Tense
10	1.8	37.0	126.9	110	Modal
11	3.5	43.0	110.0	110	Lax
12	2.1	30.6	102.0	106	Modal
13	2.5	34.0	94.4	127	Modal
14	1.5	33.3	98.0	154	Modal
15	0.8	28.6	102.1	84	Slight vocal fry
16	0.5	25.0	250.0	45	Vocal fry
17	13.3	35.1	87.7	344	Falsetto
18	4.3	43.6	80.7	213	Falsetto
19	6.8	41.7	104.2	137	Breathy
20	10.0	44.8	86.2	200	Breathy
21	2.0	37.7	127.5	126	Normal
22	2.6	42.5	116.8	102	Low F0
23	5.1	41.9	93.4	190	Low F0
24	1.5	45.0	129.5	131	Medium F0
25	9.9	32.1	75.9	288	High F0
26	2.7	40.7	102.0	129	Low level
27	10.5	57.1	97.0	249	Low level
28	1.9	45.0	127.2	127	Medium level
29	3.7	51.2	111.2	258	Medium level
30	1.6	37.7	140.5	132	High level
31	4.6	51.0	116.2	131	Breathy
32	8.1	48.3	93.9	254	Breathy
33	1.3	39.5	170.1	128	Pressed

Table 4.2: *LF parameters and associated voice qualities used for the synthetic data set: 1–6 are obtained from our own research, 7–11 are taken from (van Dinther et al., 2005), 12–21 from (Childers and Lee, 1991) and 22–33 from (Karlsson and Liljencrants, 1996) (R parameters in % and F_0 in* Hz*)*

4.2 Evaluation with real vowels

For the evaluation with real data, we used a small data set was recorded for the main purpose of performing voice quality analysis. The examples in this corpus consist of the 5 different Spanish vowels (/a/, /e/, /i/, /o/, /u/) uttered in isolation by a female professional speaker, with different voice qualities: modal, rough, creaky and falsetto. The main characteristics of each of these qualities will be detailed in Chapter 7, our main interest here is that the utterances were sustained vowels (roughly 2–3 seconds long), and thus suitable for our validation purposes.

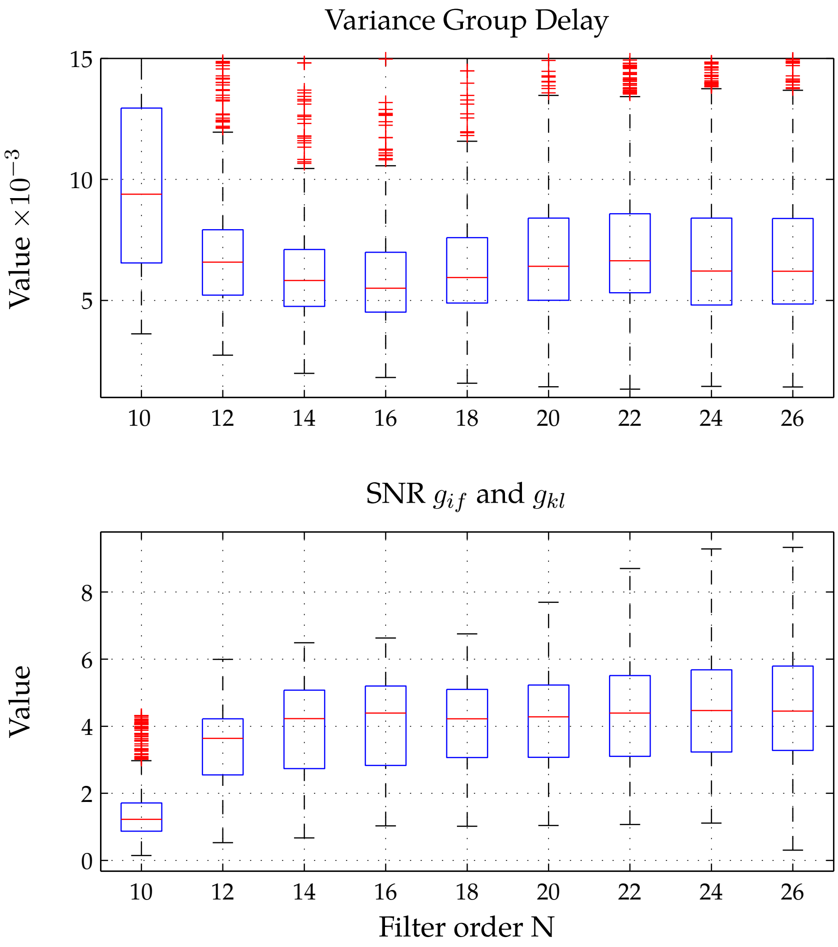


Figure 4.5: *Variance of the Group Delay (top) and SNR between g_{if} and g_{kl} (bottom) for various filter orders.*

It is well known that evaluating a glottal extraction algorithm in a real world scenario is difficult, due to the lack of reference (Bäckström et al., 2005). One possible way is to compute the averaged SNR (dB) between g_{kl} and g_{if}, since this gives an idea of how well it approximates a idealized glottal waveform. We also use the Group Delay (GD) function of the glottal waveform, since it has been shown to perform well (Alku et al., 2005). We chose to minimize the variance of the GD, as ideally it should be close to zero if all the formants have been removed in the inverse-filtering process. In order to select the optimal filter order for the vocal tract, the algorithm (without LF parametrization) was run using several filter orders (from $N = 8$ until $N = 24$, even orders only). We found that when the filter order increased, the SNR between the estimated and the parametrized glottal waveforms also increased (Fig. 4.5, bottom part), and seemed to stabilize from 16 onwards. Visual inspection of the resulting waveforms showed that orders higher than 18 produced sub-optimal glottal waveforms, since the return phase (defined by $TL(z)$) was being over-estimated. This resulted in non-existing closed phases, which should not happen for modal voices. By observing the box-plots for the variance of the GD, we then found that the optimal filter order was 16. The length of the OLA window was set to 3, with independent amplitudes for each cycle, after observing that this resulted in a higher continuity of the estimated glottal parameters. Figure 4.6 shows an example of the smooth evolution of the estimated LF parameters for the vowel /e/ in modal phonation, as expected since it was being sustained.

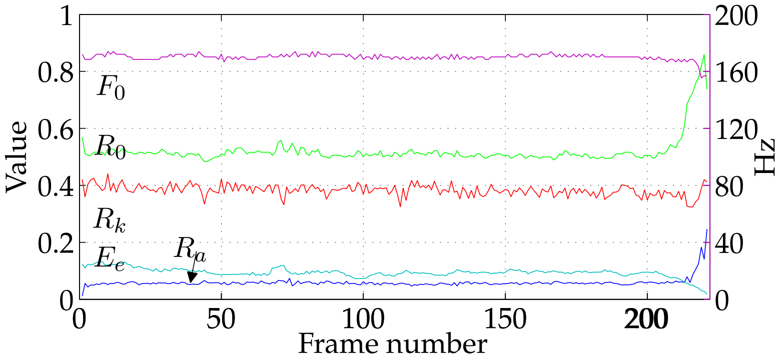


Figure 4.6: *Estimated R parameters, E_e amplitude (left axis) and fundamental frequency F_0 in HZ (right axis) for the vowel /e/ in modal phonation.*

Table 4.3 presents the SNR results (between g_{if} a g_{lf}) for the real data set, using the optimal filter order and OLA lengths determined before. The SNR needs to be taken cautiously as an absolute quality measure, since it considers the original aspiration noise present in the speech as parametrization error, thus resulting in a degradation of the SNR performance. For this reason, we have also performed the SNR evaluation using low-pass filtered versions of g_{if} and g_{lf}, that should reflect more closely the real parametrization performance (the results are shown in parenthesis). The filtering has been done by means of a convolution with a Blackman window of length 7.

As we can see, the algorithm performs well in most cases, with SNR values of more

	Voice phonation							
	Modal		Rough		Creaky		Falsetto	
/a/	11.45	(12.38)	9.12	(10.16)	12.91	(14.10)	13.12	(16.58)
/e/	15.50	(16.55)	7.05	(7.93)	14.31	(15.59)	9.52	(11.61)
/i/	10.22	(11.22)	6.05	(7.57)	9.75	(10.44)	2.19	(4.62)
/o/	9.52	(12.50)	8.09	(9.20)	12.67	(14.10)	5.24	(9.77)
/u/	11.21	(12.45)	4.56	(6.27)	9.42	(10.73)	10.00	(12.86)

Table 4.3: *SNR between inverse-filtered and LF glottal waveforms (in parenthesis, SNR computed using low-pass filtered versions of both waveforms)*

than 10 dB, even with the higher F_0 values (around 350 Hz) representative of the falsetto voice (the exception being /i/ in this case). This ability to work well with very short cycle lengths is an advantage of our method over traditional closed-phase inverse-filtering methods, since the later often degrade due to the small amount of speech samples present during the closed phase of the glottis. As expected, the SNR results for the rough phonation are lower than for other modes, since this phonation usually entails higher cycle-to-cycle variability and larger amounts of turbulences (as will be explained in Chapter 7), not totally accounted for by the Blackman low-pass filtering.

4.3 Resynthesis evaluation

We performed an evaluation aimed at evaluating the resynthesis performance of our algorithm. We decided to test two different methods in this evaluation. The first one is the algorithm explained in the previous chapter, where parametric models are adopted for the three components of the speech production model (vocal tract, glottal waveform and residual). The speech is synthesized according to the procedure detailed in Section 3.7. For the second method, we used the whitened residual waveform from Section 3.6.1 without parametrization. The speech was synthesized by first constructing the glottal LF waveform, adding the whitened residual, and filtering the resulting noisy glottal waveform with the vocal tract. This would allow us to evaluate the contribution of the glottal parametrization and residual parametrization in a separate way. We compared our two methods against the STRAIGHT channel vocoder (Kawahara et al., 1999), a high-quality analysis/synthesis algorithm which is widely regarded as the reference method for vocoder-like techniques. It uses a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction.

We conducted an on-line listening test to evaluate the subjective quality of the resynthesized speech as follows. We used the corpus that we recorded during the TC-STAR project (Bonafonte et al., 2006b). This corpus was recorded in a professional studio using high quality recording settings (96 kHz and 24 bits/sample), two microphones (membrane and close-mouth) and a laryngograph. In this test, only the membrane microphone and the laryngograph signals were used, after being downsampled to 16 kHz and 16 bits/sample. We resynthesized 4 sentences for each of the 3 methods we were comparing, so the evaluators were presented with a total of 12 sentences. These sentences contained utterances from four different speakers (two female and two male). The test is a standard mean opinion score (MOS) test, where the participants were asked to first listen carefully to each of the samples, and then rate each sentence using a 5 point scale:

1. Bad (distortion very annoying)
2. Poor (distortion annoying)
3. Fair (distortion slightly annoying)
4. Good (distortion perceptible but not annoying)
5. Excellent (distortion imperceptible)

The utterances were presented in a random order, so the users were unaware of the origin of the examples they were evaluating. The subjects were asked to use headphones if possible, and were allowed to listen to each sample as many times as necessary.

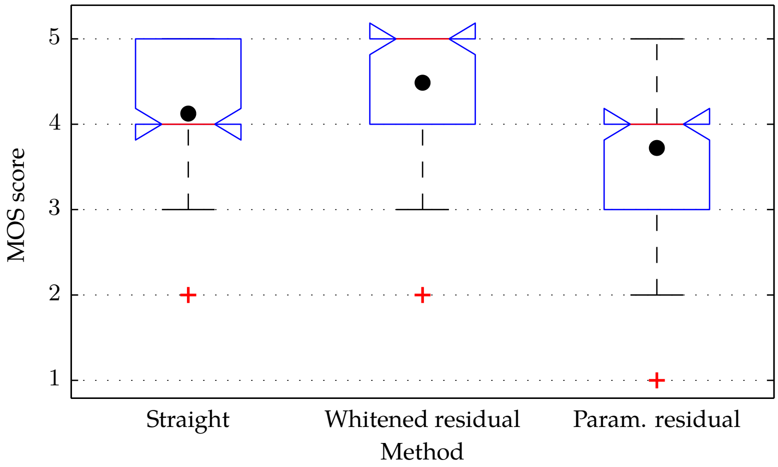


Figure 4.7: *MOS scores of the resynthesis evaluation test for the method of reference (Straight) and our two proposed methods (using resampled and parameterized residual).*

.

The results of the test are presented in figure 4.7 using boxplots. On each box, the central line is the median, the box edges are the 25th and 75th percentiles, and the whiskers extend to the furthest data points not considered outliers (marked with crosses). The black dot inside the box represents the mean value of the data. As we can see, our method using the whitened residual results in the highest rated synthetic speech, with a median MOS score of 5 and most of the ratings inside the $[45]$ range. The performance of the STRAIGHT methods is also very high, the median being 4 and also with the majority of scores in the $[45]$ range. The performance of these two methods was expected to be high, since we are not yet performing any modification (in terms of duration or pith), and in this case they act as a nearly distortion-free reconstruction algorithm. The complete parametrization results in a score slightly worse than that of the STRAIGHT method. While the median value is also 4, the boxplot indicates that most of the scores are in the $[34]$ range, as can be seen by the median indicator. The results of the test are satisfactory, proving that our decomposition and reconstruction algorithm can produce synthetic speech of acceptable quality similar to that of reference algorithms. However, the performance degradation observed with our second method (using the full parametrization of the residual) indicates that this is an aspect that requires improvement and further investigation.

4.4 Conclusions

In this chapter we have presented the evaluation of the parametrization algorithm detailed in chapter 3. We have created a synthetic corpus using glottal features reported in the literature and from our own research to be use as reference for the objective evaluation of the extracted LF glottal measures. As we have seen, the algorithm was able to estimate all the values with a low error for different values of additive noise, with R_a being as expected the parameter where the error figures were higher, since it is widely reported as the most difficult to estimate. A small corpus of real speech utterances containing sustained vowels was also recorded for the purpose of this work. It has been used to validate the parametrization method by means of continuity plots of the different LF parameters for a whole utterance, and to validate the election of the vocal tract filter order by means of both SNR plots between the inverse-filtered and the parametrized glottal waveforms, and glottal quality measures (variance of the group delay). To end the evaluation, we have conducted an online listening test to rate the quality of the resynthesis capabilities. We have proposed two methods, the difference being the treatment of the parametrization error or residual, and compared them to the vocoder method STRAIGHT, widely accepted as the standard reference in this area. The results show that the algorithm perform well when using the whitened residual without parametrization, but its performance degrades when using the fully parametrized residual. Although it is still rated in the Fair-Good range, further investigation would be required in this area to increase the quality of the synthetic speech. In next chapter we will present the algorithm to perform prosody modification, necessary for the inclusion of the proposed analysis and synthesis algorithm in speech processing applications like speech synthesis or voice conversion.

CHAPTER 5

Prosody modification

In the previous two chapters we have explained in detail the analysis and synthesis algorithm we propose for the decomposition of the speech signal into its source-filter components. We have evaluated the resynthesis capabilities both in objective and subjective terms, using artificial data and online MOS tests. We will detail now the techniques developed for performing prosody modification of the speech samples, since this would allow our analysis/synthesis algorithm to be applied to speech processing tasks such as speech synthesis or voice conversion. The later task has also been studied as part of this thesis, as we will see in Chapter 6. We will start by first explaining in Section 5.1 the modifications of the parameter set that are required for prosody modification. Then we will present in Section 5.2 two techniques that we have developed for this purpose based on feature vector selection and interpolation.

5.1 Parametrization modification

Since the prosody modification techniques we have developed use interpolation of the feature vectors, we will explain here the feature set transformations required for a correct interpolation of each of the parameters.

5.1.1 Vocal tract

Up to this point, we have represented the vocal tract by means of LPC as an all-pole filter of the form:

$$V(z) = \frac{1}{A(z)} \quad \text{where} \quad A(z) = 1 - \sum_{k=1}^{N} a_k \, z^{-k} \tag{5.1}$$

Since the LPC parameters a_k used in this representation are not well-suited for interpolation (Kondoz, 2004) (among other reasons, the interpolation of two stable filters is not guaranteed to be stable), we will use the Line Spectral Frequencies (LSF) as an alternative representation better suited to our needs (Itakura, 1975). The polynomial $A(z)$ from (5.1) is decomposed as:

$$P(z) = A(z) + z^{-(N+1)}\, A(-z), \tag{5.2}$$

$$Q(z) = A(z) - z^{-(N+1)}\, A(-z). \tag{5.3}$$

The LSF coefficients lsf_k are the roots of these two polynomials, which have too main characteristics: they are interspersed, and they are located on the unit cycle. Thus, the final parameter set used for prosody modification is then:

$$\theta_{\mathbf{vt}} = (lsf_1, lsf_2, \cdots, lsf_N)'. \tag{5.4}$$

Figure 5.1 shows the results of interpolating the LSF coefficients between two vocal tract filters: the source filter was extracted from the utterance of the Spanish $/a/$ phoneme in *high* phonation, and the target one corresponds to the Spanish $/o/$ phoneme in modal phonation.

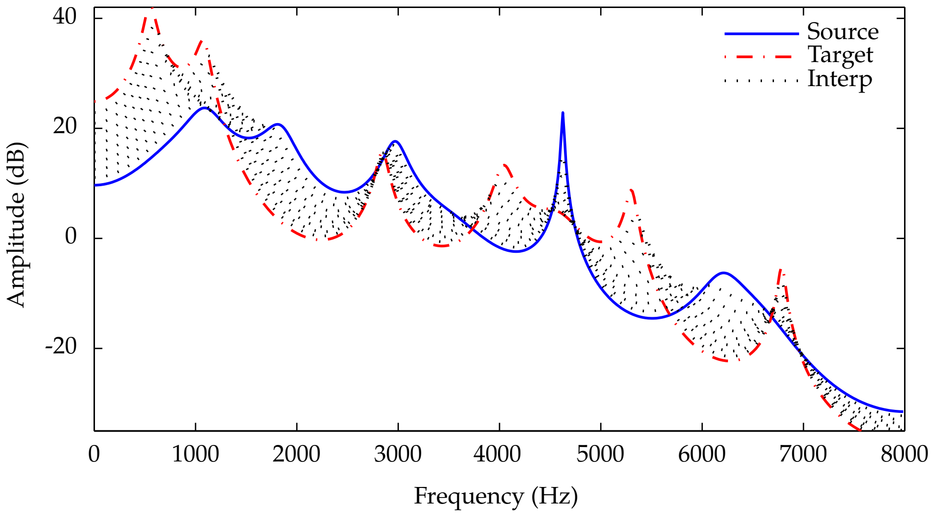


Figure 5.1: *Interpolation of LSF parameters*

5.1.2 Glottal waveform

Recall from Section 3.5 that our LF parameter set consists, for each glottal cycle, of:

- T_0: length of the glottal cycle, or equivalently, the fundamental frequency or pitch: $F_0 = \frac{1}{T_0}$
- T_p, T_e and T_a: absolute time measures, ranging from 0 to the glottal cycle duration T_0
- E_e: LF amplitude.

This time-based representation was needed during the parametrization phase, since the beginning of the glottal cycles was one of the parameters included in the optimization process. However, they are absolute values and thus not suitable for interpolation, as needed during prosody modification. Instead, we will use the *extended parameter set* that we introduced in Section 2.2.2:

$$R_a = \frac{T_a}{T_0}, \tag{5.5}$$

$$R_g = \frac{T_0}{2\,T_p}, \tag{5.6}$$

$$R_k = \frac{T_e - T_p}{T_p}. \tag{5.7}$$

The final parameter set used for prosody modification is then:

$$\theta_{\mathbf{lf}} = (R_a, R_g, R_k, E_e, F_0)', \tag{5.8}$$

from which all the needed parameters can be derived. An illustration of the interpolation properties of the extended LF parameter set is shown in Figure 5.2.

For unvoiced frames, this vector contains all zeros, since there is no glottal component.

5.1.3 Residual

We have developed two methods to modify the prosody of the residual or aspiration noise. First, we explain a method based on the resampling og the whitened residual waveform to achieve the desired modification. Then our second method is described, which uses our full parametrized residual and frame interpolation.

Residual resampling

The first method uses resampling of the residual waveform after the whitening step to achieve the desired change in F_0. The signal is resampled on a period-by-period basis to obtain the appropriate cycle duration (i.e., target F_0), in a similar way to that in Rao and Yegnanarayana (2006). The resampling is performed by concatenating an interpolator

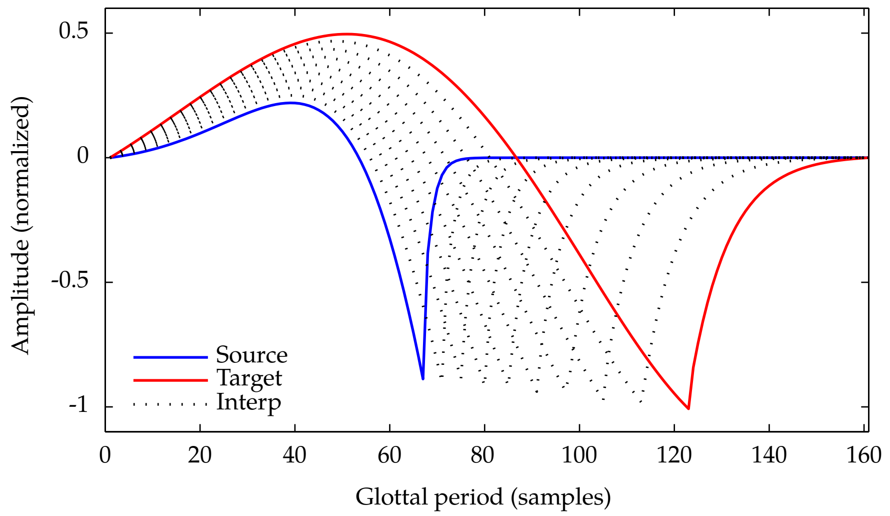


Figure 5.2: *Interpolation of LF parameters*

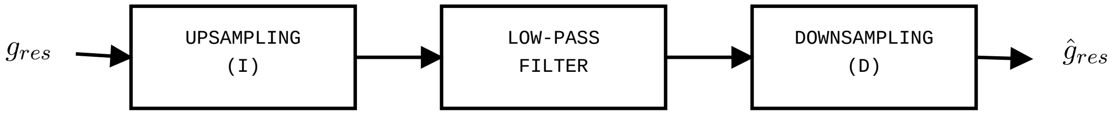


Figure 5.3: *Proposed scheme to modify the prosody of the residual using resampling.*

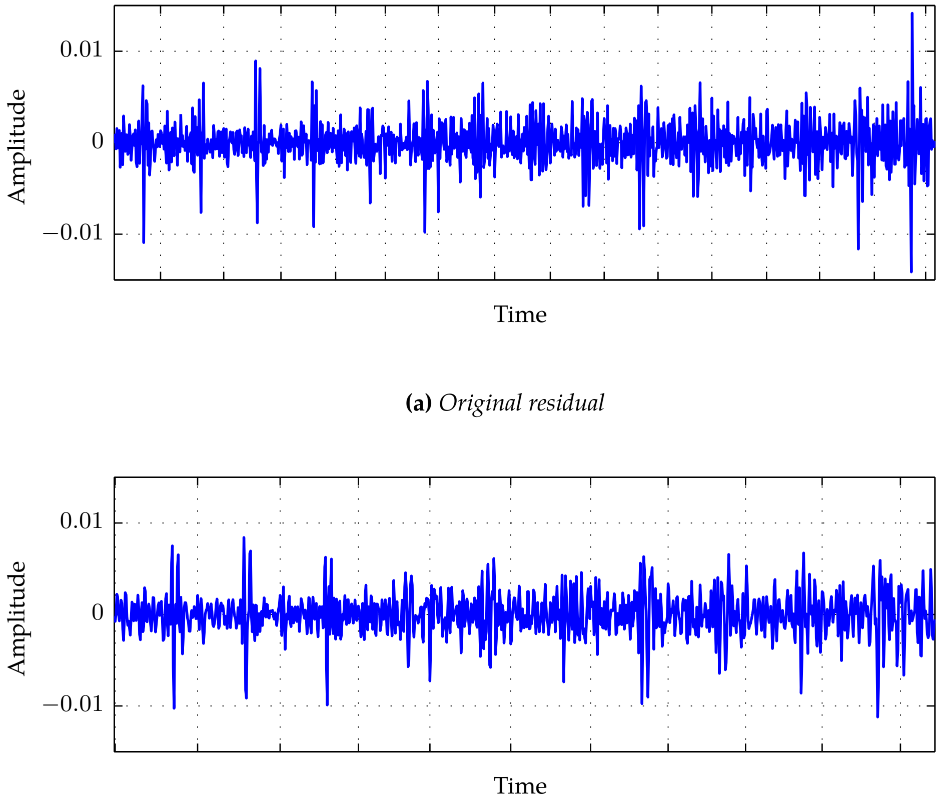
with a decimator, with a low-pass filter to avoid aliasing (Proakis and Manolakis, 1996), as shown in figure 5.3. The frequency response of the low-pass filter is:

$$H(\omega) = \begin{cases} I, & 0 \leq |\omega| \leq \min \frac{\pi}{I}, \frac{\pi}{U} \\ 0, & otherwise \end{cases} \tag{5.9}$$

where I and D are the interpolation and decimation factors respectively. The desired duration is achieved by applying the frame selection method detailed in section 5.2.

Figure 5.4 shows a segment of the original (top) and the resampled (bottom) residuals, after reducing F_0 by 70%. The vertical dotted lines indicate the limits of the glottal cycles, longer in the case of the resampled residual as corresponds to the prosody modification we are performing. As we can see, the relative position of the noise bursts inside each glottal cycle is preserved after modifying the prosody, which is important to preserve the naturalness of the synthesized speech.

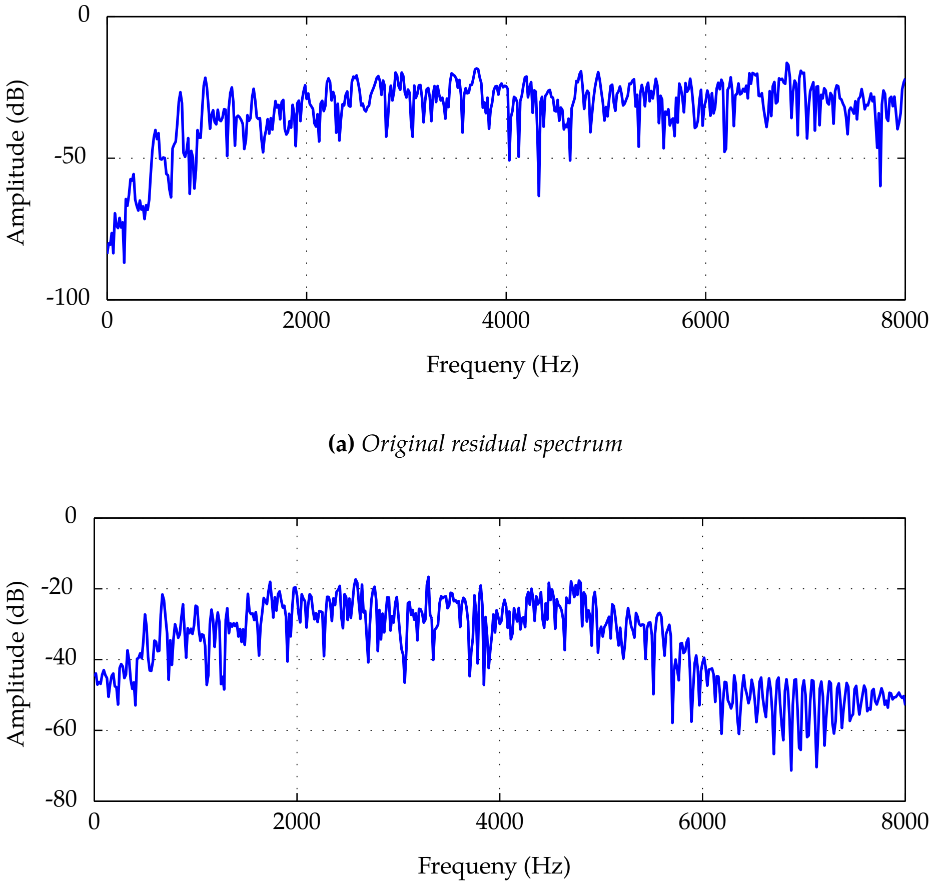
Resampling a signal results in a compression/expansion of the frequency spectrum (Proakis and Manolakis, 1996), and this effect can be observed in Figure 5.5. The top figure shows the spectrum of the original residual segment shown in Figure 5.4a, computed



(a) *Original residual*

(b) *Resampled residual*

Figure 5.4: *Segment of original and resampled residual after reducing F_0 by 70%. The vertical dotted lines indicate the limits of the glottal cycles.*



(a) *Original residual spectrum*

(b) *Resampled residual spectrum*

Figure 5.5: *Spectrum of original and resampled residual after reducing F_0 by 70%.*

using a Hamming window and a FFT of 1024 samples. Similarly, the bottom figure shows the spectrum of the resampled residual from Figure 5.4b. Notice how the spectrum of the resampled signal has been compressed with respect to the original one. This has an impact on the quality of the resulting synthetic speech, although as we will see in the evaluation from Section 5.3, the quality using this resampling technique is higher than when using the full residual parametrization from next section.

Parametrized residual

As we have seen in Section 3.6.2, we are modeling the aspiration noise using a synthetic envelope to modulate high-pass filtered Gaussian noise. The set of parameters chosen for the envelope's parametrization (eq. (3.56)) is already suitable for prosody modification, since it consists of signal levels (b_{lvl}, w_{lvl}^1 and w_{lvl}^2), and temporal parameters ($w_c^1, w_l^1, w_r^1, w_c^2, w_l^2, w_r^2$) that are already normalized by the fundamental period (i.e., range $[0, 1]$). Figure 5.6 shows the effect of interpolating between two sets (source and target) of noise envelope parameters. Thus, the parameter set is:

$$\theta_{\mathbf{res}} = (b_{lvl}, w_{lvl}^1, w_c^1, w_l^1, w_r^1, w_{lvl}^2, w_c^2, w_l^2, w_r^2)'. \tag{5.10}$$

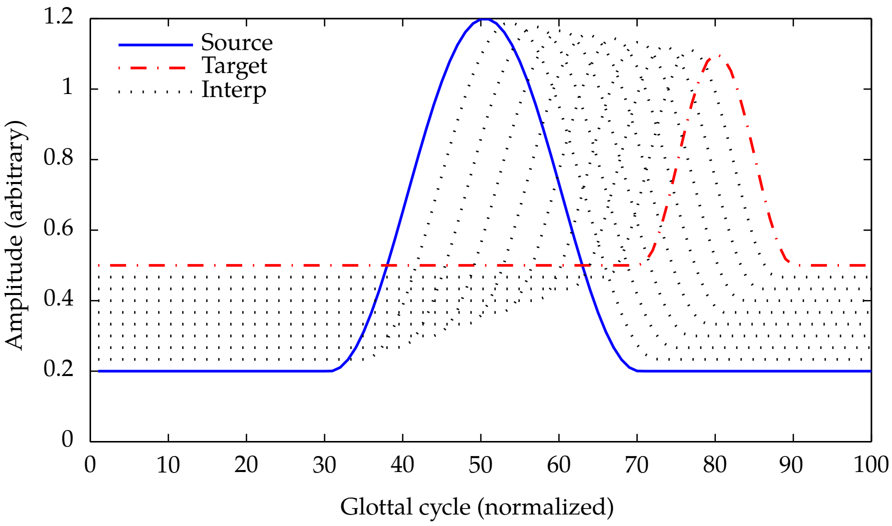


Figure 5.6: *Interpolation of noise envelope parameters*

For unvoiced periods, we are using white Gaussian noise to excite the vocal tract filter. Accordingly, equation (5.10) contains only the noise level b_{lvl}:

$$\theta_{\mathbf{res}} = (b_{lvl}, 0, 0, 0, 0, 0, 0, 0, 0)'. \tag{5.11}$$

The complete parameter set used for prosody modification is then built using equations (5.8), (5.4) and (5.10):

$$\theta = \begin{pmatrix} \theta_{lf} \\ \theta_{vt} \\ \theta_{res} \end{pmatrix}. \tag{5.12}$$

5.2 Time- and pitch-scale modification

We propose two separate methods for prosody modification, depending on which residual is used (parametrized or whitened waveform). The idea is to generate a new sequence of glottal epochs (GOIs) following the duration and modification restrictions, similar to the windowed signal selection technique using in traditional overlap-and-add algorithms (Moulines and Charpentier, 1990).

5.2.1 Modification using the parametrized residual

In this method all the components of our speech production model are parametrized, and each parameter vector is associated with its corresponding GOI. Suppose the duration is to be modified such as the duration of the synthesized signal D_s such as:

$$D_s = \alpha \cdot D_a \tag{5.13}$$

where D_a is the duration of the analysis signal, and α the modification factor. Let us adopt a similar approach for modifying the pitch:

$$F_{0,s} = \beta \cdot F_{0,a}, \tag{5.14}$$

where $F_{0,a}$ and $F_{0,s}$ are the analysis and synthesis fundamental frequencies, and β is the modifying factor. Rewriting eq. (5.14) in terms of the pitch period:

$$P_s = \frac{P_a}{\beta}, \tag{5.15}$$

where P_a and P_s are the analysis and synthesis pitch periods. Given the original sequence of analysis GOIS $goi_a(i)$, we generate the new sequence of synthesis gois $goi_s(j)$ as:

$$goi_s(j+1) = goi_s(j) + \frac{P_a(j)}{\beta}, \tag{5.16}$$

where $P_a(j)$ is computed by linear interpolation using the *virtual* analysis instant $goi'_a(j)$ resulting from projecting $goi_s(j)$ into the analysis time-line as follows. Let us assume that the jth synthesis cycle falls in between analysis cycles i and $i+1$. Let $goi'_a(j) = \frac{goi_s(j)}{\alpha}$

be the correspondence of the synthesis instant j in the original time-scale. In this case, $goi_a(i) < goi'_a(j) < goi_a(i+1)$ and $goi_a(i+1) < goi'_a(j+1) < goi_a(i+2)$, as shown in Figure 5.7. Then $P_a(j)$ in eq. (5.16) and the parameter vector θ_s^j associated to $goi_s(j)$ can be computed as:

$$P_a(j) = \gamma_j \, P_a(i) + (1-\gamma_j) P_a(i+1) \tag{5.17}$$

$$\theta_s^j = \gamma_j \, \theta_a^i + (1-\gamma_j)\theta_a^{i+1} \tag{5.18}$$

where:

$$\gamma_j = \frac{goi_a(i+1) - goi'_a(j)}{\delta_s^j/\alpha}, \tag{5.19}$$

and $\delta_s^j = goi_s(j+1) - goi_s(j)$ is the duration of the synthesis cycle j.

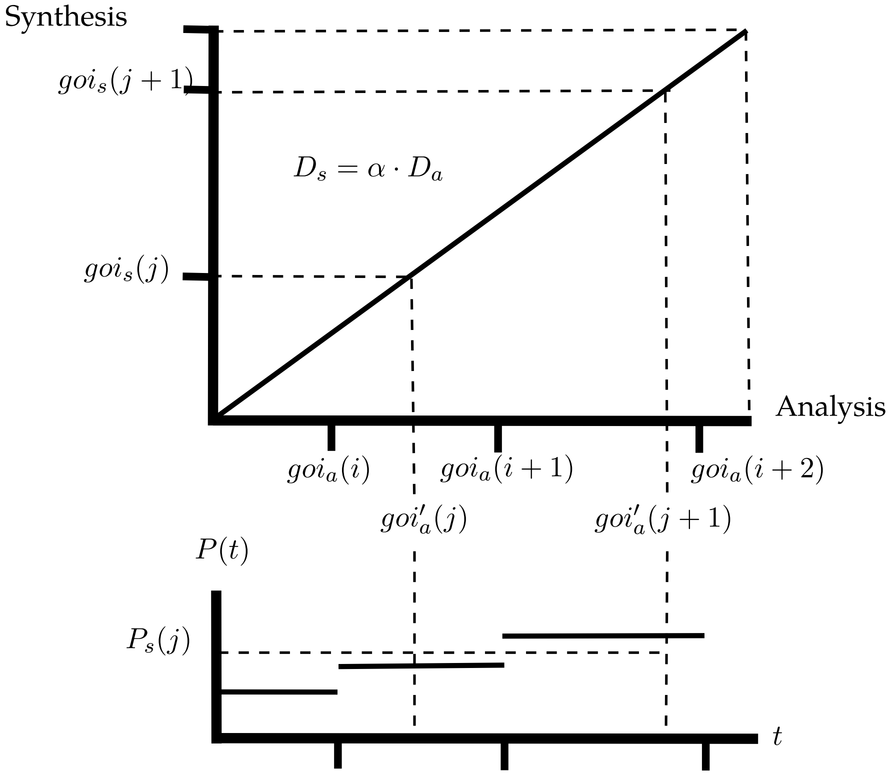


Figure 5.7: *Prosody modification scheme. The synthesis epochs $goi_s(j)$ are derived from the analysis epochs $goi_a(i)$ following the pitch/time modification function. In the example, the new duration D_s is 25% shorter than the original D_a ($\alpha = 0.75$). The pitch is kept unmodified in this example. The pitch $P_s(j)$ of the synthesis frame j is computed by averaging the original pitch of the corresponding analysis frames (the frame mapping is done using the* virtual *epochs $goi'_a(j)$). After Huang et al. (2001)*

5.2.2 Modification using the whitened residual

A procedure similar to that used in the previous case is used to define the synthesis GOIs $goi_s(j)$ and associate to them a parameter vector, consisting only on the vocal tract LSFs and glottal LF parameters, and a residual waveform. The main difference is that instead of interpolating between adjacent frames as in eq. (5.18), the closest frame is selected instead (effectively setting γ_j to 0 or 1) in eq. (5.19). The reason is that are dealing with whole residual waveforms, and due to their noisy nature, linear interpolation would result in noticeable artifacts (Moulines and Charpentier, 1990). Once the corresponding residual waveform is selected, its duration is modified using the algorithm explained in Section 5.1.3 to achieve the desired fundamental period duration.

5.3 Evaluation

In order to evaluate the quality of our proposed two prosody modification algorithms, we conducted an on-line listening test comparing them to two other high-quality methods. Thus, the test consisted of samples generated using:

- proposed source-filter model with resampled residual,
- proposed source-filter model with parametrized residual,
- STRAIGHT (Kawahara et al., 1999), a Vocoder-based analysis-synthesis algorithm, as we did in the resynthesis experiments (Section 4.3),
- the pitch-synchronous, overlap-and-add method used in our speech synthesizer Ogmios (Bonafonte et al., 2006a).

The corpus used for this evaluation was recorded during the TC-STAR project (Bonafonte et al., 2006b), the same used in the resynthesis experiments from the previous chapter (Section 4.3). This corpus was recorded in a professional studio using high quality recording settings (96 kHz and 24 bits/sample), two microphones (membrane and close-mouth, only the former was used in this work) and a laryngograph. As before, the signals were downsampled to 16 kHz and 16 bits/sample, and utterances from four speakers (two female and two male) were used. Since we wanted to study the effects of the specific duration or F_0 changes, we generated utterances modifying only one of them while keeping the other one untouched. Four modifications were evaluated, corresponding to typical values used in speech synthesis:

- shortened duration (factor of 0.8), original F_0,
- lengthened duration (factor of 1.2), original F_0,

- decreased F_0 (factor of 0.8), original duration,
- increased F_0 (factor of 1.2), original duration.

For each of the four modifications, two examples were provided for each of the four different methods, resulting in a total of 32 utterances for this evaluation. The participants in the test were asked to listen carefully to each sample using headphones, and to rate each of them using the typical 5 point MOS scale:

1. Bad (distortion very annoying)
2. Poor (distortion annoying)
3. Fair (distortion slightly annoying)
4. Good (distortion perceptible but not annoying)
5. Excellent (distortion imperceptible)

The utterances were presented in a random order, so the users were unaware of the origin of the examples they were evaluating. The subjects were allowed to listen to each sample as many times as necessary.

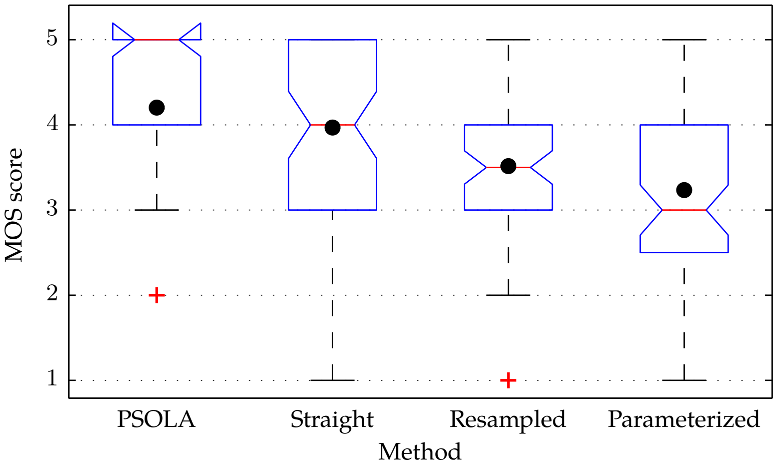


Figure 5.8: *Overall MOS scores of the prosody modification evaluation test for the methods of reference (PSOLA and Straight) and our two proposed methods (using resampled and parameterized residual).*

The overall results of the test are presented in figure 5.8 using boxplots. On each box, the central line is the median, the box edges are the 25th and 75th percentiles, and the whiskers extend to the furthest data points not considered outliers (marked with crosses).

The ranking of the different methods in terms of MOS score is PSOLA, STRAIGHT, resampled residual and parametrized residual. All four methods surpass the acceptance threshold (MOS score 3, Fair), with the first two obtaining an average rate of 4, Good. As we can see, our proposed method using the resampled residual is rated higher than that using the parametrized residual, although only by a small margin. This was a bit surprising, because according to our informal internal evaluations and the results of the resynthesis test from the previous chapter (Section 4.3), we would have expected it to rate similar to the STRAIGHT algorithm. Our second method using the completely parametrized residual is rated just above Fair, as could be expected given the results of the resynthesis test.

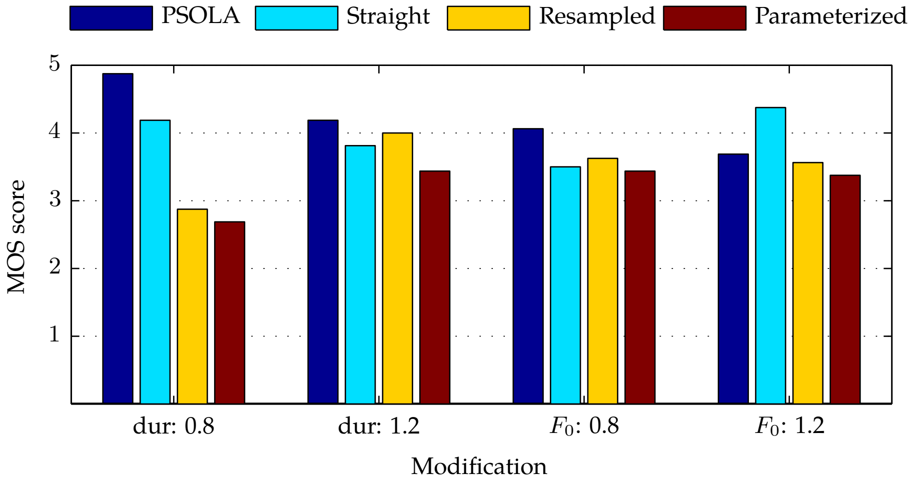


Figure 5.9: *Detailed MOS scores of the prosody modification evaluation test for the methods of reference (PSOLA and Straight) and our two proposed methods (using resampled and parameterized residual). The results are presented individually for each of the modifications: shortened (dur: 0.8) and lengthened (dur: 1.2) durations, and decreased (f0: 0.8) and increased (f0: 1.2) fundamental frequency.*

In figure 5.9 we show the individual mean results for each of the four prosody modifications. As we can see, PSOLA is the method of preference in all cases except when increasing the F_0, where STRAIGHT gets the first place. Interestingly, while there is no change in preference order with respect to our two proposed methods when shortening the duration or increasing the fundamental frequency, when increasing the signal duration the residual resampling method performs better than STRAIGHT, and when decreasing the fundamental frequency both our methods outperform it. Shortening the duration seems to be the main drawback of our algorithms, since both of them get the worst rates, more than 1 point in the MOS scale.

5.4 Conclusions

We have presented here two techniques for performing prosody modifications as required for speech processing tasks such as speech synthesis or voice conversion. The first method uses frame selection and interpolation of the complete parametrized set (vocal tract using LSF, glottal waveform using extended LF parameter set, and residual using synthetic envelopes) to achieve the desired changes in duration or prosody. The second method uses frame selection of the vocal tract and glottal parameters, and waveform resampling of the residual waveforms to perform the modifications. The performance of this second method is slightly better than that of the full parametrization, although in both cases their ratings fall in the Fair-Good range, below those of well established methods PSOLA or STRAIGHT. Together with the resynthesis results, we can conclude that although the performance of the algorithms is acceptable, more work would be needed to get the methods on the same level as current reference techniques like PSOLA or STRAIGHT, particularly in the analysis and synthesis of the residual. As we will see next in Chapter 6, we integrated our full parametrized algorithm into an existing voice conversion system with good results.

CHAPTER 6

Voice Conversion

Voice conversion (VC) consists of transforming the voice from one speaker (*source*) so that it is perceived as belonging to a different given speaker (*target*). For instance, this is sometimes needed in speech translation systems, where the system produces speech in a language unknown to the source speaker. In this case, the targe speaker is the synthetic voice resulting from the text-to-speech synthesis, which is modified to sound like what the source speaker would in the target language (Pérez and Bonafonte, 2006). The aim of this work was to investigate whether existing voice conversion (VC) methods would benefit from a more accurate speech production representation as the one we propose. The objective of a VC system is to transform the voice of a certain speaker (*source speaker*) so that it is perceived as belonging to another given speaker (*target speaker*).

Our intention was to evaluate the effect of our speech parametrization paradigm in a VC context, isolating it from other aspects of the problem (such as alignment and conversion function training). To do so, we included our parametrization in the VC system developed in our group for the European project TC-STAR [1] (Duxans, 2006; Duxans et al., 2006). We trained and tested the system using the same training and testing data sets used during the project's evaluation campaigns. This way we could compare the new results to those presented at the time, which were already available. This reference system is based on CART and GMM, and uses acoustic (LP) and phonetic characteristics.

Although newer VC paradigms (e.g., Yutani et al. (2009)) have been proposed which improve the conversion performance, we are here only interested in evaluating the impact of the signal parametrization by itself. We can also expect newer methods to benefit from a better parametrization in a similar manner.

First, we will provide a concise overview of the state-of-the-art in VC systems (Sec-

[1] Technology and Corpora for Speech to Speech Translation `http://www.tcstar.org`

tion 6.1). Then we will proceed to explain the VC method we are using in this work, the baseline parametrization, and the modifications needed to integrate our new parametrization in Section 6.2. The subjective evaluation that we performed and its results are explained in Section 6.4.

6.1 Background

As we have explained, a voice conversion systems takes the voice of a source speaker, and modifies it so it sounds like a given target speaker. The transformation of the acoustics features of the source speaker into those similar to the target speaker is done by means of mapping functions. Systems working with spectral features may use mapping codebooks (Abe, 1991), techniques based on a pre-clustering with non-overlapping acoustic classes (Sündermann and Höge, 2003; Turk and Arslan, 2003), Continuous probabilistic transform functions (Chen et al., 2003; Kain and Macon, 1998; Lee et al., 2002; Stylianou et al., 1998; Toda et al., 2001a,b; Ye and Young, 2003), Hidden Markov Model related conversions (Mori and Kasuya, 2003; Tamura et al., 2001) or Parametric mappings (Ho et al., 2002; Rentzos et al., 2003; Slifka and Anderson, 2002). A different approach has been proposed that performs a prediction of the residual signal using the vocal tract, instead of its transformation (Sündermann et al., 2005). Kain (2001) proposed a residual generation method based on the linear combination of codebook entries associated to the LSF vectors. Ye and Young (2004) proposed a simple residual selection technique in which the residual was associated to the closest target LSF vector. There are a few systems transforming prosodic features. One group of algorithms used a similar approach to that of spectrum mapping (Tamura et al., 2001; Turk and Arslan, 2003). A different method was presented in Ceyssens et al. (2002), where multi-parameter pitch contours were transformed using stochastic mappings. The transformation of the residual signal has also been tackled in several studies, since without it no good similarity can be achieved. Duxans (2006) presented a conversion system based on the Linear Prediction (LP) model of speech production, using phonetic information based on CART and GMM, to be used as a post-processing stage in a test-to-speech system. New prediction and selection techniques were presented to deal with the residual of the LP analysis. Duxans extended the method to take advantage of the unlimited amount of source data that a TTS can generate, and the phonetic information inherently available at its output. Erro (2008) studied the use of the harmonic plus stochastic model for intra- and cross -lingual voice conversion tasks. A combination of statistical Gaussian mixture models with a novel technique for frequency warping (Weighted Frequency Warped) was proposed in Erro et al. (2010b), improving the methods developed fro the TC-STAR European project that obtained excellent results during the international VC evaluation campaigns. Erro et al. also presented a new iterative technique for frame alignment which allowed the VC method to be applied to cross-lingual VC and other tasks without the requirement of parallel data for the training stage. In order to overcome some of the problems inherent to the GMM systems (mainly due to the excessive smoothing induced by the statistical averaging), Toda et al. (2007) proposed a spectral conversion method using a maximum likelihood estimation of the parameter trajectories, whith two main advantages over traditional frame-based systems: it takes into account the correlation that exists between feature frames, and it reduces the oversmoothing effect by taking into consideration the global variance as part of the conversion. With a similar purpose, Qiao and Minematsu (2009) used a mixture of probabilistic linear regressions to estimate the mapping function between two feature spaces, overcoming some of the problems found in GMM systems as a result of the introduced over-smoothing. Yutani et al. (2009) used multi-space probability distribution models to simultaneously model the spectrum and the F_0, resulting in much better converted quality than traditional approaches based on GMM. Voice conversion contributions focusing on the voice-source are scarce. Some of the works listed above presented preliminary works dealing with residual transformation (Sündermann et al., 2005; Ye and Young, 2004), although no specific voice-source model was adopted. Childers and Ahn (1995) presented a voice conversion system using a glottal excitation waveform. In their work, however, no parametrization of the glottal source is performed, and vector quantization techniques commonly used in speech coding algorithms are employed instead. Mori and Kasuya (2003) introduced a voice conversion system based on the ARX speech production model, although no details on glottal modifications are given. A GMM-based VC system using the LF for the glottal contribution was presented in del Pozo (2008); del Pozo and Young (2008) with good results, using different GMM for spectral and voice source conversion.

6.2 Reference system

Our VC algorithm (Duxans, 2006) uses a classification and regression tree (CART) to classify the vocal tract data into phonetic categories. For each category, a transformation function is build using a standard Gaussian mixture model (GMM). To build such a system, we first need to parametrize and align both source and target voices. Then we build a transformation function using the aligned data to convert one set of parameters into the other. A figure depicting our adopted reference VC system is shown in Figure 6.1.

We will first review the estimation of the GMM parameters and derivation of the transformation function, and then we will explain the use of decision trees in our voice conversion algorithm.

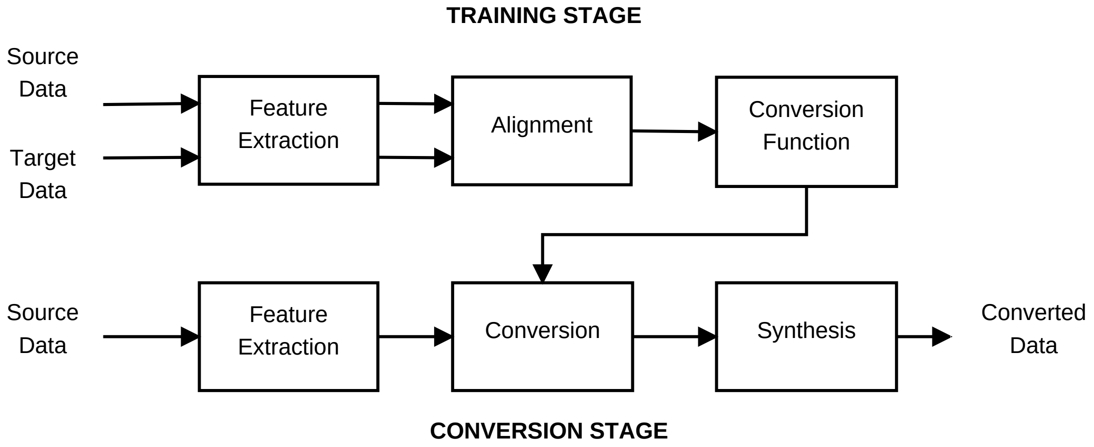


Figure 6.1: *General Voice Conversion Scheme.*

6.2.1 Transformation function for each leaf

Let $\mathcal{X} = \{\mathrm{x}_n\}$ and $\mathcal{Y} = \{\mathbf{y}_n\}$ be the set of source and target data frames respectively ($n = 1 \ldots N$). The transformation function $F(\mathbf{x})$ is defined by means of regression analysis, where the regression function is formulated as a weighted sum of linear models (Kain, 2001):

$$\hat{\mathbf{y}} = F(\mathbf{x}) = E[\mathbf{y}|\mathbf{x}] = \sum_{m=1}^{M} p(c_m|\mathbf{x})\ (\mathbf{W}_m\,\mathbf{x} + \mathbf{b}_m)\,. \tag{6.1}$$

In the equation above, $p(c_m|\mathbf{x})$ represents the weight of the mth component of the regression and corresponds to the posterior probability of the input data belonging to that particular class, $\mathbf{W}_m$ is a transformation matrix and $\mathbf{b}_m$ is a bias vector.

The components of the regression $F(\mathbf{x})$ in (6.1) are obtained from a GMM model fitted to the combined source-target data set:

$$\mathcal{Z} = \begin{Bmatrix} \mathcal{X} \\ \mathcal{Y} \end{Bmatrix} = \begin{Bmatrix} \mathbf{x_1} & \mathbf{x_2} & \cdots & \mathbf{x_N} \\ \mathbf{y_1} & \mathbf{y_2} & \cdots & \mathbf{y_N} \end{Bmatrix} = \begin{Bmatrix} \mathbf{z_1} & \mathbf{z_2} & \cdots & \mathbf{z_N} \end{Bmatrix}. \tag{6.2}$$

A GMM is a probability density function represented using a mixture of M Gaussian densities:

$$\begin{aligned} p(\mathbf{z}_n|\theta) &= \sum_{m=1}^{M} w_m \mathcal{N}(\mathbf{z}_n|\mu_m, \mathbf{\Sigma}_m) \\ &= \sum_{m=1}^{M} w_m \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_m|^{1/2}}\, e^{-\frac{1}{2}(\mathbf{z}_n-\mu_m)^t\mathbf{\Sigma}_m^{-1}(\mathbf{z}_n-\mu_m)}, \end{aligned} \tag{6.3}$$

where w_m is the weight of each component in the distribution, such that $\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0\ \forall m$. The mean vector μ_m and the covariance matrix $\mathbf{\Sigma}_m$ of the combined data can be written in terms of those of the source and target data sets as:

$$\mathbf{\Sigma}_m = \begin{pmatrix} \mathbf{\Sigma}_m^{XX} & \mathbf{\Sigma}_m^{XY} \\ \mathbf{\Sigma}_m^{YX} & \mathbf{\Sigma}_m^{YY} \end{pmatrix}, \tag{6.4}$$

$$\mu_m = \begin{pmatrix} \mu_m^X \\ \mu_m^Y \end{pmatrix}, \tag{6.5}$$

where μ_m^X, μ_m^Y, $\mathbf{\Sigma}_m^{XX}$ and $\mathbf{\Sigma}_m^{YY}$ are the mean vectors and covariance matrices of the source and target data respectively, and $\mathbf{\Sigma}_m^{XY}$ and $\mathbf{\Sigma}_m^{YX}$ are the cross-covariance matrices between source-target and target-source data respectively. Using these, we build the transformation function (6.1) using the regression:

$$F(\mathbf{x}) = E[\mathbf{y}|\mathbf{x}] = \sum_{m=1}^{M} p(c_m|\mathbf{x}) \left(\mu_m^Y + \mathbf{\Sigma}_m^{YX} \left(\mathbf{\Sigma}_m^{XX}\right)^{-1} \left(\mathbf{x} - \mu_m^X\right) \right), \tag{6.6}$$

where

$$p(c_m|\mathbf{x}) = \frac{w_m \mathcal{N}\left(\mathbf{x}|\mu_m^X, \mathbf{\Sigma}_m^{XX}\right)}{\sum_{p=1}^{M} w_p \mathcal{N}\left(\mathbf{x}|\mu_p^X, \mathbf{\Sigma}_p^{XX}\right)}. \tag{6.7}$$

From equations (6.1) and (6.6), it follows that:

$$W_m = \mathbf{\Sigma}_m^{YX} \left(\mathbf{\Sigma}_m^{XX}\right)^{-1} \tag{6.8}$$

$$b_m = \mu_m^Y - \mathbf{\Sigma}_m^{YX} \left(\mathbf{\Sigma}_m^{XX}\right)^{-1} \mu_m^X. \tag{6.9}$$

For each of the densities in the GMM, we need to obtain its weight w_m, mean vector μ_m and covariance matrix $\mathbf{\Sigma}_m$. Thus, the complete set of parameters to be estimated is $\theta = \{\theta_1, \ldots, \theta_M\}$, where $\theta_m = [w_m, \mu_m, \mathbf{\Sigma}_m]$.

The GMM parameters are obtained by maximizing their likelihood given the input data (Duda et al., 2000). The maximum-likelihood estimation problem consists in estimating the set of parameters θ defining the density function $p(\mathbf{z}|\theta)$ of a certain data set $\mathcal{Z}$, drawn from this distribution. Assuming the data vectors $\mathcal{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_N\}$ independent and identically distributed (i.i.d), the *likelihood* of the parameters given the data is:

$$\mathcal{L}(\theta|\mathcal{Z}) = p(\mathcal{Z}|\theta) = \prod_{n=1}^{N} p(\mathbf{z}_n|\theta). \tag{6.10}$$

It is usually easier to work with the logarithmic expression of eq. (6.10), the *log–likelihood*:

$$\log \mathcal{L}(\theta|\mathcal{Z}) = \sum_{n=1}^{N} \log p(\mathbf{z}_n|\theta). \tag{6.11}$$

Since the data set $\mathcal{Z}$ is fixed, we can think of the above expression as a function of the parameters θ. Our goal then is to find the optimal parametrization θ^* that maximizes $\log \mathcal{L}$:

$$\theta^* = \arg\max_{\theta} \log \mathcal{L}(\theta|\mathcal{Z}). \tag{6.12}$$

Substituting eq. (6.3) into eq. (6.11), the resulting log-likelihood expression L that needs to be optimized is given by:

$$L = \log \mathcal{L}(\theta|\mathcal{Z}) = \sum_{n=1}^{N} \log \left(\sum_{m=1}^{M} w_m \, \frac{1}{(2\pi)^{d/2} |\mathbf{\Sigma}_i|^{1/2}} \, e^{-\frac{1}{2}(x-\mu_i)^t \mathbf{\Sigma}_i^{-1} (x-\mu_i)} \right). \tag{6.13}$$

There are a number of techniques to perform the optimization, the most widely used being the Expectation-Maximization (EM) algorithm (Duda et al., 2000), an iterative method that operates in two steps:

- *Expectation* step (E-step), where it calculates the expected value of the log-likelihood function given the current estimate for the model parameters. Compute for $m = 1 \dots M$:

$$\gamma_{mn}^{(l)} = \frac{w_m^{(l)} \, \mathcal{N}\left(\mathbf{z}_n \mid \mu_m^{(l)}, \mathbf{\Sigma}_m^{(l)}\right)}{\sum_{o=1}^{M} w_o^{(l)} \, \mathcal{N}\left(\mathbf{x}_n \mid \mu_o^{(l)}, \mathbf{\Sigma}_o^{(l)}\right)} \qquad n = 1 \dots N \tag{6.14}$$

$$n_m^{(l)} = \sum_{i=1}^{n} \gamma_{mn}^{(l)}, \tag{6.15}$$

  where $\gamma_{mn}^{(l)}$ is the estimated probability that the nth sample was generated by the mth Gaussian component at the lth iteration, and $n_m^{(l)}$ represents the total weight of the mth Gaussian in the mixture.

- *Maximization* step (M-step), where the parameter estimations are updated maximizing the expected log-likelihood from the E-step. Calculate the new estimates for

$j = 1 \ldots k$:

$$w_m^{(l+1)} = \frac{n_m^{(l)}}{n}, \tag{6.16}$$

$$\mu_m^{(l+1)} = \frac{1}{n_m^{(l)}} \sum_{n=1}^{N} \gamma_{mn}^{(l)} \, \mathbf{z}_n, \tag{6.17}$$

$$\boldsymbol{\Sigma}_m^{(l+1)} = \frac{1}{n_m^{(l)}} \sum_{n=1}^{N} \gamma_{mn}^{(l)} \left(\mathbf{x}_n - \mu_m^{(l+1)}\right) \left(\mathbf{x}_n - \mu_m^{(l+1)}\right)'. \tag{6.18}$$

After each iteration of the M-step, the new log-likelihood $L^{(l+1)}$ is computed using (6.13) and the convergence of the algorithm checked:

$$L^{(l+1)} = \sum_{n=1}^{N} \log \left(\sum_{m=1}^{M} w_m^{(l+1)} \mathcal{N}\left(\mathbf{z}_n \mid \mu_m^{(l+1)}, \boldsymbol{\Sigma}_m^{(l+1)}\right) \right). \tag{6.19}$$

The algorithm is iterated until the relative increase of the log-likelihood is smaller than a certain threshold δ:

$$\Delta L = \frac{L^{m+1} - L^m}{|L^m|} \leq \delta. \tag{6.20}$$

The EM algorithm needs a good initialization in order to minimize the possibility of converging to a local minimum. We will use the k-means algorithm (Duda et al., 2000) to divide the data into M clusters, each of them defined by its centroid c_m for $m = 1 \ldots M$. The initial estimations $w_m^{(0)}, \mu_m^{(0)}, \boldsymbol{\Sigma}_m^{(0)}$ can then be estimated using two the coarse estimate of the sample n being generated by the mth Gaussian provided by the k-means algorithm. This is done by setting γ_{mn} in (6.14) as:

$$\gamma_{mn}^{(-1)} = \begin{cases} 1, & \mathbf{z}_n \text{ is in cluster } m, \\ 0, & \text{otherwise.} \end{cases} \tag{6.21}$$

The initial values for $w_m^{(0)}$, $\mu_m^{(0)}$ and $\boldsymbol{\Sigma}_m^{(0)}$ are computed using equations (6.15) – (6.18). After obtaining the initialization point for the EM algorithm, we compute the initial value of the log-likelihood function:

$$L^{(0)} = \sum_{n=1}^{N} \log \left(\sum_{m=1}^{M} w_m^{(0)} \mathcal{N}\left(\mathbf{x}_n \mid \mu_m^{(0)}, \boldsymbol{\Sigma}_m^{(0)}\right) \right), \tag{6.22}$$

which is then used to initialize the EM iterations.

6.2.2 Decision tree based voice conversion

Decision trees allow us to work with numerical data (such as spectral and glottal features) as well as categorical data (such as phonetic features) when building an acoustic model. Tables 6.1, 6.2 and 6.3 list the allophones and the characteristics used in this work, for vowels, glides and consonants respectively. They can be summarized as follows:

- category: `consonant`, `glide`, `vowel`,
- point of articulation (consonants): `alveolar`, `bilabial`, `dental`, `interdental`, `labiodental`, `palatal`, `velar`,
- manner of articulation (consonants): `affricate`, `approximant`, `fricative`, `lateral`, `nasal`, `plosive`, `tap`, `trill`,
- height (vowels): `close`, `mid_close`, `mid_open`, `open`, `schwa`,
- backness (vowels): `back`, `center`, `front`,
- voicing: `unvoiced`, `voiced`.

In our system, we are using binary trees, i.e., the trees are used to classify data into phonetic categories using yes/no questions (e.g., is `bilabial` true?, is `mid_open` true?, is `consonant` true?). A sample binary CART constructed using these characteristics is shown in Figure 6.2.

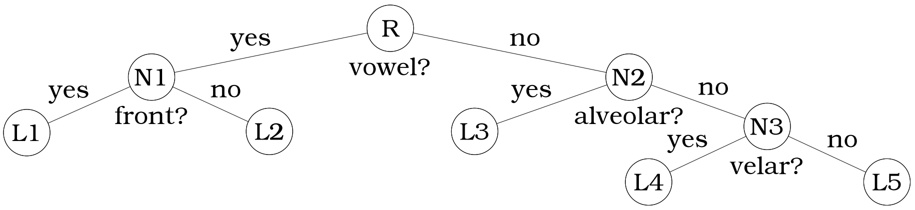


Figure 6.2: *Sample binary CART consisting of a root node (R), three intermediate nodes (N1, N2 and N3) and five leaves or categories (L1 to L5). In the example, the data is divided into five categories using four binary phonetic characteristics (vowel, front, alveolar and bilabial).*

We will proceed to describe now the procedure to grow the trees. The available training data is divided into two sets: the training set and the validation set. This is important in order to avoid over-fitting issues, where the trees are so dependent on the training data that they lose the ability to generalize (i.e., perform well with previously unseen data). A joint GMM based conversion system is estimated from the training set for the parent node t (the root node in the first iteration) as explained before. We can then calculate an error

			Vowels	
		Height	Backness	Voicing
SAMPA code	@	schwa	center	voiced
	a	open	center	voiced
	e	mid_close	front	voiced
	E	mid_open	front	voiced
	i	close	front	voiced
	o	mid_close	back	voiced
	O	mid_open	back	voiced
	u	close	back	voiced

Table 6.1: *List of vocalic allophones and phonetic characteristics.*

			Glides	
		Height	Backness	Voicing
SAMPA code	j	close	front	voiced
	uw	close	back	voiced
	w	close	back	voiced
	y	close	front	voiced

Table 6.2: *List of glides allophones and phonetic characteristics.*

index $E(t)$ for all the elements of the training set belonging to that node:

$$E(t) = \frac{1}{|t|} \sum_{n=1}^{|t|} D(\tilde{\mathbf{y}}_{\mathbf{n}}, \mathbf{y}_n), \tag{6.23}$$

where $|t|$ is the number of frames in the node t, $\mathbf{y}$ is a target frame and $\tilde{\mathbf{y}}$ its corresponding converted frame. $D(\tilde{\mathbf{y}}, \mathbf{y})$ is a measure of the distance between target and converted frames that depends on the type of parameter, as will be explained later.

We then continue by evaluating all the possible questions q of the set Q at node t. The set Q is formed by binary questions of the form *is* $\{\tilde{\mathbf{y}} \in A\}$, where A represents a phonetic characteristic of the frame $\tilde{\mathbf{y}}$. For each valid question q, two child nodes (t_L and t_R) are populated: the left descendant node t_L is formed by all the frames which fulfill the question and the right t_R node by the rest. A question q is considered valid if a minimum number of frames fulfill it.

For each child node, a joint GMM conversion system is estimated, and the error figures $E(t_L, q)$ and $E(t_R, q)$ for the training vectors corresponding to the child nodes t_L and t_R obtained from the question q are calculated. The increment of the accuracy for the

		Consonants		
		Articulation point	Articulation manner	Voicing
SAMPA code	b	bilabial	plosive	voiced
	B	bilabial	approximant	voiced
	d	dental	plosive	voiced
	D	dental	approximant	voiced
	dz	alveolar	affricate	voiced
	dZ	palatal	affricate	voiced
	f	labiodental	fricative	unvoiced
	g	velar	plosive	voiced
	G	velar	approximant	voiced
	J	palatal	nasal	voiced
	jj	palatal	fricative	voiced
	k	velar	plosive	unvoiced
	l	alveolar	lateral	voiced
	L	palatal	lateral	voiced
	m	bilabial	nasal	voiced
	n	alveolar	nasal	voiced
	N	velar	nasal	voiced
	p	bilabial	plosive	unvoiced
	r	alveolar	tap	voiced
	R	alveolar	tap	voiced
	rr	alveolar	trill	voiced
	s	alveolar	fricative	unvoiced
	S	palatal	fricative	unvoiced
	t	dental	plosive	unvoiced
	T	interdental	fricative	unvoiced
	ts	alveolar	affricate	unvoiced
	tS	palatal	affricate	unvoiced
	x	velar	fricative	unvoiced
	z	alveolar	fricative	voiced
	Z	palatal	fricative	voiced

Table 6.3: *List of consonantic allophones and phonetic characteristics.*

question q at the node t can be calculated as:

$$\Delta(t,q) = E(t) - \frac{E(t_L,q)|t_L| + E(t_R,q)|t_R|}{|t_L| + |t_R|}. \tag{6.24}$$

The question q^* corresponding to the maximum increment is selected:

$$q^* = \arg\max_q \Delta(t,q), \tag{6.25}$$

and the node is split only if the validation set accuracy for question q^* also increases. In this case, nodes t_L and t_R are added to list of nodes to be split, and question q^* is removed from the subset of questions Q available to the resulting sub-tree. The tree is grown until there is no node candidate to be split or a maximum size (i.e., number of nodes) is reached. In order to avoid over-fitting to the training data, the tree needs to be pruned by using either a pre-pruning or post-pruning approach (Duxans, 2006). Pre-pruning consists of using a validation dataset to compute the accuracy increment of the optimal question, and splitting the node only when it is greater than zero. The main drawback here is that each splitting decision is taken locally, without considering future splits ("the horizon effect"). Post-pruning avoids this problem by starting with a complete tree and operation bottom to top by recombining leafs and nodes to obtain smaller subtrees with better accuracy. This is done independently for each source–target combination to find the optimal tree corresponding to each case.

New source vectors are classified into leafs according to their phonetic features by the decision tree, and then converted according to the GMM based system belonging to its leaf. This is only applied to voiced segments, unvoiced segments are used unmodified. In order to rate the performance of the conversion, and to compare the three different set of parameters, we will use the *Performance Index* as an objective measure of the conversion's performance:

$$P = 1 - \frac{D(\hat{\mathcal{Y}}, \mathcal{Y})}{D(\hat{X}, \mathcal{Y})}, \qquad (6.26)$$

where $\mathcal{X}$ is the set of source vectors, $\mathcal{Y}$ is the set of target vectors, $\hat{\mathcal{Y}}$ is the set of converted vectors, and $D(\cdot,\cdot)$ is the distance measure used for the conversion function training. The performance index P ranges from 0, when $\hat{\mathcal{Y}} = \mathcal{X}$ (i.e., no transformation or change at all), to 1, when $\hat{\mathcal{Y}} = \mathcal{Y}$ (i.e., completely successful transformation) [2]

6.2.3 Baseline

The baseline parametrization uses line spectral frequencies (LSF) to model the vocal-tract, derived using linear prediction analysis. The distance $D(\tilde{\mathbf{y}}, \mathbf{y})$ used to compute the error index from eq. 6.23 is the mean of the Inverse Harmonic Mean Distance (IHMD) (Laroia et al., 1991):

$$\begin{aligned} D(\tilde{\mathbf{y}}, \mathbf{y}) &= \sqrt{\sum_{p=1}^{P} c(p)(\tilde{y}(p) - y(p))^2}, & (6.27) \\ c(p) &= \frac{1}{w(p) - w(p-1)} + \frac{1}{w(p+1) - w(p)}, & (6.28) \end{aligned}$$

[2] It should be noted that P is not confined to the range $[0, 1]$, although it usually falls within its limits. For extremely unsuccessful conversions, it may result in negative values.

with $w(0) = 0$, $w(P+1) = \pi$ and $w(p) = \tilde{y}(p)$ or $w(p) = y(p)$ so that $c(p)$ is maximized (p is the vector dimension), weights more the mismatch in spectral picks than the mismatch in spectral valleys when working with LSF vectors.

In this system no specific model for the residual is assumed, working with the full waveform instead. To complete the conversion from the source speaker to the target speaker, a target residual signal is predicted from the converted LSF envelopes. During training, a parallel database of LSF vectors and residual waveforms using all the available target data is constructed, mapping each LSF vector into its corespondent residual. During conversion, the converted LSF vector is compared to the LSF vectors in the database, and the residual associated to the closes one is selected. Prior to the final concatenation a smoothing step is performed on the selected residual waveforms to minimize the introduction of artifacts (Duxans, 2006).

6.3 Voice conversion using source characteristics

So far we have explained the details of the voice conversion system using CART and GMM our proposed parametrization is going to be integrated into, and we have presented the previous LP-derived parametrization we are going to use as baseline. In this section we address the requirements of the source features and the model training methods we have use to grow the different trees. First we will present the corpus used during these VC experiments.

6.3.1 Language resources

As part of the TC-STAR project, UPC produced the language resources for supporting the evaluation of English/Spanish voice conversion. Four bilingual speakers English/Spanish recorded approximately 200 phonetically rich sentences in each language, from which 150 are available for training. The sentences were recorded using a mimic style to facilitate the alignment, using high quality settings (96 kHz, 24 bits/sample), and three channels were available (membrane microphone, close-mouth microphone and laryngograph), as explained in Bonafonte et al. (2006b). For this work, the following data set was used:

- only the Spanish subset was used,
- membrane microphone, downsampled to 16 kHz, 16 bits/sample,
- two female (speakers F1 and F2) and two male (speakers M2 and M1) voices,
- four different source-target pairs have been trained (F1 → M1, F1 → F2, M2 → M1, M2 → F2).

6.3.2 Proposed parametrization

For the purpose of voice conversion we have used the same parametrization we proposed in the previous chapter for performing prosody modifications:

- vocal tract: N LSF coefficients $\theta_{\mathbf{vt}} = (lsf_1, lsf_2, \cdots, lsf_N)'$ (Section 5.1.1),
- glottal waveform: LF parameters and fundamental frequency $\theta_{\mathbf{lf}} = (R_a, R_g, R_k, E_e, F_0)'$ (Section 5.1.2),
- residual: reduced set of parameters using one symmetric window from eq. (3.59), and the parametrization technique from Section 5.1.3 with one symmetric window to reduce the dimensionality $\theta_{\mathbf{res}} = (b_{lvl}, w_{lvl}, w_c, w_l)'$.

The approach we follow in this work to integrate our feature set into the reference voice conversion paradigm is to treat each of the three parameter sets independently. This way, we grow an individual CART for each of the parameter sets described above (vocal tract, glottal waveform and residual), and the transformation is performed independently from each other.

Due to the different nature of the source-filter features, we use different measures to compute the distance between source, target and transformed frames. For the vocal tract, we use the same IHMD distance measure from (6.27) to compute the error index (6.23). For the glottal source and aspiration noise, we use the Euclidean distance between the converted and target vectors by setting $c(p) = 1$ in eq. (6.27) above:

$$D(\tilde{\mathbf{y}}, \mathbf{y}) = \sqrt{\sum_{p=1}^{P} (\tilde{y}(p) - y(p))^2}. \tag{6.29}$$

Since the growing procedure is performed independently for each parameter set, the resulting trees and GMM will have different size, as we will see in next section, which details the training stage.

6.3.3 Model training

During training, a small subset of the corpus was reserved for validation purposes, as we will see in Section 6.4. The purpose of this data was to select the appropriate number of mixtures per GMM to use when building the transformation function, and to perform the pruning process, as we will see next. Using the performance index as an objective measure of the quality of the conversion, we obtained numerical results that allowed us to identify training problems and performance bottlenecks. Although these measures are no substitute for the subjective MOS evaluation using previously unknown test data described in Section 6.4 and do not give an overall figure of the conversion performance, they are useful for the selection and adjustment of the different parameters.

For each of three parameter sets (vocal tract, glottal model and residual) we built a number of trees as follows. With the training data we created 5 training datasets using 5, 10, 25, 50 and 100 files respectively. For each of them, we built 7 classification trees, using a different GMM size on each of them (1, 2, 4, 8, 16, 32 and 64 mixtures). For each of the combinations of training dataset size and GMM size, three trees were obtained: one normally built, one used the pre-pruning method, and a third one using post-pruning. Each of these CART was then used to classify the validation data, and a measure of the transformation quality was obtained by computing the performing index from eq.(6.26) (using the distance measure from eqs.(6.27) or (6.29) depending on the parameter set as explained in the previous section). The figures showing the performance of all the described combinations have been placed at the end of the chapter so they do not interrupt the flow of the explanation. We will provide here detailed figures highlighting the specific issues we are dealing with. Figures 6.7 to 6.10 (pages 121 to 124) show the results for the validation data-set for the four source–target pairs (M2 $\rightarrow$ M1, M2 $\rightarrow$ F2, F1 $\rightarrow$ M1, and F1 $\rightarrow$ F2). Each figure shows the results of the vocal tract (top), glottal model (middle) and aspiration noise (bottom) conversions. On the x axis we have the number of files available in training (5, 10, 25, 50 and 100). The bars are grouped in triplets of the same color, each color indicating the number of mixtures used for the GMM. Each single-color triplet shows the results of the basic CART without pruning (left bar), with pre-pruning (middle bar) and post-pruning (right bar).

First we studied the effect of insufficient data on the overall performance and the issues derived from over-fitting the models to the training data. We also wanted to evaluate the relation between the number of mixtures in the GMM on the overall performance. Lastly, we studied the impact of the tree pruning methods, both in terms of conversion performance and tree size. Observing the general trend in the figures, the performance of the conversion improves when we increase the number of files used in training. Figure 6.3 shows this effect for the case of LF training in the F1 to F2 direction. As we can see, the more available data during the training stage, the better the performance. We can also observe that, although overfitting still occurs, its effect is greatly diminished when we increase the size of the training dataset. This was to be expected, since the larger the training data set, the more representative of the overall data set it is.

As Figure 6.4 illustrates with an example of the conversion direction F1 to F2, when we increment the number of mixtures per GMM while keeping constant the size of the training dataset (25 files in this case), the performance of the conversion system decreases. This is the result of overfitting, where the statistical modeling captures too much of the specific particularities of the training data and loses the ability to generalize and perform

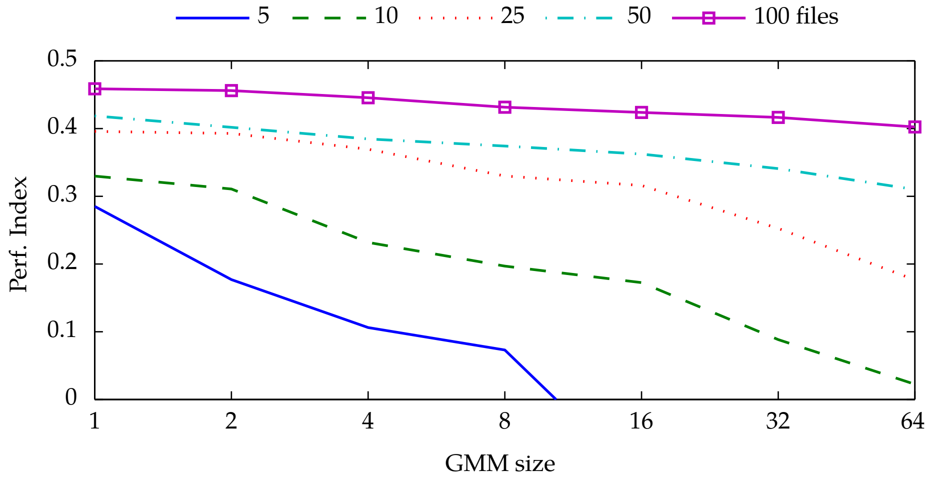


Figure 6.3: *Positive effect of increasing the training dataset size on the classification performance (LF CART, F1 → F2).*

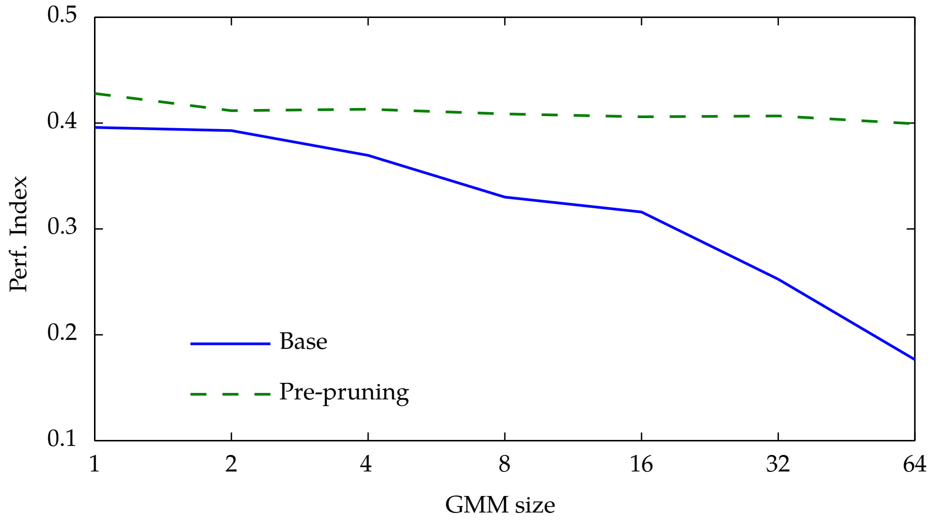


Figure 6.4: *Overfitting due to reduced size of training set and positive effect of pruning (direction F1 → F2 trained with 25 files).*

well with previously unknown data. As we can see, pruning the tree (in this particular example with the pre-pruning technique) mostly solves this problem.

The performance vastly improves when the trees are pruned, and this improvement is larger the less available training data we have (since the training set is then less representative of the overall data). There is a still a small degradation due to over-fitting when we use 25 or less files for training, but once we increase this number to 50 or 100, this effect is reduced and almost negligible in most cases. Figure 6.5 shows the relative improvement due to pre-pruning the tree for two different dataset sizes (10 and 100 files).

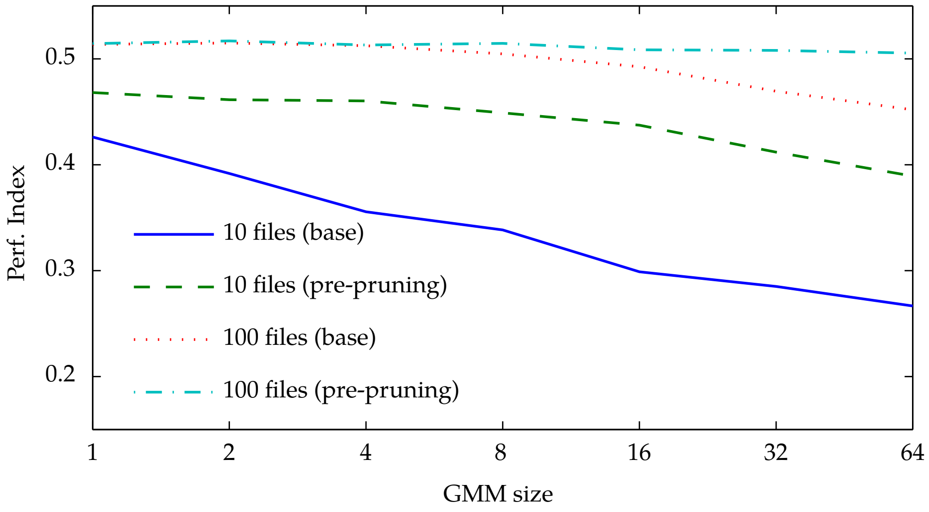


Figure 6.5: *Relative improvement of pruning on the overall performance when increasing the size of the training dataset (LSF CART, F1 → F2).*

The resulting size of the different CART can be seen in Tables 6.4 to 6.7 (pages 125 to 128). As expected, the dimensionality of the resulting trees is greatly reduced by applying either of the two pruning algorithms. Generally, pre-pruning resulted in smaller trees without an obvious impact on the conversion quality, which can be beneficial for certain applications (e.g., smaller trees require less memory and processing power). We also noticed that non-pruned CART tend to be larger (i.e., more nodes) the higher the ratio number of training files to number of mixtures is. After pruning, the tendency continues but with smaller differences.

In order to analyze the relevance of the different phonetic questions used in this work, we have included a representation of the first four levels of each CART in the form of a table. The resulting tables are in Section 6.6.3, Tables 6.8 to 6.11. This way we can gain knowledge of what phonetic characteristics are most discriminative. As we can see, although there is no clear consensus among the different conversion pairs and parametrization trees, there are some questions that would appear to dominate the first CART nodes. Among the most asked questions we find *consonant* (or the sort of complementary *vowel*), *nasal*, *close* and *back*. On a lesser degree, we could include such questions as *alveolar*, *palatal* or *lateral*. These questions seem reasonable, since they split the corpus in two big groups (vowels and consonants), and then each of them in large sections (close and back for vowels, nasal, alveolar or lateral for consonants). We can also observe that most of the GMM used for the transformation function are relatively small, having between 1 and 4 mixtures on average. This could also be seen in Figures 6.7 to 6.10.

Up to this point we have presented the selection criterion for finding the optimal CART configuration for each conversion pair and feature set. Now that we have obtained the optimal CART for each parameter and transformation pair, we will proceed with the subjective evaluation of the conversion performance using an online listening test. It is worth noting that the use of phonetic CART is also beneficial when using glottal features, as shown by the fact the trees grow also for this feature set, although phonetic information is usually more dependant on the vocal tract configuration.

6.4 Evaluation and results

As usual when evaluating VC algorithms, two metrics were used during this evaluation: one for rating the success of the transformation in achieving the desired speaker identification, and one for rating the quality. This is needed since strong changes usually achieve the desired identity at the penalty of degrading the quality of the signal. The evaluation was based on subjective rating by 14 human judges which were presented with examples from the transformed speech and the target one and had to decide on two aspects:

- the similarity of the converted and target voices using a 5-point MOS scale (1 – completely different, to 5 – identical),
- the transformed voice quality using a similar 5-points MOS scale (1 – bad, 5 – excellent).

For the reference system, in this evaluation we used the utterances generated during the last evaluation campaign of the European project TC-STAR, since the same training and testing conditions were followed. For our system, we chose the optimal CART using validation data for each pair of source target speakers and parametrization from the set of systems trained with 100 sentences. We then used them to convert the utterances belonging to the testing set. Some natural source-target and target-target examples were also presented for calibration reasons. Obviously, the participants in this evaluation ignored the origin of the samples they were being presented with.

The table below presents the results of the similarity and quality tests, for both the baseline and proposed parametrizations:

	Reference	**Proposed**	**Orig Src**	**Orig Tgt**
Quality	2.11	2.47 (+17%)	4.88	4.96
Similarity	2.87	3.02 (+5%)	1.62	4.79

As we can see, there is a noticeable improvement in terms of transformed voice quality as a result of the new features used, which rises from 2.11 to 2.47 points. The speaker identity transformation is also rated as more successful (3.02 vs 2.87). The last two columns are shown mainly for reference. We can see that the original source and target speaker voices are judged to be different (rated 1.62), while the real target–target combinations are naturally judged identical. Real samples are found to have an excellent quality.

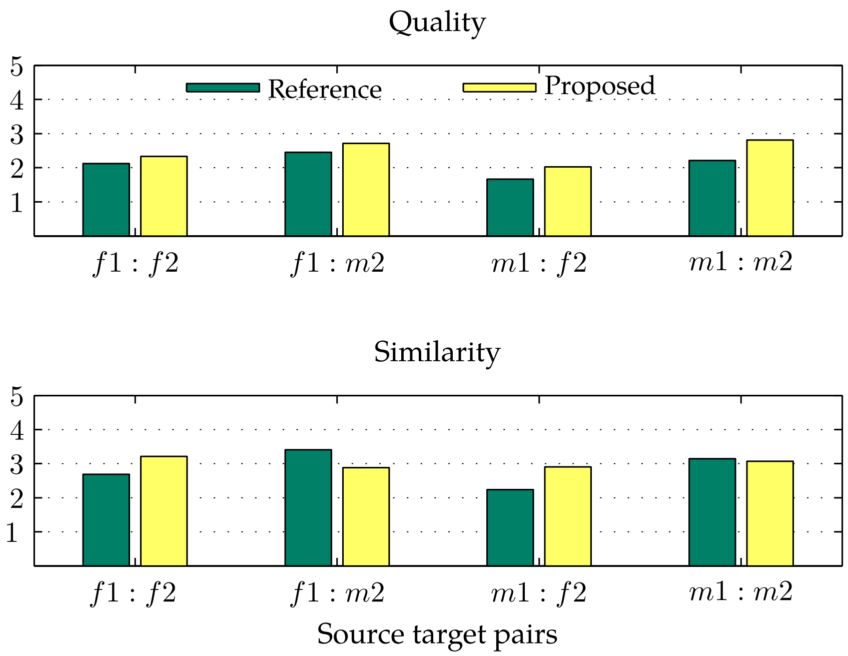


Figure 6.6: *Results of the quality and similarity MOS evaluation tests for the reference and proposed methods*

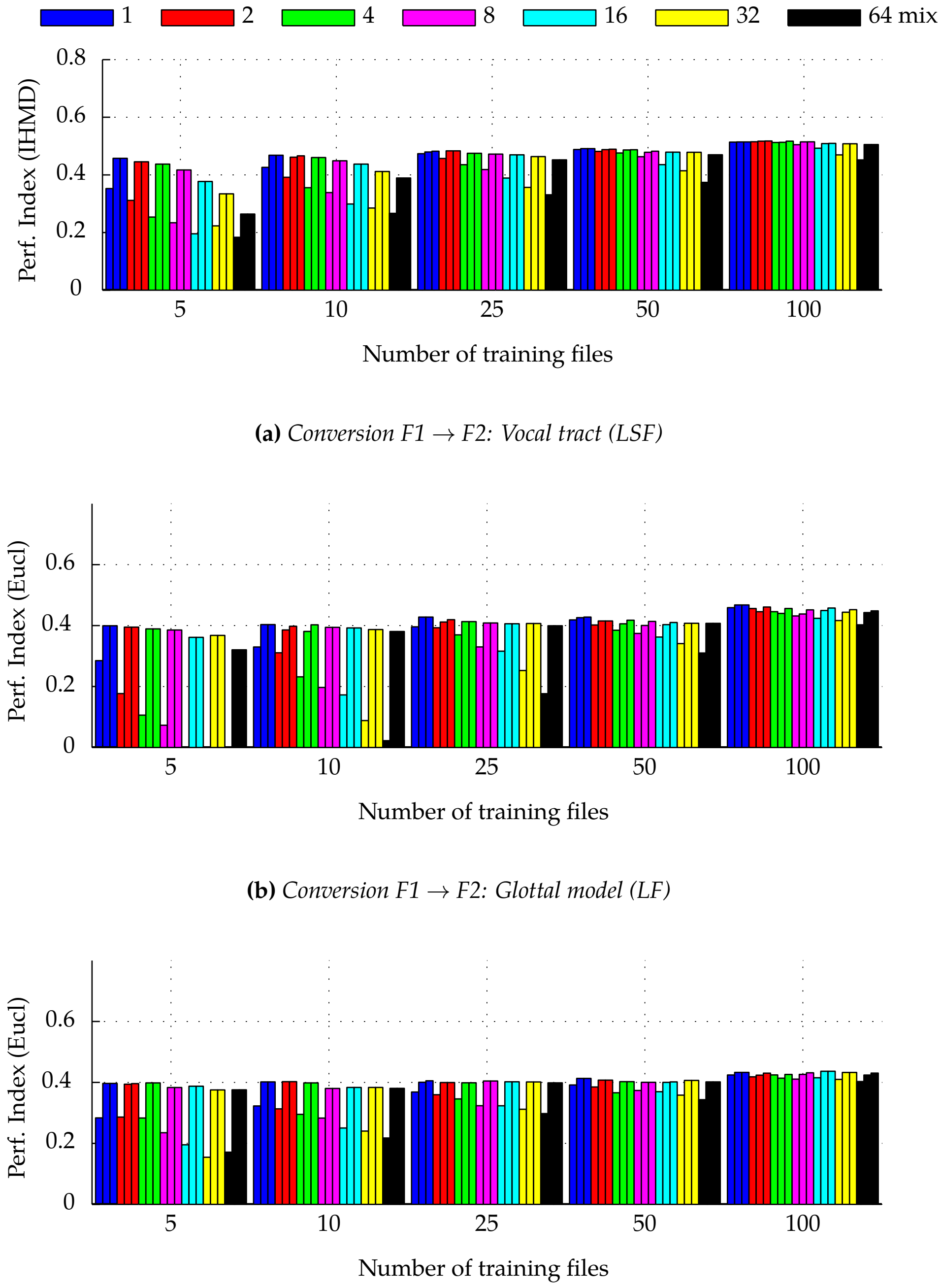
Figure 6.6 contains the results separated per source-target pairs. As we can see, the proposed parametrization results in an improved quality in all four conversion directions. In terms of similarity, we observe that the transformation towards the second female voice F2 is more successful using the proposed parametrization, whereas in the other two cases there is a slight decrease in the performance.

6.5 Conclusions

Voice conversion (VC) technology transforms the voice of a source speaker so that it is perceived as that of a target speaker. In this chapter we have studied the inclusion of glottal source characteristics in voice conversion systems (Pérez and Bonafonte, 2011). We used our previously reported glottal analysis algorithm to obtain three sets of parameters: one for the vocal tract using LSF, another for the glottal source using the LF model, and a last one for the aspiration noise using a parametrized envelope to modulate in amplitude high-pass filtered AWGN noise. To evaluate the benefits of this new parametrization in voice conversion tasks, we used a reference conversion system that estimates a linear transformation function using a joint target/source model obtained with CART and GMM. The reference system is based on the LPC model, uses LSF to represent the vocal tract and a selection technique for the residual. To include the new parametrization, we used the reference system algorithm to build a VC system for each of the three parameter sets using CART and GMM. We compared both parametrizations in the framework of an intra-lingual voice conversion task in Spanish. The tests showed that the new source/filter representation clearly improves the overall performance, both in terms of speaker identity transformation and voice quality of the converted voice. However, the quality is still poor compared to that of equivalent speech synthesis systems, and the voice conversion method would need to be further researched. However, we can conclude that voice source information can have a positive impact on voice conversion system.

6.6 Additional figures and tables

6.6.1 Performance figures

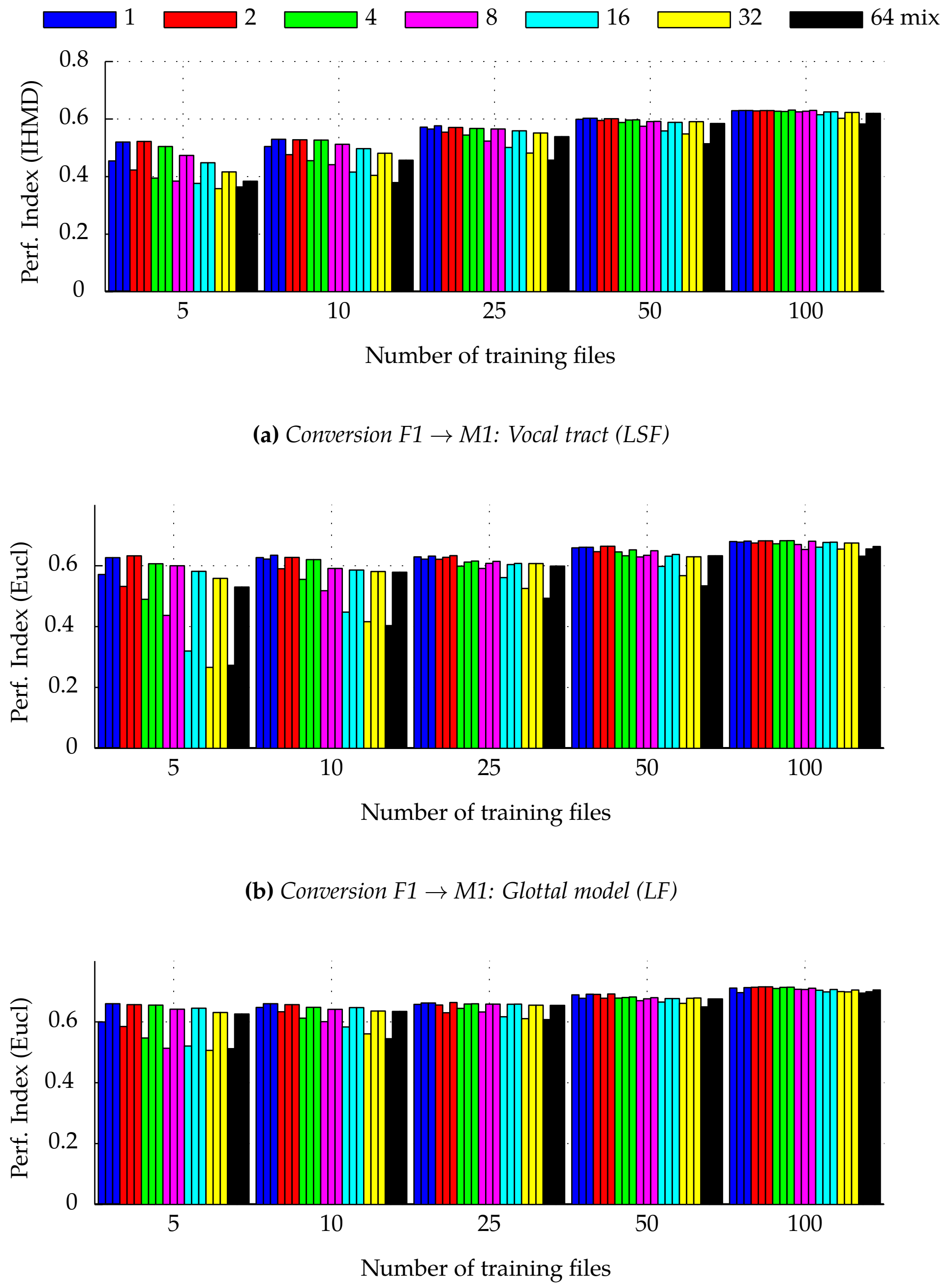


(a) *Conversion F1 → F2: Vocal tract (LSF)*

(b) *Conversion F1 → F2: Glottal model (LF)*

(c) *Conversion F1 → F2: Aspiration noise*

Figure 6.7: *Performance evaluation of the CART conversion from F1 to F2. The results are shown for the various combinations of training corpora and GMM sizes. For each number of mixtures per GMM, the performance index (from 0 worst, to 1 best) is shown in triplets of the same color (left: basic CART, middle: CART with pre-pruning, right: CART with post-pruning)*

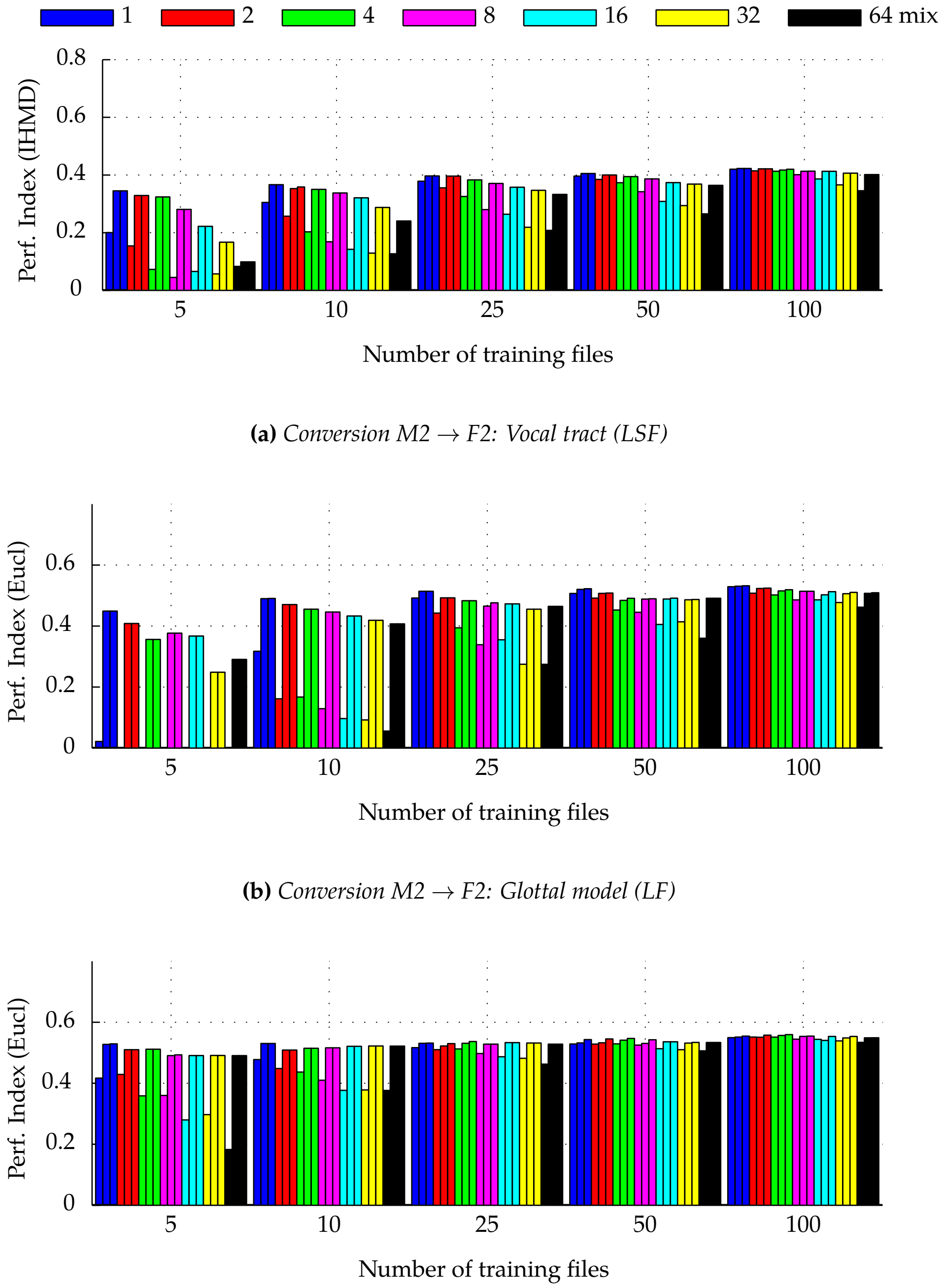


(a) *Conversion F1 → M1: Vocal tract (LSF)*

(b) *Conversion F1 → M1: Glottal model (LF)*

(c) *Conversion F1 → M1: Aspiration noise*

Figure 6.8: *Performance evaluation of the CART conversion from F1 to M1. The results are shown for the various combinations of training corpora and GMM sizes. For each number of mixtures per GMM, the performance index (from 0 worst, to 1 best) is shown in triplets of the same color (left: basic CART, middle: CART with pre-pruning, right: CART with post-pruning)*

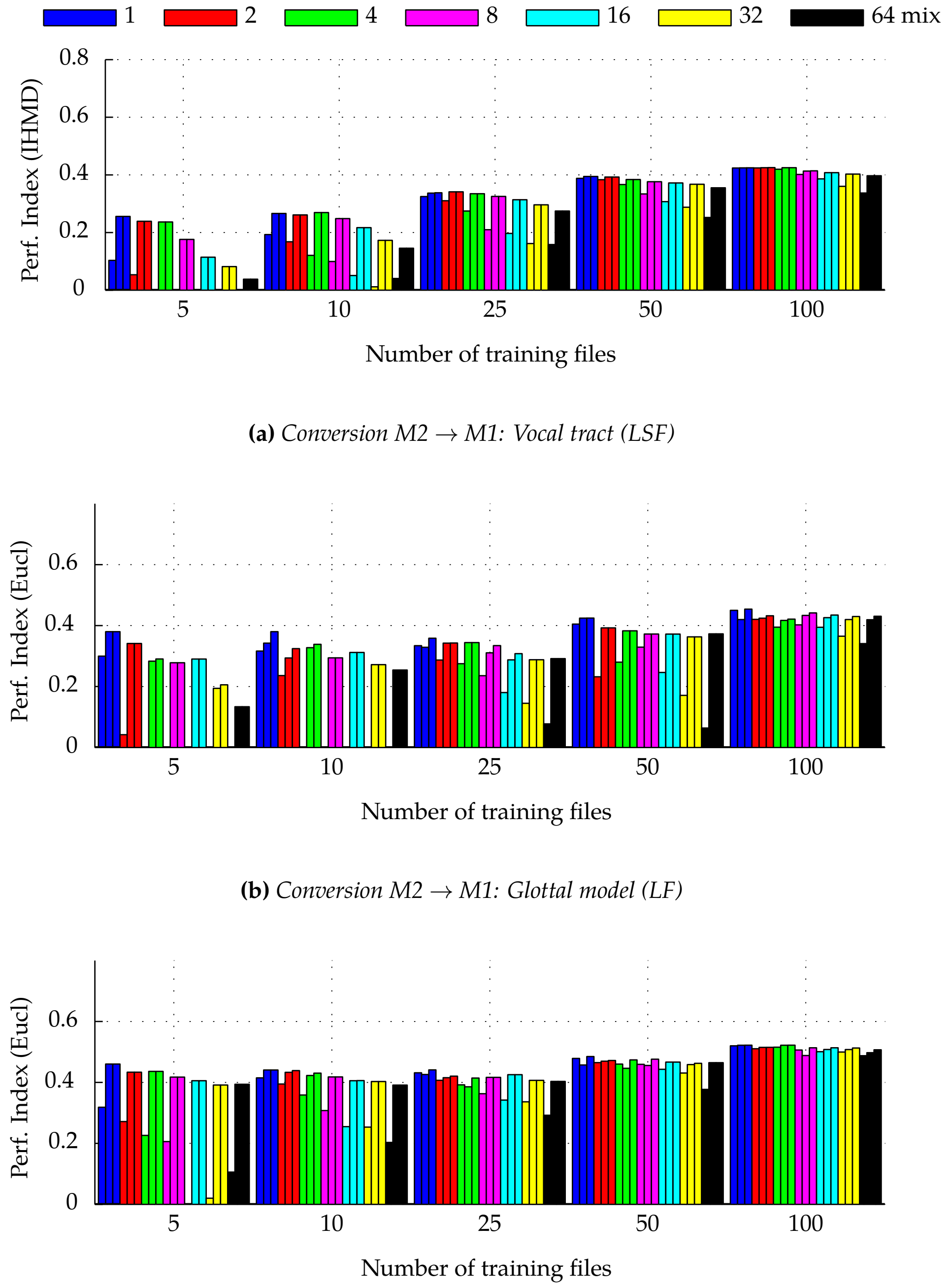


(a) *Conversion M2 → F2: Vocal tract (LSF)*

(b) *Conversion M2 → F2: Glottal model (LF)*

(c) *Conversion M2 → F2: Aspiration noise*

Figure 6.9: *Performance evaluation of the CART conversion from M2 to F2. The results are shown for the various combinations of training corpora and GMM sizes. For each number of mixtures per GMM, the performance index (from 0 worst, to 1 best) is shown in triplets of the same color (left: basic CART, middle: CART with pre-pruning, right: CART with post-pruning)*



(a) *Conversion M2 → M1: Vocal tract (LSF)*

(b) *Conversion M2 → M1: Glottal model (LF)*

(c) *Conversion M2 → M1: Aspiration noise*

Figure 6.10: *Performance evaluation of the CART conversion from M2 to M1. The results are shown for the various combinations of training corpora and GMM sizes. For each number of mixtures per GMM, the performance index (from 0 worst, to 1 best) is shown in triplets of the same color (left: basic CART, middle: CART with pre-pruning, right: CART with post-pruning)*

6.6.2 CART size after pruning

Table 6.4: *Number of CART nodes in F1 → F2 conversion for the various combinations of Gaussian mixtures and training corpora size (basic / pre–pruning / post–pruning).*

				Conversion F1 → F2		
				Vocal tract (LSF)		
Files		5	10	25	50	100
GMM	1	37/1/1	47/3/3	53/7/25	53/39/39	59/43/43
	2	39/1/1	47/1/7	53/13/13	53/13/31	57/35/43
	4	39/1/1	47/1/1	53/3/3	53/13/17	57/13/29
	8	39/1/1	47/1/1	53/1/1	53/1/7	57/23/27
	16	35/1/1	45/1/1	53/1/1	53/3/3	57/5/15
	32	29/1/1	35/1/1	49/1/1	53/1/1	57/5/5
	64	19/1/1	27/1/1	41/1/1	51/1/1	53/3/3
				Glottal model (LF)		
Files		5	10	25	50	100
GMM	1	39/3/3	47/9/9	53/15/15	53/21/27	59/25/25
	2	39/1/1	47/1/11	53/1/13	49/5/5	59/5/27
	4	39/1/1	47/1/7	53/7/7	55/1/19	59/3/21
	8	39/1/1	47/1/1	53/3/3	53/1/11	53/9/31
	16	35/1/1	45/1/1	53/3/3	53/1/5	57/7/17
	32	29/1/1	35/1/1	49/3/3	53/3/3	51/5/17
	64	19/1/1	27/1/1	41/1/1	51/3/3	55/1/19
				Residual		
Files		5	10	25	50	100
GMM	1	39/3/3	49/3/3	53/9/27	53/13/13	59/19/19
	2	37/3/9	47/3/3	49/3/3	51/11/11	51/7/31
	4	37/3/3	45/3/3	49/3/3	51/3/3	49/3/39
	8	39/3/3	47/3/3	53/5/5	53/3/3	59/5/27
	16	35/1/1	43/3/3	53/5/5	53/3/13	57/23/23
	32	25/1/1	33/3/3	49/3/3	53/5/5	57/15/15
	64	17/1/1	29/1/1	41/3/3	51/3/3	53/3/23

Table 6.5: *Number of CART nodes in F1 → M1 conversion for the various combinations of Gaussian mixtures and training corpora size (basic / pre–pruning / post–pruning).*

				Conversion F1 → M1		
				Vocal tract (LSF)		
Files		**5**	**10**	**25**	**50**	**100**
GMM	**1**	37/1/1	43/9/9	51/7/25	53/33/33	55/47/47
	2	39/1/1	43/1/1	51/19/19	53/21/21	55/39/39
	4	39/1/1	43/1/1	51/5/9	53/13/17	55/19/35
	8	39/1/1	43/1/1	51/1/1	53/5/13	55/17/27
	16	35/1/1	41/1/1	51/1/1	51/1/1	55/9/17
	32	25/1/1	35/1/1	49/1/1	51/1/1	53/3/3
	64	17/1/1	23/1/1	37/1/1	49/1/1	51/1/1
				Glottal model (LF)		
Files		**5**	**10**	**25**	**50**	**100**
GMM	**1**	37/1/1	43/3/29	51/17/37	53/35/35	55/35/43
	2	39/1/1	43/5/5	51/9/19	53/35/35	55/35/35
	4	39/1/1	43/3/3	51/3/9	53/7/41	55/35/39
	8	39/3/3	43/3/3	51/7/17	53/3/25	55/1/31
	16	35/1/1	41/1/1	51/3/15	51/3/23	55/25/31
	32	23/1/1	33/1/1	49/1/1	51/1/1	51/23/23
	64	17/1/1	23/1/1	37/3/3	49/1/1	51/3/17
				Residual		
Files		**5**	**10**	**25**	**50**	**100**
GMM	**1**	39/5/5	43/11/11	51/15/15	53/15/33	55/15/37
	2	39/7/7	43/7/7	51/1/23	53/13/35	53/31/31
	4	39/7/7	43/7/7	51/11/15	47/15/35	55/27/31
	8	39/3/3	41/5/5	51/9/9	51/9/25	55/17/35
	16	33/7/7	41/7/7	51/9/13	51/11/25	55/15/35
	32	25/3/3	31/3/3	49/9/9	51/11/19	53/15/23
	64	17/3/3	27/5/5	37/9/9	49/13/13	51/15/23

Table 6.6: *Number of CART nodes in M2 → F2 conversion for the various combinations of Gaussian mixtures and training corpora size (basic / pre–pruning / post–pruning).*

				Conversion M2 → F2		
				Vocal tract (LSF)		
Files		**5**	**10**	**25**	**50**	**100**
GMM	**1**	37/1/1	47/7/7	55/23/23	55/35/35	59/39/39
	2	37/3/3	47/1/7	55/13/13	55/21/21	59/31/31
	4	37/1/1	47/1/1	55/9/9	55/17/17	59/19/29
	8	37/1/1	47/1/1	55/3/3	55/5/5	59/11/17
	16	35/1/1	43/1/1	55/3/3	55/1/1	57/9/9
	32	25/1/1	35/1/1	47/1/1	53/1/1	55/7/7
	64	17/1/1	29/1/1	41/1/1	51/1/1	53/3/3
				Glottal model (LF)		
Files		**5**	**10**	**25**	**50**	**100**
GMM	**1**	37/1/1	47/7/13	55/17/17	55/15/27	59/25/35
	2	37/1/1	47/5/5	51/5/5	53/11/15	59/13/31
	4	39/3/3	47/3/3	53/5/5	55/1/9	59/11/31
	8	39/1/1	49/1/1	53/1/5	53/1/9	59/7/7
	16	33/1/1	43/3/3	55/1/1	49/1/7	57/1/11
	32	23/3/3	33/1/1	49/5/5	53/1/5	57/5/13
	64	15/1/1	27/1/1	39/1/1	53/1/1	53/3/11
				Residual		
Files		**5**	**10**	**25**	**50**	**100**
GMM	**1**	39/1/7	47/1/1	55/1/29	55/3/35	59/19/39
	2	33/5/5	47/5/5	51/1/29	55/3/33	59/15/43
	4	37/1/1	49/3/3	51/3/15	55/7/27	59/15/35
	8	37/1/5	49/3/3	55/3/3	55/3/33	57/11/21
	16	33/3/3	43/1/1	53/1/1	55/3/3	59/3/33
	32	25/1/1	35/1/1	47/1/1	53/3/29	57/7/27
	64	17/1/1	23/1/1	41/1/1	51/3/3	53/11/15

Table 6.7: *Number of CART nodes in M2 → M1 conversion for the various combinations of Gaussian mixtures and training corpora size (basic / pre–pruning / post–pruning).*

				Conversion M2 → M1		
				Vocal tract (LSF)		
Files		**5**	**10**	**25**	**50**	**100**
	1	31/1/1	43/1/1	55/21/29	53/35/35	55/47/51
	2	33/1/1	43/1/1	53/11/11	53/27/27	55/37/43
	4	33/1/1	43/1/1	53/3/3	53/11/11	55/25/29
GMM	**8**	31/1/1	45/1/1	53/3/3	53/5/5	57/15/25
	16	31/1/1	37/1/1	51/1/1	53/3/3	57/9/9
	32	19/1/1	33/1/1	41/1/1	51/1/1	55/3/3
	64	9/1/1	23/1/1	33/1/1	43/1/1	51/1/1
				Glottal model (LF)		
Files		**5**	**10**	**25**	**50**	**100**
	1	31/1/1	43/1/23	55/1/27	53/15/23	55/1/45
	2	27/1/1	43/1/25	55/23/27	51/1/1	55/3/33
	4	31/1/7	45/3/9	53/9/9	53/1/1	57/3/31
GMM	**8**	31/1/1	45/3/3	53/1/7	49/1/1	53/3/19
	16	31/1/1	37/1/1	51/1/13	53/1/1	55/5/31
	32	19/1/5	29/3/3	41/5/5	53/1/1	55/1/11
	64	9/1/1	21/1/1	31/1/1	43/1/1	53/1/5
				Residual		
Files		**5**	**10**	**25**	**50**	**100**
	1	31/3/3	45/3/3	55/3/23	55/5/23	57/37/41
	2	31/3/3	39/13/19	51/9/27	55/25/39	57/33/37
	4	31/3/3	43/5/11	55/1/27	51/3/21	57/33/41
GMM	**8**	33/3/3	43/3/3	53/5/5	53/7/21	57/3/21
	16	29/3/3	37/1/5	53/5/5	53/11/11	57/17/27
	32	19/3/3	31/1/1	41/3/3	53/5/23	55/21/31
	64	9/1/1	21/3/3	33/1/1	45/5/5	53/9/23

6.6.3 CART phonetic questions

Table 6.8: *Final CARTs used for the F1 → F2 conversion of vocal tract, glottal model and residual (only the first 4 levels are shown). Nodes kept after the pruning are in bold; the dashes represent empty nodes*

Vocal tract (LSF)

CART: 43 nodes (leaves: 22) – GMM: 2 mixture(s)

consonant	**nasal**	**bilabial**	—
			alveolar
		dental	**approximant**
			alveolar
	close	**glide**	**back**
			front
		front	—
			center

Glottal model (LF)

CART: 25 nodes (leaves: 13) – GMM: 1 mixture(s)

center	—	—	—
			—
		—	—
			—
	mid_close	*back*	—
			—
		nasal	**bilabial**
			close

Residual

CART: 23 nodes (leaves: 12) – GMM: 16 mixture(s)

vowel	*close*	*back*	—
			—
		mid_close	*front*
			—
	nasal	**bilabial**	—
			alveolar
		fricative	**palatal**
			alveolar

Table 6.9: *Final CARTs used for the F1 → M1 conversion of vocal tract, glottal model and residual (only the first 4 levels are shown). Nodes kept after the pruning are in bold; the dashes represent empty nodes*

Vocal tract (LSF)

CART: 35 nodes (leaves: 18) – GMM: 4 mixture(s)

alveolar	**lateral**	—	—
			—
		nasal	—
			fricative
	back	**close**	*vowel*
			—
		nasal	**bilabial**
			mid_close

Glottal model (LF)

CART: 39 nodes (leaves: 20) – GMM: 4 mixture(s)

back	*mid_close*	—	—
			—
		vowel	—
			—
	consonant	**voiced**	**fricative**
			fricative
		glide	—
			close

Residual

CART: 31 nodes (leaves: 16) – GMM: 2 mixture(s)

consonant	**nasal**	**palatal**	—
			bilabial
		plosive	*voiced*
			alveolar
	close	**back**	**glide**
			glide
		back	—
			open

Table 6.10: *Final CARTs used for the M2 → F2 conversion of vocal tract, glottal model and residual (only the first 4 levels are shown). Nodes kept after the pruning are in bold; the dashes represent empty nodes*

Vocal tract (LSF)
CART: 39 nodes (leaves: 20) – GMM: 1 mixture(s)

consonant	**nasal**	**palatal**	—
			alveolar
		alveolar	**lateral**
			dental
	close	**back**	*glide*
			glide
		back	—
			center

Glottal model (LF)
CART: 35 nodes (leaves: 18) – GMM: 1 mixture(s)

close	**back**	**glide**	—
			—
		glide	—
			—
	consonant	**nasal**	**bilabial**
			trill
		back	—
			center

Residual
CART: 35 nodes (leaves: 18) – GMM: 4 mixture(s)

vowel	**close**	**back**	—
			—
		open	—
			back
	nasal	*bilabial*	—
			alveolar
		fricative	*palatal*
			back

Table 6.11: *Final CARTs used for the M2 → M1 conversion of vocal tract, glottal model and residual (only the first 4 levels are shown). Nodes kept after the pruning are in bold; the dashes represent empty nodes*

Vocal tract (LSF)

CART: 43 nodes (leaves: 22) – GMM: 2 mixture(s)

consonant	**alveolar**	**lateral**	—
			nasal
		palatal	**nasal**
			bilabial
	back	**mid_close**	—
			glide
		mid_close	—
			close

Glottal model (LF)

CART: 45 nodes (leaves: 23) – GMM: 1 mixture(s)

consonant	**nasal**	*bilabial*	—
			palatal
		unvoiced	**alveolar**
			lateral
	close	**back**	**glide**
			glide
		back	—
			center

Residual

CART: 41 nodes (leaves: 21) – GMM: 1 mixture(s)

consonant	**lateral**	**alveolar**	—
			—
		nasal	**palatal**
			plosive
	back	**close**	**glide**
			—
		close	**glide**
			center

CHAPTER 7

Laryngeal voice quality

Voice quality (VQ) is a general term that has been used to define a broad range of vocal characteristics depending on the field of phonetics it is being used in. In a general sense, by voice quality we are referring to the set of attributes or characteristics of a particular speaker's voice that are caused by a continuous variation of the laryngeal (vocal folds) and supra-laryngeal (vocal tract) apparatus. In this work, we are concentrating only on the first type: the laryngeal voice quality.

7.1 Voice quality classification

According to the description in Laver (1980) of the laryngeal voice quality, several voice types can be differentiated (after Keller (2005)):

- Modal voice, produced by a moderate adductive tension, medial compression and longitudinal tension (see Fig. 7.1).

- Falsetto voice, characterized by a high adductive tension, large medial compression and high longitudinal tension. It is an alternative to modal voice.

- Whispery voice, defined by a low adductive tension, moderate to high medial compression and variable longitudinal tension. It produces a triangular opening of vocal folds of variable size. This voice type can be combined with modal or falsetto voice.

- Creaky voice (a.k.a. *vocal/glottal fry* or *laryngealization*) is a type of phonation in which the arytenoid cartilages in the larynx are drawn together (see Fig. 7.3). The vocal folds are compressed rather tightly, becoming relatively slack and compact, forming a large, irregularly vibrating mass. The frequency of the vibration is very low and the airflow through the glottis is very slow.

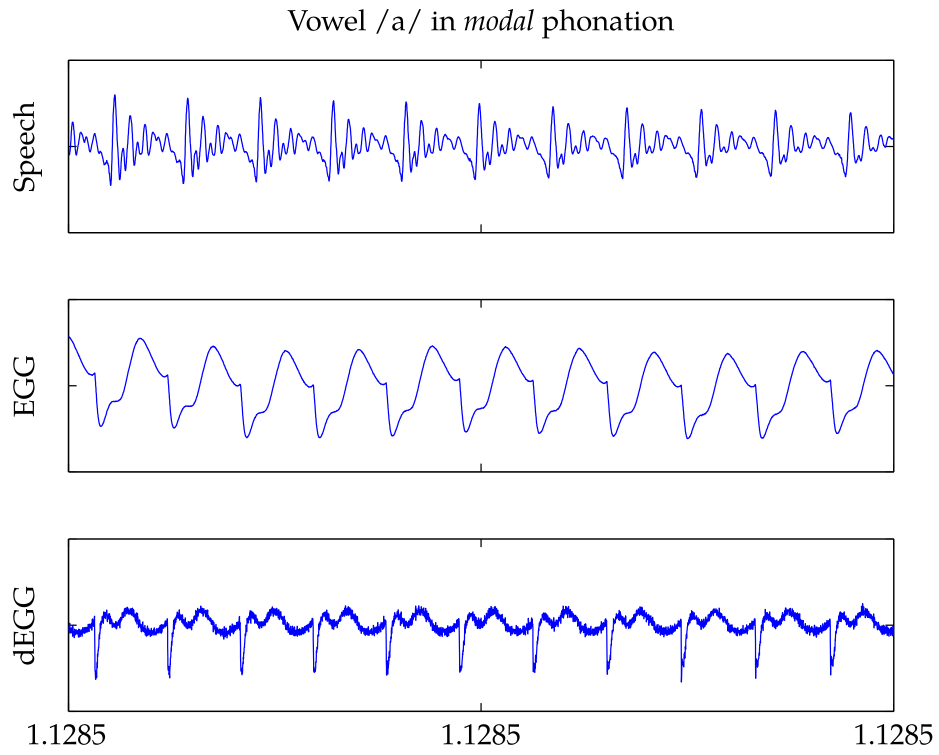


Figure 7.1: *Speech, laryngograph (EGG) and differentiated laryngograph (dEGG) waveforms corresponding to the vowel /a/ uttered in* modal *phonation*

- Rough voice (a.k.a. *ventricular*, *pressed* or *harsh*), characterized by the production of speech sounds (typically vowels) with a constricted laryngeal cavity (see Fig. 7.4). This generally involves some kind of epiglottal co-articulation. The ventricular folds (the false vocal cords) are used to damp the glottis. For voiced sounds, this is similar to what happens when someone who is lifting a heavy load talks. For voiceless sounds, the example would be clearing one's throat.

- Breathy voice (or *murmured*), defined by the vocal folds being held apart while they are vibrating. A larger volume of air escapes between them, producing an audible noise. This phonation can be heard as an allophone to the English sound /h/ between vowels, as in *behold*.

One of the main goals of researchers on voice quality is the automatic acquisition of reliable voice source measures connected with the human production system. For instance, the European Center of Excellence in Speech Synthesis (ECESS) included the possibility of using voice quality information in modular speech synthesis systems (Pérez et al., 2006). Depending on the application and on the amount of data to be processed, it may be sufficient to use manual methods requiring operator interactivity. However, for many

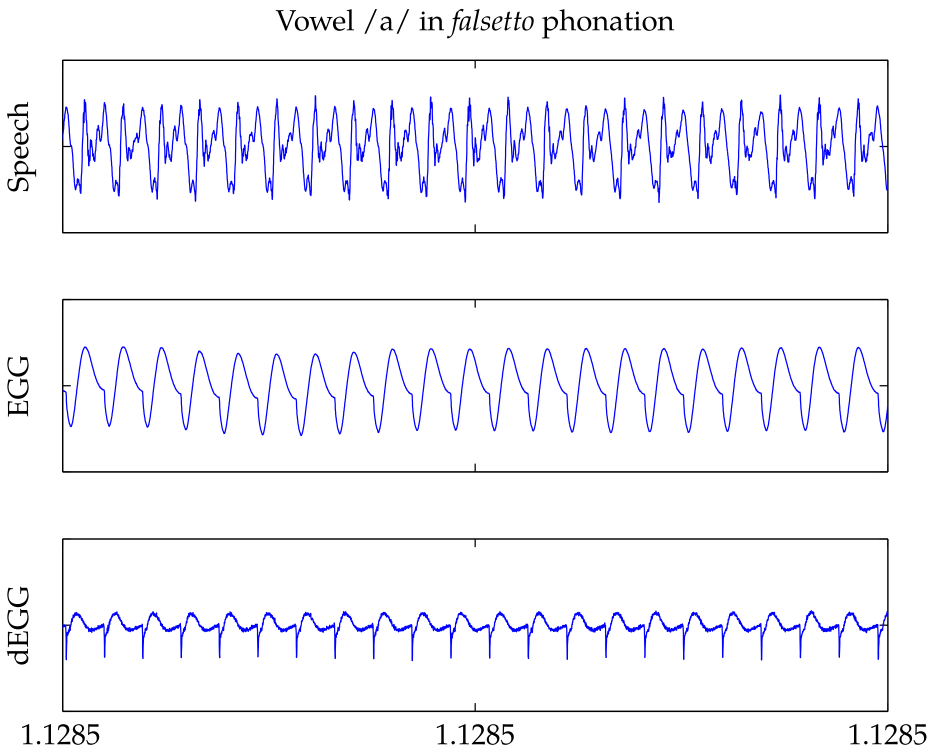


Figure 7.2: *Speech, laryngograph (EGG) and differentiated laryngograph (dEGG) waveforms corresponding to the vowel /a/ uttered in* falsetto *phonation*

current applications there is need to process large amounts of data, thus requiring automatic methods performing with a fair degree of accuracy. The more immediate measures that can be directly obtained from the estimated glottal waveform are:

- Open Quotient (OQ): ratio of open phase duration to the fundamental period,
- Speed Quotient (SQ): ratio of the glottal opening phase to the closing phase,
- Closing Quotient (ClQ): ratio of the glottal closing phase to the fundamental period.

According to several reports, time-based measures are often prone to contain errors. Fant and Lin (1998) presents a study of the relation between glottal parameters and spectral properties. Their work compresses analyzes of both subglottal coupling effects and covariant formant bandwidths and narrow band processing techniques benefiting of the frequency-domain representation. They conclude that the t_a parameter of the LF model, controlling the abruptness of the closing phase (and thus the spectral tilt of the glottal flow), can be more reliably estimated using frequency-domain techniques. Other authors present similar conclusions (Alku and Vilkman, 1996; Arroabarren and Carlosena, 2003b; Childers and Lee, 1991).

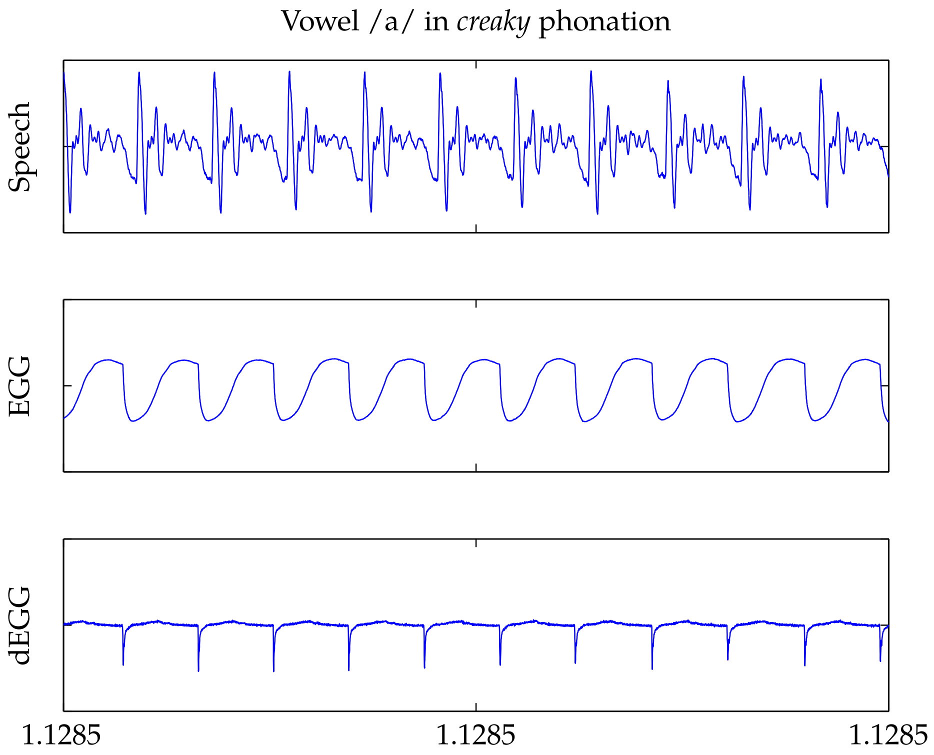


Figure 7.3: *Speech, laryngograph (EGG) and differentiated laryngograph (dEGG) waveforms corresponding to the vowel /a/ uttered in* creaky *phonation*

Many reported works suggest that amplitude-based measures present less difficulties than time-derived ones (Alku et al., 2002; Gobl and Chasaide, 2003a). Fant et al. (1994) present a new parameter R_d, a normalized version of T_d to the fundamental period (with a scale factor such that R_d equals T_d in milliseconds for a fundamental frequency of 110 Hz):

$$R_d = 1000(U_p/E_e)(f_0/110) = (T_d/T_0)/0.11. \tag{7.1}$$

This new R_d parameter can be used to derive default predictions of the remaining parameters using correlations obtained with regression methods (Fant, 1995, 1997; Fant et al., 1994). Alku and Vilkman (1996) introduce an equivalent parameter: the amplitude quotient (AQ) defined as

$$AQ = f_{ac}/d_{peak}, \tag{7.2}$$

where d_{peak} is the maximum negative amplitude of the differentiated glottal flow and f_{ac} is the amplitude of the glottal flow pulse (equivalent to E_e and U_p in (7.1) above). A normalized version of the AQ parameter is presented in Alku (2003):

$$NAQ = AQ/T_0 = 0.11\,R_d. \tag{7.3}$$

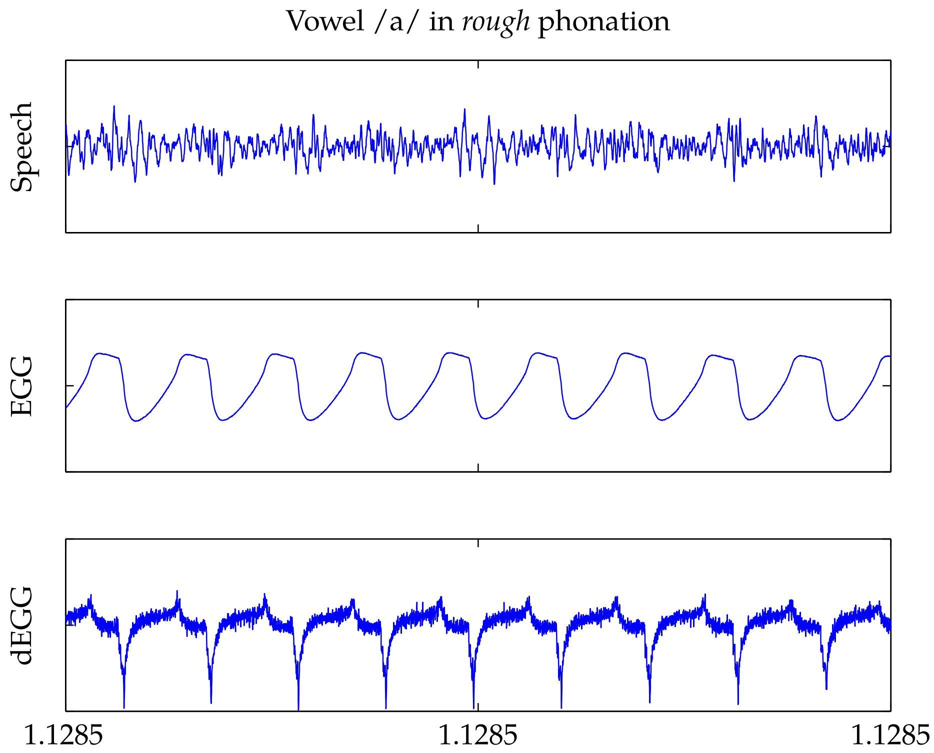


Figure 7.4: *Speech, laryngograph (EGG) and differentiated laryngograph (dEGG) waveforms corresponding to the vowel /a/ uttered in* rough *phonation*

The work of Alku and his colleagues suggests that these parameters are well suited for discrimination of the tense-lax voice qualities (Alku et al., 2002; Alku and Vilkman, 1996; Mokhtari and Campbell, 2003). Gobl and Chasaide (2003a) add to this parameter set some amplitude-based approximations of the glottal skewness and the open quotient. The intention is to extend the analysis to the differentiation of other voice types. The experiments show ambiguous results and no conclusions can be extracted. They state their intention of investigating the use of the extended set of LF parameters proposed in Fant (1995), Fant and Kruckenberg (1996) and Fant (1997). Airas and Alku (2007) present a study of the performance of 21 different glottal flow parameters using statistical methods. The authors conclude that, in terms of expressing phonation type changes, NAQ (7.3) and AQ (7.2) outperform the rest, including R_d. Although this may seem strange at first, due to its direct relation with NAQ as expressed in (7.3), Airas and Alku argue that the difference may reside on the origin of the estimations: whereas NAQ is directly estimated from the inverse-filtered glottal waveform, R_d is obtained from the fitted parametric LF model.

Campbell and Mokhtari (2003) have been working with large databases of conversational speech, studying the role of voice-quality in everyday speech [1]. They report that a significant correlation exists between the NAQ parameter and some speaker dependent characteristics (interlocutor, speech-act and speaking-style).

Mokhtari et al. (2003) propose a statistical approach using Principal-Component Analysis (PCA) for modeling the (derivative) glottal volume-velocity waveform. The database presented by Laver (1980), containing sentences uttered with different phonation types, is used in the experiments. Acoustic processing is performed by means of an unsupervised algorithm that detects centers of measurement reliability (details in Mokhtari and Campbell, 2003). The authors report that four principal components are needed to achieve discrimination among the thirteen voice qualities. They are interpreted in terms of duration of the fundamental period, convexity/concavity of the open phase, speed of the opening and closing phases.

Childers and Lee (1991) presents a study of the laryngeal vocal quality of four different voice types (modal, breathy, falsetto and vocal fry). They report that, among the different characteristics of the glottal flow, four factors are important for discerning among the studied voice types: glottal pulse width and skewness, noise component and the duration of the closing phase. First, the closure instants of the glottis are located using the peaks of the prediction error. These epochs are then used to perform a closed-phase LP analysis for each pitch period. The glottal volume-velocity waveform is obtained by inverse filtering the speech signal with the estimated vocal tract filters and integrating the result. The method is validated using a cascade formant synthesizer (Klatt, 1980) excited with waveforms generated using the LF model (Fant et al., 1985). The reported listening tests show that the aforementioned factors are sufficient for synthesizing the four voice types (a new three-poles glottal flow model is proposed for generating the synthetic voice types). The research is a continued and an extended article is presented in Childers and Hu (1994) using some of the previous results. The authors use statistical multiple linear regression analysis and others techniques to determine which LF parameters are more significant for each voice type. As a result, simple rules are introduced to synthesize each type using the formant synthesizer and the LF model as presented in (Fant et al., 1985; Klatt and Klatt, 1990).

A new parameter for the parametrization of the spectrum of the glottal flow, the Parabolic Spectral Parameter (PSP), is introduced in Alku et al. (1997). The PSP value is computed from the width a of a parabolic curve $y = a \cdot t^2 + b$ to the lower part of the glottal spectrum. A least squares error criterion is followed to obtain the best parabolic estimation. The a value is also computed from the hypothetical DC-flow, modeled using a rectangular window of length to the fundamental period, to obtain a_{max}. Thus, the

[1] Details of the recording process can be found in Campbell (2002).

definition of PSP is:

$$PSP = \frac{a}{a_{max}}. \tag{7.4}$$

The authors argue that with the new method the dependency on the correct identification of the harmonics is avoided. Moreover, PSP analysis allow the comparison of voices with different fundamental frequencies, since the main spectral dependency is with the spectral decay (tilt). Further studies by Arroabarren and Carlosena (2003a) comparing the NAQ and PSP parameters using the LF model conclude that in spite of the different definition, both parameters are equivalent. They conclude that the NAQ is a more reliable measure than the PSP, since the later depends on the accuracy of the glottal flow period extraction.

For our work on voice quality analysis, we will use the LF parameters resulting from our parametrization algorithm R_a, R_g, R_k and fundamental frequency F_0. They are related the standard open quotient OQ and asymmetry coefficient or speed quotient SQ measures explained before as:

- $OQ = \frac{T_e}{T_0} = \frac{1+R_k}{2\,R_g}$,
- $SQ = \frac{T_p}{T_e - T_p} = \frac{1}{R_k}$.

This relation will allow us to compare our results with those previously reported in the literature.

7.2 Voice quality analysis using glottal measures

A small corpus was recorded to study glottal extracted parameters across different voice qualities or phonation modes. A professional female speaker was asked to produce sustained vowels for each of the following voice qualities: modal, falsetto, rough and creaky. The recordings for breathy and whispery were discarded for this experiment due to the lack of useful laryngograph information, necessary for our decomposition algorithm. For each of the modes, the five Spanish vowels were recorded in isolation, averaging 1.5 seconds long. The sentences were recorded in a professional studio using a high-quality membrane microphone, at 96 kHz and 24 bits/sample. It was later reduced to 16 kHz and 16 bits/sample for the experiment. The corpus was then analyzed using our decomposition and parametrization algorithm as explained in Chapter 3. As a result, 16 LSF coefficients representing the vocal tract, 4 LF parameters (R_a, R_g, R_k and E_e) for the glottal source (plus the fundamental frequency F_0), and 4 parameters for the aspiration noise or residual (b_{lvl}, w_{lvl}, w_c and w_l) were obtained.

Figures 7.5 and 7.6 below show boxplots of the four different LF and residual parameters respectively, for five vowels uttered in modal (i.e., normal), falsetto, rough and creaky modes. Prior to analyzing and comparing the different voice qualities as produced by our particular speaker with those reported in the literature, one must consider that the differences between the different voice qualities are often small. We are not dealing with extreme or pathological phonation types and, in principle, a regular speaker with a normal apparatus should be able to produce and control each of them (Laver, 1980). Out analysis focuses on the glottal source features, and its impact on the production of each voice quality, and residual or aspiration noise characteristics will only be mentioned when being particularly important.

The rough or harsh voice quality is usually very difficult to analyze, due to the large variations from cycle-to-cycle, as can be seen in figure 7.5. All the parameters have larger boxes (the bottom and the top represent the 25^{th} and 75^{th} percentile respectively) in the boxplots. The duration of the open phase is generally much shorter than in the modal case, as shown in the open quotient figure 7.5f. The fundamental frequency is approximately 5% lower than its modal counterpart, and the asymmetry coefficient R_k (figure 7.5b) is on average larger, but with higher variance. Our results generally agree with those used in Gobl and Chasaide (2003b) to study the relationship between voice quality and the communication of emotion. The values for the open quotient were approximately 40%, lower than for modal voices (60%). The fundamental frequency used was approximately the same in both cases. The main differences are in the values for R_k and R_a. Gobl and Chasaide report using higher speed quotient SQ (i.e., lower R_k) for the harsh voices, as opposite to our findings. Furthermore, they use less attenuation for the higher frequencies in the case of harsh voices, which translates into more abrupt closures of the glottis. In our case we have the opposite case: rough voices usually have higher R_a values; in other words, the closure in the rough case is smoother than in the modal case. We can also observe in figure 7.6 that there is more aspiration noise present than in other modes, specially modal. The values for the levels b_{lvl} and w_{lvl} in figs. 7.6a and 7.6b are higher in this voice quality.

Creaky voice is usually characterized by irregular pulse amplitudes and F_0, which is usually lower than in modal voice. Our analysis shows that the open quotient (figure 7.5f) is similar, but slightly lower, to that corresponding to modal phonation, which seems to agree with the majority of studies (Gobl, 1989; Gobl and Chasaide, 2003b; Ní Chasaide and Gobl, 1997). On the other hand, Childers and Lee (1991) reports shorter pulses for creaky voices. In general, R_a is found to be lower than in the modal case, although in our case the values are very similar, slightly higher in the creaky case (figure 7.5c). This may be to the double pulsing phenomenon explained below. Regarding the asymmetry coefficient R_k shown in figure 7.5b, our results seem to confirm those previously reported in the literature: the pulses in creaky phonation have lower R_k values, and are thus more skewed.

One of the main characteristics of creaky voice is its diplophonia or double pulsing: each consecutive glottal pulse is different, with every second pulse looking similar, although not identical (Gobl, 1989; Karlsson and Liljencrants, 1996). Due to the averaging nature of the results presented in figure 7.5, this phenomenon is not clearly reflected there, since it is caused by a cycle-to-cycle variation. This diplophonia can be clearly seen in figure 7.7, where several waveforms corresponding to the vowel /a/ in creaky phonation are shown. Observing the laryngograph waveforms in the upper figures, the double pulses are clearly defined there. In the lower portion of the figure, the corresponding sections of the inverse-filter and parametrized glottal waveforms are plotted and the same effect can be clearly seen. There we see that the larger pulses (larger amplitude E_e) have shorter return phases, as opposite to the shorter pulses, that show smoother return phases and lower amplitudes. This coincides with the results reported in Ní Chasaide and Gobl (1997). The KLSYN88 formant synthesizer (Klatt and Klatt, 1990) contains a parameter addressing this particular phenomenon, allowing the user to alter every second pulse, reducing its fundamental frequency and amplitude (Gobl and Chasaide, 2003b).

For the falsetto voice, as expected, the values of the fundamental frequency F_0 (figure 7.5e) are higher than in the modal phonation (approximately $330\,\mathrm{Hz}$, whereas they normally are $170\,\mathrm{Hz}$ in the modal case). We can also see that the open quotient OQ (figure 7.5f) is larger than for the modal case in four of the five analyzed vowels. As figure 7.5a shows, the asymmetry coefficient R_k is clearly larger (and has a higher variability) for the falsetto voice quality. These results are in general agreement with those reported in the literature. Childers and Lee (1991) reported a tendency towards larger pulse widths (i.e., open quotient), less pulse skewing (i.e., more symmetrical pulses) the and relatively smooth and progressive closure. Kaburagi et al. (2007) reports similar results, but their method tends to underestimate the open quotient, probably due to using a different model for the glottal waveform (a polynomial model derived from the Rosenberg (1971) one).

The following list summarizes our findings (the changes are noted with respect to the modal phonation):

- rough phonation: larger variations of all parameters, shorter O_q, slightly lower F_0, higher R_k, higher R_a, higher noise levels

- creaky phonation: irregular pulse amplitudes and F_0, diplophonia (first pulse: large E_e, low R_a; second pulse: low E_e, large R_a), similar O_q, lower R_k,

- falsetto phonation: higher F_0, higher O_Q, higher R_k (larger variability), slightly higher R_a (larger variability).

7.3 Experiments

To complement the analytical study of the voice quality corpus detailed in the previous section, we conducted two experiments, first addressing here the issue of whether our voice source parametrization algorithm is able to accurately capture the characteristics of each of the voice qualities, and then exploring the utility of our parametric model in the field of emotion recognition. The next sections present the details of the experimental procedures.

7.3.1 Voice quality identification

This experiment was design to assess whether or not our glottal parametrization algorithm was able to capture all the details of the different voice qualities. Ideally, all the VQ information should be contained in the LF and aspiration noise parameters, since we are dealing with laryngeal characteristics. We designed an experiment aimed at assessing whether our glottal parametrization algorithm was capturing the characteristics of the different phonation types as predicted by the theory. For each of the non-modal VQ (rough, creaky and falsetto), each of the five vowels analyzed in the previous section were resynthesized using the extracted glottal LF and residual parameters, but using the vocal tract filter computed in the modal case. The filter frames were interpolated or discarded in order to obtain the appropriate number for each mode. The set of samples used in the test was:

- utterances representing the modal, rough, creaky and falsetto VQ, resynthesized using the *original* vocal tract,
- modified utterances representing the non-modal qualities (rough, creaky and falsetto), resynthesized using the vocal tract corresponding to the same vowel in *modal* phonation.

The mixing of modified and normally resynthesized utterances was done for calibration purposes.

An on-line test was then conducted in which 10 test participants were asked to listen and classify 15 utterances according to the perceived VQ (modal, rough, creaky or falsetto). As a reference, a table with the original vowels as uttered by the professional speaker in the different modes was included. Each participant was asked to familiarize her- or himself with the examples prior to completing the test, and was allowed to listen to them again during the evaluation. The participants ignored the origin of the utterances.

We have separated the results of the two sets of utterances in two tables. The following table shows the confusion matrix of the different VQ when the evaluators were presented with the resynthesized utterances:

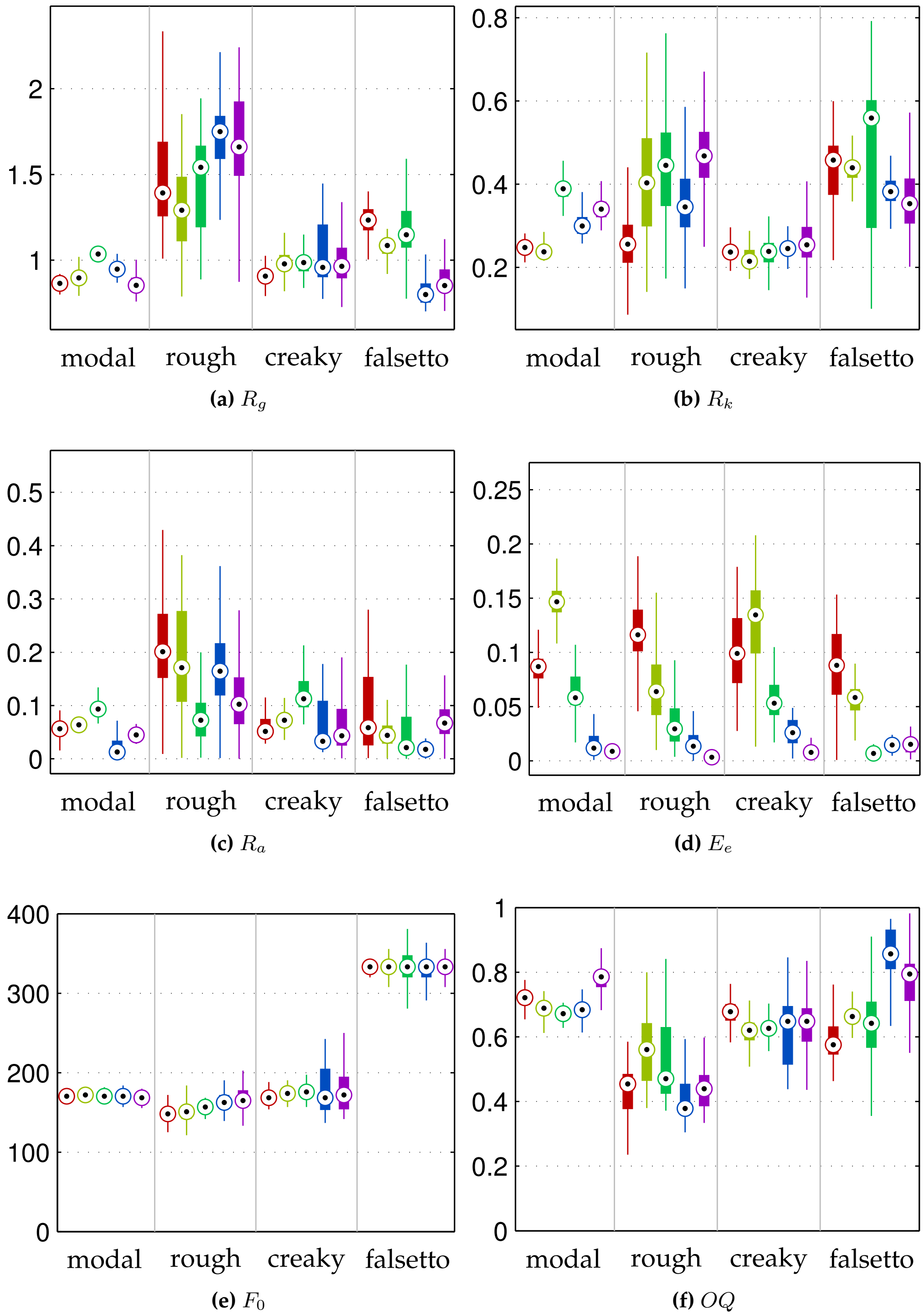
		Selected VQ			
		Modal	Rough	Creaky	Falsetto
Actual VQ	Modal	83.33	16.67	0.00	0.00
	Rough	0.00	90.00	10.00	0.00
	Creaky	0.00	5.00	95.00	0.00
	Falsetto	0.00	5.00	0.00	95.00

These results can be used as reference when analyzing the results of the modified utterances:

		Selected VQ			
		Modal	Rough	Creaky	Falsetto
Actual VQ	Rough	5.00	95.00	0.00	0.00
	Creaky	5.00	10.00	85.00	0.00
	Falsetto	0.00	5.00	0.00	95.00

As we can see, with the resynthesized utterances the confusion between modal and rough is relatively high, and the listeners confused them in almost 17% of the cases. As expected from other reported studies, the confusion between rough and creaky is higher than in the other two cases (rough–falsetto, creaky–falsetto). These two qualities present some similarities in terms of aspiration noise levels and irregularities, that make them sometimes difficult to discern.

As the falsetto mode can be easily discerned, among other things due to the higher pitch involved, we devised a second experiment specifically addressing this mode. The idea was to see whether a simple change of pitch, arguably the single most important characteristic of this mode, would suffice to transmit the same voice quality or if, on the contrary, more glottal information would be needed to convey it. The experiment was implemented using a modified ABX test. The test subjects were presented with the five vowels uttered in falsetto mode. For each of the examples, two options were given: one corresponding to the modal utterance with modified pitch, and the other to the modified utterance used in the previous experiment (glottal source and residual extracted from the falsetto utterance, and vocal tract obtained from the modal one). The test subjects were asked to select which option (A or B) was closest to the original (X). The results of this test show that 98% of the time the test subjects preferred the source-modified utterance over the pitch-modified one. This clearly indicates that modifying the pitch only is not enough to convey the characteristics of the falsetto voice.



(a) R_g

(b) R_k

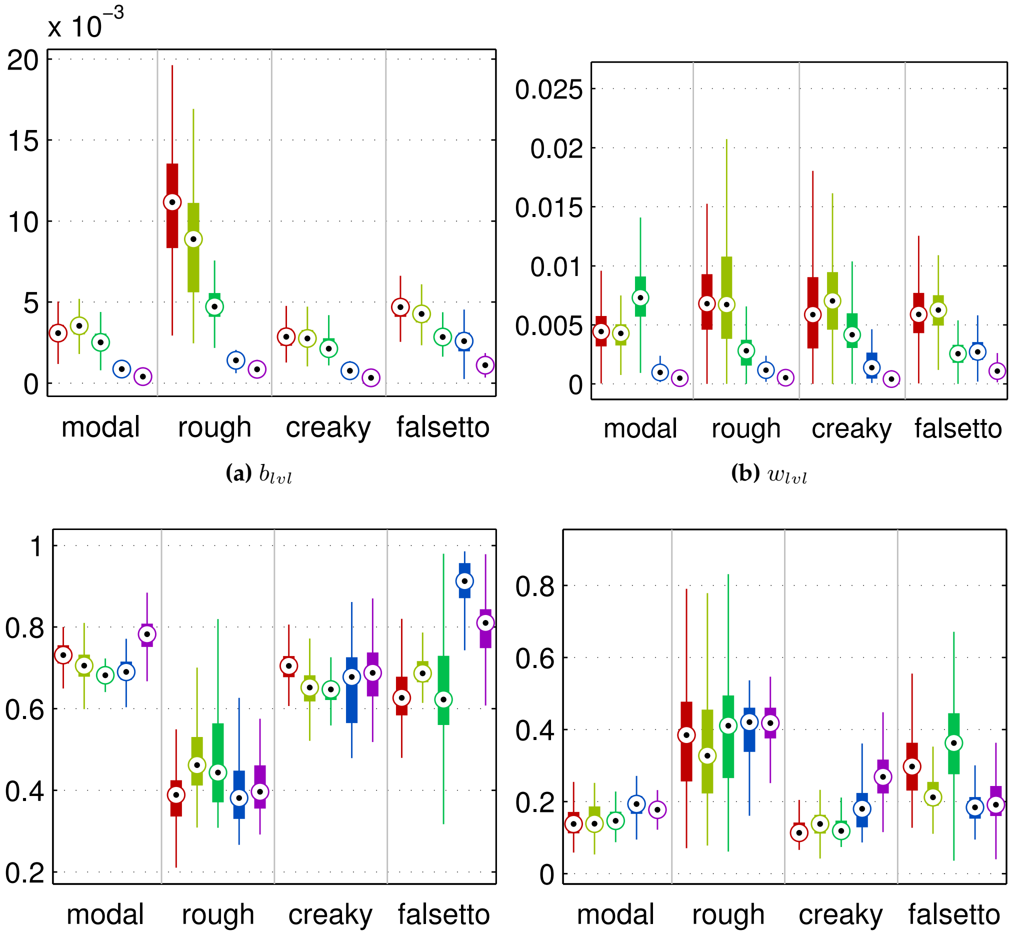
(c) R_a

(d) E_e

(e) F_0

(f) OQ

Figure 7.5: *Boxplot of the LF parameters (R_a, R_g, R_k and E_e), the fundamental frequency (F_0) and the open quotient (OQ) for five isolated vowels uttered with different voice qualities (modal, rough, creaky and falsetto). From left to right: /a/ (red), /e/ (yellow), /i/ (green), /o/ (blue) and /u/ (purple).*



(a) b_{lvl} **(b)** w_{lvl}

(c) w_c **(d)** w_l

Figure 7.6: *Boxplot of the residual parameters for five isolated vowels uttered with different voice quality (modal, low F_0, high F_0 (falsetto), creaky and rough). From left to right: /a/ (red), /e/ (yellow), /i/ (green), /o/ (blue) and /u/ (purple).*

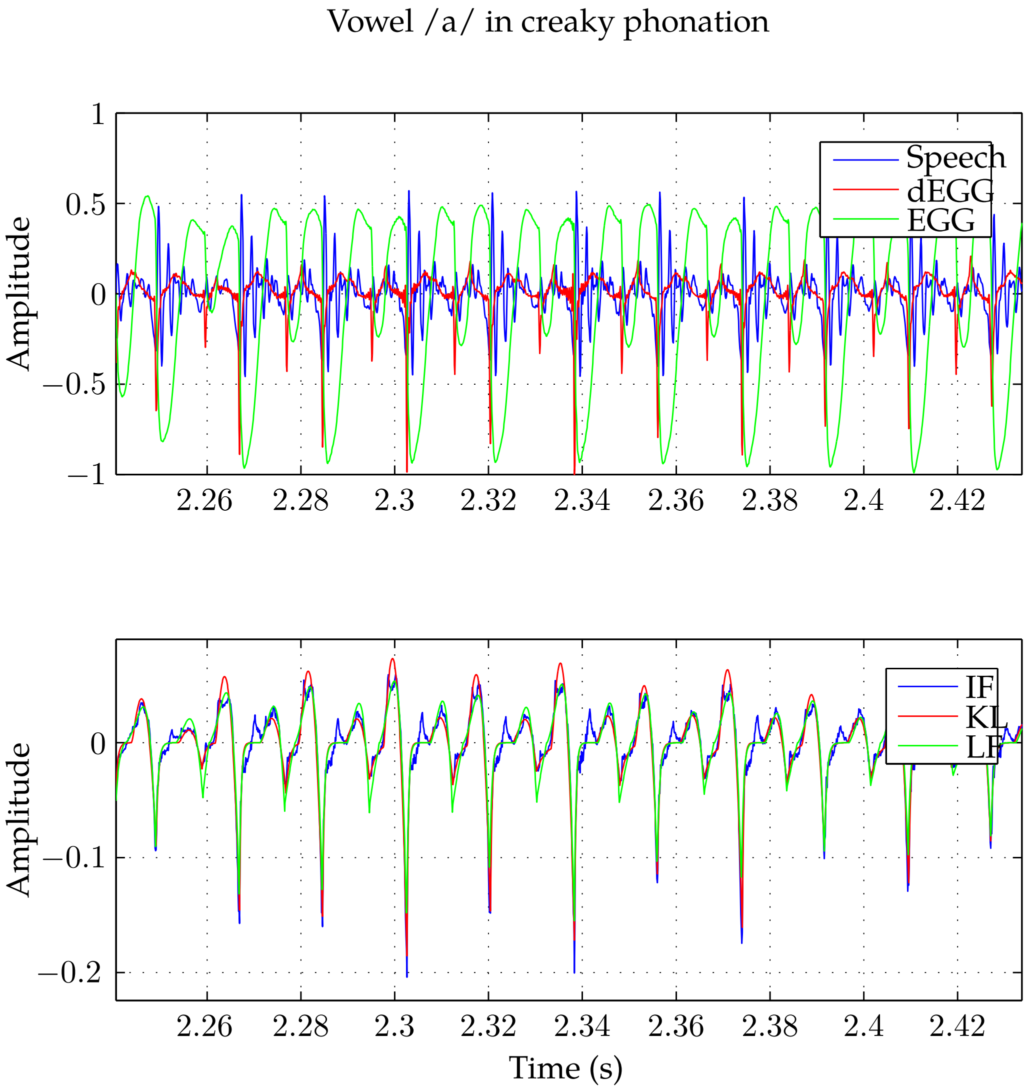


Figure 7.7: *Upper plot: speech, laryngograph (EGG) and differentiated laryngograph (dEGG) waveforms corresponding to the vowel /a/ in creaky phonation. Lower plot: inverse-filtered (IF) and parametrized KLGLOTT88 (KL) and LF glottal waveforms.*

7.3.2 Emotion classification

The recognition of emotions from voice, although a young field in comparison with the more established fields of automatic speech recognition or speaker identification, has been gaining considerable interest in the last decade. Emotion recognition is being researched for a wide spectrum of applications, for instance human-machine interfaces or human-robot communication. The reliability of emotion classification systems is highly dependent on the set of features extracted from the speech signal. Initial efforts focused on speaker-dependent emotion recognition, using a small set of prosodic features to train statistical classifiers using HMM or GMM with good results (Amir, 2001; Jiang and Cai, 2004; Nogueiras et al., 2001; Schuller et al., 2003). When working on speaker-independent emotion recognition using real emotions recorded in a natural environment, as opposed to studio generated (i.e., acted) emotions, hundreds and even thousands of features are used. Current focus is on the selection and reduction of the feature set, and on the combination of the different features (Lee et al., 2011; Luengo et al., 2010; Lugger and Yang, 2007). In general there is no clear consensus on the best set of parameters and how to combine them, since the requirements and definition of what *best* means highly depends on the task at hand. On a similar note, different studies on emotion recognition use different definitions of what constitutes an emotion, making the method comparison difficult and often meaningless. Recent efforts like the 2009 Interspeech Emotion Challenge (Schuller et al., 2009) are aimed at providing a common framework within which different classification methods and parametrizations could be tested and compared.

The relation between voice quality and the transmission of affect or emotion has been long since established by researchers in the field of expressive speech and transmission of affect (Cahn, 1990; Ryan et al., 2003; Schröder, 2001) and some authors have named it "the 4^{th} prosodic feature", the first three being pitch, power and duration (Campbell and Mokhtari, 2003). The differences in speaker affect can be achieved by changing the voice quality, deeming fundamental frequency alone ineffective (Gobl et al., 2002), and although certain voice qualities have been found to be associated to specific affective attributes, no direct one-to-one mapping has been found (Gobl and Chasaide, 2000, 2003b). Burkhardt and Sendlmeier (2000) presented an emotion-synthesizer based on the formant synthesizer *KLSYN88* (Klatt and Klatt, 1990), including specific VQ settings related to the phonation type, used to modify the voicing source. They later conducted perceptual experiments showing that the different types were distinguished by the listeners, leading to an emotional impression agreeing with the reported literature (Burkhardt, 2009). Recent work in emotion recognition has shown that adding voice quality derived features to the set of spectral and prosodic features improves the accuracy of the classification (Lugger and Yang, 2006, 2007; Schuller et al., 2003).

We will present now our work in the field of speaker-dependent emotion classification using voice quality information derived from glottal measures. For these experiments, we used the INTERFACE Emotional Speech Synthesis Database (Tchong et al., 2000), a database designed for the study of emotional speech and the analysis of emotion characteristics for the purpose of speech synthesis. This same database was used in Nogueiras et al. (2001) to conduct preliminary studies on emotion recognition using prosodic features with good results. The database contains recordings in four different languages (English, French, Slovenian and Spanish), two speakers for each language (one female and one male). The speakers were professional actors simulating neutral style and six emotions, as defined in the MPEG-4 standard (anger, disgust, fear, joy, sadness and surprise). The Spanish corpus contains approximately 5000 utterances and includes numbers, isolated words, sentences in affirmative, exclamatory and interrogatory form and paragraphs. The utterances were recorded in two sessions in our recording facilities at the university. The original sampling frequency was 32 kHz and the signals were later downsampled to 16 kHz for the purposes of this work. As usual with our studio recordings, a two channel setup was used to obtain both the speech, using a high-quality electrodynamics microphone AKG 320, and the laryngograph signals.

For the purpose of this work, we have used 450 samples of each of the above emotions to the GMM training, containing a balanced number of samples of both speakers and sessions. We reserved 10% of this training corpus for validation purposes (i.e., choosing the optimal set of model parameters for each emotion and classifier). The remaining utterances of the database (approx. 1900) have been all used for evaluating the performance of the emotion classifiers.

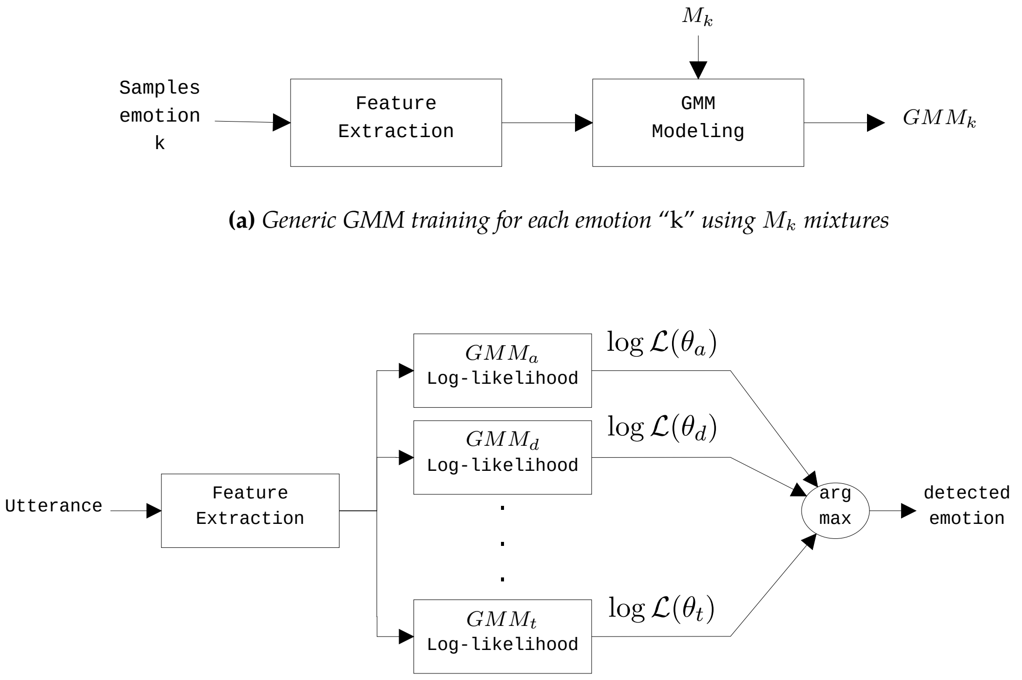
The emotion classifier is built by individually modeling each of the emotions using a GMM trained with features extracted from the speech utterances corresponding to that particular emotion. As we have seen in Chapter 6, Section 6.2, a GMM models a probability density function using a mixture of M Gaussian densities:

$$\begin{aligned} p(\mathbf{x}_n|\theta) &= \sum_{m=1}^{M} w_m \, \mathcal{N}(\mathbf{x}_n|\mu_m, \mathbf{\Sigma}_m) \\ &= \sum_{m=1}^{M} w_m \, \frac{1}{(2\pi)^{d/2}|\mathbf{\Sigma}_m|^{1/2}} \, e^{-\frac{1}{2}(\mathbf{x}_n-\mu_m)^t \mathbf{\Sigma}_m^{-1}(\mathbf{x}_n-\mu_m)}, \end{aligned} \tag{7.5}$$

where $\mathbf{x}_n$ is the feature data, w_m is the weight of each component in the distribution ($\sum_{m=1}^{M} w_m = 1$ and $w_m \geq 0 \ \forall m$), μ_m is the mean vector, $\mathbf{\Sigma}_m$ the covariance matrix and $\theta = \{\theta_1, \ldots, \theta_M\}$, where $\theta_m = [w_m, \mu_m, \mathbf{\Sigma}_m]$, is the complete set of parameters to be estimated. The estimation is performed as before, using the Expectation-Maximization algorithm from eqs.(6.14)–(6.18) so that the the *log–likelihood* of the model parameters given the input data $\mathcal{X} = \{\mathbf{x_1}\,\mathbf{x_2}\,\cdots\,\mathbf{x_N}\}$ is maximized:

$$\log \mathcal{L}(\theta|\mathcal{X}) = \sum_{n=1}^{N} \log p(\mathbf{x}_n|\theta). \tag{7.6}$$

For this work we have opted for a standard GMM classifier in which a GMM per emotion has been trained using all the available training data for each emotion, resulting in a total of 7 GMM. To classify a new utterance, the log-likelihood of each of the 7 GMM given the utterance is computed using eq.(7.6) and the model resulting in the maximum value is selected as detected emotion. The whole procedure is depicted in Figure 7.8: the training stage for a generic emotion denoted with the letter k is illustrated in the upper part, and the classification process at the bottom.



(a) *Generic GMM training for each emotion "k" using M_k mixtures*

(b) *Generic GMM classification of unknown samples*

Figure 7.8: *Block diagram of the training and testing schemes for emotion classification using GMM (only three emotions shown for illustration purposes)*

The described procedure corresponds to a standard generic GMM classification paradigm, independent of the signal parametrization used for the feature extraction. Since our purpose with this work is to evaluate the effect of adding voice quality information in emotion classification, we have built a classification system using both spectral and prosodic features to be used as baseline system. We have then constructed new classifiers extending this baseline parametrization with glottal features, and we have compared their performance with respect to the baseline. For the baseline system, we have used 20 MFCC representing the spectrum and fundamental frequency information ($\log F_0$ and $\Delta \log F_0$). The MFCC parameters are extracted at a constant frame rate of 5 ms, using a Hamming window and frames of 25 ms. The fundamental frequency F_0 is computed independently using our glottal parametrization algorithm as the inverse of the glottal period length.

The idea now is to add glottal information to the baseline classifier and evaluate its performance. For this work, we have used the LF parameters resulting from our analysis to compute the following three glottal measures:

- open quotient $OQ = \frac{T_e}{T_0} = \frac{1+R_k}{2\,R_g}$,
- speed quotient $SQ = \frac{T_p}{T_e - T_p} = \frac{1}{R_k}$,
- basic shape parameter $R_d = \frac{1}{0.11}\ (0.5 + 1.2\,R_k)\left(\frac{R_k}{4\,R_g} + R_a\right)$.

We have then combined these features to train the following set of classifiers:

1. base, baseline classifier using standard mel-cepstral coefficients and fundamental frequency information (F_0 and ΔF_0).
2. base+oq, adding the open quotient (OQ) to the baseline parametrization,
3. base+loq, adding the logarithmic open quotient ($\log OQ$),
4. base+sq, adding the speed quotient (SQ),
5. base+lsq, adding the logarithmic speed quotient ($\log SQ$),
6. base+rd, adding the LF model shape parameter (R_d),
7. base+lrd, adding the logarithmic LF model shape parameter ($\log R_d$),
8. base+oq+sq+rd, adding the combination of OQ, SQ and R_d,
9. base+loq+lsq+lrd, adding the combination of $\log OQ$, $\log SQ$ and $\log R_d$.

To combine the different sets of features we use an early fusion approach, concatenating them in new feature vectors to train a single model. We follow this approach because our classification problem is well-conditioned and of moderate size, using a relatively large database of acted emotional speech, recorded in a controlled environment by two professional actors. More complex classification tasks (e.g., dealing with bigger databases, natural emotions occurring in spontaneous human or man-machine interactions, multimodal recordings or larger number of speakers) would require a different and much larger number of features and accordingly different ways of combining them and reducing the problem dimensionality (Lee et al., 2011; Luengo et al., 2010; Lugger and Yang, 2008). Since the MFCC analysis is performed at a constant rate, using fixed-length frames, there is no direct correspondence between glottal cycles and MFCC frames. In order to add the necessary glottal information, each MFCC frame is projected into the glottal analysis timeline, to find the overlapping glottal frames. Figure 7.9 illustrates this procedure with an example in which MFCC frame m partially overlaps with glottal cycles k, $k+1$ and $k+2$. In this particular case, the glottal feature vector for MFCC frame m, $\hat{\theta}_m$, is constructed as the weighted average of the feature vectors corresponding to glottal cycles k, $k+1$ and $k+2$:

$$\hat{\theta}_m = \frac{\sum_{l=k}^{k+2} \omega_l \, \theta_l}{100}, \tag{7.7}$$

where θ_l is the glottal vector corresponding to the l-th cycle, and the weights ω_l correspond to the percentage of MFCC frame overlapping with the l-th glottal cycle, so that $\sum_{l=k}^{k+2} \omega_l = 100$.

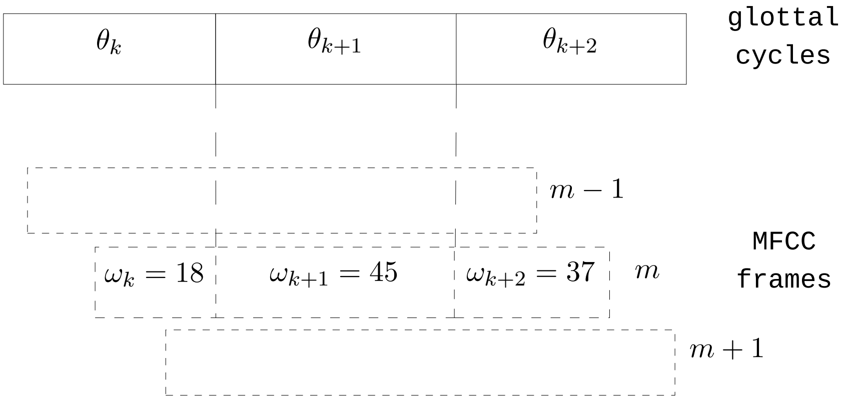


Figure 7.9: *Addition of glottal information to MFCC frame m, using a weighted average of the parameters corresponding to glottal cycles k, $k+1$ and $k+2$, using weights ω_k, ω_{k+1} and ω_{k+2} respectively*

Prior to the evaluation with the test data, the optimal size of the GMM has been individually determined for each system using a subset of the training data kept for validation purposes. Figure 7.10 shows the performance of the GMM classifier using the different parametrizations on the validation corpus, for different number of mixtures in the GMM. For each of the parametrizations, the optimal number of mixtures in the GMM was selected as the one resulting in lower classification error using the validation corpus.

Once the size of each GMM is determined, we proceed to rate the different systems by evaluating its classification performance using a previously unseen corpus of test data. Figure 7.11 contains the results of the evaluation for all the systems. Figure 7.11a shows

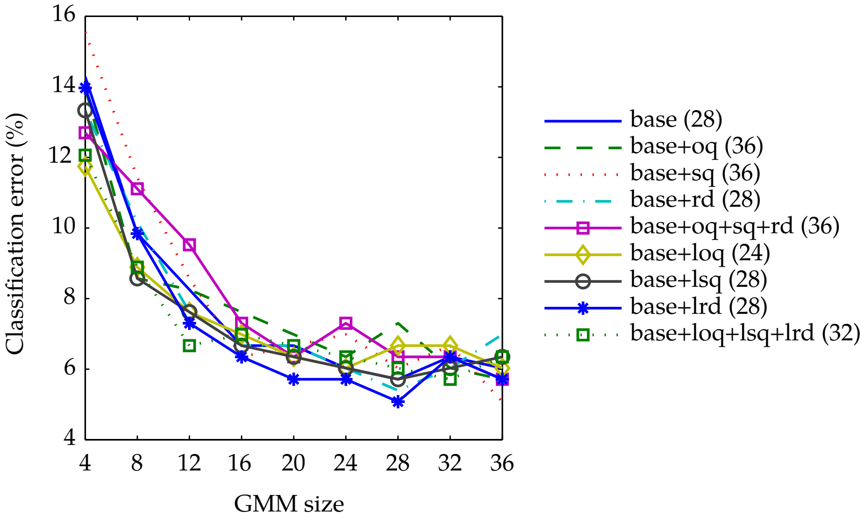
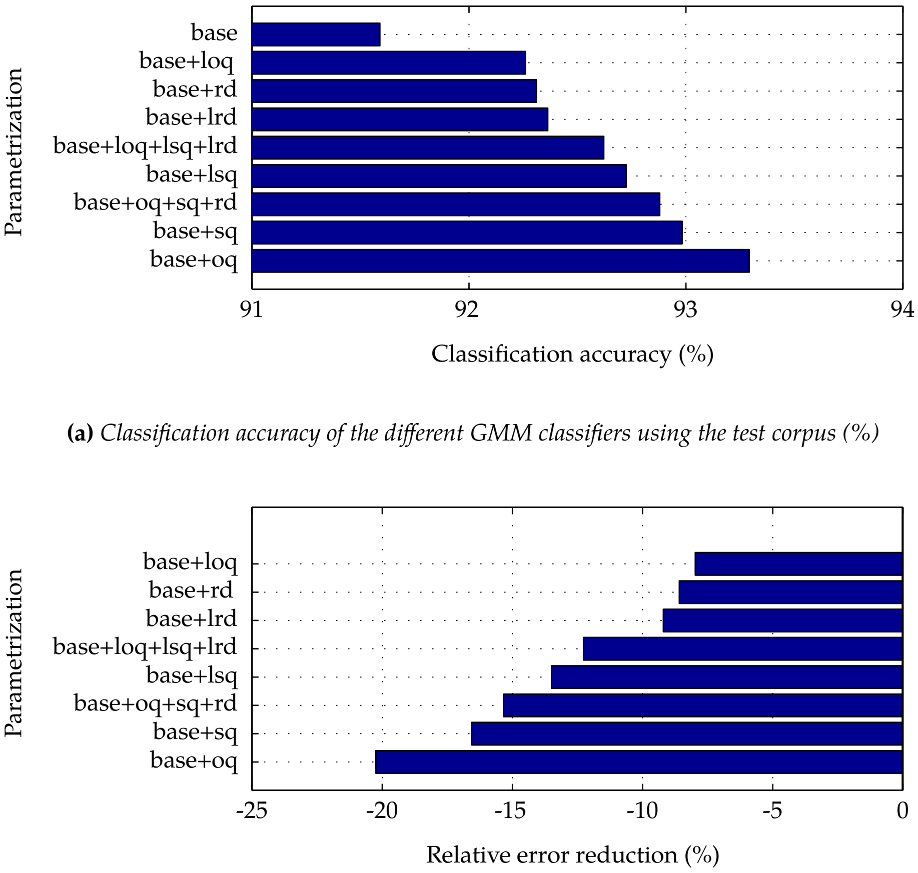


Figure 7.10: *Emotion classification error for the validation dataset of the GMM systems for a different number of mixtures. base: baseline system (mel-cepstral coefficients plus pitch information,* $\log F_0$ *and* $\Delta \log F_0$*), oq: open quotient OQ (loq:* $\log OQ$*), sq: speed quotient SQ (lsq:* $\log SQ$*), rd: LF shape parameter* R_d *(lrd:* $\log R_d$*). In parenthesis: optimal number of mixtures in the GMM.*

the overall accuracy of the different systems ordered from worst (top) to best (bottom). In figure 7.11b we have the relative reduction of the error with respect to the the baseline system (0 in the x axis). As we can see, adding glottal information to the baseline parametrization improves the performance of the classification. The best classification system is the one using the open quotient OQ in linear form. It achieves an overall accuracy of 93.2% and a relative error reduction of more than 20% w.r.t. the baseline system. As we can see, using the log-scaled versions of the VQ parameters does not have a clear advantage, with linear parameters outperforming them except in the R_d case, where it slightly improves the accuracy, although the difference is small. The combination of all three parameters results in a good overall accuracy and relative error reduction, but the classifier is outperformed both by the OQ and SQ alone.

Table 7.1 shows the test results categorized by system and emotion. As we can see, most of the systems follow the same pattern. Sadness, surprise and the neutral state are correctly classified in more than 95% of the cases by all systems. The emotion most systems struggle to identify is disgust, which gets a correct decision 86.5% of the time on average. Anger and joy follow in terms of difficulty, with an average of 88.5% across all the systems, and fear achieves an average of 90% accuracy. As we can see, adding voice quality related parameters has a larger impact on the classification of disgust, joy and surprise.

It is interesting to see which emotions are more often mistaken by which ones. For



(a) *Classification accuracy of the different GMM classifiers using the test corpus (%)*

(b) *Relative classification error reduction with respect to the baseline system (0)*

Figure 7.11: *Ranking of the different GMM parametrizations using the optimal GMM size, evaluated on the test dataset.*

Table 7.1: *Correct classification rate of the different systems for each emotion*

		Emotion						
		Anger	Disgust	Fear	Joy	Surprise	Sadness	Neutral
Parametrization	base	89.05	85.46	89.79	86.59	95.64	97.37	97.51
	base+loq	87.59	87.94	89.08	89.13	98.18	97.74	96.44
	base+rd	89.42	84.40	90.14	89.86	97.45	97.37	97.86
	base+lrd	87.59	88.65	89.44	87.32	96.73	98.87	98.22
	base+loq+lsq+lrd	85.77	87.23	92.25	88.77	98.55	98.87	97.15
	base+lsq	88.32	86.52	91.20	88.41	98.91	98.50	97.51
	base+oq+sq+rd	89.05	86.17	91.90	89.49	98.91	98.50	96.44
	base+sq	90.51	85.11	91.90	87.68	99.64	97.37	98.93
	base+oq	88.69	88.65	90.85	90.94	98.55	98.50	97.15

Table 7.2: *Confussion matrix for the baseline classification system (in percentage)*

		Detected emotion						
		Anger	Disgust	Fear	Joy	Surprise	Sadness	Neutral
Actual emotion	Anger	**89.05**	0.73	0.00	7.30	2.55	0.36	0.00
	Disgust	0.71	**85.46**	4.61	3.19	1.77	3.55	0.71
	Fear	0.00	1.06	**89.79**	1.76	6.34	1.06	0.00
	Joy	2.17	1.09	0.72	**86.59**	8.70	0.36	0.36
	Surprise	0.00	0.00	0.36	4.00	**95.64**	0.00	0.00
	Sadness	0.00	0.00	2.63	0.00	0.00	**97.37**	0.00
	Neutral	0.00	0.00	0.00	0.00	0.00	2.49	**97.51**

Table 7.3: *Confussion matrix for the base+oq (best) classification system (in percentage)*

		Detected emotion						
		Anger	Disgust	Fear	Joy	Surprise	Sadness	Neutral
Actual emotion	Anger	**88.69**	1.46	0.00	7.30	2.55	0.00	0.00
	Disgust	0.00	**88.65**	4.61	1.77	1.42	2.84	0.71
	Fear	0.00	0.70	**90.85**	1.06	6.34	1.06	0.00
	Joy	0.00	0.00	1.09	**90.94**	7.25	0.36	0.36
	Surprise	0.00	0.00	0.36	1.09	**98.55**	0.00	0.00
	Sadness	0.00	0.00	1.50	0.00	0.00	**98.50**	0.00
	Neutral	0.00	1.07	0.00	0.00	0.00	1.78	**97.15**

Table 7.4: *Confussion matrix for the subjective evaluation presented in Nogueiras et al. (2001) (in percentage)*

		Perceived emotion						
		Anger	Disgust	Fear	Joy	Surprise	Sadness	Neutral
Actual emotion	Anger	**66.41**	3.91	1.56	10.94	3.91	1.56	11.72
	Disgust	1.56	**82.81**	3.91	0.78	2.34	1.56	7.03
	Fear	0.78	3.91	**80.47**	0.78	10.16	3.12	0.78
	Joy	5.47	1.56	0.00	**89.84**	1.56	0.00	1.56
	Surprise	0.78	2.34	12.50	2.34	**78.91**	0.78	2.34
	Sadness	5.47	4.69	0.00	15.62	1.56	**69.53**	3.12
	Neutral	1.56	3.12	0.78	1.56	0.78	0.00	**92.19**

this reason, we have computed the confusion matrices for the baseline and optimal systems, presented in tables 7.2 and 7.3 respectively. Table 7.4 summarizes the results of the emotion informal subjective evaluation carried on in Nogueiras et al. (2001) during their experiments with HMM and prosodic features for emotion recognition. As we can see, the listeners could correctly identify with high accuracy the utterances belonging to either the neutral state or the joy emotion. The majority of the problems occurred with anger, mostly misclassified as either joy or neutral, and sadness, mostly mistakenly recognized as joy or, in fewer cases, as either anger or disgust. The baseline classifier performs very well with the surprise-sadness-neutral set, achieving accuracies over 95% for all of them. This is similar to our best proposed classifier, which also has high classification accuracy for surprise, sadness and neutral, outperforming the baseline system in the first two cases by approximately 3% and 1.2% in absolute accuracy. Most of the problems are with anger, often misclassified as joy, and disgust, labeled as fear in less than 5% of the cases, although in both cases the accuracy is very good (close to 89%). The baseline system performs worse in this case (disgust), distributing the classification mistakes between the fear-joy-sadness trio. The other two emotions, fear and joy, are often labeled as belonging to the surprise class by both the baseline and proposed systems.

7.4 Conclusions

In this section we have presented our study of the voice source characteristics associated to the different laryngeal voice qualities (modal, rough, creaky and falsetto) and its relation to different emotions (anger, disgust, fear, joy, surprise and sadness). The chapter has been divided in two main parts: voice quality analysis and emotion recognition.

We have first analyzed a small corpus of sustained vowels recorded using a female professional speaker in Spanish uttered with different phonations, and analyzed each of the source parameters (LF and residual) by means of boxplots. As we have seen, the results mostly correlate with those previously reported in the literature, with some differences that may be attributed to the reduced size of the speech corpus and the fact that contains utterances of only one speaker. We have then presented an online evaluation designed to asses whether the source parametrization algorithm was successfully capturing the laryngeal characteristics of the different phonation types. As we have seen, the results of the test are very satisfactory and show that our analysis/synthesis algorithm succeeds in the task of capturing the different VQ.

We have then evaluated the performance of an automatic emotion classifier using glottal measures. The classification is performed by statistical GMM classifiers trained for each emotion using different features. We have compared our parametrization to a baseline system using spectral (MFCC) and prosody (F_0 and $\log F_0$) characteristics. The results of the test using an emotional database of almost 5000 utterances and two speakers were very satisfactory, showing a relative error reduction of more than 20% with respect to the baseline system. The accuracy of the different emotions detection was also high, improving the results of previously reported works using the same database. Overall, we can conclude that the glottal source parameters extracted using our algorithm have a positive impact in the field of automatic emotion classification.

CHAPTER 8

Conclusions and further work

The objective of this dissertation was to study and develop techniques to decompose the speech signal into its two main components: voice source and vocal tract. We wanted to explore the utility of this model in different areas of speech processing: speech synthesis, voice conversion or emotion detection among others. Thus, we have studied different techniques for prosodic and spectral manipulation. Our efforts have been focused on the automatic extraction and parametrization of the glottal source from high-quality speech databases used in the fields of speech synthesis and voice conversion. In the next section the main conclusions from the studies carried out in the previous chapters are presented. To end the dissertation, we will outline some lines of investigation that continue the work presented in this dissertation.

8.1 Conclusions

The focus of this thesis was the extraction of the glottal source information. We used a speech production model in which the glottal flow produced by the vibrating vocal folds goes through the vocal (and nasal) tract cavities and its radiated by the lips. According to this model, a source-filter decomposition should be possible, in which the two main contributions (from the glottis and the vocal tract) could be independently analyzed and modified.

Most of our efforts were focused on the analysis and synthesis algorithm, a key element on the posterior applications. The analysis methodology and algorithms have been detailed in Chapter 3. Traditional inverse filtering approaches based on closed-phase linear prediction have the problem of having to work only with samples corresponding to the glottal closed phase, which in some cases can be quite short. This results in a large amount of noise present in the estimated inverse-filtered waveforms, which poses further problems when parameterizing them. We overcome this problem by using a parametric model for the glottal waveform and thus including the glottal open phase in the estimation. As a result of the source-filter decomposition, we not only obtain a better (i.e., less noisy) inverse-filtered glottal estimation, but also a first parametrization using the KLGLOTT88 model. This parametrization is then used to initialize the LF estimation, a crucial step due to the non-linear nature of the required optimization using least-squares, that should otherwise need to be done by performing measures on the noisy glottal estimation. A parametric model for the residual comprising the aspiration noise was also proposed as part of the parametrization, and was estimated from the glottal residual after the LF modeling. The original residual was first filtered using a high-pass whitening filter to eliminate some artifacts, its envelope was extracted using the Hilbert transform and then it was parametrized using our proposed function for the synthetic envelope.

As part of this work, we recorded a small database consisting on the five Spanish vowels uttered in isolation and sustained for approximately 1.5–2 seconds, in order to study the stable part of the utterance. The vowels were recorded in four different phonations (modal, rough, creaky and falsetto) and were later also used in our study of voice quality. In order to validate the accuracy of the parametrization algorithm, we designed a synthetic corpus using LF glottal parameters reported in the literature, complemented with our own results from the vowel database. Synthetic was generated using the glottal information, synthetic noise added at different SNR and a vocal tract constructed with known formant central frequencies and bandwidths, and was subsequently analyzed using our source-filter decomposition algorithm. Since the parameters used in synthesis were known a priori, they were used as reference to compute the parametrization error with the estimated parameters. The results show that our method gives satisfactory results in a wide range of glottal configurations and at different levels of SNR.

We also conducted an on-line evaluation in which the quality of the resynthesized speech was rated by a group of listeners by means of a MOS test. We proposed two methods for this evaluation: one consisting of the parametrized residual explained before, and another using the whitened residual waveform. A third method (STRAIGHT) was included as reference. Our method using the whitened residual compared favorably to this reference, achieving high quality ratings (Good-Excellent). Our full parametrized system scored lower than the other two ranking in third place, but still higher than the acceptance threshold (Fair-Good). We are quite satisfied with these results, since both parametrizations are suited for different applications.

Next we proposed two methods for prosody modification, one for each of the residual representations explained above. The first method used our full parametrization system and frame interpolation to perform the desired changes in pitch and duration. Since the fundamental frequency is already part of this representation, changing the pitch involved only updating the corresponding parameter. Frame interpolation was then used to obtain the desired duration. The second method used the residual waveform and a frame selection technique to generate a new sequence of frames to be synthesized. The selected residual waveforms were then resampled to obtain the target pitch duration. Both options were again evaluated, using two standard algorithms as reference (STRAIGHT as in the resynthesis test, and the PSOLA-like algorithm as implemented in our speech synthesizer Ogmios). The results showed that the resampling method outperformed the parametrization one, scoring very similarly to the two reference methods. Our full parametrized system scored lower than the other three, but still higher than the acceptance threshold (Fair-Good).

Our speech production model was incorporated into an existing voice conversion (VC) system to evaluate the specific impact of our proposed parametrization on the conversion performance. The system used for VC uses CART to incorporate phonetic features into the conversion model, dividing the data into phonetic classes prior to the conversion function training, which is performed by means of GMM. This conversion system was developed in our group as part of the TC-STAR project, and it has participated in several evaluation campaigns. In order to obtain meaningful comparisons of the results, the same testing conditions were replicated and used with our parametrization model. We used the full parametrized residual for the voice conversion experiments. Three independent CART were trained, one for each of the production model parameter sets (glottal pulse, vocal tract and aspiration noise or residual). The optimal configuration of each CART (phonetic features, number of nodes, size of the GMM) was individually determined using a small subset of the training dataset kept for validation purposes. Once all the models were trained, voice conversion was performed on the test utterances, and the resulting waveforms were compared with those obtained with the original VC system. The comparison was performed using an online MOS evaluation, in which both the speaker similarity and the utterance quality were rated. The results showed that the evaluators preferred our method over the original one, rating it with a higher score in the MOS scale in both quality and similarity. Nevertheless, the quality of both methods needs to be improved, since the ratings were still relatively low. Further research is needed in this regard.

As part of this dissertation, we conducted some research in the field of voice quality analysis and identification. First the main studies and methodologies in this field were reviewed and the more relevant findings were presented, concentrating on the importance of the different glottal pulse parameters in each voice quality. For this purpose, we recorded a small database consisting of the five Spanish vowels, uttered in isolation and sustained for two to three seconds each. The database contains samples of four different voice qualities (modal or normal, rough, creaky and falsetto). Each of them was analyzed using our decomposition and parametrization algorithm, and boxplot of the glottal and residual parameters were produced. The LF parameters were compared with those reported in the literature, and we found them to generally agree with previous findings. Some differences existed, but they could be attributed to the difficulties in comparing voice qualities produced by different speakers. We also conducted an experiment aimed at assessing whether our glottal parametrization algorithm was able to capture all the information related to the laryngeal voice quality. For each of the non-modal VQ (rough, creaky and falsetto), each of the five vowels analyzed in the previous section were resynthesized using the extracted glottal LF and residual parameters, but using the vocal tract filter computed in the modal case. The filter frames were interpolated or discarded in order to obtain the appropriate number for each mode. An on-line test was then conducted in which 10 test participants were asked to listen and classify 15 utterances according to the perceived VQ (modal, rough, creaky or falsetto). As a reference, a table with the original vowels as uttered by the professional speaker in the different modes was included. Each participant was asked to familiarize her- or himself with the examples prior to completing the test, and was allowed to listen to them again during the evaluation. The participants were able to correctly identify each of the voice qualities in the majority of the cases, thus validating our glottal parametrization algorithm. The results showed that the different voice qualities were. As expected from other reported studies, the confusion between rough and creaky is higher than in the other two cases (rough–falsetto, creaky–falsetto). These two qualities present some similarities in terms of aspiration noise levels and irregularities, that make them sometimes difficult to discern.

We have also evaluated the performance of an automatic emotion classifier using voice quality related glottal measures. The classification was performed using statistical GMM models trained for each emotion using different features. We the experiments on emotion recognition we have used the Spanish subset of the INTERFACE Emotional Speech Synthesis Database, consisting of almost 5000 utterances uttered by two professional speakers, who simulated the six emotions defined in the MPEG-4 standard (anger, disgust, fear, joy, sadness and surprise) plus the neutral state. We have compared our parametrization to a baseline system using spectral (MFCC) and prosody (F_0 and $\log F_0$) characteristics. We tried different features, both individually and combined (open quotient, speed quotient and LF basic shape parameter). The optimal size of each GMM was individually determined for each feature and emotion using a subset of the training corpus kept apart for validation purposes. The results of the test were very satisfactory, showing a relative error reduction of more than 20% with respect to the baseline system. The accuracy of the different emotions detection was also high, improving the results of previously reported works using prosodic features on the same database. Overall, we can conclude that the glottal source parameters extracted using our algorithm have a positive impact in the field of automatic emotion classification.

8.2 Further work

Improvements of the parametrization algorithm One of the main causes of the degradation of the synthetic speech is the relatively simple model used to represent the aspiration noise present in the original signal. A different paradigm should be followed to improve the quality of the resulting speech, perhaps using codebooks of prototypical residual waveforms (Kim, 2003) or applying some of the residual selection and smoothing techniques used in voice conversion (Duxans, 2006). The algorithm should also be extended to work with speech-only databases, since this would allow its use on a wider range applications. The main problem is the location of glottal epochs, although the same epoch optimization procedure could be combined with initial estimation procedures using the speech signal alone (Drugman et al., 2009a).

Voice conversion Our work in the field of voice conversion (VC) was aimed at studying the performance of our parametrization model compared to traditional Linear Prediction (LP) models. The results of the listening tests show that the model works, outperforming traditional LP models. However, the VC system used for the comparison was based on GMMs, a relatively successful technique with some drawbacks, like the oversmoothing of the converted spectrum (Stylianou et al., 1998; Toda et al., 2001b). Newer conversion paradigms exist that overcome these problems (Qiao and Minematsu, 2009; Toda et al., 2007; Villavicencio et al., 2009; Yutani et al., 2009), and they could be extended with our parametrization paradigm. Furthermore, our prosody modification experiments show that the resampled residual performs better than its parametrized counterpart. New approaches based on residual selection techniques could be adopted to incorporate the former method into VC.

Expressive speech The field of expressive speech is a young and promising one, and is expected to play a key role in the near future. Traditional speech synthesis paradigms do not cope well with the incorporation of emotion and its associated higher variability in terms of prosody and voice quality (VQ). The relation between VQ and the transmission of affect has been proved in several studies (Campbell and Mokhtari, 2003; Gobl et al., 2002; Gobl and Chasaide, 2000). Our work on emotion recognition shows promising results for the inclusion of glottal features in automatic emotion classifiers. Given the unsupervised nature of our source-filter decomposition and parametrization algorithms, they could be applied to the analysis of larger emotional speech databases. Rules could then be extracted linking the variability of the source parameters to the corresponding degree of emotion.

Work in this area already shows promising results (Burkhardt, 2009; Mokhtari et al., 2003; Ryan et al., 2003).

HMM-based speech synthesis HMM-based speech synthesis is statistical parametric method based on hidden Markov models. HMMs are used to simultaneously model the fundamental frequency, the vocal tract and the duration of speech. A maximum likelihood criterion is used to synthesize the speech waveforms (Tokuda et al., 2000). The voice source is usually only represented with the fundamental frequency, and current work focused on extending this paradigm using better models shows promising results (Cabral et al., 2007; Lanchantin et al., 2010; Raitio et al., 2011). Our group is currently working on integrating HTS into our speech synthesizer Ogmios (Bonafonte et al., 2006a). A continuation work would be to replace the standard source parametrization with our improved model.

Bibliography

Abe M (1991). A segment-based approach to voice conversion. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 765–768. Salt Lake City, Utah, USA. 102

Airas M and Alku P (2007). Comparison of multiple voice source parameters in different phonation types. In *Proc. of INTERSPEECH*, 1410–1413. Antwerpen, Belgium. 137

Akande O O and Murphy P J (2005). Estimation of the vocal tract transfer function with application to glottal wave analysis. *Speech Communication*, 46:15–36. 17

Alku P (1992). Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering. *Speech Communication*, 11:109–118. 17

Alku P (2003). Parameterisation methods of the glottal flow estimated by inverse filtering. In *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, 81–87. Geneva, Switzerland. 136

Alku P, Airas M, Bäckström T and Pulakka H (2005). Group delay function as a means to assess quality of glottal inverse filtering. In *Proc. the of European Conference on Speech Communication and Technology*, 1053–1056. Lisbon, Portugal. 22, 82

Alku P, Airas M and Story B (2004). Evaluation of an inverse filtering technique using physical modeling of voice production. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 497–500. Jeju Island, Korea. 17, 21

Alku P, Bäckström T and Vilkman E (2002). Normalized amplitude quotient for parametrization of the glottal flow. *Journal of the Acoustical Society of America*, 112(2):701–710. 136, 137

Alku P, Strik H and Vilkman E (1997). Parabolic spectral parameter – A new method for quantification of the glottal flow. *Speech Communication*, 22:67–79. 138

Alku P and Vilkman E (1996). Amplitude domain quotient for characterization of the glottal volume velocity waveform estimated by inverse filtering. *Speech Communication*, 18:131–138. 135, 136, 137

Amir N (2001). Classifying emotions in speech: a comparison of methods. In *Proc. the of European Conference on Speech Communication and Technology*, 127–131. Aalborg, Denmark. 147

Arroabarren I and Carlosena A (2003a). Glottal source parameterization: a comparative study. In *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, 29–34. Geneva, Switzerland. 139

Arroabarren I and Carlosena A (2003b). Glottal spectrum based inverse filtering. In *Proc. the of European Conference on Speech Communication and Technology*, 57–60. Geneva, Switzerland. 20, 135

Arroabarren I and Carlosena A (2003c). Unified analysis of glottal source spectrum. In *Proc. the of European Conference on Speech Communication and Technology*, 1761–1764. Geneva, Switzerland. 9

Bäckström T, Airas M, Lehto L and Alku P (2005). Objective quality measures for glottal inverse filtering of speech pressure signals. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 1, 897–900. Philadelphia, Pennsylvania, USA. 22, 23, 81

Bonafonte A, Agüero P D, Adell J, Pérez J and Moreno A (2006a). Ogmios: The upc text-to-speech synthesis system for spoken translation. In *TC-Star Workshop on Speech to Speech Translation TC-STAR 2006*. Barcelona,Spain. 96, 162

Bonafonte A, Höge H, Kiss I, Moreno A, Ziegenhain U, van den Heuvel H, Hain H, Wang X S and Garcia M N (2006b). TC-STAR: Specifications of language resources and evaluation for speech synthesis. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 311–314. Genoa, Italy. 83, 96, 112

Bouzid A and Ellouze N (2009). Voice source parameter measurement based on multiscale analysis of electroglottographic signal. *Speech Communication*, 51(9):782 – 792. Special issue on non-linear and conventional speech processing - NOLISP 2007. 38

Boyd S and Vandenberghe L (2004). *Convex Optimization*. Cambridge University Press. 31

Bozkurt B, Doval B, D'Alessandro C and Dutoit T (2005). Zeros of Z-Transform (ZZT) representation with application to source-filter separation in speech. *IEEE Signal Processing Letters*, 12(4):344–347. 20, 21

Bozkurt B and Dutoit T (2003). Mixed-phase speech modeling and formant estimation, using differential phase spectrums. In *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, 21–24. Geneva, Switzerland. 20

Burkhardt F (2009). Rule-based voice quality variation with formant synthesis. In *Interspeech*, 2659–2662. 147, 162

Burkhardt F and Sendlmeier W F (2000). Verification of acoustical correlates of emotional speech using formant-synthesis. In *Proc. of the ISCA Workshop Speech and Emotion*. Belfast. 2, 147

Cabral J P, Renals S, Richmond K and Yamagishi J (2007). Towards an improved modeling of the glottal source in statistical parametric speech synthesis. In *Proc. of the 6th ISCA Workshop on Speech Synthesis*. Bonn, Germany. 2, 17, 162

Cahn J (1990). The generation of affect in synthesized speech. *J Am Voice I/O Soc*, 8. 147

Campbell N (2002). Recording techniques for capturing natural, every-day speech. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 2029–2032. Las Palmas, Spain. 138

Campbell N and Mokhtari P (2003). Voice quality: the 4th prosodic dimension. In M Solé, D Recasens and J Romero, eds., *Proc. of the International Congress of Phonetic Sciences*, 2417–2420. Barcelona. 137, 147, 161

Ceyssens T, Verhelst W and PWambacq (2002). On the construction of a pitch conversion system. In *European Signal Processing Conference*. 102

Charpentier F and Moulines E (1988). Text-to-speech algorithms based on FFT synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 667–670. New York, New York, USA: IEEE. 2

Chen Y, Chu M, Chang E, Liu J and Liu R (2003). Voice conversion with smoothed GMM and MAP adaptation. In *European Conference on Speech Communication and Technology*. 102

Childers D G and Ahn C (1995). Modeling the glottal volume-velocity waveform for three voice types. *Journal of the Acoustical Society of America*, 97(1):505–518. 2, 12, 20, 25, 103

Childers D G and Hu H T (1994). Speech synthesis by glottal excited linear prediction. *Journal of the Acoustical Society of America*, 96(4):2026–2036. 2, 138

Childers D G and Lee C K (1991). Vocal quality factors: Analysis, synthesis, and perception. *Journal of the Acoustical Society of America*, 90(5):2394–2410. 17, 20, 25, 29, 34, 36, 73, 80, 135, 138, 140, 141

Coleman T F and Li Y (1996). An interior trust region approach for nonlinear minimization subjet to bounds. *SIAM J Optimization*, 6(2):418–445. 19, 58

Cook P R (1991). *Synthesis of the Singing Voice Using a Waveguide Articulatory Vocal Tract Model*. Ph.D. thesis, Stanford University. 24

del Pozo A (2008). *Voice Source and Duration Modelling for Voice Conversion and Speech Repair*. Ph.D. thesis, University of Cambridge.

del Pozo A and Young S (2008). The linear transformation of LF glottal waveforms for voice conversion. In *Proc. of INTERSPEECH*, 1457–1460. Brisbane, Australia. 12, 72, 103

Ding W and Campbell N (1997). Optimising unit selection with voice source and formants in the CHATR speech synthesis system. In *Proc. the of European Conference on Speech Communication and Technology*, 537–540. Rhodes, Greece. 2, 12, 18, 19

Ding W and Campbell N (1998). Determining polarity of speech signals based on gradient of spurious glottal waveforms. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 857–860. Seattle, Washington, USA. 38

Ding W, Campbell N, Higuchi N and Kasuya H (1997). Fast and robust joint estimation of vocal tract and voice source parameters. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1291–1294. Munich, Germany. 18

Doval B and d'Alessandro C (1997). Spectral correlates of glottal waveform models: an analytic study. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1295–1298. Munich, Germany. 9, 10, 15

Doval B and d'Alessandro C (1999). The spectrum of glottal flow models. Tech. rep., Notes et Documents LIMSI 99-07, 22p. 9

Doval B and d'Alessandro C (2006). The spectrum of glottal flow models. *Acta Acustica united with Acustica*, 92:1026–1046. 9

Doval B, d'Alessandro C and Henrich N (2003). The voice source as a causal/anticausal linear filter. In *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, 15–20. Geneva, Switzerland. 20

Drugman T, Bozkurt B and Dutoit T (2009a). Complex cepstrum-based decomposition of speech for glottal source estimation. In *Proc. of INTERSPEECH*. Brighton, United Kingdom. 20, 21, 161

Drugman T, Bozkurt B and Dutoit T (2012). A comparative study of glottal source estimation techniques. *Computer Speech & Language*, 26(1):20 – 34. 20, 21

Drugman T, Dubuisson T and Dutoit T (2009b). On the mutual information between source and filter contributions for voice pathology detection. In *Proc. of INTERSPEECH*, 1463–1466. Brighton, United Kingdom. 2

Duda R, Hart P and Stork D (2000). *Pattern Classfication*. John Wiley & Sons, 2nd edn. 26, 105, 106, 107

Dutoit T (1997). *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Publishers. 2

Dutoit T and Leich H (1992). Improving the TD-PSOLA text-to-speech synthesizer with a specially designed mbe re-synthesis of the segments database. *EUSIPCO 92, Proceedings of*, I:343–346. 2

Dutoit T and Leich H (1993). MBR-PSOLA: Text-to-speech synthesis based on an MBE resynthesis of the segments database. *Speech Communication*, 13(34):167–184. 2

Duxans H (2006). *Voice Conversion applied to Text-to-Speech systems*. Ph.D. thesis, Universitat Politècnica de Catalunya. 101, 102, 103, 111, 112, 161

Duxans H, Erro D, Pérez J, Diego F, Bonafonte A and Moreno A (2006). Voice conversion of non-aligned data using unit selection. In *Proceedings of the TC-Star Workshop on Speech to Speech Translation*. Barcelona, Spain. 101

Edwards J A and Angus J A S (1996). Using phase-plane plots to assess glottal inverse filtering. *Electronic Letters*, 32(3):192–193. 22

El-Jaroudi A and Makhoul J (1991). Discrete all-pole modeling. *IEEE Transactions on Signal Processing*, 39(2):411–423. 17

Erro D (2008). *Intra-lingual and Cross-lingual Voice Conversion Using Harmonic Plus Stochastic Models*. Ph.D. thesis, Universitat Politècnica de Catalunya. 102

Erro D, Moreno A and Bonafonte A (2010a). INCA algorithm for training voice conversion systems from non-parallel corpora. *IEEE Transactions on Audio, Speech and Language Processing*, 18(5):944–953. 102

Erro D, Moreno A and Bonafonte A (2010b). Voice conversion system based on weighted frequency warping. *IEEE Transactions on Audio, Speech and Language Processing*, 18(5):922–931. 102

Eysholdt U, Tigges M, Wittenberg T and Pröschel U (1996). Direct evaluation of high-speed recordings of vocal fold vibrations. *Folia Phoniatrica et Logopaedica*, 48:163–170. 15

Fant G (1970). *Acoustic theory of speech production*. The Hague: Mouton, 2nd edn. 2, 5

Fant G (1995). The LF-model revisited. transformations and frequency domain analysis. Tech. Rep. 2-3:119–1, STL-QPSR. 12, 13, 15, 136, 137

Fant G (1997). The voice source in speech communication. *Speech Communication*, 22:125–139. 15, 136, 137

Fant G and Kruckenberg A (1996). Voice source properties of the speech code. *TMH-QPSR*, 4:45–56. 137

Fant G, Kruckenberg A, Liljencrants J and Båvegård M (1994). Voice source parameters in continuous speech. transformation of lf-parameters. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 1451–1454. Yokohama, Japan. 136

Fant G, Liljencrants J and Lin Q (1985). A four-parameter model of glottal flow. *STL-QPSR*, 4:1–13. 7, 8, 138

Fant G and Lin Q (1998). Frequency domaing interpretation and derivation of glottal flow parameters. *STL-QPSR*, 2–3. 20, 135

Fourcin A (2000). Voice quality and electrolaryngography. In R D Kent and M J Ball, eds., *Voice Quality Measurement*, 285–306. San Diego: Singular Thompson Learning. 38

Frölich M, Michaelis D and Strube H W (2001). SIM – simultaneous inverse filtering and matching of a glottal flow model for acoustic speech signals. *Journal of the Acoustical Society of America*, 110(1):479–488. 18

Fu Q and Murphy P (2003). Adaptive inverse filtering for high accuracy estimation of the glottal source. In *ITRW on Non-Linear Speech Processing (NOLISP 03)*. Le Croisic, France. 20

Fu Q and Murphy P (2004). A robust glottal source model estimation technique. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 81–84. Jeju Island, Korea. 19

Funaki K, Miyanaga Y and Tochinai K (1997). A time varying armax speech modeling with phase compensation using glottal source model. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 1299–1302. Munich, Germany. 19

Funaki K, Miyanaga Y and Tochinai K (1998). On subband analysis based on glottal-armax speech model. In *Third ESCA/COSCOSDA International Workshop on Speech Synthesis*. Jenolan Caves, Blue Mountains, Australia. 19

Gobl C (1989). A preliminary study of acoustic voice quality correlates. *STL-QPSR*, 30(4):9–22. 12, 140, 141

Gobl C (2003). *The Voice Source in Speech Communication*. Ph.D. thesis, KTH. 12

Gobl C, Bennett E and Chasaide A N (2002). Expressive synthesis: how crucial is voice quality? In *Proceedings of the IEEE Workshop on Speech Synthesis*. 147, 161

Gobl C and Chasaide A N (2000). Testing affective correlates of voice quality through analysis and resynthesis. In *Proc. of the ISCA Workshop Speech and Emotion*. Belfast. 147, 161

Gobl C and Chasaide A N (2003a). Amplitude-based source parameters for measuring voice quality. In *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, 151–156. Geneva, Switzerland. 136, 137

Gobl C and Chasaide A N (2003b). The role of voice quality in communicating emotion, mood and attitude. *Speech Communication*, 40:189–212. 2, 140, 141, 147

Gómez-Vilda P, Fernández-Baillo R, Rodellar-Biarge V, Lluis V N, Álvarez-Marquina A, Mazaira-Fernández L M, Martínez-Olalla R and Godino-Llorente J I (2009). Glottal source biometrical signature for voice pathology detection. *Speech Communication*, 51(9):759 – 781. 2

Granqvist S, Hertegaard S, Larsson H and Sundberg J (2003). Simultaneous analysis of vocal fold vibration and transglottal airflow; exploring a new experimental set-up. In *TMH-QPSR*, vol. 45, 35–46. Speech, Music and Hearing, KTH, Stockholm, Sweden. 15, 38

Gutiérrez-Arriola J, Hsiao Y, Montero J, Pardo J and Childers D (1998). Voice conversion based on parameter transformation. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*. Sydney, Australia. 2

Hanson H M (1997). Glottal characteristics of female speakers: Acoustic correlates. *Journal of the Acoustical Society of America*, 101(1):466–481. 12

Hanson H M and Chuang E S (1999). Glottal characteristics of male speakers: Acoustic correlates and comparison with female data. *Journal of the Acoustical Society of America*, 106(2):1064–1077. 12

Henrich N, d'Alessandro C, Doval B and Castellengo M (2004). On the use of the derivative of electroglottographic signals for characterization of nonpathological phonation. *Journal of the Acoustical Society of America*, 115(3). 15, 38

Hermes D J (1991). Synthesis of breathy vowels: some research methods. *Speech Commun*, 10(5-6):497–502. 25

Hirano M (1981). *Clinical Examination of Voice*. New York: Springer. 15

Ho C, Rentzos D and Vaseghi S (2002). Formant model estimation and transformation for voice morphing. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 2149–2152. Denver, Colorado, USA. 102

Holmberg E B, Hillman R E and Perkell J S (1988). Glottal airflow and transglottal air pressure measurements for male and female speakers in soft, normal, and loud voice. *Journal of the Acoustical Society of America*, 84(2):511–529. 16

Holzrichter J F, Burnett G S, Ng L C and Lea W A (1998). Speech articulator measurements using low power em-wave sensors. *Journal of the Acoustical Society of America*, 103:622–625. 15

Huang X, Acero A and Hon H W (2001). *Spoken Language Processing. A Guide to Theory, Algorithm, and Systems Development*. Prentice Hall PTR. 95

Itakura F (1975). Line spectrum representation of linear predictor coefficients of speech signals. *Journal of the Acoustical Society of America*, 57(S1):S35–S35. 88

Jiang D N and Cai L H (2004). Speech emotion classification with the combination of statistic features and temporal features. In *IEEE International Conference on Multimedia and Expo (ICME)*. 147

Kaburagi T, Kawai K and Abe S (2007). Analysis of voice source characteristics using a constrained polynomial representation of voice source signals (l). *Journal of the Acoustical Society of America*, 121(2):745–748. 141

Kain A (2001). *High resolution voice transformation,*. Ph.D. thesis, OGI School of Science and Engineering. 102, 104

Kain A and Macon M W (1998). Spectral voice conversion for text-to-speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 285–288. Seattle, Washington, USA. 102

Karlsson I (1985). Glottal wave forms for normal female speakers. *STL-QPSR*, 26(1):31–36. 16

Karlsson I (1988). Glottal waveform parameters for different speaker types. *STL-QPSR*, 29(2-3):061–067. 34

Karlsson I and Liljencrants J (1996). Diverse voice qualities: models and data. *STL-QPSR*, 2:143–146. 73, 80, 141

Kawahara H, Masuda-Katsuke I and de Cheveigné A (1999). Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based f0 extraction: possible of a repetitive structure in sounds. *Speech Communication*, 27(3–4):187–207. 83, 96

Keller E (2005). The analysis of voice quality in speech processing. In *Lecture Notes in Computer Science*, vol. 3445, 54–73. Springer-Verlag. 133

Kim Y E (2003). *Singing Voice Analysis/Synthesis*. Ph.D. thesis, MIT. 12, 18, 26, 29, 35, 72, 161

Klatt D H (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, 67(3):971–995. 2, 9, 25, 138

Klatt D H and Klatt L C (1990). Analysis, synthesis, and perception of voice quality variations among female and male talkers. *Journal of the Acoustical Society of America*, 87(2). 8, 9, 12, 19, 25, 138, 141, 147

Kondoz A M (2004). *Digital speech: coding for low bit rate communication systems*. John Wiley & Sons. 88

Krishnamurthy A K and Childers D G (1986). Two-channel speech analysis. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4). 38

Lanchantin P, Degottex G and Rodet X (2010). A HMM-based speech synthesis system using a new glottal source and vocal-tract separation method. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 4630–4633. Dallas, Texas, USA. 162

Laroche J, Stylianou Y and Moulines E (1993). HNS: Speech modification based on a harmonic + noise model. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, 550–553. Minneapolis, Minnesota, USA: IEEE. 2

Laroia R, Phamdo N and Farvardin N (1991). Robust and efficient quantization of speech LSP parameters using structured vector quantizers. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 641–644. Toronto, Ontario, Canada. 111

Larsson H, Hertegård S, Lindestad P Å and Hammarberg B (2000). Vocal fold vibrations: High-speed imaging, kymography, and acoustic analysis: A preliminary report. *The Laryngoscope*, 110(12):2117–2122. 15

Laver J (1980). *The Phonetic Description of Voice Quality*. Cambridge: Cambridge University Press. 133, 138, 140

Lecluse F, Brocaar M and Verschuure J (1975). The electroglottography and its relation to glottal activity. *Folia Phoniatrica et Logopaedica*, 17:215–224. 15

Lee C C, Mower E, Busso C, Lee S and a S N (2011). Emotion recognition using a hierarchical binary decision tree approach. *Speech Communication*, 53:1162–1171. 147, 151

Lee K, Doh W and Youn D (2002). Voice conversion using low dimensional vector mapping. *IEICE Trans on Information and System*, 8:1297–1305. 102

Liljencrants J (1967). The ove iii speech synthesizer. *STL-QPSR*, 8:2–3. 2

Lin Q (1990). *Speech Production Theory and Articulatory Speech Synthesis*. Ph.D. thesis, KTH. 13

Lu H L (2002). *Toward a high-quality singing synthesizer with vocal texture control*. Ph.D. thesis, Stanford University. 12, 18, 25, 29, 30, 31, 35, 56, 71, 72

Luengo I, Navas E and Hernáez I (2010). Feature analysis and evaluation for automatic emotion identification in speech. *IEEE Transactions on Multimedia*, 12(6):490–501. 147, 151

Lugger M and Yang B (2006). Classification of different speaker groups by means of voice quality parameters. In *ITG Fachtagung Sprachkommunikation*. 147

Lugger M and Yang B (2007). The relevance of voice quality features in speaker independent emotion recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 17–20. Honolulu, Hawaii, USA. 147

Lugger M and Yang B (2008). *Speech Recognition*, chap. Psychological Motivated Multi-Stage Emotion Classification Exploiting Voice Quality Features. InTech. 151

Marasek K (1997). Egg and voice quality. Online tutorial, Universität Stuttgart. 38, 39

Mathews M V, Miller J E and E E David J (1961). Pitch synchronous analysis of voiced sounds. *Journal of the Acoustical Society of America*, 33(2):179–186. 16

Matthews B, Bakis R and Eide E (2006). Synthesizing breathiness in natural speech with sinusoidal modelling. In *Proc. of INTERSPEECH*, 1790–1793. Pittsburgh, PA, USA. 25

McAulay R and Quatieri T (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 34(4):744–754. 2

Mehta D and Quatieri T (2005). Synthesis, analysis, and pitch modification of the breathy vowel. *Applications of Signal Processing to Audio and Acoustics, 2005 IEEE Workshop on*, 199–202. 25

Miller R L (1959). Nature of the vocal cord wave. *Journal of the Acoustical Society of America*, 31:667–677. 16

Mokhtari P and Campbell N (2003). Automatic measurement of pressed/breathy phonation at acoustic centres of reliability in continuous speech. *IEICE Trans on Information and Systems*, E-86-D(3):574–583. 137, 138

Mokhtari P, Pfitzinger H R and Ishi C T (2003). Principal componets of glottal waveforms: towards parameterisation and manipulation of laryngeal voice-quality. In *Voice Quality: Functions, Analysis and Synthesis (VOQUAL'03)*, 133–138. Geneva, Switzerland. 138, 162

Monsen R B and Engebretson A M (1977). Study of variations in the male and female glottal wave. *Journal of the Acoustical Society of America*, 62(4):981–993. 16

Moore E and Torres J (2006). Improving glottal waveform estimation through rank-based glottal quality assessment. *Proc of INTERSPEECH*. 23

Moore E and Torres J (2008). A performance assessment of objective measures for evaluating the quality of glottal waveform estimates. *Speech Communication*, 50(1):56–66. 22, 23

Mori H and Kasuya H (2003). Speaker conversion in ARX-based source-formant type speech synthesis. In *Proc. the of European Conference on Speech Communication and Technology*, 2421–2424. Geneva, Switzerland. 2, 102, 103

Moulines E and Charpentier F (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467. 2, 94, 96

Ní Chasaide A and Gobl C (1997). Voice source variation. In W Hardcastle and J Laver, eds., *The Handbook of Phonetic Sciences*, 427–461. Blackwell. 140, 141

Nogueiras A, Moreno A, Bonafonte A and Mariño J B (2001). Speech emotion recognition using hidden markov models. In *Proc. the of European Conference on Speech Communication and Technology*, 2679–2682. Aalborg, Denmark. xx, 147, 148, 155

Ohtsuka T and Kasuya H (2000). An improved speech analysis-synthesis algorithm based on the autoregressive with exogenous input speech production model. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 787–790. Beijing, China. 18, 19

Orlikoff R F (1991). Assessment of the dynamics of vocal fold contact from the electroglottogram: Data from normal male subjects. *J Speech Hear Res*, 34(5):1066–1072. 38

Pinto N B, Childers D G and Lalwani A L (1989). Formant speech synthesis: Improving production quality. *IEEE Transactions on Acoustics, Spech and Signal Processing*, 37(12):1870–1887. 2

Plumpe M, Quatieri T and Reynolds D (1999). Modeling of the glottal flow derivative waveform with application to speaker identification. *Speech and Audio Processing, IEEE Transactions on*, 7(5):569–586. 12, 17

Proakis J G and Manolakis D G (1996). *Digital Signal Processing. Principles, algorithms and applications*. Prentice Hall International Editions. 21, 35, 41, 90

Pérez J and Bonafonte A (2005). Automatic voice-source parameterization of natural speech. In *Proc. the of European Conference on Speech Communication and Technology*, 1065–1068. Lisbon, Portugal. 72

Pérez J and Bonafonte A (2006). GAIA: A common framework for the development of speech translation techonologies. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 2550–2553. Genoa, Italy. 101

Pérez J and Bonafonte A (2009). Towards robust glottal source modeling. In *Proc. of INTERSPEECH*, 68–71. Brighton, United Kingdom. 72

Pérez J and Bonafonte A (2011). Adding glottal source information to intra-lingual voice conversion. In *Proc. of INTERSPEECH*. Florence, Italy. 119

Pérez J, Bonafonte A, Hain H, Keller E, Breuer S and Tian J (2006). ECESS inter-module interface specification for speech synthesis. In *Proc. of the International Conference on Language Resources and Evaluation (LREC)*, 303–306. Genoa, Italy. 134

Qiao Y and Minematsu N (2009). Mixture of probabilistic linear regressions: A unified view of gmm-based mapping techiques. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:3913–3916. 103, 161

Raitio T, Suni A, Yamagishi J, Pulakka H, Nurminen J, Vainio M and Alku P (2011). HMM-based speech synthesis utilizing glottal inverse filtering. *IEEE Transactions on Audio, Speech, and Language Processing*, 19:153–165. 162

Rao K and Yegnanarayana B (2006). Prosody modification using instants of significant excitation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(3):972 – 980. 89

Rentzos D, Vaseghi S, Yan Q, Ho C and Turajlic E (2003). Probability models of formants parameters for voice conversion. In *Proc. the of European Conference on Speech Communication and Technology*, 2405–2408. Geneva, Switzerland. 102

Rosenberg A E (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, 49(2B):583–590. 8, 16, 141

Rothenberg M (1973). A new inverse-filtering technique for deriving the glottal air flow waveform during voicing. *JASMAN*, 53(6):1632–1645. 16

Ryan C, Chasaide A N and Gobl C (2003). Voice quality variation and the perception of affect: Continuous or categorical? In *Proc. of the 15th ICPhs*. Barcelona. 147, 162

Schröder M (2001). Emotional speech synthesis: A review. In *Proc. the of European Conference on Speech Communication and Technology*, 561–564. Aalborg, Denmark. 147

Schuller B, Rigoll G and Lang M (2003). Hidden markov model-based speech emotion recognition. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 401–404. Hong Kong, Hong Kong. 147

Schuller B, Steidl S and Batliner A (2009). The INTERSPEECH 2009 emotion challenge. In *Proc. of INTERSPEECH*, 312–315. Brighton, United Kingdom. 147

Slifka J and Anderson T (2002). Speaker modification with LPC pole analysis. In *IEEE Benelux Signal Processing Symposium*, 65–68. 102

Sondhi M M (1975). Measurement of the glottal waveform. *Journal of the Acoustical Society of America*, 57:228–232. 16

Sondhi M M and Resnik J R (1983). The inverse problem for the vocal tract: Numerical methods, acoustical experiments, and speech synthesis. *Journal of the Acoustical Society of America*, 73:958–1002. 16

Story B H and Titze I R (1995). Voice simulation with a body-cover model of the vocal folds. *Journal of the Acoustical Society of America*, 97(2):1249–1260. 21

Strik H (1998). Automatic parametrization of differentiated glottal flow: Comparing methods by means of synthetic flow pulses. *Journal of the Acoustical Society of America*, 103(5):2659–2669. 19, 58, 75

Strik H and Boves L (1994). Automatic estimation of voice source parameters. In *Proc. of the International Conference on Spoken Language Processing (ICSLP)*, 155–158. Yokohama, Japan. 58

Strik H, Cranen B and Boves L (1993). Fitting a LF-model to inverse filter signals. In *3rd European Conference on Speech Communication and Technology, Proceedings of the*, 103–106. 56, 58, 59, 60

Sturmel N, d'Alessandro C and Doval B (2007). A comparative evaluation of the zeros of z transform representation for voice source estimation. In *Proc. of INTERSPEECH*, 558–561. Antwerp, Belgium. 21

Stylianou Y (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on Spech and Audio Processing*, 9(1):21–29. 2

Stylianou Y, Cappé O and Moulines E (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Spech and Audio Processing*, 6(2):131–142. 102, 161

Sundberg J and Gauffin J (1978). Waweform and spectrum of the glottal voice source. *STL-QPSR*, 19(2–3):35–50. 16

Sündermann D, Bonafonte A, Ney H and Höge H (2005). A study on residual prediction techniques for voice conversion. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 13–16. Philadelphia, Pennsylvania, USA. 102, 103

Sündermann D and Höge H (2003). VTLN-based cross-language voice conversion. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 676–681. 102

Tamura M, Masuko T, Tokuda K and Kobayashi T (2001). Text-to-speech synthesis with arbitrary speaker's voice from average voice. In *Proc. of the EUROSPECH*. 102

Tchong C, Toen J, Kacic Z, Moreno A and Nogueiras A (2000). Emotional speech synthesis database recordings. Tech. rep., Tech. Rep. IST-1999-No 10036-D2, INTERFACE Project. 148

Titze I R (2002). Regulating glottal airflow in phonation: Application of the maximum power transfer theorem to a low dimensional phonation model. *Journal of the Acoustical Society of America*, 111(1):367–376. 21

Titze I R and Story B H (2002). Rules for controlling low-dimensional vocal fold models with muscle activation. *Journal of the Acoustical Society of America*, 112(3):1064–1076. 21

Titze I R, Story B H, Burnett G C, Holzrichter J F, Ng L C and Lea W A (2000). Comparison between electroglottography and electromagnetic glottography. *Journal of the Acoustical Society of America*, 107(1):581–588. 15

Toda T, Black A and Tokuda K (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8):2222–2235. 103, 161

Toda T, Saruwatari H and Shikano K (2001a). High quality voice conversion based on gaussian mixture model with dynamic frequency warping. In *Proc. of the EUROSPECH*. 102

Toda T, Saruwatari H and Shikano K (2001b). Voice conversion algorithm based on gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 841–844. Salt Lake City, Utah, USA. 102, 161

Tokuda K, Matsumura H and Kobayashi T (1994). Speech coding based on adaptive mel-cepstral analysis. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 197–200. Adelaide, South Australia, Australia. 19

Tokuda K, Yoshimura T, Masuko T, Kobayashi T and Kitamura T (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1315–1318. Istanbul, Turkey. 162

Turk O and Arslan L (2003). Voice conversion methods for vocal tract and pitch contour modification. In *Proc. the of European Conference on Speech Communication and Technology*, 2845–2848. Geneva, Switzerland. 102

van Dinther R, Veldhuis R and Kohlrausch A (2005). Perceptual aspects of glottal-pulse parameter variations. *Speech Communication*, 46(1):95–112. 73, 74, 80

Veldhuis R (1998). A computationally efficient alternative for the liljencrants–fant model and its perceptual evaluation. *Journal of the Acoustical Society of America*, 103(1):566–571. 8

Villavicencio F, Robel A and Rodet X (2009). Applying improved spectral modeling for high quality voice conversion. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 0:4285–4288. 161

Švec J and Schutte H (1996). Videokymography: High-speed line scanning of vocal fold vibration. *Journal of Voice*, 10:201–205. 15

Wong D, Markel J and Jr A G (1979). Least squares glottal inverse filtering from the acoustic speech waveform. *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, 27(4):350–355. 17

Ye H and Young S (2003). Perceptually weighted linear transformations for voice conversion. In *Proc. the of European Conference on Speech Communication and Technology*, 2409–2412. Geneva, Switzerland. 102

Ye H and Young S (2004). High quality voice morphing. In *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 9–12. Montreal, Quebec, Canada. 102, 103

Yuan Y X (2000). A review of trust region algorithms for optimization. In J Ball and J Hunt, eds., *ICM99: Proceedings of the Fourth International Congress on Industrial and Applied Mathematics*, 271–282. Oxford University Press. 58

Yutani K, Uto Y, Nankaku Y, Lee A and Tokuda K (2009). Voice conversion based on simultaneous modelling of spectrum and f0. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 0:3897–3900. 101, 103, 161