SPECTRAL PROCESSING OF THE SINGING VOICE

A DISSERTATION SUBMITTED TO THE
DEPARTMENT OF INFORMATION AND COMMUNICATION TECHNOLOGIES
FROM THE POMPEU FABRA UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR PER LA UNIVERSITAT POMPEU FABRA

Alex Loscos
2007

THESIS DIRECTION

Dr. Xavier Serra
Department of Information and Communication Technologies
Universitat Pompeu Fabra, Barcelona

This research was performed at the Music Technology Group of the Universitat Pompeu Fabra in Barcelona. Primary support was provided by YAMAHA corporation and the EU project FP6-IST-507913 SemanticHIFI http://shf.ircam.fr/

Dipòsit legal: B.42903-2007
ISBN: 978-84-691-1201-4

Abstract

This dissertation is centered on the digital processing of the singing voice, more concretely on the analysis, transformation and synthesis of this type of voice in the spectral domain, with special emphasis on those techniques relevant for music applications.

The digital signal processing of the singing voice became a research topic itself since the middle of last century, when first synthetic singing performances were generated taking advantage of the research that was being carried out in the speech processing field. Even though both topics overlap in some areas, they present significant differentiations because of (a) the special characteristics of the sound source they deal and (b) because of the applications that can be built around them. More concretely, while speech research concentrates mainly on recognition and synthesis; singing voice research, probably due to the consolidation of a forceful music industry, focuses on experimentation and transformation; developing countless tools that along years have assisted and inspired most popular singers, musicians and producers. The compilation and description of the existing tools and the algorithms behind them are the starting point of this thesis.

The first half of the thesis compiles the most significant research on digital processing of the singing voice based on spectral domain, proposes a new taxonomy for grouping them into categories, and gives specific details for those in which the author has mostly contributed to; namely the sinusoidal plus residual model Spectral Modelling Synthesis (SMS), the phase locked vocoder variation Spectral Peak Processing (SPP), the Excitation plus Residual (EpR) spectral model of the voice, and a sample concatenation based model. The second half of the work presents new formulations and procedures for both describing and transforming those attributes of the singing voice that can be regarded as voice specific. This part of the thesis includes, among others, algorithms for rough and growl analysis and transformation, breathiness estimation and emulation, pitch detection and modification, nasality identification, voice to melody conversion, voice beat onset detection, singing voice morphing, and voice to instrument transformation; being some of them exemplified with concrete applications.

Resumen

Esta tesis doctoral versa sobre el procesado digital de la voz cantada, más concretamente, sobre el análisis, transformación y síntesis de este tipo de voz basándose e dominio espectral, con especial énfasis en aquellas técnicas relevantes para el desarrollo de aplicaciones musicales.

El procesado digital de la voz cantada se inició como área de investigación a medianos del siglo pasado, cuando se generaron las primeras melodías humanas sintéticas a partir de la investigación que se llevaba a cabo en el campo del procesado del habla. Aunque ambos campos de investigación, voz y canto, tienen puntos en común, presentan evidentes e importantes diferencias entre sí, no sólo en cuanto a las propiedades del sonido fuente objeto de estudio, sino también en cuanto al tipo de aplicaciones a las que se orientan. Más concretamente, mientras las tecnologías del habla se concentran principalmente en tareas de reconocimiento y síntesis, las tecnologías del canto se concentran, seguramente debido a la consolidación de una gigantesca industria musical, en la experimentación y transformación; generando un sinfín de herramientas que a lo largo de los años han asistido e inspirado a los más conocidos cantantes, músicos y productores. La recopilación de estas herramientas y los algoritmos subyacentes consolidan el punto de inicio de este trabajo.

La primera mitad de la tesis compila los trabajos de investigación más significativos en torno al procesado de la voz cantada basados en dominio espectral, propone una nueva taxonomía para agruparlos en categorías, y da detalles específicos de aquellas tecnologías en las que el autor ha contribuido; que son el modelo sinusoidal más residuo Spectral Modeling Síntesis (SMS), la variante del vocoder de fase bloqueada Spectral Peak Processing (SPP), el modelo espectral de voz Excitation plus Residual (EpR), y un modelo basado en la concatenación de muestras. La segunda mitad de la tesis incluye, entre otros, algoritmos para el análisis y la generación de desórdenes vocales como rugosidad, ronquera, o voz aspirada, detección y modificación de la frecuencia fundamental de la voz, detección de nasalidad, conversión de voz cantada a melodía, detección de los golpes de voz, mutación de voz cantada, y transformación de voz a instrumento; ejemplificando algunos de éstos en aplicaciones concretas.

Resum

Aquesta tesi doctoral versa sobre el processament digital de la veu cantada, més concretament, sobre l'anàlisi, transformació i síntesi d'aquets tipus de veu en el domini espectral, amb especial èmfasi en aquelles tècniques rellevants per al desenvolupament d'aplicacions musicals.

El processament digital de la veu cantada es va iniciar com a tòpic de recerca a mitjans del segle passat, arrel de les primeres melodies humanes sintètiques generades a partir de la recerca que es duia a terme en el camp del processament de la parla. Encara que ambdós camps de recerca, parla i cant, tenen punts en comú, presenten diferències evidents a l'hora que importants entre sí, no tan sols en quant a les propietats del so font objecte d'estudi, sinó també en quant al tipus d'aplicacions a les que s'orienten. Més concretament, mentres les tecnologies de la parla es concentren en tasques de reconeixement i síntesi, les tecnologies del cant es concentren, segurament degut a la consolidació d'una indústria musical molt poderosa, en l'experimentació i la transformació, generant una infinitat d'eines que, al llarg dels anys, han assistit i inspirat als cantants, músics i productors més coneguts de l'escena. La recopilació d'aquestes eines i els algorismes subjacents consoliden el punt d'inici d'aquest treball.

En aquest context, la primera meitat de la tesi compila els treballs de recerca més significatius en torn al processament de la veu cantada basats en el domini espectral, proposa una nova taxonomia per agrupar-los en categories, i dóna detalls específics d'aquelles tecnologies en les que l'autor ha contribuït; que són el model de sinusoides més residu Spectral Modeling Síntesis (SMS), la variant del vocoder de fase bloquejada Spectral Peak Processing (SPP), el model espectral de veu Excitation plus Residual (EpR), i un modelo basat en la concatenació de mostres. La segona meitat de la tesis inclou, entre d'altres, algorismes per l'anàlisi i la generació de desordres vocals como ara rugositat, ronquera, o veu aspirada, detecció i modificació de la freqüència fonamental de la veu, detecció de nasalitat, conversió de veu cantada a melodia, detecció de cops de veu, mutació de veu cantada, i transformació de veu a instrument; exemplificant alguns d'aquests algorismes en aplicacions concretes.

Acknowledgments

Thanks to Dr. Xavier Serra for all his support.

Thanks to Jordi Bonada for his valuable advice and guidance.

Thanks to all my colleagues Salvador Gurrera, Cristina Garrido, Joana Clotet, Maarten de Boer, Lars Fabig, Oscar Mayor, Merlijn Blaauw, Esteban Maestre, Marc Vinyes, Jordi Janer, Emilia Gutierrez, Perfecto Herrera, Fabien Gouyon, Sebastian Streich, Bee Suan Ong, Enric Guaus, Oscar Celma, Jaume Ortola, Pau Arumi, Xavier Amatriain, David Garcia, Ramon Loureiro, Carlos Atance, Nicolas Wack, Thomas Aussenac, Markus Koppenberger, Pedro Cano, Gunnar Holmberg, Jose Pedro Garcia, Hendrik Purwins, Rafael Ramirez, Amaury Hazan, Alfonso Perez, Ricard Marxer, Bram de Jong, Hugo Solís, Günter Geiger, Martin Kaltenbrunner, Sergi Jorda, Marcos Alonso, Hideki Kenmochi, Takuya Fujishima, Yasuo Yoshioka, Hiraku Kayama, Yuji Hisaminato.

Thanks to Dr. Wang Ye for his kind invitation and friendship.

Thanks to all friends and family, very especially to Emma.

Contents

Abstract .... **v**

Resumen .... **vii**

Resum .... **ix**

Acknowledgments .... **xi**

Chapter 1: Introduction **1**

1.1 Motivation 1

1.2 Personal background 2

1.2.1 As an engineer 2

1.2.1.1 Elvis 2

1.2.1.2 SMSPerformer 3

1.2.1.3 Daisy 3

1.2.1.4 Rosa 4

1.2.1.5 Vocal Processor and ComboVox 4

1.2.1.6 Semantic HIFI 5

1.2.1.7 ESPRESSO 5

1.2.1.8 AudioScaner 6

1.2.2 As a musician 6

1.3 Aims and contributions 7

1.4 Summary of the PhD work 8

Chapter 2: Technologies for the singing voice **9**

2.1 Introduction 9

2.2 Singing voice tools 9

2.2.1 Singing voice tools 10

2.2.1.1 Analysis and Representation 10

2.2.1.2 Coding 13

2.2.1.3 Synthesis 14

2.2.1.4 Transformation 17

2.2.1.4.1 Software 17

2.2.1.4.2 Hardware 24

2.2.2 Singing voice in music production 26

2.2.3 Running Out of Melodies? 28

2.3 Voice production 31

2.3.1 Anatomy of the voice 31

2.3.2 The source-filter model 33

2.3.3 Voice Production Acoustics 35

2.3.4 Singing versus Speech 36

2.4 Voice technologies 39

2.4.1 Voice mechanics technologies 39

2.4.1.1 Singing robots 39

2.4.1.2 Digital waveguides 40

2.4.2 Voice timbre technologies 41

2.4.2.1 Vocoder 41

2.4.2.1.1 Phase vocoder 42

2.4.2.1.2 Phase locked vocoder 44

2.4.2.2 Spectral peak processing 44

2.4.2.2.1 Spectrum segmentation 45

2.4.2.2.2 Maximally flat phase alignment 46

2.4.2.3 Linear predictive coding 47

2.4.2.3.1 Code excited linear prediction 47

2.4.2.4 Sinusoidal based technologies 48

2.4.2.4.1 Sinusoidal model 48

2.4.2.4.2 Sinusoidal plus residual model 49

2.4.2.4.3 Transients 51

2.4.2.5 Excitation plus resonances voice model 51

2.4.2.5.1 EpR source filter 51
2.4.2.5.2 EpR vocal tract filter 53
2.4.2.5.3 EpR phase alignment 54
2.4.2.5.4 Implementation of the EpR filter 55
2.4.3 Pressure waves technologies 56
2.4.3.1 Glottal pulse model 56
2.4.3.2 Overlap and add 58
2.4.3.2.1 Synchronous overlap and add 58
2.4.3.2.2 Pitch synchronous overlap and add 58
2.4.3.3 Formant synthesis 59
2.4.3.4 Concatenative 59
2.4.3.4.1 Concatenation of sound pieces 59
2.4.3.4.2 Phase concatenation 60
2.4.3.4.3 Spectral shape concatenation 62

Chapter 3: Spectral based description of the singing voice 65
3.1 Introduction 65
3.2 Voice excitation descriptors 65
3.2.1 Amplitude 65
3.2.2 Pitch 68
3.2.3 Vibrato and tremolo 69
3.2.4 Rough and growl 69
3.2.4.1 The growl observation 70
3.2.5 Pulse regularity 71
3.2.6 Breathiness 72
3.3 Timbre descriptors 74
3.3.1 Whawactor: real time timbre variation estimation 76
3.3.1.1 'Cepstrum': MFCC's variation 77
3.3.1.2 'LPC': LPC roots 78
3.3.1.3 'Slope': Low-band Harmonics Slope 79
3.3.1.4 'Centroid': Low-band Spectral Centroid 80
3.3.1.5 'Area': Low-band Spectral Weighted Area 80
3.3.1.6 Evaluation of the descriptors 81

3.3.2 Nasality 83

3.4 Musical meaningful parameters 86

3.4.1 Melody 86

3.4.2 Singing voice onset 87

3.5 Conclusions 90

Chapter 4: Techniques for the transformation of the singing voice 91

4.1 Introduction 91

4.2 Vocal tract transformations 91

4.2.1 Equalization filters 92

4.2.2 Timbre mapping 92

4.3 Voice excitation transformations 94

4.3.1 Roughness 94

4.3.2 Growl 97

4.3.3 Breathiness 98

4.3.4 Whisper 99

4.3.5 Dynamics 100

4.3.6 Pitch transposition 101

4.3.6.2 Peak picking 104

4.3.6.3 Phase alignment 105

4.3.6.4 Partial dependent frequency scaling 107

4.4 High level transformations 108

4.4.1 Pitch based high level transformations 108

4.4.1.1 Intonation 108

4.4.1.2 Pitch discretization to the temperate scale 109

4.4.1.3 Vibrato 109

4.4.2 Identity transformations 110

4.4.2.1 Gender change 110

4.4.2.2 Age transformations 111

4.4.2.3 Exotic alienations 111

4.4.3 Multiplication 112

4.4.3.1 Harmonizer 112

4.4.3.2 Choir 113

4.4.4 Voice enhancement ... 114

4.4.4.1 The preprocess and the unvoiced cycle ... 116

4.4.4.2 Feeding the synthesis engine ... 116

4.4.4.3 Healthy pitch and timbre ... 117

4.5 Morph and control, voice driven synthesis ... 118

4.5.1 Singing voice conversion ... 119

4.5.1.1 Front end parameterization of the singing voice ... 120

4.5.1.2 Alignment ... 121

4.5.1.3 Morph and synthesis ... 123

4.5.2 Voice orchestration ... 124

4.5.2.1 Non-real time processes: instrument database creation ... 124

4.5.2.2 Real time processes: voice analysis, morph and instrument synthesis ... 124

4.6 Conclusions ... 125

Chapter 5: Conclusions and future work ... 127

5.1 Author's contribution ... 127

5.2 Conclusions and future work ... 128

References ... 129

Annex A ... 147

A.1 Journal ... 147

A.2 Book chapters ... 148

A.3 Conferences ... 148

Annex B ... 155

Chapter 1

Introduction

1.1 Motivation

"Men sang out their feelings long before they were able to speak their thoughts."
--Otto Jespersen, *Language, Its Nature, Development and Origin*

The most common belief in today's anthropologic research is that the very first sound uttered by a human being was sung. Primitive utterances are presumed to have originated by mimicking the sounds heard in nature such as the singing of the birds of the roaring of a beast. These utterances, which had more in them of sophisticated vocal music than of monotonous spoken language, were, at first, purely exclamative containing no sense or thought at all. At what point fellow-creatures realized such singing could communicate ideas and feelings cannot be established, but it probably meant the major step in the creation of language, and preceded any other embryo of musical expression.

In this context it is said the voice is the original music instrument and the presence of singing in each and every human culture suggests its inception may arise from an inner craving of the individual. If so, one might presume such nature has somehow remained underlying all along millions of years of singing evolution and differentiations in a way that, still today, sung voice is considered to be the most fascinating, expressive, and powerful instrument.

Among all elements catalyzing the evolution of the singing voice, this work aims to uncover a spur placed at the very last link of the chain: digital technology. Sound digitalization and transformation has played a critical roll in spotlighting and massifying the singing voice discipline over the last few years. Although traditionally subordinated to digital speech technology and mainly supported by music industry, digital singing voice technology has achieved to become a specialty its own while moving on towards other industries such as videogaming or media production.

1.2 Personal background

Two profiles blend to configure the author's relevant personal background: the digital signal processing engineer and the musician.

1.2.1 As an engineer

The author received the B.S and M.S. degrees in Telecommunication Engineering from Catalunya Polytechnic University, Barcelona, Spain, in 1997 and 1999 respectively. In 1997 he joined the Music Technology Group (MTG) of the Universitat Pompeu Fabra as a junior researcher and developer. Since then he has worked in voice processing/recognition, digital audio analysis/synthesis and transformations, and statistical digital signal processing and modeling. Almost all of the projects he has been involved in the MTG attend to the singing voice and Yamaha Corporation.

1.2.1.1 Elvis

Elvis was a collaboration project with Yamaha. The aim of the project was to develop an automatic singing voice impersonator for the Japanese karaoke customers. The system had to be able to change some voice characteristics in real-time in order to make the karaoke user's singing resemble a famous pop-rock singer (as for example Elvis) while performing. To do so, a frame-based real-time singing voice analysis-transformation-synthesis engine was implemented.

The analysis-transformation-synthesis engine was built on top of the Spectral Modeling Synthesis (SMS), an analysis by synthesis technique based on the sinusoidal plus residual decomposition of the sound. The analysis parameterized the voice, the transformation step modified the parameters obtained in the analysis, and the synthesis step generated the synthetic voice out of a modified parameterization.

The main research done for Elvis project has been published in Loscos et al. (1999), Cano et al, (1999,2000), and Boer et al. (2000).

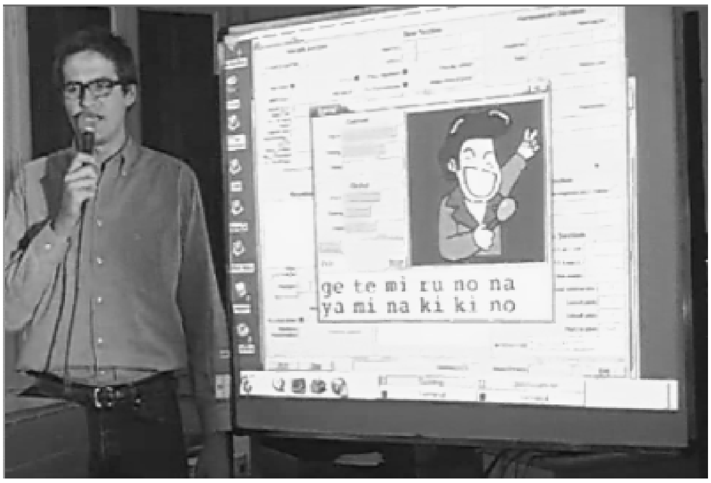


Figure 1: The author demonstrating the prototype

1.2.1.2 SMSPerformer

SMSPerformer, see Loscos and Resina (1998) was an MTG inside project. The aim of the project was to come out with a graphical interface for the real-time SMS synthesis engine that could work from already SMS analyzed sounds. Althought SMSPerformer was originally designed to test in real-time the voice impersonator synthesis and morph parameters, the sound manipulation capabilities of the application made it to be used by the argentinian composer Ricardo Ventura to perform his electro acoustic piece "Desnudo de mujer en sofá azul".

1.2.1.3 Daisy

Daisy was a collaboration project with Yamaha. The aim of the project was to develop a singing voice synthesizer in which the user would input the lyrics and the notes of a vocal melody and obtain a synthetic performance of a virtual singer. To synthesize such performance the system concatenates a chain of elemental synthesis units. These units are obtained by transposing and time-scaling samples from singers databases. These databases are created out of recording, analyzing, labeling and storing singers performing in as many different musical and phonetic contexts as possible.

Based on Daisy's research, Yamaha released a product named Vocaloid. Vocaloid was presented at the 114th Audio Engineering Society (AES) Convention in Amsterdam in March 2003 and had a big media impact leaded by The New York Times article "Could I Get That Song in Elvis, Please?". The Vocaloid synthesizer was Nominee for the European IST (Information Society Technologies) Prize 2005 and is the choice of leading artists including Mike Oldfield.

The research carried out for Daisy is mainly described in Bonada and Loscos (2003), Bonada et al. (2003), and Bonada et al (2001).

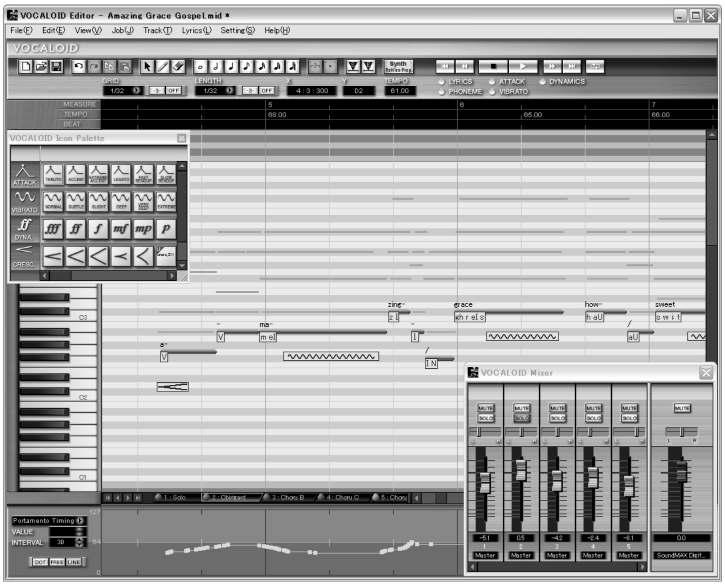


Figure 2: Screenshot of Yamaha's commercial application Vocaloid

1.2.1.4 Rosa

Rosa was a collaboration project with Spanish leader in telecommunication services in Spain and Latin America, Telefónica. The aim of the project was to evaluate speech voice quality obtained using MTG's synthesis techniques at that time and study the feasibility of porting them to Telefonica's automatic text-to-speech system. This evaluation was based on comparing the quality of the synthetic voices obtained from both Telefonica's and MTG's synthesis engines. The evaluation was carried out through a web test questionnaire based on ITU-T P.85 standard. The results of the evaluation revealed MTG's synthesis was perceived as "good" as Telefonica's highest quality synthesis.

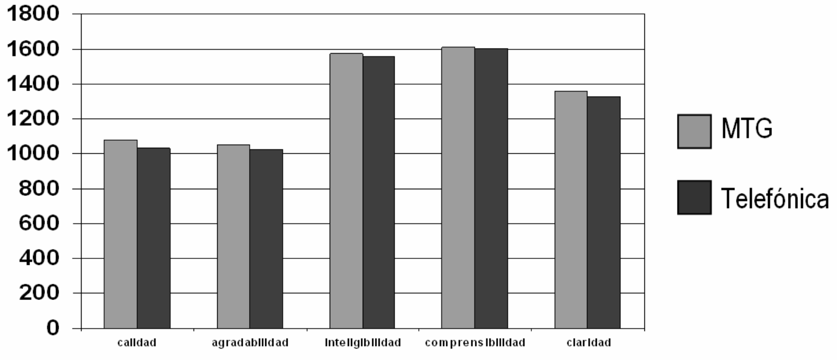


Figure 3: Quality assessment results for a test performed by over 80 people.

1.2.1.5 Vocal Processor and ComboVox

The aim of these two projects was the implementation of a real-time singing voice specific effect processor able to modify attributes such as timbre, vibrato, intonation, tuning, breathiness, roughness, harmonies, and others. Vocal Processor was a collaboration project with YAMAHA and partial research results can be found in Amatriain et al. (2001, 2002, 2003). ComboVox was a project with Pinnacle Systems. ComboVox was included in the remarkably popular video editing software Studio 10 release in the form of an exclusive plug-in. While Vocal Processor was intended to be a flexible tool featuring multiple controls over all available transformations, ComboVox was not as versatile and based its transformations on presets.

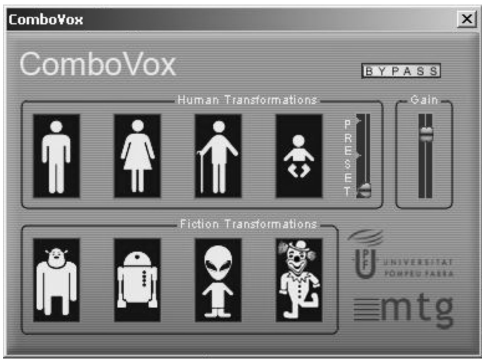


Figure 4: ComboVox graphical interface

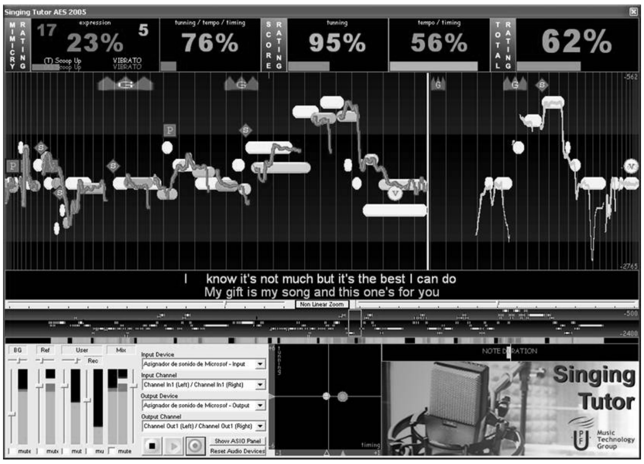
1.2.1.6 Semantic HIFI

Semantic HIFI (FP6-IST-507913) was a European project on the sixth framework programme for networked audiovisual systems and home platforms. The goal of the project was to develop a new generation of HIFI systems, offering new functionality for browsing, interacting, rendering, personalizing and editing musical material. The Music Technology Group was in charge of the Performing work-packages, for which we developed a set of interaction and transformation tools for the user to forge, customize, play with and create the media at his / her own taste.

Figure 5: Thomas Aussenac playing guitar and Whawactor, see Loscos and Aussenac (2005), one of the outcomes of the Performing work-package.

1.2.1.7 ESPRESSO

ESPRESSO was a collaboration project with Yamaha. The goal of the project was to build system for an automatic expert evaluation of a singing performance. The evaluation had to assess different facets of the performance (timbre, voice character, tuning, vibratos, portamentos, energy, and etcetera) from two different points of view, as a performance on its own, and as mimicry of a professional singer performance.



I know it's not much but it's the best I can do
My gift is my song and this one's for you

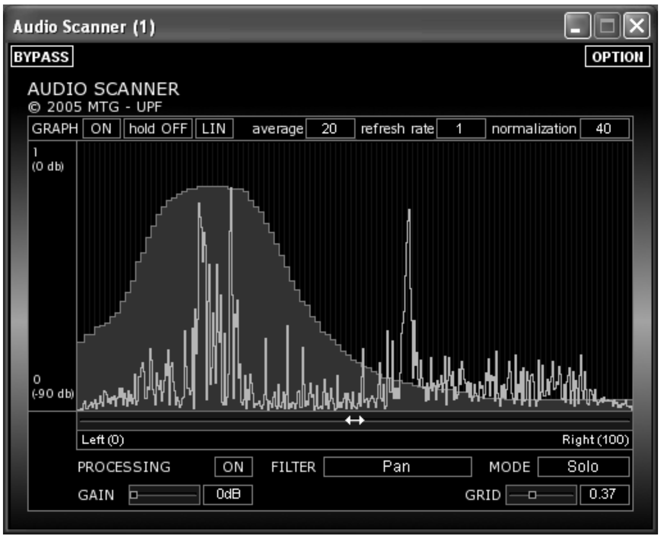
Input Device: Asignador de sonido de Microsof - Input
Input Channel: Channel In1 (Left) / Channel In1 (Right)
Output Device: Asignador de sonido de Microsof - Output
Output Channel: Channel Out1 (Left) / Channel Out1 (Right)
Show ASIO Panel
Reset Audio Devices

Figure 6: Screenshot of the Singing Tutor, ESPRESSO's prototype

1.2.1.8 AudioScaner

AudioScanner was an internal MTG project. The aim of the project was to process a single audio object of a mix independently of the rest of the sound. With this we could modify the sound of a violin in an ensemble recording or we could apply transformations to the lead voice in a whole band recording.

Two different approaches were taken in this area: multi-pitch estimation plus spectral peaks transformation, and frequency domain stream separation based on stereo panoramic.



GRAPH ON hold OFF LIN average 20 refresh rate 1 normalization 40

PROCESSING ON FILTER Pan MODE Solo

GAIN 0dB GRID 0.37

Figure 7: Screenshot of the Audio Scanner VST plug-in, see Vinyes et al. (2006)

1.2.2 As a musician

The author's career as a musician, singer, songwriter and producer dates from 1996. Since then he has been involved in several pop-rock music projects with which he has released 3 albums, and 3 EPs. Some of his music has been licensed to radio and television commercials adverts and breaks, and some has been included as soundtrack in international distribution movies and well-known TV serials. Televisions such as MTV, TVE, 40TV or TVC, and radio stations such as RNE, Catalunya Radio or Los 40 Principales have frequently broadcasted singles, concerts and video-clips of the bands he performs with. Some of these bands have international distribution along Europe, EEUU, and Asia and most of the author's music is available at internet sites such as mp3.com, emusic, cdbaby, or allmusic.

This first hand experience has proven to bee remarkably useful in order to dig up an inside view of specific aspects of current music production business. Information such as how do studio engineers record, which material do they work with, how do they evaluate new products, how much money companies invest on production, or how much production is typically put in lead voices, turns to be extremely valuable for any engineer doing research and development in this specific area.

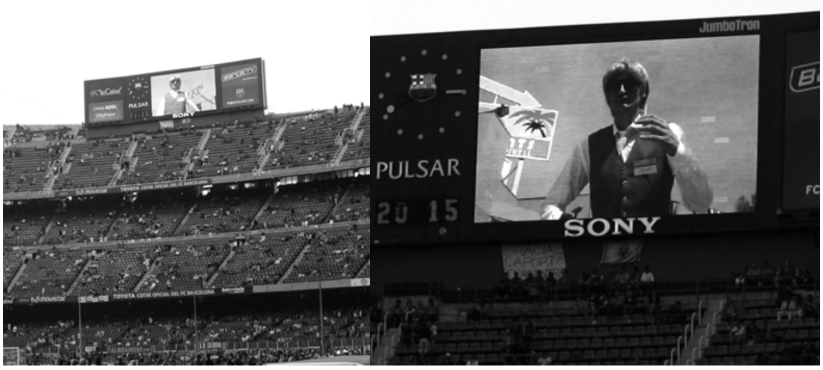


Figure 8: A video-clip of one of the author's songs screened in F.C. Barcelona stadium Camp Nou.

1.3 Aims and contributions

Plenty of research has already been carried out on the subject of the singing voice. A considerable number of theses that examine different aspects of it have been published; going from 70's Miller research on separating voice from orchestra to restore Enrico Caruso recordings, see Miller (1993), stepping on the 90's first singing voice synthesis approaches with Cook's digital waveguides, Cook (1990), Macon's sinusoidal model Macon et al. (1997), and Lomax (1997) and Meron (1999) neural network controlled, and reaching 2000's with additional synthesis and transformation methods and models, see Hui-Ling Lu (2002), Kob (2002), Uneson (2003), Kim (2003), Thibault (2004), Lee (2005), physiological, acoustic and aesthetic investigations, see Ternström (1989), Prame (200), Henrich (2001), Thomasson (2003), Arroabaren (2004), and music information retrieval relate, see Mellody (2001), Carre (2002), Gerhard (2003).

The Author's main contributions in the digital signal processing of the singing voice include:

- an extensive compilation of current digital transformations that can be specifically applied to the singing voice
- an exhaustive compilation of main voice transformation devices and applications available as of today
- a compilation of voice specific descriptors, presenting new and alternative formulations and estimation algorithms for some of them
- a new taxonomy proposal for those technologies that address digital signal processing of the voice
- new transformations concepts and algorithms in frequency domain
- a proposal for the challenges of future research on singing voice
- more than ten patents in the field
- a set of algorithms that have been deployed in industrial applications: transformations that feature in Yamaha's Vocaloid, or ComboVox plug-in for Pinnacle's Sutdio10.
- dissemination of the voice transformation: installations in Barcelona and Madrid main Science Museums (CosmoCaixa)

1.4 Summary of the PhD work

This PhD starts presenting the author (motivations, and research background) and the work itself (context, previous related works and main contributions).

Chapter 2 summarizes different previous knowledge related with the singing voice. It is expound as an introduction to understand the challenges of singing voice processing and starts chronicling the emergence and evolution of voice transformation recourse along record production history, from the very first tape delayed voice to latest digital software. Different technologies are exposed according to a new taxonomy and special attention is given to those technologies in which the author has been mostly involved as a researcher.

Chapter 3 gives an overview of automatic computation of voice specific attributes, mostly based on spectral techniques. Starting with voice excitation and vocal tract attributes, the chapter concludes with musical meaningful descriptors.

Chapter 4 comprises transformations that cope with those attributes presented in Chapter 3 and, in a similar way, it presents them from to lower to higher levels of abstraction. Although transformations on this chapter stand on models and parameterizations of well-know approaches, some of them include improvements that represent significant contributions to the original technique. Such is the case of the roughness effect using the sinusoidal model or the transposition based on phase locked vocoder.

Last chapter, Chapter 5 concludes the work giving an overview of the contributions of the author and presenting a sketch of the future work guidelines.

Chapter 2

Technologies for the singing voice

2.1 Introduction

This chapter gathers concepts and views that are required to understand different facets of singing voice analysis, modeling, coding, transforming, and synthesis.

As it has already been pointed out by Kim (2003) singing voice coding and synthesis tend to converge. The explanation is that, despite their different goals, both share the same challenge, which is the quest of a parametric representation of a voice production model. Such convergence was originally perceived by Dudley, and was included in the MPEG-4 audio coding standard, see Scheirer and Kim (1999), as a synthesis-based coding named structured audio coding, see Vercoe (1998). The theory of diffuse boundaries between voice coding and synthesis topics is totally reasonable and certainly extendable to gather analysis, modeling, and transformation as well. Analysis and modeling, in fact, can be considered inner components of coding; and transformation, when meaningful, calls for an abstract model of the voice which enables significant control.

A compilation of available tools related with the singing voice in different areas of application show the "where". A historic overview of singing in music production and a statistical study of melodies frame the chapter on its context and introduce the "why". A description of the voice organ and production physiology tells us the "what". And the outline of technologies for modeling, analyzing and synthesizing the singing voice establishes the "how".

2.2 Singing voice tools

As of 2006, among all different applications and industries around the singing voice, it has been the ones on music that have lead to the germination, sprouting and multiplication of technologies. Technical and artistic music production habits and requirements have originated a forceful industry around singing voice processing. At the present time most popular songs very seldom hold less than four voice effect layers. The means and possible reasons that brought voice musical production up to this point are exposed in this section

2.2.1 Singing voice tools

If we define tool as any piece of software or hardware that has been coded to perform digital signal processing, a few tools that are specific to the singing voice exist in the areas of analysis and representation, coding, transformation and synthesis.

2.2.1.1 Analysis and Representation

Although many tools exist for the analysis and representation of generic audio (see figure 9), fewer exist for the specific analysis and representation of speech. The tools covering the speech field come mainly from three different areas of application: medical, artistic, and voice quality assessment.

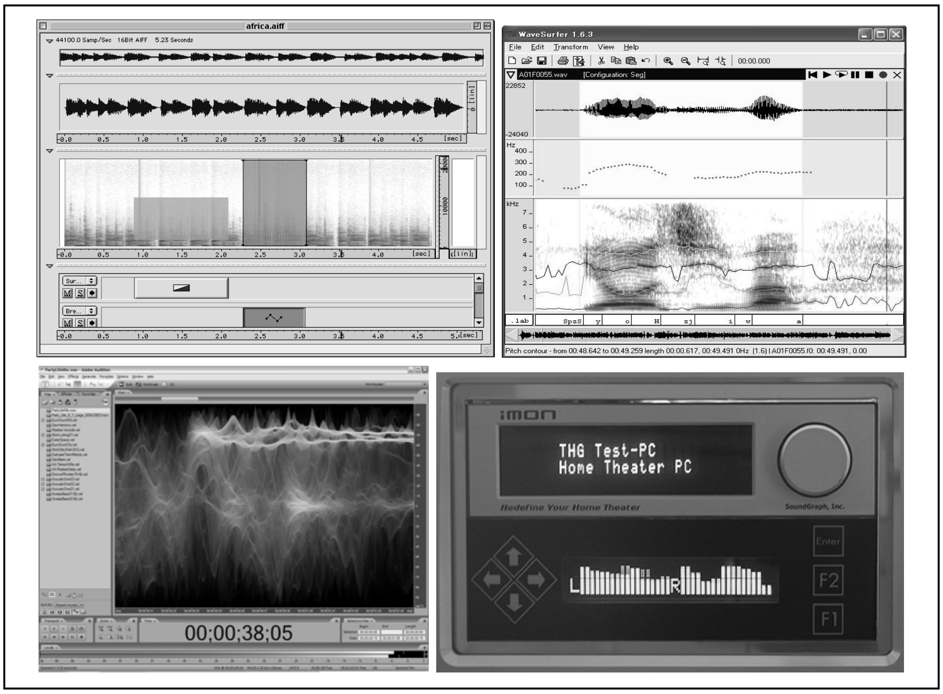

africa.aiff
44100.0 Samp/Sec 16Bit AIFF 5.23 Seconds
WaveSurfer 1.6.3
File Edit Transform View Help
A01F0055.wav [Configuration: Seg]
00;00;38;05
iMON
THG Test-PC
Home Theater PC
Redefine Your Home Theater
SoundGraph, Inc.
Enter
F2
F1

Figure 9: (up left) IRCAM's AudioSculpt[1], (up right) KTH WaveSurfer[2], (down left) Spectral Pan and Phase plug-in from Adobe Audition[3], and (down right) double channel spectrogram from Soundgraph's iMon[4].

[1] http://forumnet.ircam.fr/349.html

[2] http://www.speech.kth.se/wavesurfer

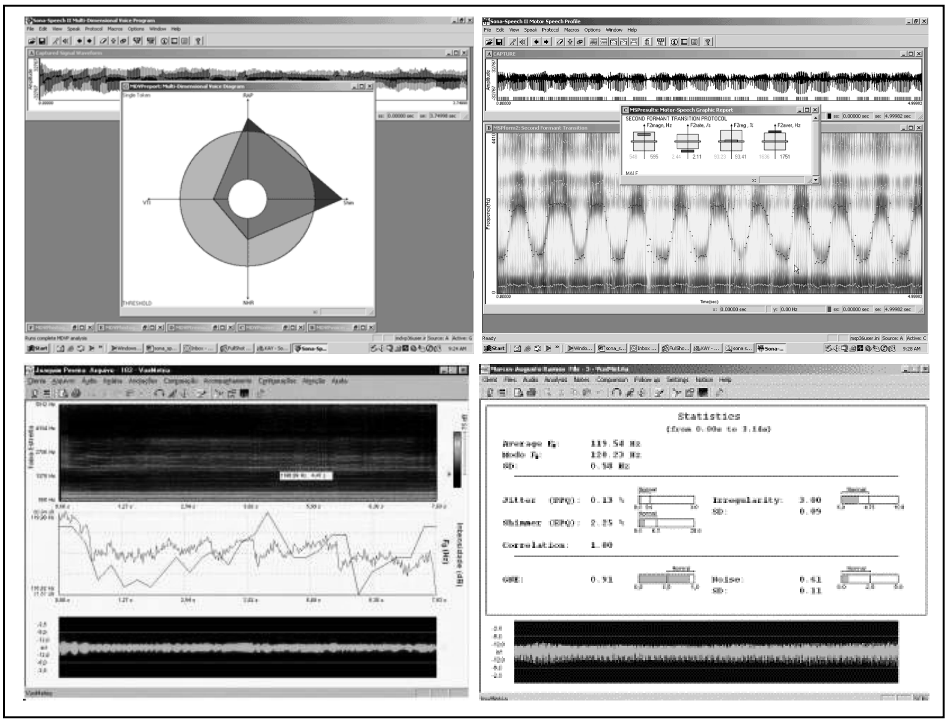
[3] http://www.adobe.com/products/audition/

[4] http://www.soundgraph.com/pr/pr_11.asp

Those tools under the medical category belong mainly to voice pathology detection and voice recovery training and evaluation (see Figure 10).

Those tools under the artistic category can not be grouped under a main functionality. These tools (see Figure 11) are used for completely different purposes such as augmented reality (2Messa di Voice) or video remix by beatboxing ("ScrAmBlEd? HaCkZ!").

And finally there are those under the voice quality assessment category. These are essential tools in the quality supervision process of any telecommunications system (see Figure 12).


Sona-Speech II Multi-Dimensional Voice Program
Sona-Speech II Motor Speech Profile
SECOND FORMANT TRANSITION PROTOCOL
THRESHOLD
Time(sec)
Ready
Start
9:28 AM

Statistics
(from 0.00s to 3.16s)

Average F0: 119.54 Hz
Modo F0: 120.23 Hz
SD: 0.58 Hz

Jitter (PPQ): 0.13 %
Irregularity: 3.80
SD: 0.09
Shimmer (EPQ): 2.25 %
Correlation: 1.00

GNE: 0.91
Noise: 0.61
SD: 0.11

Figure 10: Screen samples of "*Multi Dimensional Voice Program*"[5] (MDVP) (up left and right) and Voxmetria[6] (down left and right). Up left shows MDVP graph, providing a useful snapshot of the client's voice quality. Client performance is plotted against normative thresholds (darkest) that are built into the software. Up right displays second formant transitions, which are frequently diminished in patients who neutralize vowels. Down right shows Voxmetria spectrogram trace in narrow-band with fundamental frequency and intensity plot chart. Down left displays the acoustic analysis values of fundamental frequency, jitter, shimmer, irregularity, noise rate and GNE.

[5] http://www.kayelemetrics.com/Product Info/CSL Options/5105/5105.htm
[6] http://www.ctsinformatica.com.br/voxmetria.htm

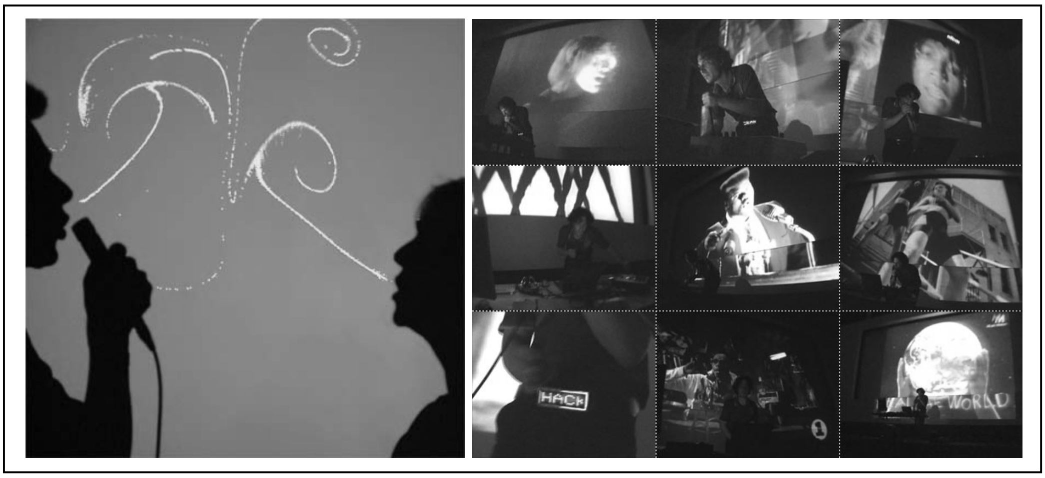


Figure 11: (left) 2Messa di Voice[7], and (right) ScrAmBlEd? HaCkZ![8]

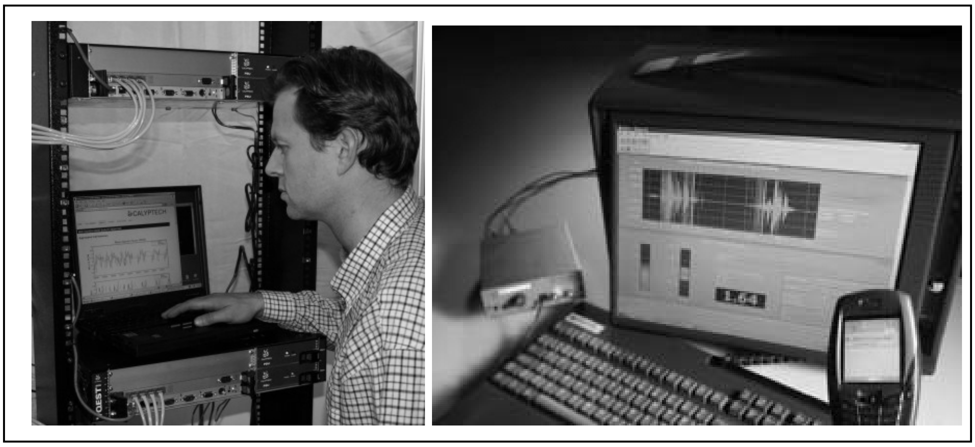


Figure 12: (left) "*Voice Quality Analysis*"[9] and (right) Opera[10]

But which are the tools that exist specifically for the analysis and representation of the singing voice? If we group them according to their purpose we can assume singing tools exist in the fields of education, entertainment and artistic production.

Tools in the education field (see Figure 13) respond to the quest for new ways and methods to educate people in the art of singing. Usually these tools plot in real time a score based representation of the analysis of the voice fundamental frequency. Applications such as these can be found in music schools or city science museums.

Tools for the entertainment (see Figure 14) focus mainly as well in the analysis of the voice fundamental frequency, together with the analysis of the energy. In this field voice is used as a game controller and singing performance is analyzed to drive the game and score the user. One of the main researchers in this field is Perttu Hämäläinen, see

[7] http://tmema.org/messa/messa.html

[8] http://www.popmodernism.org/scrambledhackz/

[9] http://www.calyptech.com/products/vqa/

[10] http://www.opticom.de/products/opera.html

Hämäläinen (2004), from the Telecommunications Software and Multimedia Laboratory in Finland.

Those analysis and representation tools used singing voice artistic music production are always attached to transformation tools.

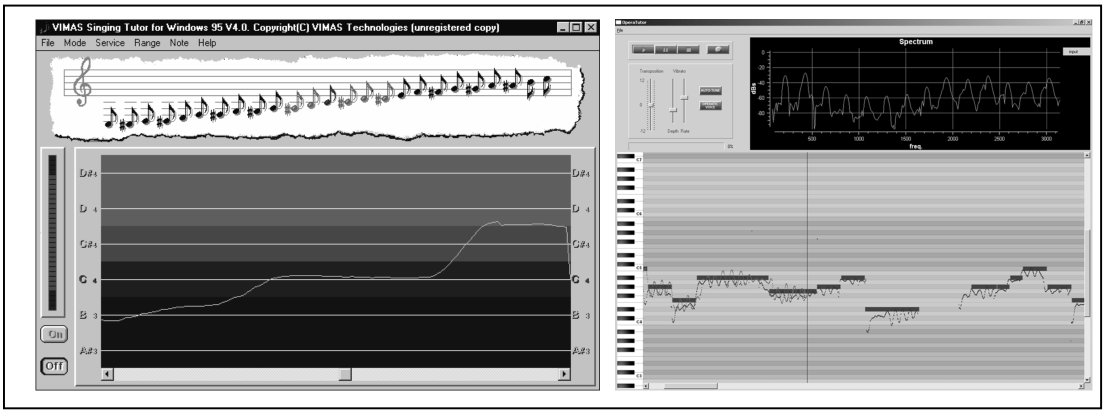
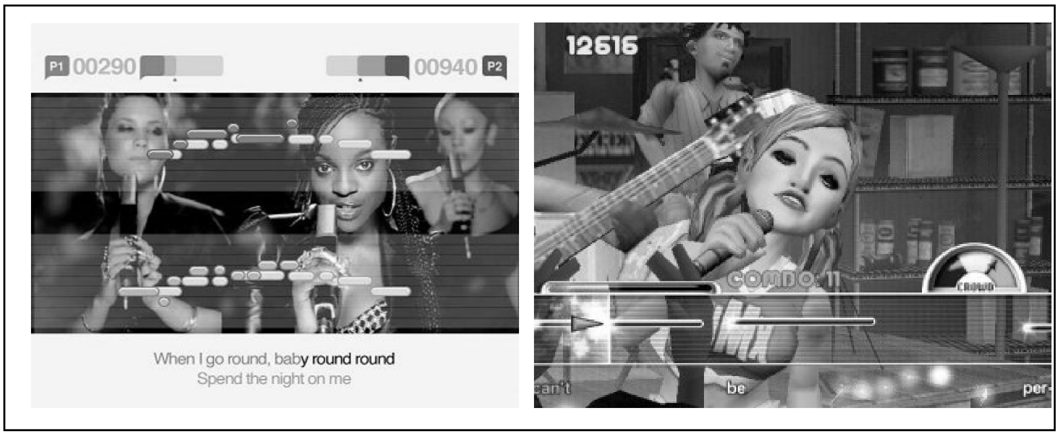

VIMAS Singing Tutor for Windows 95 V4.0. Copyright(C) VIMAS Technologies (unregistered copy)
File Mode Service Range Note Help
D#4
D 4
C#4
C 4
B 3
A#3
On
Off
Spectrum
freq.

Figure 13: Screenshots of (left) Singing Tutor[11] and (right) OpenDrama[12]



When I go round, baby round round
Spend the night on me

Figure 14: Screenshots of (left) SingStar[13] and (right) Karaoke Revolution[14]

2.2.1.2 Coding

The field of speech coding has a long history and represents one of the most active research areas in voice processing. Nevertheless there is only one relevant reference so far that tackles the coding from a singing voice exclusive point of view, see Kim (2003).

Kim proposed two main modifications on the CELP coder (see section 2.4.2.3.1) inside MPEG-4's structured audio scheme with the aim of improving the quality of coded singing for individual singers. On one side, in order to compensate the voiced / unvoiced mixture difference between speech and singing (see section 2.3.4), the proposal includes

[11] http://www.vimas.com/ve_str.htm

[12] http://www.iua.upf.es/mtg/opendrama/

[13] http://www.singstargame.com

[14] http://www.karaokerevolution.net/

a recomputation of the stochastic codebook using the residuals of real singing recordings after self-excitation coding. On the other side, in order to avoid the CELP pitch tracker finding multiples of the fundamental, the proposal also included an adaptation of the pitch analysis algorithm to the singing voice.

Although remarkably current coding schemes used in today's speech air communication systems are inefficient for singing voice transmission, one should expect to have singing enabled coders very soon as more transparent quality communication protocols are being used massively in the network.

2.2.1.3 Synthesis

Singing voice synthesis, together with voice transformation, has attracted attention from both the academy and the industry. Create a completely natural artificial singing engine is challenging topic for researchers and being able to use it is appealing for any open-minded musician.

Most probably, the first synthesizer ever specifically designed for the singing voice is the Music and Singing Synthesis Equipment (MUSSE) developed at the Royal Institute of Technology (KTH) in Stockholm, see Larson (1977). Since then, different virtual singing tools have been implemented and coexist today. Next, an overview of them is presented.

Chant[15]
Chant was developed at IRCAM by Xavier Rodet and Yves Potard; a review of the system can be found in Rodet et al. (1985). Chant uses an excitation resonance model. For each resonance, a basic response simulated with Formant Wave Functions (FOF) is generated, see Rodet (1984), is associated. Chant produces the resulting sound by adding the FOF corresponding to each formant for a given pseudo-periodic source. Chant synthesis results impressive in some cases although it is said they require from tedious manual tuning of parameters.

SPASM[16]
SPASM (see Figure 15) was developed by Perry Cook at Stanford University. The synthesizer is based on physical models and uses a waveguide articulatory vocal tract model and is described in Cook (1990,1992)

Lyricos[17]
Lyricos was developed by a team lead by Michael Macon at the Georgia Institute of Technology, see Macon et al. (1997). The synthesizer is based on an Analysis-by-Synthesis Overlap-Add sinusoidal model (ABSOLA), see George and Smith (1992).

Although Lyricos project was abandoned a few years ago, a tool called Flinger[18] was developed based on Macon's singing synthesizer and on the Festival speech

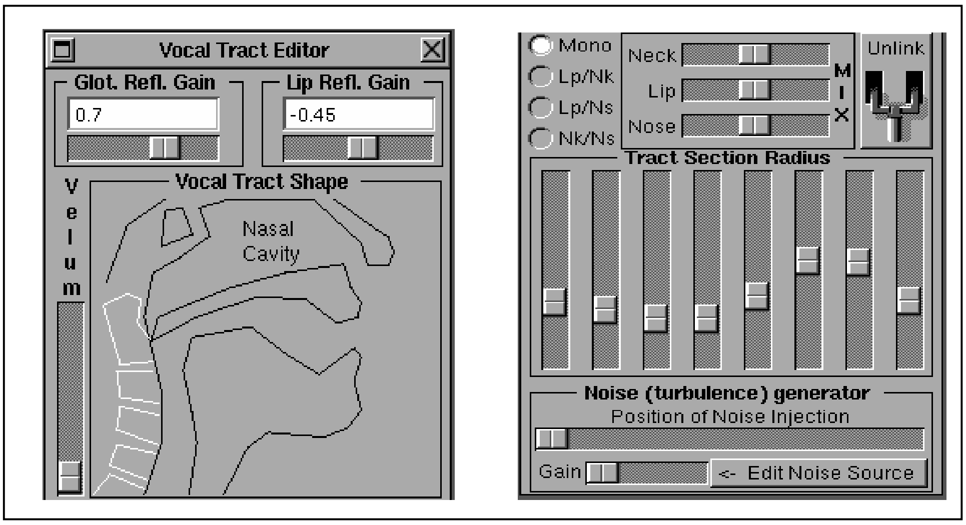
[15] http://musicweb.koncon.nl/ircam/en/artificial/chant.html

[16] http://www.cs.princeton.edu/~prc/SingingSynth.html

[17] http://www.lyricos.org/

[18] http://cslu.cse.ogi.edu/tts/flinger/

synthesis system, a general multi-lingual speech synthesizer developed at the Centre for Speech Technology Research of the University of Edinburgh.



Glot. Refl. Gain: 0.7

Lip Refl. Gain: -0.45

Neck

Lip

Nose

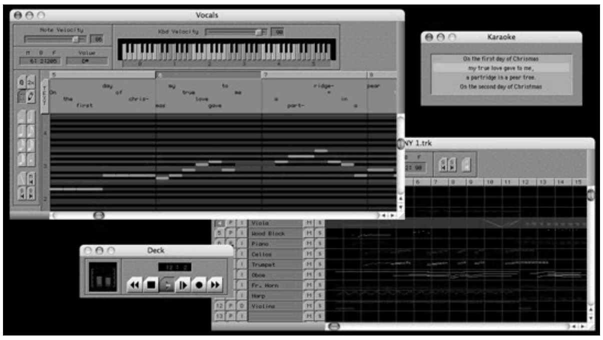
Position of Noise Injection

Gain

Figure 15: Perry Cook's SPASM user interface

VocalWriter[19]

Vocal Writer is a shareware singing synthesizer software for Macintosh. The interface looks like a typical MIDI sequencer interface with some functions added on top to allow the user edit the lyrics and shape the voice character and expression with controls such as brightness, vibrato, breath and others.



On the first day of Christmas
my true love gave to me,
a partridge in a pear tree.
On the second day of Christmas

Figure 16: Vocal Writer's user interface

MaxMBROLA[20]

MaxMBROLA is a real-time singing synthesizer for Max MSP based on the MBROLA speech synthesizer. Max MSP is a graphical environment for music, audio, and multimedia originally developed by Miller Puckette at IRCAM.

MBROLA, see Dutoit and Leich (1993), is a free multilingual speech synthesizer based on diphones concatenation available in Brazilian Portuguese, Breton,

[19] http://www.kaelabs.com/

[20] http://tcts.fpms.ac.be/synthesis/maxmbrola/

British English, Dutch, French, German, Romanian, Spanish, and Swedish. Both MBROLA and its singing extension MaxMBROLA have been developed by the group Théorie des Circuits et Traitement du Signal in Faculté Polytechnique de Mons, Belgium.

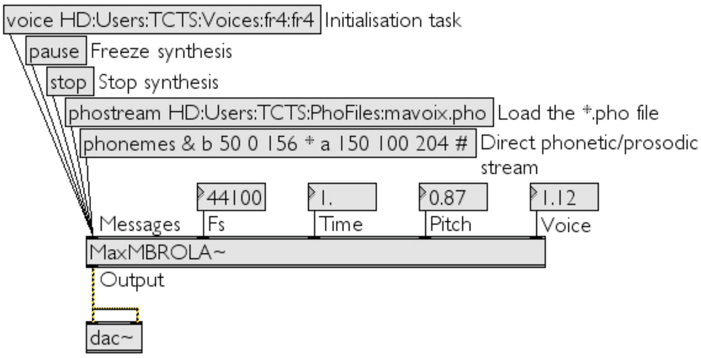

voice HD:Users:TCTS:Voices:fr4:fr4
Initialisation task
pause
Freeze synthesis
stop
Stop synthesis
phostream HD:Users:TCTS:PhoFiles:mavoix.pho
Load the *.pho file
phonemes & b 50 0 156 * a 150 100 204 #
Direct phonetic/prosodic stream
44100
1.
0.87
1.12
Messages
Fs
Time
Pitch
Voice
MaxMBROLA~
Output
dac~

Figure 17: A usage example of a MaxMBROLA instance

Harmony's Virtual Singer[21]

Harmony’s virtual singer is an opera-like singing synthesizer from Myriad Software. Its main attributes are the wide amount of languages that the synthesizer supports, the sound-shaping control (timbre and intonation), and the RealSinger function, which allows defining a Virtual Singer voice out of recordings of the user’s own voice.

Figure 18: Virtual Singer user interface

Vocaloid[22]

Vocaloid has been released as a product by Yamaha and is based on the research of the MTG, see Bonada et al. (2003). The synthesizer is based on the concatenation of real singer recorded diphonemes, after having them transposed and time scaled to fit the score specified by the user. More details on Vocaloid’s voice model and transformation and concatenation techniques can be found in later sections.

As of today, Vocaloid can be considered without any doubt the best singing voice synthesizer for popular music.

[21] http://www.myriad-online.com/vs.htm

[22] http:// www.vocaloid.com

SMARTTALK[23]

SMARTTALK 3.0 is a text to voice conversion engine that supports an additional singing feature. The synthesizer score is specified by allocating notes and words of a song in staff notation.

Figure 19: Smarttalk player interface

2.2.1.4 Transformation

High quality voice transformation is required in many different scenarios: broadcasters want to stress the low deep character of their male presenters; dubbers want to reach an effortless kid voice, journalists need to hide the voice of anonymous witnesses in their documentaries, etcetera.

For singing voice, the use cases are about sound surgery to customize people's vocal personality: musical producers want to tune the singer's vocal performance, performers want to harmonize their voice in real-time, Japanese karaoke users want to sound like Elvis, studio engineers want to emulate voice double-tracking to save time, and electronic artists want to alienate their voices with a unique effect.

Next we present an extensive overview of available applications for voice transformation. The list is divided in software applications (usually as plug-ins), and hardware.

2.2.1.4.1 Software

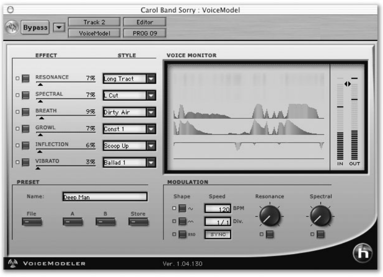
The majority of available solutions for voice transformation are software products as:

Voice Modeler[24]

Voice Modeler is a Powercore plug-in from TC-Helicon based on Overlap and Add technologies. Voice Modeler can tweak the apparent size or gender of the singer and to add growl, breath, specific vibrato styles, and inflection.

[23] http://www.oki.com/jp/Cng/Softnew/English/sm.htm

[24] http://www.soundonsound.com/sos/feb04/articles/tcvoice.htm



RESONANCE 7% Long Tract
SPECTRAL 7% L Cut
BREATH 9% Dirty Air
GROWL 7% Const 1
INFLECTION 6% Scoop Up
VIBRATO 3% Ballad 1
Name: Deep Man
Speed 120 BPM
Div. 1/1

Figure 20: Voice Modeler user interface

Voice Machine[25]
Voice Machine was released by Steinberg under a license of TC-Helicon technologies. The plug-in is backing choir oriented and allows the user to create up to four additional voices by triggering them via MIDI using the keyboard in real-time or drawing MIDI note events in a sequencer program.

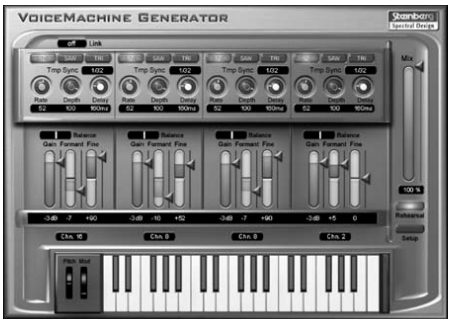


Figure 21: Voice Machine user interface

Vocal Rack[26]
Vocal Rack was released by Yamaha and assembles a set of vocal processing features including high pass filter, compressor, harmonic enhancer, three-band equalization, de-esser, gate and delay. None of the transformations available in Vocal Rack can be considered voice meaningful but generic and adapted to a singing voice input.

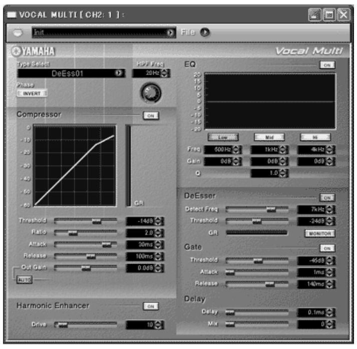


Figure 22: Vocal Rack user interface

[25] http://messe.harmony-central.com/Musikmesse01/Content/Steinberg/PR/VoiceMachine.html
[26] http://www.etcetera.co.uk/products/YAM072.shtml

Decca Buddy[27]

Decccca Budy was released by Akai under as a harmony generator. Just like voice machine it can generate up to four harmony voices and harmonies can be pre-programmed or can be played via MIDI in real time.

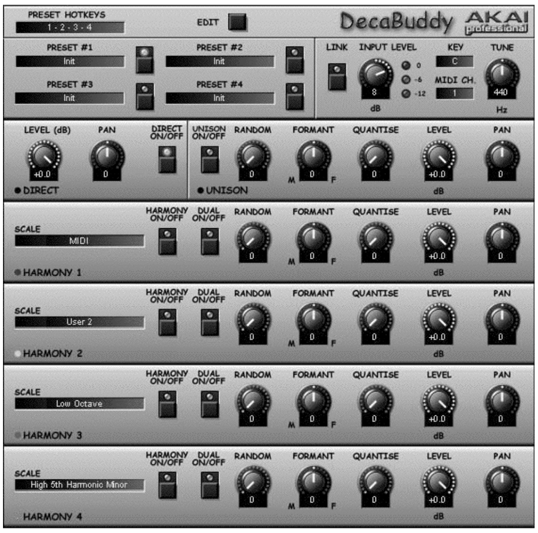


Figure 23: Decca Buddy user interface

Octavox[28]

Octavox was released by Eventide as a plug-in for Prootols. Octavox allows up to eight voice diatonic pitch shifting with individual delay adjustment and pan controls for each of them. A randomizer is included for control modulation in order to emulate the unsynchronized natural character of the synthetic voices. The user interface includes a score like control to draw the desired resultant melodies.

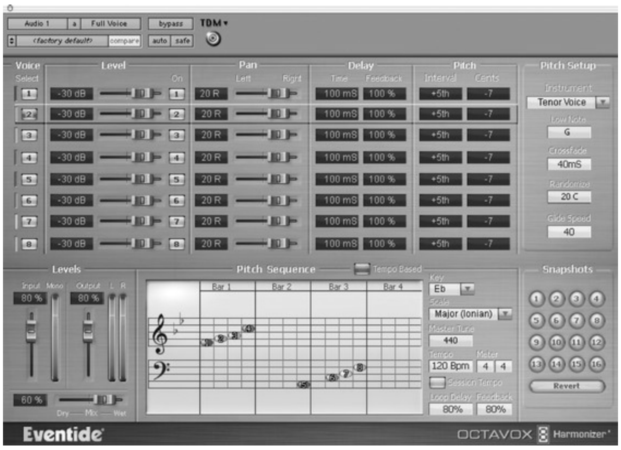


Figure 24: Octavox user interface

[27] http://www.macmusic.org/softs/view.php/lang/EN/id/127/

[28] http://www.pluginz.com/product/12489?vendor=383

PurePitch[29]

PurePitch was released by Sound Toys as a plug-in for Prootools. According to Sound Toys, PurePitch was the first pitch-shifter that could shift vocal formants (see section 4.2) offering enhanced control, naturalness, and flexibility on their vocal effects: harmony, unison, doubling, and gender and age change.

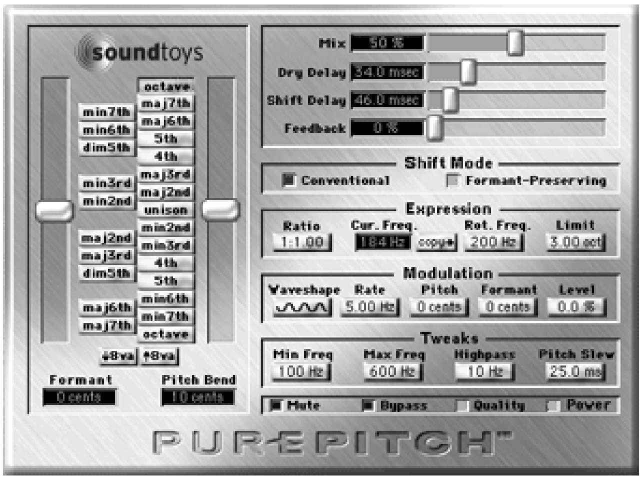

Mix 50 %
Dry Delay 34.0 msec
Shift Delay 46.0 msec
Feedback 0 %
Shift Mode: ☒ Conventional □ Formant-Preserving
Expression: Ratio 1:1.00, Cur. Freq. 184 Hz, Rot. Freq. 200 Hz, Limit 3.00 oct
Modulation: Rate 5.00 Hz, Pitch 0 cents, Formant 0 cents, Level 0.0 %
Tweaks: Min Freq 100 Hz, Max Freq 600 Hz, Highpass 10 Hz, Pitch Slew 25.0 ms
Formant 0 cents
Pitch Bend 10 cents
☒ Mute ☒ Bypass □ Quality □ Power

Figure 25: PurePitch user interface

THROAT[30]

THROAT, Antares' vocal toolkit, processes vocals through a physical model of the human vocal tract. THROAT can shape the characteristics of the modeled vocal tract as well as can modify the voice's glottal waveform or globally stretch, shorten, widen or constrict the modeled vocal tract. THROAT's breathiness controls can add variable frequency noise to the model, resulting in a range of vocal effects from subtle breathiness to full whisper. The plug-in also allows effects for voice doubling, choir, de-esser, and dynamic vocal compression.

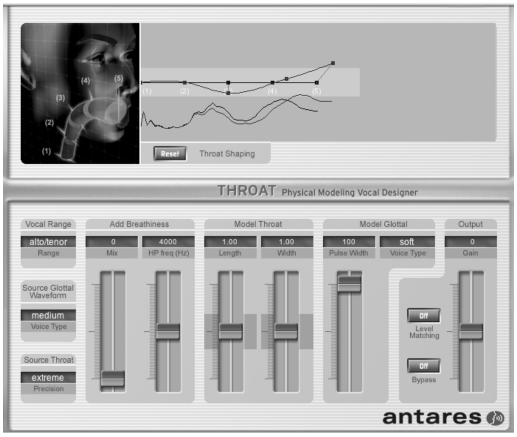

Vocal Range: Range alto/tenor
Source Glottal Waveform: Voice Type medium
Source Throat: Precision extreme
Add Breathiness: Mix 0, HP freq (Hz) 4000
Model Throat: Length 1.00, Width 1.00
Model Glottal: Pulse Width 100, Voice Type soft
Output: Gain 0
Level Matching Off
Bypass Off

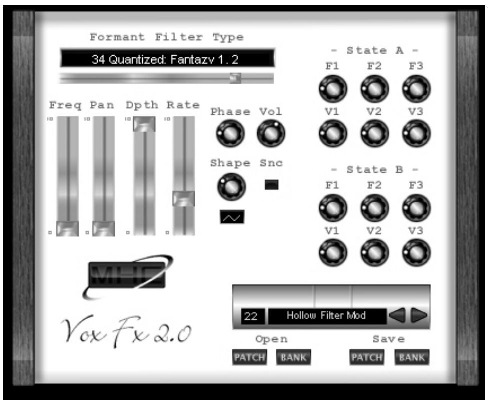
Figure 26: THROAT user interface

[29] http://www.soundtoys.com/products/PurePitch/

[30] http://www.antarestech.com/products/avox.shtml

Vox FX[31]

Vox FX, released by MHC is a plug-in base on formant filter presets. Each filter is supposed to define a certain vocal personality and can be combined with others to create new effects. Filters change the energy calibration on the different formant spectral bands.



Formant Filter Type
34 Quantized: Fantazv 1. 2

Figure 27: Vox FX user interface

Vokator[32]

Vokator is a vocoder based plug-in released by Native Instrument's.

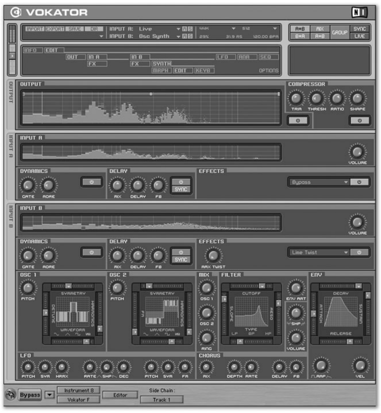


Figure 28: Vokator user interface

AV Voice Changer[33]

AV Voice Changer is a stand alone application for real time voice transformation in environments such as voice chat, voice-over-IP, and online gaming. Voice Changer allows transformation on two vocal attributes: pitch and timbre.

[31] http://www.mhc.se/software/plugins/voxfx/

[32] http://www.soundonsound.com/sos/sep03/articles/vokator.htm

[33] http://www.audio4fun.com/

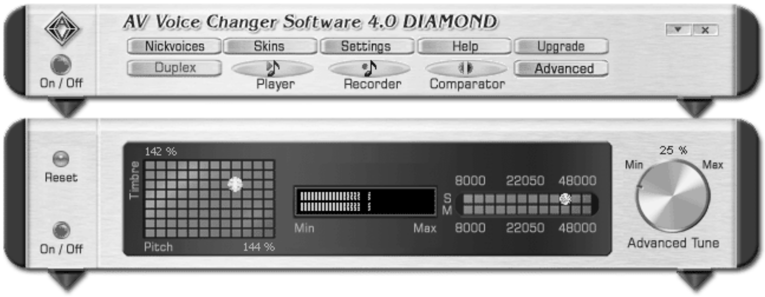


Figure 29: Voice Changer main menu screenshot

Clone Ensemble[34]

Clone Ensemble is a plug-in for choir emulation. Clone Ensemble can generate a room full of up to 32 replicas of the input singer in unison or octaves. Gender change is also possible only in vocals.

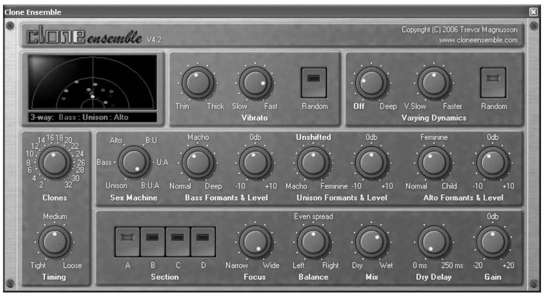


Figure 30: Clone Ensemble user interface

Voxengo's Voxformer[35]

Voxformer is a multi-functional vocal channel strip plug-in featuring compression, de-esser, equalization, vintage saturation and others. Just like Voice Machine, none of the transformations available in Voxformer are truly voice specific.

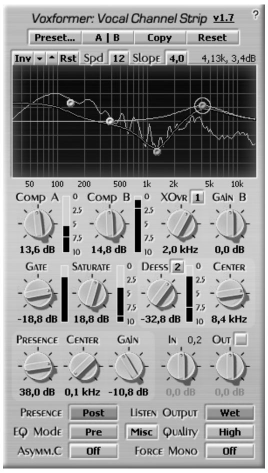


Figure 31: Voxformer user interface

[34] http://www.cloneensemble.com/

[35] http://www.voxengo.com/product/voxformer/

Antares Autotune[36]

Autotune from Antares is one of the most popular ever plug-ins for tuning the singing voice. It was the first product to offer high quality pitch correction in real time using key and scale controls. It also allows fine tuning of the pitch envelope in an off line mode. Nearly all recording studios in the world use Antares Autotune for postprocessing the vocals. In fact the name of the product, Autotune, has given name to a popular effect described in section 4.4.1.2.

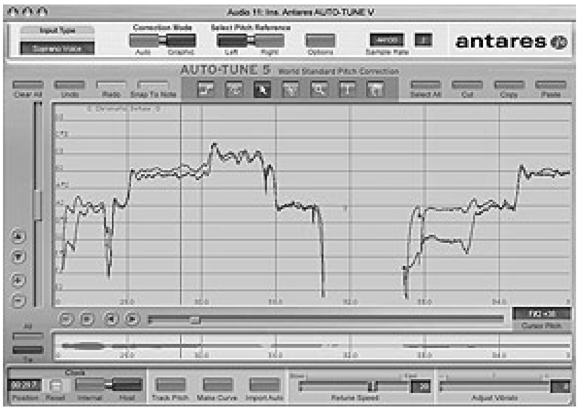


Figure 32: Autotune user interface

Melodyne[37]

Celemony's Melodyne is a multitrack audio recording and editing that allows the processing of the files in a musically intuitive way. Melodyne represents tracks splicing the waveform view into notes, placing each note at the corresponding quantized note, and drawing on top the pitch envelope. The user can transform melodies by changing the duration or the height of each note in a very intuitive way and with natural sounding results. Although Melodyne deals with any kind of pseudo-harmonic audio source, it is widely used in vocals postproduction.

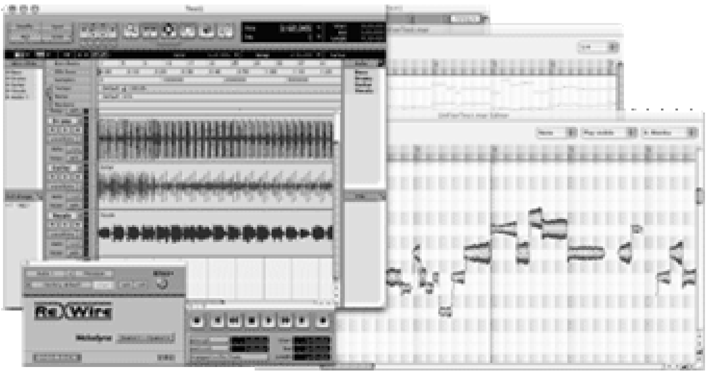


Figure 33: Melodyne multiple displays

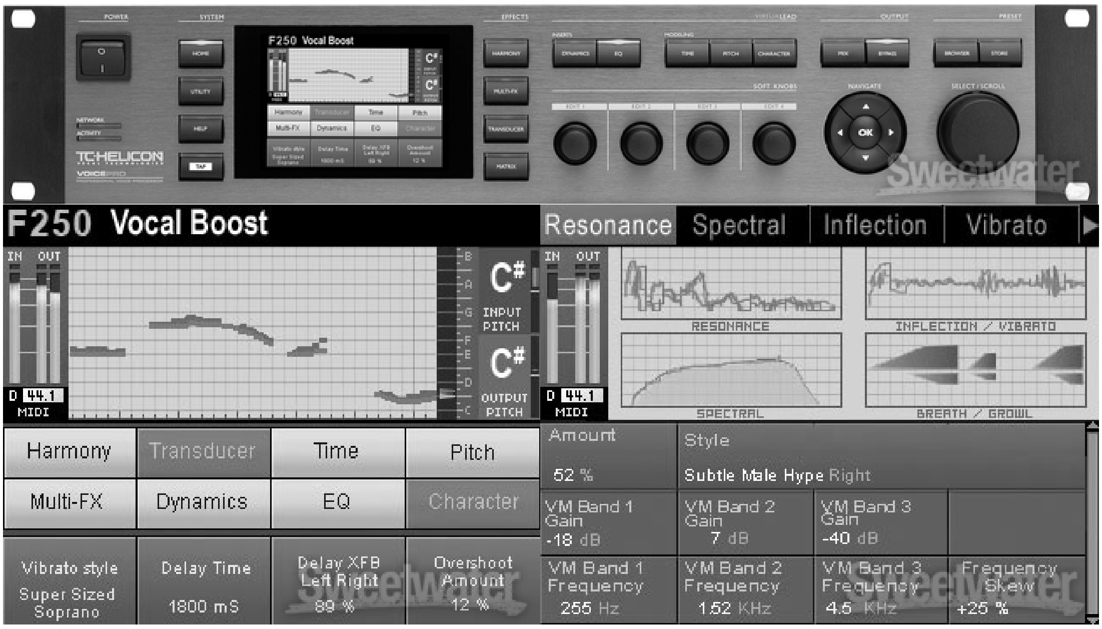
[36] http://www.antarestech.com/products/auto-tune5.shtml

[37] http://www.celemony.com/melodyne/

2.2.1.4.2 Hardware

Hardware for the transformation of the voice include:

VoicePro[38]
Considered by most the richest voice transformation tool in the market, the VoicePro from TC Helicon provides the tools necessary for vocal production including harmony generation, intonation correction, pitch and time manipulation, doubling, formant changing and classic and special effects.



F250 Vocal Boost

Vibrato style: Super Sized Soprano
Delay Time: 1800 mS
Delay XFB Left Right: 89 %
Overshoot Amount: 12 %

Amount: 52 %
Style: Subtle Male Hype Right
VM Band 1 Gain: -18 dB
VM Band 2 Gain: 7 dB
VM Band 3 Gain: -40 dB
VM Band 1 Frequency: 255 Hz
VM Band 2 Frequency: 1.52 KHz
VM Band 3 Frequency: 4.5 KHz
Frequency Skew: +25 %

Figure 34: VoicePro front panel and screen views

Voice Transformer[39]
Voice Transformer by Boss allows you to shape vocal timbre, tone, pitch, and formants.

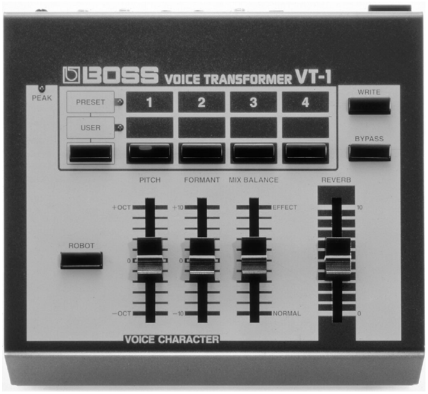


Figure 35: Voice Transformer pedal

[38] http://www.tc-helicon.com/
[39] http://www.zzounds.com/item--BOSVT1

Digitech Vx400[40]

Vx400 vocal effects processor offers the choice of multiple vocal character selections, modeling of 16 different pro microphones.

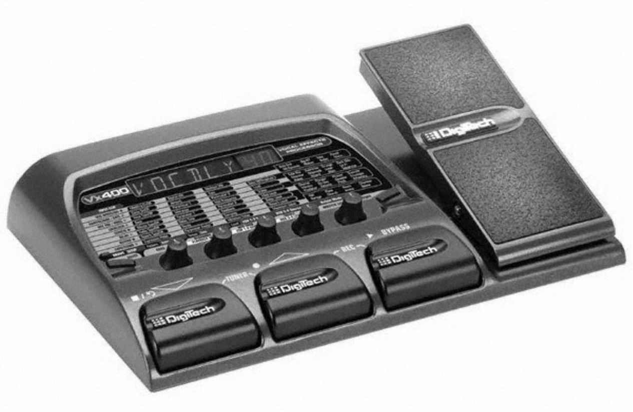


Figure 36: Digitech Vx4000 pedal

VP-70[41]

Roland's VP-70 Voice/instrument Processor contains four pitch shifters and a pitch-to-MIDI conversion circuitry. Pitch shifters can be controlled in real time by the internal memory or by external MIDI messages.

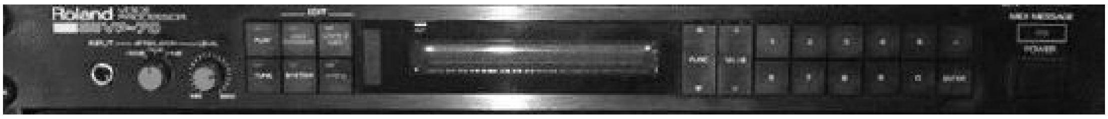


Figure 37: Voice processor rack

PLG100-VH[42]

Yamaha's PLG100-VH provides harmony and vocoder effects to your voice. Harmonizer can give up to three-part harmonies and can be controlled by internal memory programs or by real-time MIDI external messages. The card also includes gender change and vibrato.

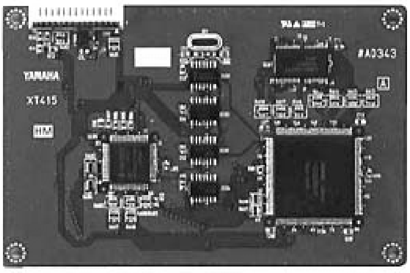


Figure 38: PLG100-VH extension card

[40] http://www.music123.com/Digitech-Vx400-i91029.music

[41] http://www.sonicstate.com/synth/roland_vp70.cfm

[42] http://www.yamaha-europe.com/yamaha_europe/uk/10_musical_instruments/70_synthesizer/40_plugin_boards/10_plugin_boards/10_no_series/PLG100_VH/

2.2.2 Singing voice in music production

Perry Cook defines technology as “any intentionally fashioned tool or technique”. According to such definition, our brief chronicle restricts to those technologies in which the tool derives from magneto-mechanical and electronic devices and leaves out singing techniques, primitive cupped hands amplification, drugs, castration and many others.

In 1948 Bing Crosby gave his friend Lester Polfus, most popularly known as Les Paul, one of the first production units of the Ampex Model 200 reel-to-reel tape recorder. Paul managed to modify the tape adding additional recording and playback heads to be able to simultaneously record a new track while monitoring the playback of a previously recorded one. With the hacked Ampex, Les Paul achieved what could be considered the first artificial effect applied to voice: doubling, an effect that remains popular today. Using Paul’s invention, Patti Page recorder her own “Confess”, a song that required one singer to reply another, and its released was announced as ‘Patti Page and Patti Page’.

After Paul’s experiments, Ampex released the first commercial multitrack recorder in 1955. Such machines allowed having one microphone exclusively dedicated to the singer track. Accordingly, the microphone would not any longer be placed at two feet from the singer to collect all band sound, but could be placed at two inches of the mouth. This was a key issue for successful recordings of soft singing artists such as Bing Crosby or Nat ‘King’ Cole.

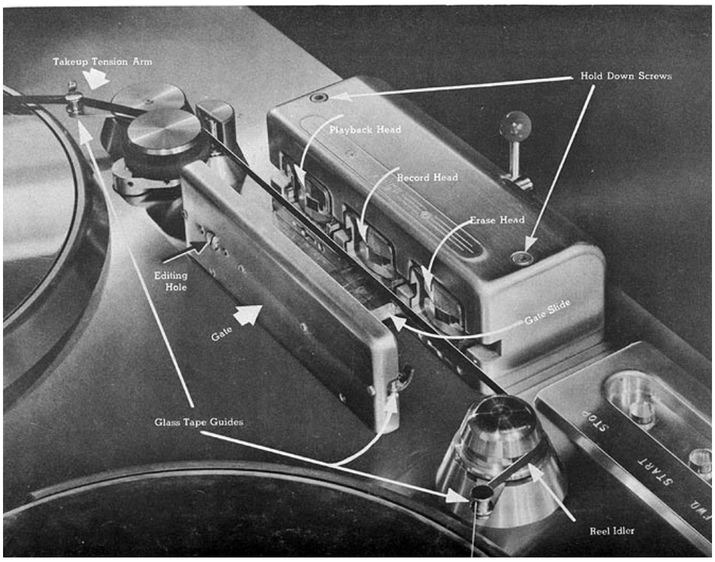


Figure 39: Model 200A head block and tape path, scanned from the original manual

Les Paul kept on experimenting with the tape recorder and he soon after came up with the ‘echo / delay’ effect, by placing the playback head behind the record head. With this, time delay could be adjusted by modifying the distance between heads; the closer the heads, the shorter the time delay. Such effect was first applied to a voice in a recording of Paul’s wife, Mary Ford, performing “How high the moon” in 1950 but the effect, on its shorter time delay version, called ‘slap back echo’, became popular by the hand of Elvis Presley, who used it as a voice effector in many of his early Sun Record recordings.

A few years later, in the late 1950’s, early 1960’s, another voice effect became popular under the name of ‘echo chamber’, which originally consisted in artificially using the acoustics of an architectural space by placing an emitting loudspeaker and a recording microphone on different points of the chamber. Most famous echo chambers include those in Atlantic Studios, the stairwell in Columbia Records, and EMI’s air conditioning pipe system.

In April 1966, Abbey Road Studios engineer Ken Townshend invented Automatic Double Tracking (ADT) using linked tape recorders to automatically create a synchronized duplicate of a lead vocal or an instrument. ADT allowed modulations on the delay that was being applied to the track and gave birth to those effects today know as ‘flanger’, ‘chorus’ and ‘vibrato’. Most famous band in those days, The Beatles, was delighted with Townshend's invention and used it routinely thereafter. One of the first instances of the flanging effect, named ‘flanger’ after John Lennon, on a commercial pop recording was the Small Faces' 1967 single “Itchykoo Park” but significant examples of ‘flanger’, ‘chorus’ or ‘vibrato’ effects applied to the lead singing voice can be found in Beatles’ ”Tomorrow never knows” from 1966 and “Blue Jay Way” from 1967, or in Bowie’s “The man who sold the world” from 1970.

During the 60’s and 70’s decade, music production was vastly impregnated by the exploration and experimentation that was up in the air those days in most of the artistic disciplines. Based on technologies that preceded the integrated circuit, different types of effects were tried and successfully established as vocal transformations. Limiters and compressors were applied in extreme configurations in Beach Boys “All I want to do” from 1969 and in most of the garage rock production, in which even sometimes shouting was used to achieve microphone saturation sound as in The Stooges or MC5. Reversed vocals were occasionally used as background ambient and major delays were applied to lead voice as in Beatles’ “Being for the benefit of Mr. Kite” in 1967 or Pink Floyd’s “Dark side of the moon” in 1973.

Also in the 70’s an engine named Vocoder, originally developed in 1939 by Homer.W. Dudley at Bell Laboratories for research in compression schemes to transmit voice over copper phone lines, turned into one of the most celebrated voice effects. The Vocoder (Voice Operated reCOrDER) measured spectrum energy levels along time via a bank of narrow band filters. These energy envelopes were afterwards supplied to a feedback network of analytical filters energized by a noise generator to produce audible sounds. Werner Meyer-Eppler, then the director of Phonetics at Bonn University, recognized the relevance of the machines to electronic music after Dudley visited the University in 1948, and used the vocoder as a basis for his future writings which in turn became the inspiration for the German "Electronische Musik" movement. Most significant examples of vocoder being applied to singing voice are Walter Carlos “Timesteps” from 1971’s Clockwork Orange soundtrack and Kraftwerk’s “Man Machine” from 1978

Once proven semiconductor devices could perform vacuum tubes function, mid-20th-century technology advancements in semiconductor device fabrication allowed the integration of large numbers of tiny transistors into integrated circuit (IC) chips. The chips, with all their components, are printed as a unit by photolithography and not constructed a transistor at a time. The IC's mass production capability, reliability, scalability and low price prompted the use of standardized ICs in place of designs using discrete transistors which quickly pushed vacuum tubes into obsolescence. Among the most advanced integrated circuits are the microprocessors, which control everything from computers to cellular phones. And ever since the apparition of computers of new generation there has been a wide proliferation of software solutions that are available for music production of vocals at whatever quality. New software started releasing digital replicas of each and every preceding voice effect but almost immediately after established as the source of a vast number of new vocal transformations. These new transformations, which in fact are the main subject of study of this work, are most of the times implemented on top of models and meaningful parameterizations of the voice. To name some, a step-like pitch quantification effect has become very popular in the last few years with the name of 'auto-tune' since Cher and Roy Vedas used it for the first time in "Believe" and "Fragments of live" respectively in 1998. Other alienating voice effects such as morphing voice with tapping guitar or mp3-alike sound can be heard in the pop-electronic music released by contemporary artist Daft Punk.

Powerful computers, reasonably good and cheap sound cards, and simple availability of digital audio software have spread, democratized, and in some ways, globalized music production up to the point where bedroom musicians all around the world produce their different style songs with latest cutting edge technologies and using powerful voice enhancement and transformation software.

2.2.3 Running Out of Melodies?

In the late 80's, according to the author hypothesis, music industry demand of good selling artists was not being covered at the rate it was required/dictated by record companies ambitions. It is not easy to find outstanding composers with an extraordinary voice. Some of the industry players tried to solve the problem with handicapped talent artists whose main showcased attribute or singularity was not their music but their beauty, popularity, addictions, dancing skills, or others. World hit songs such as Macarena from Los del Rio or Aqua's Barbie Girl drew button-down melodies but were fun to dance. At the same time some companies tried to replace artistic shortcomings with compensating technologies that could generate new sounds, tune melodies, enhance vocal performance, place beats at tempo, or create original sonic spaces, bringing about a talent shift that moved creativity from the artist to the artistic producer. As of today, voice production has reached a top of sophistication in which neither an inborn unique voice nor an original melody is required for success anymore. Artists such as Madonna or Enrique Iglesias do not inherit exceptional vocal qualities but they are good-looking and moreover, their vocal performances are very carefully and complexly produced.

Moreover, an important issue arises in the discussion: no matter how you represent a melody, if the representation is quantized, the number of possible finite melodies is finite. This could mean that the later you take the chance to create a melody; the more difficult it will be to come up with an original tune that does not resemble any of the preceding. Such extreme consideration has a relevant importance when the analysis focuses on western commercial popular music, where melodies are usually repeated every minute the longer, and fitted to dodecaphonic temperate scale, to certain tempo and key, and to people's ear pleasantness. Only in EEUU one hundred albums of such kind of music are released every day [ ] and in the UK alone, tens of thousands of new songs are published every year. In this context, we can assume melodic differentiation has become a task harder and harder to achieve along years. Probably somehow aware of this fact, music companies and artists have unbalanced the established weights to focus their differentiation on the sound and not anymore on the score, increasing enormously the emphasis of sound manipulation in musical production.

Both of the aforementioned hypotheses (lack of talent compensation and western melody saturation) complement each other when trying to reveal what caused vocal production to be enthroned in the way it is nowadays. Besides, the hypotheses do not exclude simultaneous evolutions in which artists took profit of new sound manipulation and enhancement tools to create new artistic canons, but they discard them as a major development force.

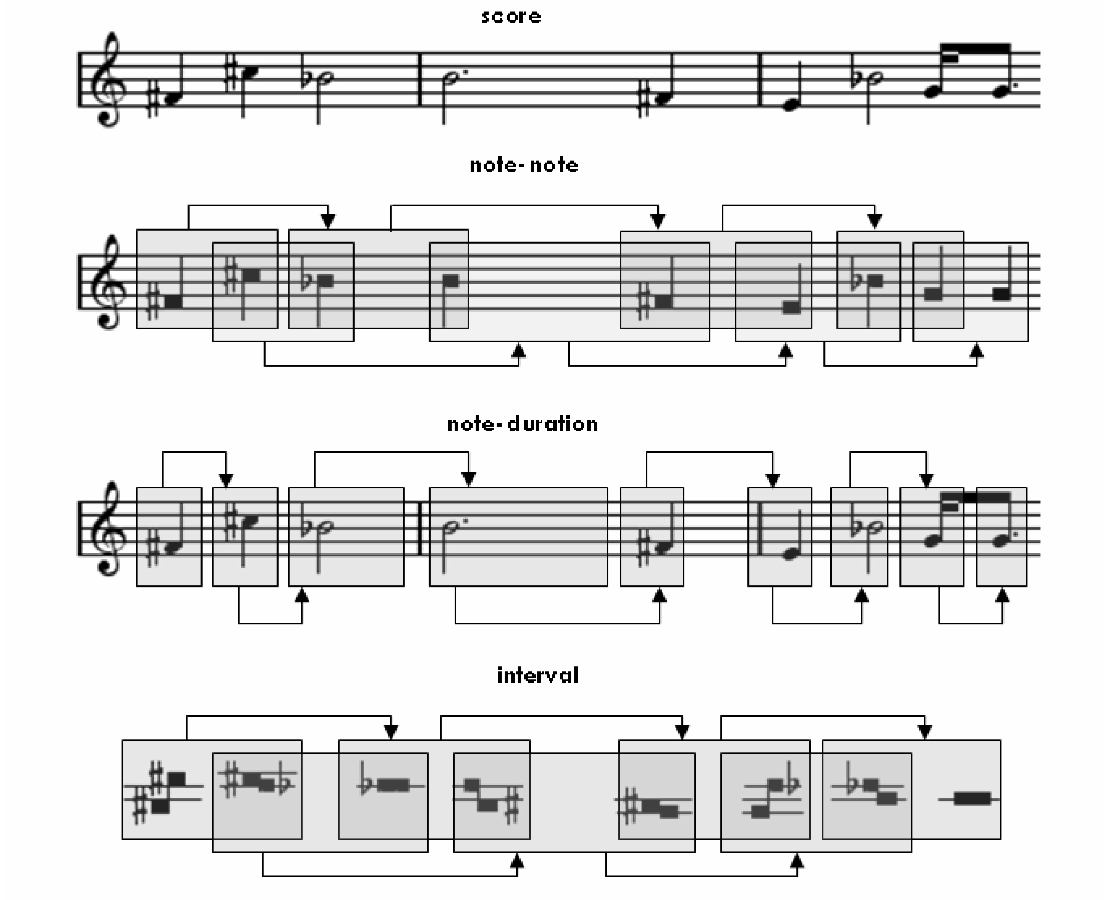


Figure 40: Representation of the different definitions of node and link

In order to prove the consistency of the western melody saturation hypothesis, we carried out a set of experiments for the study of melody exploitation in western popular music using complex network analysis tools.

Researchers divide melodies into at least six components: pitch (the notes in the melody), musical intervals between the notes, key, contour (how the melody rises and falls), rhythm (the relative lengths of notes and silences), and tempo (the speed at which a melody is played).

For the experiments three different lexicons were defined by means of note duplets, note and duration pairs, and interval duplets (see Figure 40). For the note-note network, we consider notes as interacting units of a complex system. Accordingly notes are nodes and a link exists between notes if they co-occur in a music piece. The connection of this set of notes allows constructing an unlimited number of melodies of which the collection of melodies of our corpus is a small sample. Analogous procedures are used to create note-duration and interval complex networks.

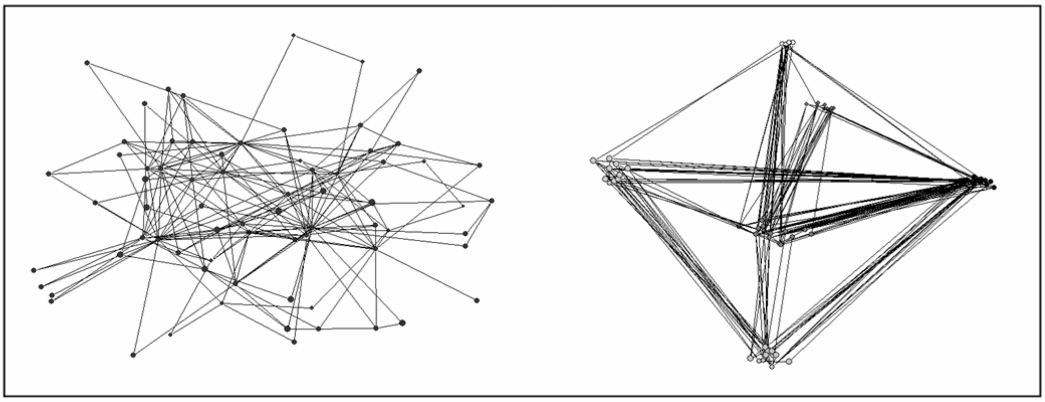


Figure 41: Left figure displays a melody by Michael Jackson using the note-duration representation. Right figure displays the same network manually arranging the nodes corresponding to the same note.

The analysis were run over a music corpus consisting of over 13000 western contemporary music pieces (in MIDI format) covering a broad spectra of music styles and reveals that music shares properties with other complex networks, namely the small world phenomenon as well as a scale-free degree distribution.

Zanette reported that music follows the Zipf's law, see Zipf (1949), the frequency of notes decays as a power function of its rank. It suggested that this process was due, similarly to human language, to the formation of context; a note was more likely to appear on a musical piece if it had previously appeared. However, the music networks obtained from our analysis display complex patterns different from the language, at least at this representation level, revealing differences in the two systems.

Some music network properties, such as non trivial correlation $K_{nn}(k)$ (average degree of the site neighbors of one site whose degree is k) and the clustering coefficient $C(k)$, are not explained by the rich get richer mechanism alone. We compared the network resulting from the interaction of tri-phonemes, syllables and words in Cervantes' "*El Quijote*" and it was found that music, opposed to language, shows a clustering coefficient that increases as a function of degree, and that music networks are assortative as opposed to language dissortativeness. A network is considered assortative when $K_{nn}(k)$ increases along k, that is when nodes with a certain degree of connectivity tend to connect with others with similar degree, a typical attribute of social networks.

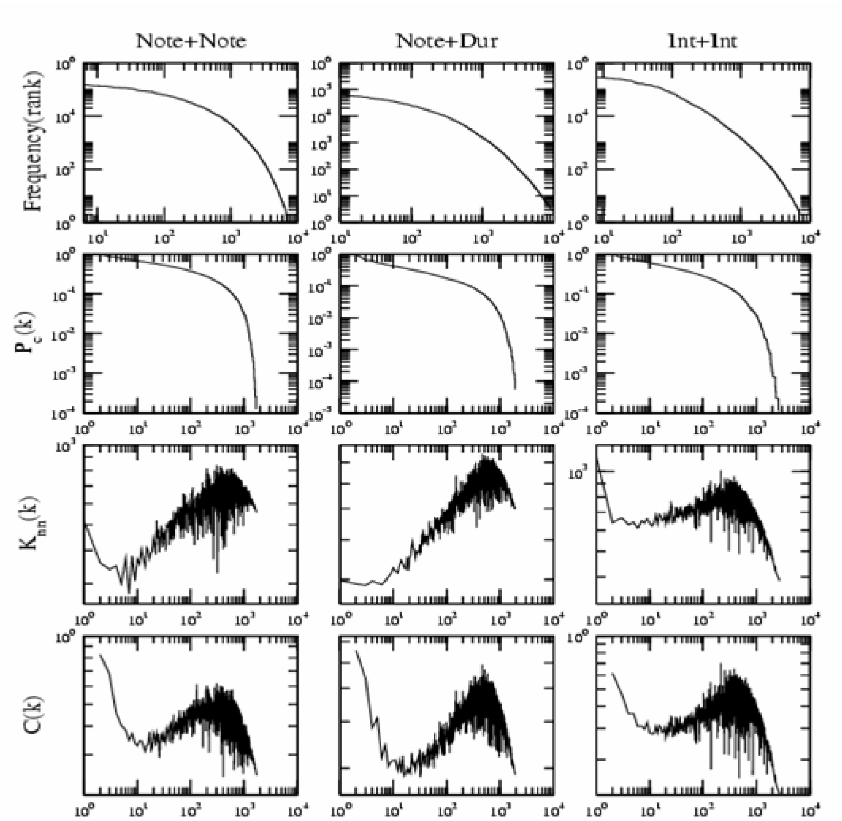


Figure 42: Plots of *C*(*k*), *Knn*(*k*), *Pc*(*k*) and *frequency*(*rank*) for all three different lexicons (note-note, note-duration, and intervals)

Just like social networks, where a friend of someone with a lot of friends will most probably have a lot of friends his own, the properties of the musical networks tells us, whatever lexicon we choose to the define the nodes, that most frequent melodic atoms belong to a very crowded and interconnected community while rare links belong to the ghettos.

Which complex network model is the most appropriate to explain the discovered properties is something that still remains uncertain. However, results prove the consistency of the western melody saturation hypotheses presented in this section.

2.3 Voice production

Despite all singularities that differentiate singing from speech, both vocal utterances originate from the same production mechanism.

2.3.1 Anatomy of the voice

Voice production begins when air is inhaled. The inhalation expands the volume of the lungs as air rushes in and causes the diaphragm (the muscle below them) to lower. When exhale, the air in the lungs is pushed out by muscle force of the rig cage lower and excites the vocal mechanism through the bronchi, trachea, and larynx.

When the adductor muscles (the "*vocal cord closers*") are activated (vocal folds are tensed), the airflow resistance causes them to vibrate. Air then bursts through the closed vocal cords. The so called Bernoulli Effect makes the vocal chords being blown apart and sucked backed together hundreds of times per second chopping the airflow into quasi-periodic pulses. This sound, created at the level of the vocal cords, is then shaped by muscular changes in the pharynx (throat) and oral cavity (including the lips, tongue, palate, and jaw) to create speech.

When the vocal folds are relaxed, in order to produce a noise, the airflow either must pass through a constriction in the vocal tract and thereby become turbulent, producing so-called unvoiced sound, or it can build up pressure behind a point of total closure within the vocal tract (e.g. the lips), and when the closure is opened, the pressure is suddenly released, causing a brief transient sound.

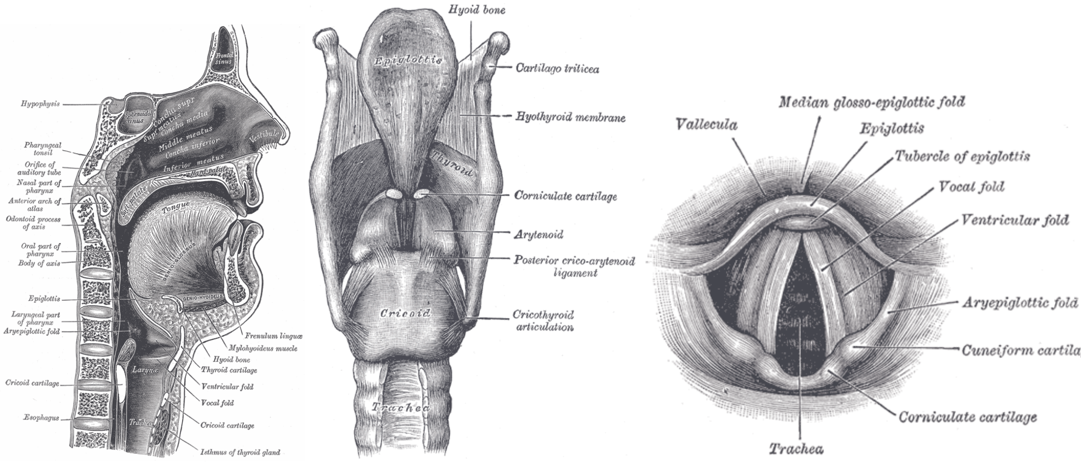


Figure 43: (left) mouth and neck section view, (middle) posterior view of the ligaments of the larynx, and (right) laryngoscopic view of the vocal folds

The larynx is a tube shaped structure comprised of a complex system of muscle, cartilage, and connective tissue. The larynx is suspended from the hyoid bone, which is significant in that it is the only bone in the body that does not articulate with any other bone. The larynx houses the vocal cords, two elastic bands of tissue (right and left) that form the entryway into the trachea (airway). Above and to the sides of the true vocal cords are the false vocal cords, or ventricular cords. The false vocal cords do not usually vibrate during voicing, but are often seen coming together (adducting) in individuals with muscle tension dysphonia, a common voice disorder characterized by excessive muscular tension with voice production. The true vocal cords open (abduct) when we are breathing and close (adduct) during voicing, coughing, and swallowing.

The third unpaired cartilage of the larynx, the epiglottis, can invert to direct food and liquid into the esophagus and to protect the vocal cords and airway during swallowing. The arytenoids paired cartilages are a point of attachment for the vocal cords and thus allow the opening and closing movement of the vocal cords necessary for respiration and voice.

There are two primary groups of laryngeal muscles, extrinsic and instrinsic. The extrinsic muscles are described as such because they attach to a site within the larynx and to a site outside of the larynx (such as the hyoid bone, jaw, etc.). The intrinsic muscles include the interarytenoid, lateral cricoarytenoid, posterior cricoarytenoid, cricothyroid, and thyroarytenoid (true vocal cord) muscles. All of the intrinsic muscles are paired (that is, there is a right and left muscle) with the exception of the transverse interarytenoid. All of the intrinsic laryngeal muscles work together to adduct (close) the vocal cords with the exception of the posterior cricoarytenoid, which is the only muscle that abducts (opens) the vocal

2.3.2 The source-filter model

The most common approach to model the voice production system is based on a source-filter decomposition assumption. According to the source filter conception, the voice excitation can be rather voiced, unvoiced, or a combination of both. The unvoiced excitation corresponds to the turbulent airflow that arises from the lungs and the voiced excitation corresponds to the glottal pulses that originate the vocal fold vibrations. The voice filter is characterized by a set of resonances called formants that have their origin in the aforementioned voice organs lengths and shapes: trachea, esophagus, larynx, pharyngeal cavity, velum, hard palate, jaw, nasal cavity, nostril, lip, tongue, teeth, and oral cavity).

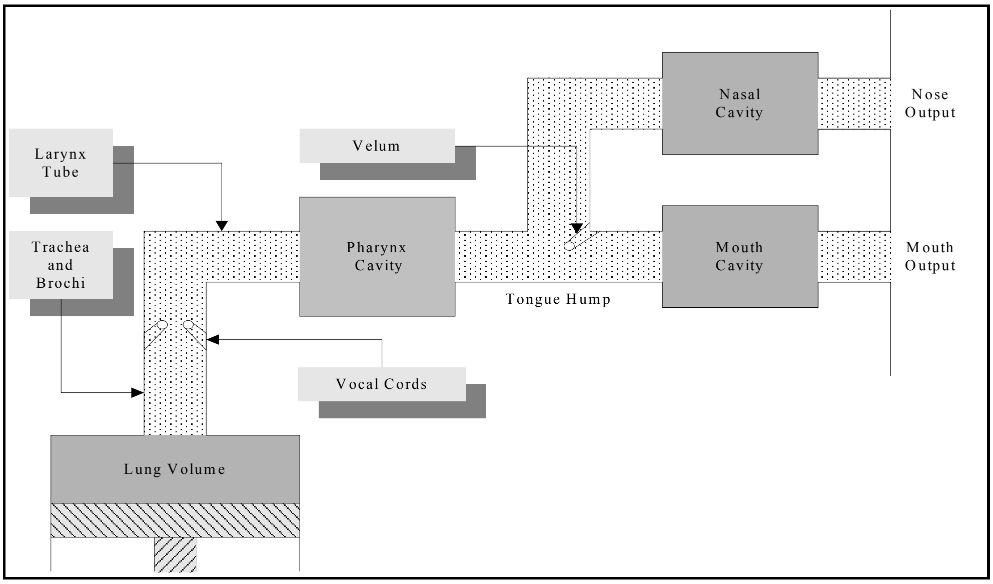


Figure 44: Schematic representation of the speech production system after Flanagan (1972)

The source-filter acoustic theory of speech production states voice can be modeled as an acoustic source signal (noise for unvoiced and a periodic pulse-train for voiced) filtered by a dynamic filter that emulates the vocal tract (supra laryngeal filter), see Fant (1982). On top of that, the linear speech production model, Flanagan (1972), assumes for the voiced speech *S(z)* that the source *U(z)* is a delta train and the filter is the cascade of a glottal pulse filter *G(z)*, the vocal tract filter *V(z)*, and the lip radiation filter *L(z)* as represented in Figure 45.

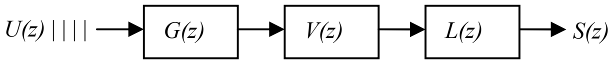


Figure 45: Voiced speech production diagram

According to the source-filter theory of speech production lip radiation *L(z)* can be modeled as a derivative operator that applies to the produced acoustic signal, meaning the derivative of the glottal flow is the effective excitation of the vocal tract, see Fant (1982). Lip radiation filter *L(z)* is usually given as:

$$L(z) = 1 - \mu \cdot z^{-1} \text{ with } \mu \approx 1 \text{ and } \mu < 1 \qquad (1)$$

and the glottal-pulse filter *G(z)* as:

$$G(z) = \frac{1}{\left(1 - e^{-2\pi \cdot f_c \cdot T_s} \cdot z^{-1}\right)^2} \qquad (2)$$

where T_s is the sampling period and f_c the cut-off frequency, which is usually set to around 100 Hz.

When U(z) is voiced, the glottal waveform presents a typical period shape, shown in Figure 46. Glottal waveforms are usually emulated in synthesis using mathematical functions that fit the waveform behavior along time. Started by Rosenberg in 1970, see Rosenberg (1970), the most relevant model is the one by Fant, Liljencrants, and Lin, see Fant et al. (1985).

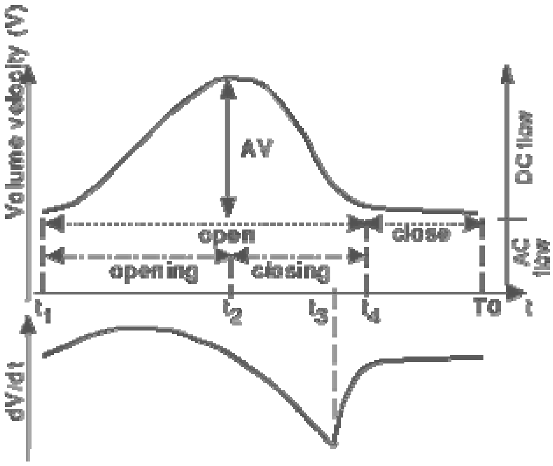


Figure 46: Hand drawn representation of a glottal waveform *U(z)* and its time derivative taken from [w] where *T0* is the duration of the period, *t*1 is the beginning of the separation of the vocal folds and onset of the airflow, *t*2 is the instant of the maximum glottal flow through the glottis with maximum amplitude *AV* , *t*3 is the moment of glottal closure, and *t*4 is the instant of complete glottal closure and when no airflow occurs. Taken from www.ims.uni-stuttgart.de/phonetik/EGG/page13.htm with permission of author

Important parameterization of the glottal flow is given by the Open Quotient and the Speed Quotient. The values of these parameters are highly correlated to different types of phonation.

The Open Quotient (OQ) indicates the duty ratio of the glottal airflow and is defined as the ratio of the time in which the vocal folds are open and the whole pitch period duration (*(t*4-*t*1*)/T0* in Figure 46). Substantial changes in the spectrum of the voice excitation imply notable ratio variations.

The Speed Quotient or skewness is an indicator of the glottal pulse asymmetry and is defined as the ratio of rise and fall time of the glottal flow ((*t*2-*t*1)*/(t*4-t_2) in Figure 46). The glottal airflow is usually skewed to the right, which means that the decrease of the airflow is faster than its increase, see Rothenberg (1981), Ananthapadmanabha (1984), and Titze (1988).

2.3.3 Voice Production Acoustics

In the spectrum, the magnitude of the harmonics gradually falls. For normal voicing this decrease is about -12 dB / octave, see Klatt (1980). When the vocal folds are relaxed, the spectrum of turbulent noise source is stochastic with a slight magnitude decrease of -6 dB / octave, see Klatt (1980). An example of this kind of source spectrum is shown in Figure3.6.

The produced source spectra *S(f)*, regardless it is voiced and unvoiced, is then modulated in amplitude in passing through the pharynx (the throat cavity), the mouth cavity, and possibly the nasal cavity. Depending on the positions of the various articulators (i.e., jaw, tongue, velum, lips, mouth), different sounds are produced. This modification originated by the different cavities and articulators is called supralaryngeal filter, see Lieberman and Blumestein (1988), because it can be modeled as a linear system with a transfer function *T(f)*, modifying the source spectrum *S(f)*.The modified spectrum *U(f)* can be written as $U(f) = S(f) \cdot T(f)$.

During the production of human speech the shape of the supralaryngeal airway and thus the transfer function *T(f)* changes continually. This acoustical filter suppresses the transfer of sound energy at certain frequencies and lets maximum energy through at other frequencies. In speech processing the frequencies at which local energy maximum may pass through the supralaryngeal airway are called formant frequencies F_i. They are determined by the length (speaker dependent) and the shape of the vocal tract (more articulation than speaker dependent). The larynx including the vocal folds and the subglottal system has only minor effects on the formant frequencies F_i, see Flanagan (1972). Different vowels owe their phonetic difference, or so-called quality, to their different formant frequencies.

As an example, Figure 47 shows an idealized transfer function of the supralaryngeal airway for the vowel [u] (the phoneme in "shoe``) including a grid of the harmonic frequencies for f_0=100 Hz (a) and f_0=400 Hz (b). The first three formant frequencies of this vowel sound are F_1=500 Hz, F_2=900 Hz, and F_3=2200 Hz. The bandwidth of each formant varies from 60 to 150 Hz. The first three formants of vowels play a major role in specifying these sounds. Higher formants exist, but they are not necessary for the perception of vowel qualities, see Flanagan (1972).

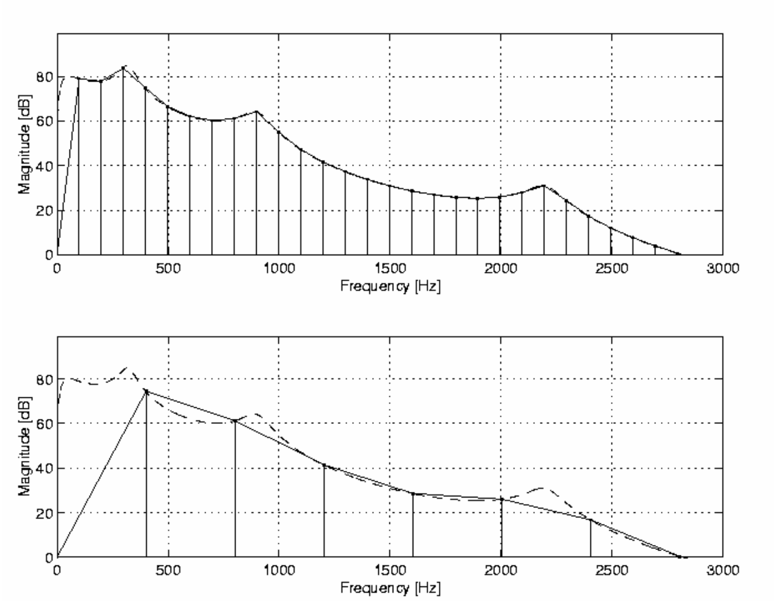


Figure 47: Idealized spectrum *P(f)* of the phoneme [u] with f_0= 100Hz (a) and f_0=400 Hz (b) including the spectral envelope of the deterministic trajectories

The resulting spectrum $U(f) = S(f) \cdot T(f)$ describes the speech signal at the end of the airways of the vocal tract. Sound energy is present at each of the harmonics of the glottal source, but the amplitude of each harmonic will be a function of *T(f)* and *S(f)*. Finally, the spectrum *U(f)* is modified by the radiation characteristic *R(f)* in a way that the resulting spectrum can be written as

$$P(f) = S(f) \cdot T(f) \cdot R(f) \tag{3}$$

with the frequency *f* in Hz. The radiation characteristic *R(f)* models the effect of the directivity patterns of sound radiation from the head as a function of frequency. This effect depends on the lip opening and can be modeled as a high-pass filter according to Klatt (1980). The resulting spectra *P(f)* of the previous examples of f_0=100 Hz (a) and f_0=400 Hz (b) can be seen in Figure 47.

2.3.4 Singing versus Speech

The most important class of voiced speech sounds is vowels. Vowels are produced by exciting an essentially fixed vocal tract shape with quasi-periodic pulses of air, caused by the vibration of the vocal folds (glottal excitation). The vocal tract has to be open enough for the pulsed air to flow without meeting any obstacle. The vocal tract shape modifies the glottal excitation and thus the vocal timbre in a manner that the appropriate vowel is perceived. If the vocal tract narrows or even temporarily closes, the airflow gives birth to a noise, so that a consonant is produced.

The spectral qualities of different vowels are mainly caused by the glottal excitation (speaker dependent) and the vocal tract transfer function (depending on the type of vowel, the language, and the speaker). As introduced in section 2.3.2 the vocal tract transfer function can be characterized by formants, which are resonant peaks in the spectrum.

According to Sundberg (1987) the higher the formant frequency, the more its frequency depends on nonarticulatory factors, such as vocal tract length and the vocal tract dimension within and around the larynx tube. Good singers position and move the length and the shape of the vocal tract in a very precise way. These variations make the formants have different amplitudes and frequencies, which we perceive as different voiced sounds.

The length of the vocal tract defined as the distance from the glottis to the lips can be modified up to a certain extent. The larynx can be raised or lowered increasing or decreasing respectively the length. Also the posture of the lips can cause fluctuations to the length of the vocal tract, if we protrude the lips we lengthen it and if we smile we reduce it. The longer the vocal tract, the lower the formant frequencies are. The shape of the vocal tract, on the other hand, is modified by what is referred as articulators. These articulators are the tools we can use to change the area function of the vocal tract and they can move in different ways and each of the movements has a different effect. A movement of any of the articulators generally affects the frequencies of all formants. Sundberg experiments have shown that the first three formants define the vowel type and differ less between different speakers, for example, between a male and a female, than the higher formants. The fourth and the fifth formant are relevant for perceiving the voice timbre, which is the personal component of a voice sound.

The effects of different articulators (e.g. the tongue) and speaker dependent physiological characteristics on the different formants can be summarized as:

- First formant: sensitive to varying the jaw opening.
- Second formant: sensitive to changes of the tongue shape.
- Third formant: sensitive to the position of the tongue, particularly to the size of the cavity that is just behind the front teeth.
- Fourth formant: depends on the vocal tract dimensions within and around the larynx tube. Thus, the speaker has no explicit control of this formant frequency.
- Fourth and fifth formant: depend on the length of the vocal tract and thus are speaker dependent.

A more detailed description of the dependency of formants on articulators or on physiological characteristics can be found in Sundberg (1987) and Lieberman and Blumestein (1988).

The representation of the spectral shape of a singing voice depends highly on the fundamental frequency f_0. If the glottal excitation is low pitched the idealized spectrum *P(f)* can be easily observed in the spectral envelope of the trajectories. But if the voiced source has a very high fundamental frequency f_0 and the harmonic frequencies of the source spectrum *S(f)* are thus very wide spaced, then *P(f)* and the spectral envelope of the trajectories differ significantly. Figure 3.10 below illustrate such problem.

The trajectories show only *P(f)* at special frequencies f_i. In the case of an ideal analysis, f_i can be written as

$$f_i = i \cdot f_0 \qquad \text{i=1,2,3, …} \qquad (4)$$

and thus the magnitudes *M(f_i)* of the trajectories are equidistant samples of the spectrum *P(f)*. It can occur that the voiced sound is too high pitched so that almost no similarity to the continuous spectrum *P(f)* can be found in the discrete spectral envelope of the sinusoidal trajectories. All these considerations show that visual formant detection from the spectral shape will only be possible when dealing with low pitched voices (approximately under 100Hz).

Although speech and singing voice sounds have many properties in common because they originate from the same production physiology, there are some important differences that make them different. In singing, the intelligibility of the phonemic message is often secondary to the intonation and musical qualities of the voice. Vowels are often sustained much longer in singing than in speech and independent control of pitch and loudness over a large range is required. A trained singer can sing an entire sequence of vowel sounds at the same pitch. These major differences are highlighted in the following brief list, which is not meant to lay claim to completeness but rather to point out some important differences between speech and singing voice processing:

Voiced/unvoiced ratio: The ratio between voiced sounds, unvoiced sounds, and silence is about 60%, 25\%, and 15\% respectively, see Cook (1990), in the case of normal speech. In singing, the percentage of phonation time can increase up to 95\% in the case of opera music. This makes voiced sounds special important in singing because they carry most of the musical information.

Vibrato: Two types of vibrato exist in singing, which differ in their production mechanism. The classical vibrato in opera music corresponds to periodic modulation of the phonation frequency. The regularity of this modulation is considered as a sign of the singer's vocal skill. It is characterized by two parameters, namely the rate and the extent of the modulation of the phonation frequency. In popular music, the vibrato is generated by variations in the subglottal pressure, implying an amplitude modulation of the voice source, see Sundberg (1987). In speech, no vibrato exists. The Vibrato adds certain naturalness to the singing voice and it is a very specific characteristic of the singing voice of a singer.

Dynamic: The dynamic range as well as the average loudness is greater in singing than in speech. The spectral characteristics of a voiced sound change with the loudness, see Sundberg (1987).

Singer's formant: This is a phenomenon that can be observed especially in the singing of male opera singers, see Sundberg (1987). The singer's formant is generated by clustering of the third, fourth, and fifth formant. The frequency separation of these formants is smaller in the sung than in the spoken vowel, resulting in one peak at approximately 2-3 kHz with a great magnitude value in the spectrum. A trained singer has the ability to move the formant frequencies so that already existing formants in speech occurring only in higher frequency regions are tuned down. As a result, the magnitude of the down-tuned formants increases towards lower frequencies due to the voice source spectrum.

Modification of vowels: Singing at a very high pitch includes that the fundamental frequency f_0 is greater than the one of the first formant. In order not to loose this characteristic resonance and thus important acoustic energy, trained singers move about the first formant in frequency to the phonation frequency. The gain of loudness increases as soon as the frequency of the first formant joins the fundamental f_0. Although the perceived vowel is slightly modified, an important ability to use a greater dynamically range is gained and thus it is a gain of musical expression.

Fundamental frequency: In speech, the fundamental frequency variations express an emotional state of the speaker or add intelligibility to the spoken words. This is called prosody and distinguishes, for example, a question from a normal statement. The frequency range of f_0 is very small compared to singing. Considering a trained singer, the minimum range is about two octaves and for excellent singers like Maria Callas it can be up to three octaves.

2.4 Voice technologies

Music technologies are usually divided into: analysis, modeling, coding and synthesis; however, as discussed in section 2.4., technologies mix all together sharing concepts and scope. If we take voice modeling as an example, the model is built to fit and thus requires irremediably from analysis, and nearly most of the times models merit is judged by means of the quality of the synthesis it can generate.

Until today, technologies have been traditionally described as whether belonging to physical models or spectral models. Physical models are defined as those which sculpt the physiological mechanisms of the source producing the sound while spectral models are those mimicking the perception of the emitted sound. Physical versus spectral models classification has been recently gradually consigned to oblivion. Their boundaries crossover severely and mainly most recent techniques embrace both points of view. Alternative taxonomies classify according to the parameterization of the technology or according to the domain they work in (time or transformed) but none of these cluster more effectively.

Nevertheless, since categorization has proven to be useful for knowledge disposition, we propose a taxonomy which organizes major technologies employed in singing voice according to what field of the voice they stand on, parameterize and mimic: the mechanics, the pressure wave or the timbre.

2.4.1 Voice mechanics technologies

2.4.1.1 Singing robots

Already in 1791, Von Kempelen managed to produce whole words and short sentences playing a machine he designed himself (see Figure 48). His machine consisted of a bellows attached to a collection of mechanical model emulating different parts of the vocal tract. The bellows emulated the action of the lungs, an oscillating reed emulated the vocal chords, a leather tube emulated the vocal tract and a holed multi piece wood box emulated the mouth and nostrils. In 1846, Joseph Faber demonstrated a mechanical device similar to Kempelen's that could produce not only ordinary and whispered speech, but could also sing "God Save the Queen". And as late as in 1937, R. R. Riesz refined previous machines with a vocal tract shape that was close to the natural.

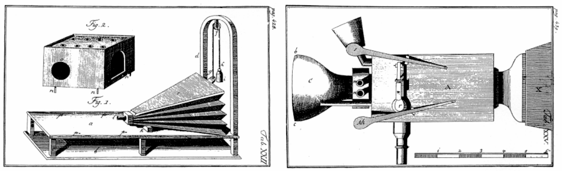


Figure 48: Drawings of Von Kempelen singing machine. Taken from www.ling.su.se/staff/hartmut/kemplne.htm with permission of authors

Soon afterwards, in 1939, Dudley's VODER appeared and pure mechanical voice production was completely left behind. Recent singing synthesis approaches such as "*The Squeeze Vox*", by Cook and Leider (20000), seem to be partially inspired in those old machines as what as control of the synthesizer refers. However, early in this century the pure robot approach resurges in Japan by the hand of Hideyuki Sawada, see Sawada and Nakamura (2004). His model consists in a compressor that forces air into the system emulating the action of the lungs, a valve that mimics the windpipe and controls the airflow, a rubber vocal chord, a flexible plastic resonance tube that shapes to emulate different vocal tract filters.

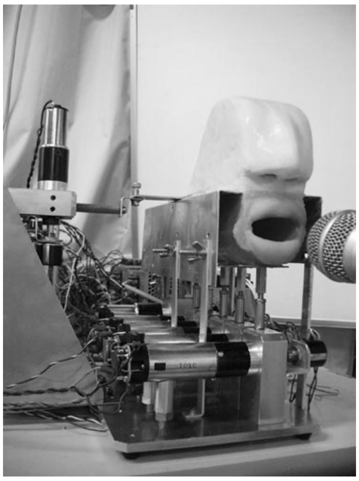


Figure 49: Picture of Sawada's mechanical voice system

2.4.1.2 Digital waveguides

The aim of digital waveguide modelling is to design a discrete-time model that behaves like a physical system. Digital waveguides use delay lines, digital filters and often nonlinear elements to emulate resonators such as a vibrating string, a thin bar or an acoustic tube. It was in fact a series of cylindrical tube sections represented by a digital ladder filter what Kelly and Lochbaum used to implement for the first time a digital physical model of the human vocal tract in 1962, see Kelly and. Lochbaum (1962). In 1985, Fant, Liljencrants and Lin improved the model with a more realistic excitation source, see Fant et al. (1985), and later in 1990 Perry Cook extended the digital waveguide model of the vocal tract to include as well nasal cavity and throat radiation in Cook (1990). Latest most relevant contribution was the insertion of variable length conical sections inside the model from Välimäki and Karjalainen (1994).

2.4.2 Voice timbre technologies

2.4.2.1 Vocoder

The Vocoder, see Dudley (1936) is an analysis synthesis technique based on the source filter model of the human voice. For an input signal, a bank of bandpass filters computes the distribution of energies along the spectrum as an estimation of the vocal tract filter; and a pitch detector determines whether the input is voiced or unvoiced and estimates a fundamental frequency for the voiced. Because variations on the parameters of the voice model (filter bank energies and pitch) were much lesser than the variations of speech itself, the Vocoder reduced the bandwidth required for speech transmission, see Dudley (1939).

Figure 50: According to Science News Letter, of January 14 1939,"*The young lady striking keys is creating a man-like voice. This new synthetic orator will "lecture" with his "electrical accent" at the New York and San Francisco world fairs. It is a compact machine resting on a small table, plus as many loudspeakers as are necessary to reach the audience.VODER, THE MACHINE THAT TALKS. No recording of any kind is used in this latest addition to the anatomy of the Mechanical Man. A girl at a keyboard controlling varying electrical currents does the trick. A pedal operated by her right foot enables the operator to make voice inflection, and synthesized speech is heard from a loudspeaker.*"

The Vocoder was the first speech analysis / synthesis technique ever and thus a huge amount of evolutions exist from the original proposal. The so called channel vocoder increased the sound quality while reducing the bandwidth required for transmission, see Gold and Rader (1967). Later, the phase vocoder improved the computational model based upon the use of the Discrete Fourier Transform. Because vocoder conception does not restrict to speech and can be applied to other “instruments” and signals, phase vocoder based technologies are used widespread in general sound effects processors nowadays, see Moorer (1978).

2.4.2.1.1 Phase vocoder

The phase-vocoder, see Flanagan and Golden (1966) is a frequency-domain technique based on the Short Time Fourier Transform (STFT). The STFT characterizes the spectral behavior of a signal *x(n)* along time and can be written as:

$$X(n,k) = \sum_{m=-\infty}^{m=\infty} x(m)\cdot h(n-m)\cdot e^{-j\frac{2\pi mk}{N}}, \qquad k=0,1,...N-1 \tag{5}$$

where *X(n, k)* is the complex time-varying spectrum with the frequency bin *k* and time index *n*. At each time index, the input signal *x(m)* is weighted by a finite length window *h(n-m)* and then computed its spectrum.

The most common understanding of the phase-vocoder technique is the filter bank interpretation. This filter bank is a parallel bank of *N* bandpass filters with the following impulse response

$$h_k(n) = h(n)\cdot e^{j\frac{2\pi nk}{N}}, \qquad k=0,1,...N-1 \tag{6}$$

which means the same filter shape is shifted along frequency by *2π/N* radians steps.

The bandpass signal of band *k*, $y_k(n)$, is obtained by filtering the input signal *x(n)* with filter $h_k(n)$ as showed in the following expression

$$\begin{aligned} y_k(n) &= x(m) * h_k(n) = \sum_{m=-\infty}^{m=\infty} x(m)\cdot h_k(n-m) \\ &= \sum_{m=-\infty}^{m=\infty} x(m)\cdot h(n-m)\cdot e^{j\frac{2\pi(n-m)k}{N}} = \\ &= e^{j\frac{2\pi nk}{N}} \cdot \sum_{m=-\infty}^{m=\infty} x(m)\cdot e^{-j\frac{2\pi mk}{N}} \cdot h(n-m) = \\ &= e^{j\frac{2\pi nk}{N}} \cdot X(n,k) \end{aligned} \tag{7}$$

and the output signal *y(n)* is the sum of all bandpass signals

$$y(n) = \sum_{k=0}^{k=N-1} x(n) * h_k(n) = \sum_{k=0}^{k=N-1} X(n,k)\cdot e^{-j\frac{2\pi nk}{N}} \tag{8}$$

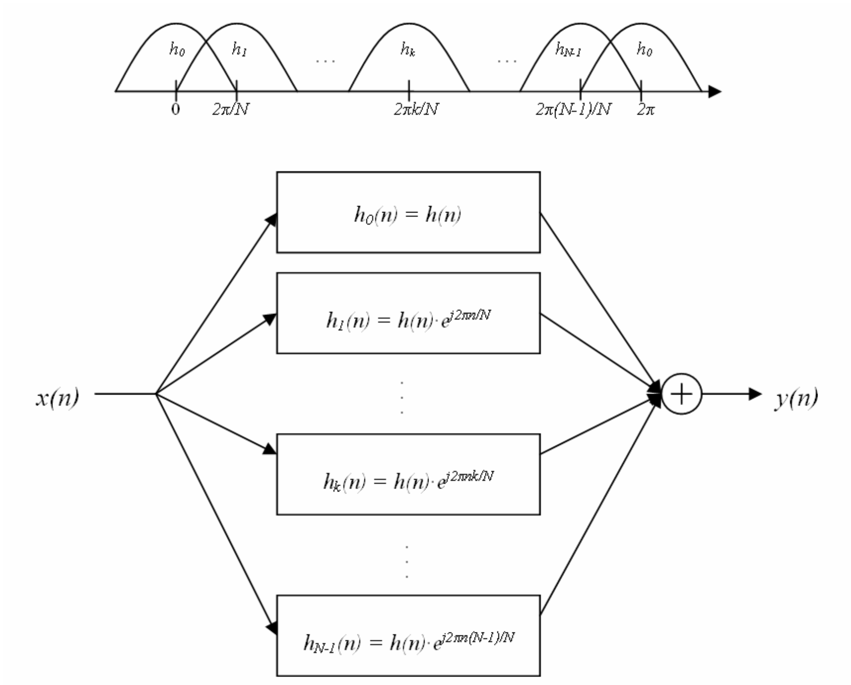

h_0
h_1
...
h_k
...
h_{N-1}
h_0
0
$2\pi/N$
$2\pi k/N$
$2\pi(N-1)/N$
2π
$h_0(n) = h(n)$
$h_1(n) = h(n)\cdot e^{j2\pi n/N}$
$h_k(n) = h(n)\cdot e^{j2\pi nk/N}$
$h_{N-1}(n) = h(n)\cdot e^{j2\pi n(N-1)/N}$
$x(n)$
$y(n)$

Figure 51: Filter bank description of the short time Fourier transform. The frequency bands on the top show the displacement of each of the bandpass filters.

Thus, for a certain time index *n*, *y(n)* will be defined by a *N* length vector containing an estimation of each band's energy

$$y_k(n) = e^{j\frac{2\pi nk}{N}} \cdot X(n,k) = |X(n,k)| \cdot e^{j\varphi_X(n,k)} \cdot e^{j\frac{2\pi nk}{N}} \tag{9}$$

so

$$\begin{gathered} |y_k(n)| = |X(n,k)| \\ \varphi_{y_k}(n) = \varphi_X(n,k) + \frac{2\pi \cdot n \cdot k}{N} \end{gathered} \tag{10}$$

Before synthesis, spectral modifications can be applied to the resulting analysis data. If we term the modified data $y_k^{\text{mod}}(n)$, the synthesis window *w(n)*, the synthesis hop size *H*, and $y_m(n)$ is the Inverse Fourier Transform of the modified short time spectra

$$y_m(n) = \frac{1}{N}\sum_{k=0}^{k=N-1} y_k^{\text{mod}}(n) \cdot e^{j\frac{2\pi nk}{N}} = \frac{1}{N}\sum_{k=0}^{k=N-1}\left(e^{j\frac{2\pi mH}{N}} \cdot X(mH,k)\right)^{\text{mod}} \cdot e^{j\frac{2\pi nk}{N}} \tag{11}$$

the resulting synthesis signal *s(n)* can be written as the overlap and add of the windowed $y_m(n)$'s

$$s(n) = \sum_{m=-\infty}^{m=\infty} w(n - m \cdot H) \cdot y_m(n) \tag{12}$$

2.4.2.1.2 Phase locked vocoder

Different evolutions of the original vocoder appeared in the 70's and 80's. The Formant Vocoder attempted to enclose formant amplitudes, frequencies, and bandwidths information in the algorithm, see Rabiner and Schafer (1978). Homomorphic Vocoder modelled voice as a lineal convolution of the vocal excitation and the vocal tract and operated both components in the cepstral (inverse log-Fourier) domain as an addition, see Oppenheim and Schafer (1989). More recently, Phase Locked Vocoder appeared as an approach to solve the phase unwrapping problem, see Puckette (1995) and Laroche and Dolson (1999).This problem is caused by the fact that the phase obtained for each bin *k* depends on a term that is multiple of *2π* and that is not the same for all bins, see equation (2.6). These phase differences are dispersed and cause synthesis artifacts when applying spectral transformations.

Out of this new vocoder family named phase-locked vocoders, the rigid phase locking algorithm proposed by Laroche and Dolson called the Identity Phase Locking is of vital relevance in this work.

The identity phase locking vocoder it is based on spectrum peak detection and segmentation and on the assumption that spectrum peaks are sinusoids. In this technique a simple peak detection algorithm detects all spectra local maximums and considers them peaks. This series of peaks are then used to chop the spectrum into regions. Each of these regions contains one peak and sets the boundaries with the neighboring regions to the middle frequency between adjacent peaks or to the lowest amplitude between the two peaks. The proposal states that only the phases of the peaks are calculated while the phases of the rest of the bins are locked to their corresponding region peak.

$$\varphi_s(k) = \varphi_s(k_p) + \varphi_a(k) - \varphi_a(k_p) \text{ where } k \in Region(k_p) \tag{13}$$

The identity locking asserts that the synthesis phases around a spectral peak are related with that peak's phase in the same way they are in the analysis. So, being k_p the bin index of the dominant peak, φ_s the synthesis phases, and φ_a the analysis phases, the phases of the bins k under the region of influence of peak k_p will be

2.4.2.2 Spectral peak processing

The Spectral Peak Processing (SPP) is a technique based on the rigid locked phase vocoders. As pointed out in previous sections, this sort of technique can be somehow related with the sinusoidal models. The main difference with the rigid locked-phase vocoder is that in the SPP not all spectrum local maximums are considered to be peaks and thus sinusoids. SPP only considers as peaks those local maximums that the analysis classifies as harmonic peaks. However, mimicking the locked phase vocoders, the SPP approach considers the spectrum compounded of spectral peak regions each of which consist of a harmonic spectral peak and its surrounding spectra which we assume to contain all the non-perfect harmonic behavior information of the corresponding harmonic peak. The goal of such technique is to preserve the convolution of the analysis window after transposition and equalization transformations.

SPP shares analysis and synthesis procedures and dataflow with the SMS except from those SMS procedures that specifically regard the residual component, which do not exist here. Moreover, the SPP analysis includes spectrum segmentation into regions out of the harmonic spectral peaks as shown in Figure 2.4.

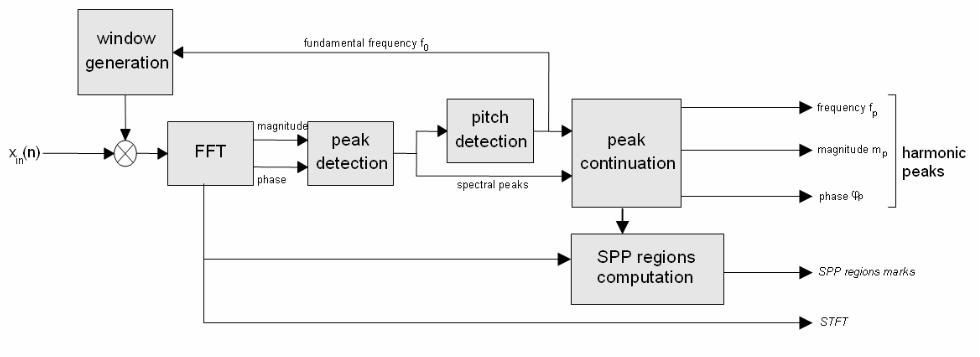


Figure 52: Block diagram of the SPP analysis process

The synthesis in the SPP includes the STFT and the SPP spectral peak regions marks aside from the frequency, magnitude and phase of the harmonic spectral peaks.

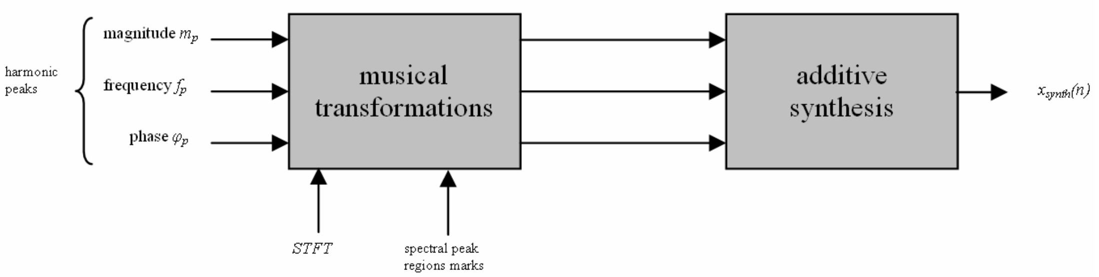


Figure 53: Block diagram of the SPP synthesis process

From now on, when talking about SPP, we will do it in the context of pseudo-harmonic monophonic sounds.

2.4.2.2.1 Spectrum segmentation

The outcome of the SPP analysis are the harmonic peaks (frequency, amplitude and phase), the STFT, and the SPP regions marks. Remember SPP divides the spectrum into a set of regions, each of which belongs to one harmonic spectral peak and its surroundings.

Two different algorithms were introduced in section 2.4.2.2.1 for the segmentation of the spectrum into regions out of peaks information. These techniques can also be applied in the SPP case. The region boundary is set to be at the lowest local minimum spectral amplitude between two consecutive peaks as illustrated in Figure 54.b. If there are no local minimums, then the boundary is set to be at the middle frequency between two consecutive peaks as illustrated in Figure 54.a.

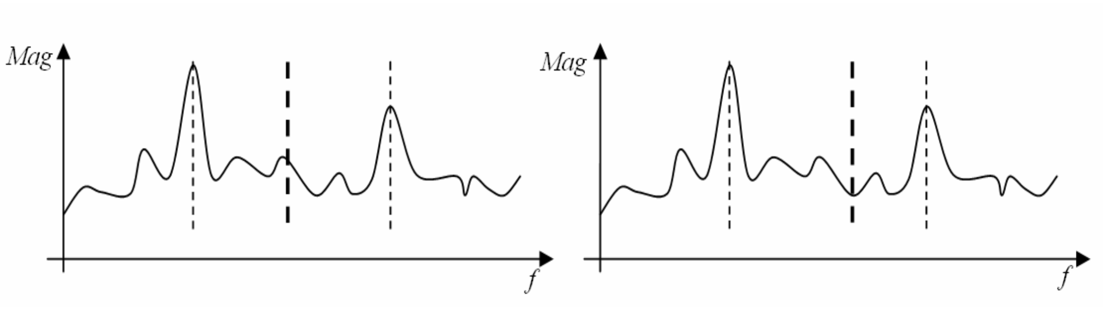


Figure 54: SPP region boundaries computation techniques: (a) mid frequency value between consecutive harmonic peaks (b) min amplitude value between consecutive harmonic peak's corresponding frequency

2.4.2.2.2 Maximally flat phase alignment

Latest implementations of phase locked vocoder incorporate phase alignment preservation. The preservation of the phase alignment in synthesis results into an improvement of the perceived synthesis quality, especially when input voice signal is transposed downwards.

Initially the phase alignment for a certain frame was computed by moving the analysis phases up to the point in which the phase of the fundamental reached zero ($\varphi'_0=0$). That was assumed to be the phase state of the beginning of the voice pulse period.

Afterwards, it was found this assumption was too vague and that the fundamental phase value φ'_0 at which voice pulses were triggered was intrinsic to the voice excitation and thus to the singer.

More precise computation methods appeared to find the fundamental phase φ'_0 value inside $[-\pi,\pi]$ that defines the optimal maximally flat phase alignment. An exemplification of such methods is:

First, define φ'_{0c} candidates from $-\pi$ to π in $2\pi/80$ steps. And for each candidate,

1. Select first harmonics, up to 2 KHz, maximum 10 harmonics
2. Apply time shift (corresponding to φ'_{0c}) to move the phases of the selected harmonics of the frame spectrum up to point in which the phase of the fundamental equals φ'_{0c}.
3. Calculate the phase average taking of the harmonics selected in (b) taking into account the phase wrapping.
4. Calculate the phase deviation. This is computed as the sum of the differences between the phases and the average in absolute value divided by the number of harmonics that contribute.
5. Select the candidate φ'_{0c} with minimum phase deviation as φ'_0
6. Apply some φ'_0 continuation in order to avoid abrupt changes (see Figure 55)

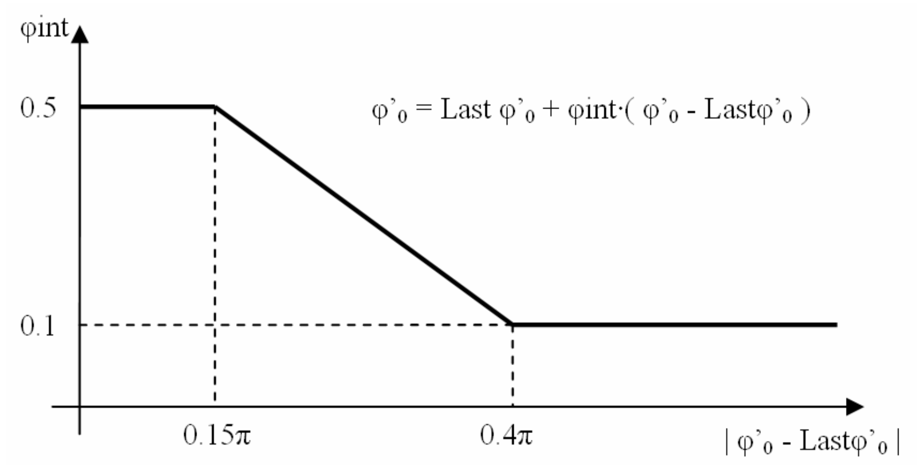


Figure 55: Continuation interpolation factor φ_{int} versus consecutive frames φ'_0 difference.

An alternative method for the estimation of the phase alignment can be found in Bonada (2004).

2.4.2.3 Linear predictive coding

Linear Predictive Coding (LPC) is an analysis synthesis technique introduced in the sixties based on the adaptive prediction of following future samples from linear combinations of certain number of past previous samples, see Atal and Hanauer (1971). The LPC parameters can be computed using the autocorrelation or covariance methods over a linear least square formulation, see Makhoul (1966) and Rabiner and Schafer (1978). The analysis problem is equivalent to the identification of the coefficients of an all pole filter that, in the case of voice, would model the vocal tract response. Matched with the excitation source model of speech, original LPC implementations for voice applications required a pitch estimator that determined the type of voice excitation (voiced or unvoiced) and the pitch value in the voiced parts. In the standard formulation of LPC, the source signals are modeled either a white noise or a pulse train.

LPC reached popularity because of its efficiency and effectivity when dealing with speech due to the speed of the analysis algorithm and the low bandwidth required to encode speech that could kept intelligible when decoded, see Schroeder (1999). Nevertheless, LPC is nowadays the basis of most speech codecs we use.

2.4.2.3.1 Code excited linear prediction

In those LPC implementations where voiced vocal excitation is modelled as a delta train, the synthesis quality degrades into some blurriness. In order to solve such problem, Code-Excited Linear Prediction (CELP) proposed a complex excitation model based on a codebook, see Schroeder and Atal (1984).

The method for modelling the excitation is based on searching through a table of candidate excitation vectors on a frame by frame basis and can be described as follows. LPC analysis is performed to obtain the filter and the residual. The residual is compared against all entries of the codebook filtered by the LPC filter. The one vector that results into the most similar residual match is chosen as excitation. Most commonly codebooks contain hundreds of vectors and may exist for both different types of excitation, in which case the voice unvoiced binary decision turns into an estimation of voiced and unvoiced vector gains.

Vector Quantization techniques were incorporated lately to improve the efficiency of this technique, seeMakhoul et al. (1985). The CELP codec can reach intelligible speech transmission rates under 4 kbits/sec and is one of the most common standards in today's speech communication devices.

LPC, when coupled with the source filter theory for processing speech signals, can provide intelligible speech synthesis at rates around 2 Kbps. At the expense of increasing the bandwidth, natural sounding can be achieved by using more complex models for the excitation such as multipulse, see Atal and Remde (1982), and Singhal and Atal (1989), regular pulse, see Kroon et al. (1986), or stochastic codeword, seeSchroeder and Atal (1985). Still, pitch-excited LPC synthesis quality can be improved through the exploitation of the residual's phase and amplitude spectrum and through the inclusion of some of the glottal phase characteristics. Models such as Glottal Excited Linear Prediction (GELP), take into account glottal features remaining in the residual and can achieve high-quality synthetic speech at low bit rates, see Childers and Hu (1994) and Hu and Wu (2000).

2.4.2.4 Sinusoidal based technologies

The sinusoidal, the sinusoidal plus residual and the sinusoidal plus residual plus transients models are presented in this section.

2.4.2.4.1 Sinusoidal model

Within the spectrum models there are the additive models, which are based on the basic idea that a complex sound can be constructed by a large number of simple sounds. These models try to represent the spectral characteristics of the sound as a weighted sum of basic components, so-called basis expansions.

$$x(n) = \sum_{j=1}^{J} \alpha_j g_j(n) \tag{14}$$

where $g_j(n)$ is the j^{th} basis function and α_j is its appropriate chosen weight. There is no restriction to their number J. A detailed discussion on the properties of these basis expansions can be found in Goodwin (1997).

When the additive synthesis is the summation of time varying sinusoidal components rooting in Fourier's theorem, which states that any periodic waveform can be modeled as a sum of sinusoids of various amplitudes and harmonic frequencies, we talk about sinusoidal modeling. The sinusoidal model and in general additive models have been under consideration in the field of computer music since its inception.

One of the widely applied models in speech coding and processing, or audio analysis-synthesis systems is the sinusoidal model as proposed in McAulay and Quatieri

(1986). At this, the signal is modeled as a sum of sinusoids with time-varying amplitudes *$A_p(t)$*, frequencies *$\omega_p(t)$*, and phases *$\theta_p(t)$*. Therefore we estimate the signal as

$$\begin{aligned}\hat{x}(t) &= \sum_{p=1}^{P} A_p(t)\sin\left[\theta_p(t)\right] \\ \theta_p(t) &= \int_0^t \omega_p(\sigma)\partial\sigma + \eta_p + \phi_p\left[\omega_p(t)\right]\end{aligned} \tag{15}$$

where P is the number of sinusoids, which can also be a function of time in some applications. Here, the index p denotes that the parameter belongs to one time-evolving sinusoid, called p^{th} partial or track. The time-varying phase *θp(t)* contains the time-varying frequency contribution, a fixed phase offset η_p, which accounts for the fact that the sinusoid will generally not be in phase, and a time-varying frequency-dependent term $\phi_p\left[\omega_p(t)\right]$ that according to the speech production model proposed by McAulay and Quatieri (1986) is determined by the phase function of the vocal tract.

Sinusoidal synthesis is accepted as perhaps the most powerful and flexible method. Because independent control of every component is available in sinusoidal synthesis, it allows the pitch and length of sounds to be varied independently, as well as it is possible to implement models of perceptually significant features of sound such as inharmonicity and roughness. Another important aspect is the simplicity of the mapping of frequency and amplitude parameters into the human perceptual space. These parameters are meaningful and easily understood by musicians. Recently, a new sinusoidal synthesis method based on spectral envelopes and Fast Fourier Transform has been developed by Rodet and Depalle (1992). Use of the inverse FFT reduces the computation cost by a factor of 15 compared to oscillators. This technique renders possible the design of low cost real-time synthesizers allowing processing of recorded and live sounds synthesis of instruments and synthesis of speech and the singing voice.

This model makes strong assumptions on the behavior of the time-evolving sinusoidal components. Because of the fact that this system is a frame-based approach, the signal parameters are assumed to be slowly varying in time compared to the analysis frame-rate. In real-world signals, such as a musical note played by a flute, this assumption is almost suited by the property of pseudo-periodicity. These sounds consist mainly of stable sinusoids, which can be perfectly modeled by Equation 15. But modeling audio signals only as a sum of sinusoids suffers from a major limitation. If the signal contains noisy or impulsive components, an excessive number of sinusoids are needed to model them. The resulting residual signal shows that these broadband processes are present in every natural sound, e.g. the sound of the breath stream in a wind-driven instrument or the sliding of the bow against a string of a cello.

2.4.2.4.2 Sinusoidal plus residual model

While most of the simplest heuristics for sinusoidal versus noise decomposition include checking inter-bin (adjacent bins) or inter-frame (consecutive frames) phase relations, see Griffith and Lim (1988) and Settle and Lippe (1994), or comparing autoregressive and minimum variance distortionless response, see Dubnov (2006), more advanced methods based on Fourier analysis employ additional steps for modeling the noisy components such as obtaining the residual part by careful subtraction of the sinusoidal components from the original signal, see Serra and Smith (1990) or using random phase modulation for broadening the spectral peaks at the sinusoidal frequencies, see Fitz and Haken (1996), Fitz et al. (2000), and Peeters and Rodet (1999). In this section some details on the deterministic-plus-stochastic model as proposed by Serra and Smith (1990) are given.

The use of the estimation of the signal $\hat{x}(t)$ in Equation 15 is to imply that the sum-of-partials model does not provide an exact reconstruction of the signal *x(t)*. Because of the fact that a sum of sinusoids is ineffective for modeling impulsive events or highly uncorrelated noise, the residual consists of such broadband processes, which correspond to musical important features, such as the turbulent streaming inside the bore of a flute. The stable sinusoids represent the main modes of the vibrating system, and the residual the excitation mechanism and non-linear behaviors. In the case of bowed strings the stable sinusoids are the result of the main modes of vibration of the strings and the noise is generated by the sliding of the bow against the string, plus by other non-linear behavior of the bow-string-resonator system. Since these features are needed to synthesize a natural sound, the additional stochastic component, the residual, should be included in the signal model, according to Serra and Smith (1990) and Serra (1997).

$$x(t) = x_{\mathrm{det}}(t) + x_{stoch}(t) \tag{16}$$

The model assumes that the sinusoids are stable partials of the sound and that each one has a slowly changing amplitude and frequency. The deterministic part of the model is the same proposed by the sinusoidal model in Equation 15.

$$\begin{aligned} x_{\mathrm{det}}(t) &= \sum_{p=1}^{P} A_p(t)\sin\left[\theta_p(t)\right] \\ \theta_p(t) &= \int_0^t \omega_p(\sigma)\partial\sigma + \eta_p + \phi_p\left[\omega_p(t)\right] \end{aligned} \tag{17}$$

By assuming that $x_{stoch}(t)$ is a stochastic signal, it can be described as filtered white noise,

$$x_{stoch}(t) = \int_0^t h(t,\tau)u(\tau)\partial\tau \tag{18}$$

where *u(τ)* is white noise and *h(t, τ)* is the response of a time varying filter to an impulse at time t . That is, the residual is modeled by the convolution of white noise with a time-varying frequency-shaping filter. The deterministic-plus-stochastic decomposition has been discussed in several later works, Rodet (1997), Ding and Qian (1997), Goodwin (1997), and Verma and Meg (1998).

Using the terms deterministic and stochastic brings up the question about the theoretical distinction between these two kinds of processes. Deterministic signals are not only restricted to the sum of sinusoids. However, in this model the class of deterministic signals considered is restricted to quasi-sinusoidal components and the stochastic part is likened to residual.

2.4.2.4.3 Transients

A main question remains unsolved in the sinusoidal plus residual model when applied to signals such as human voice: how can a multifaceted nature signal be fit into a model based on a binary classification? In order to give answer to such question a new evolution of the model was proposed in which the concept of transients was included as a new basic component. The new "*Sinusoidal + Noise + Transients*" model [1T,2T,3T,4T] bases its transient detection algorithm on the assumption that short pulses in the time domain correspond to sine-like curves in the spectrum domain, see Verma et al. (1997), Verma and Meng (1998, 2000), and Levine and Smith (1998).

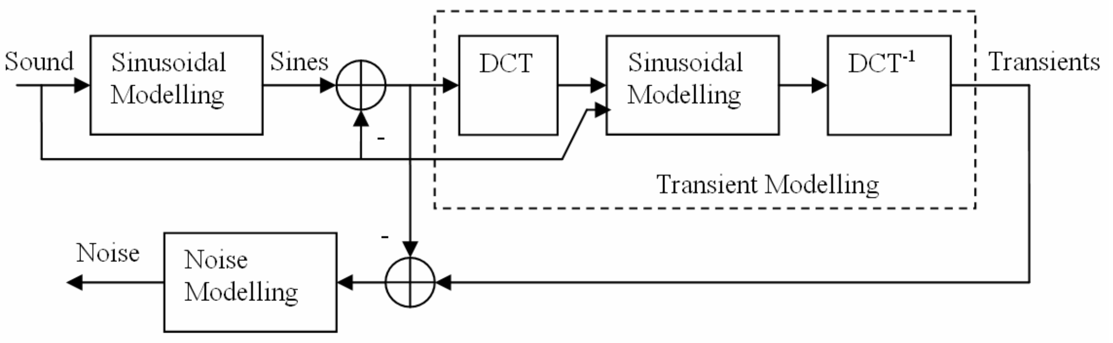


Figure 56: Block diagram of the sinusoidal + noise + transients decomposition.

The DCT blocks in Figure 56 operate the *N* point Discrete Cosine Transform, defined as:

$$C(k)=\sum_{n=0}^{N-1} x(n)\cdot\cos\left((2n+1)\cdot k/2N\right) \tag{19}$$

The DCT transforms impulses into cosines and a sum of pulses becomes a superposition of cosines. Thus, sinusoidal model is applied in the transformed domain for the estimation of the transient's components.

2.4.2.5 Excitation plus resonances voice model

The Excitation plus Resonances (EpR) voice model is based on an extension of the well known source/filter approach, see Childers (1994). The EpR filter can be decomposed in two cascade filters. The first of them models the differentiated glottal pulse frequency response and the second the vocal tract (resonance filter).

2.4.2.5.1 EpR source filter

The EpR source is modeled as a frequency domain curve and one source resonance. The curve is defined by a gain and an exponential decay as follows:

$$Source_{dB} = Gain_{dB} + SlopeDepth_{dB}\left(e^{Slope\cdot f}-1\right) \tag{20}$$

It is obtained from an approximation to the harmonic spectral shape (*HSS*) determined by the harmonics identified in the analysis

$$HSS(f) = envelope_{i=0..n-1}\left[f_i, 20\log(a_i)\right] \tag{21}$$

where *i* is the index of the harmonic, *n* is the number of harmonics, f_i and a_i are the frequency and amplitude of the i^{th} harmonic. On top of the curve, we add a second resonance in order to model the low frequency content of the spectrum below the first formant.

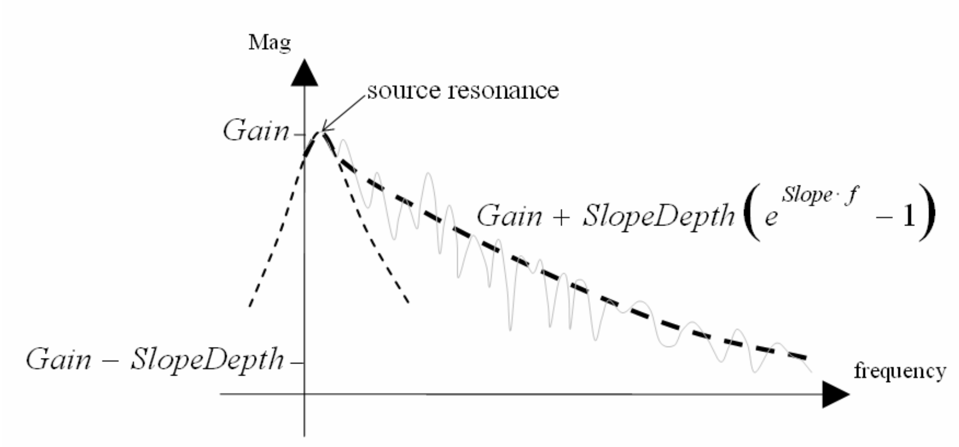


Figure 57: Representation of the EpR source resonance

The source resonance is modeled as a symmetric second order filter (based on the Klatt formant synthesizer, see Klatt (1980) with center frequency *F*, bandwidth *Bw* and linear amplitude *Amp*. The transfer function of the resonance *R(f)* can be expressed as follows

$$\begin{aligned} H(z) &= \frac{A}{1 - Bz^{-1} - Cz^{-2}} \\ R(f) &= Amp\frac{H\left(e^{j2\pi\left(0.5+\frac{f-F}{fs}\right)}\right)}{H\left(e^{j\pi}\right)} \end{aligned} \tag{22}$$

where

$$\begin{aligned} fs &= \text{Sampling rate} \\ C &= -e^{-\frac{2\pi Bw}{fs}} \\ B &= 2\cos(\pi)e^{-\frac{\pi Bw}{fs}} \\ A &= 1 - B - C \end{aligned}$$

The amplitude parameter (*Amp*) is relative to the source curve. Notice that in Equation 22 the resonance amplitude shape is always symmetrical respect to its center frequency.

2.4.2.5.2 EpR vocal tract filter

The vocal tract is modeled by a vector of resonances plus a differential spectral shape envelope. It can be understood as an approximation to the vocal tract filter. These filter resonances are modeled in the same way as the source resonance, see equation (3.6), where the lower frequency resonances are somewhat equivalent to the vocal tract formants.

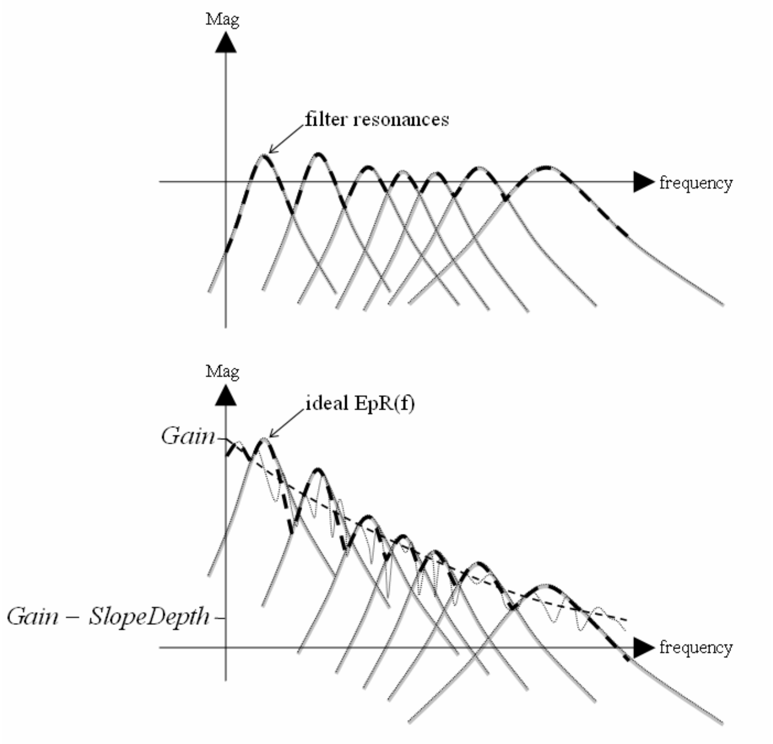


Figure 58: Representation of the EpR filter resonances

The differential spectral shape envelope actually stores the differences (in dB) between the ideal EpR model and the real harmonic spectral shape of a singer's performance. We calculate it as a 30 Hz equidistant step envelope.

$$DSS(f) = envelope_{i=0..}\left[30i, HSS_{dB}(30i) - iEpR_{dB}(30i)\right] \qquad (23)$$

Thus, the original singer's spectrum can be obtained if no transformations are applied to the EpR model.

The EpR filters for voiced harmonic and residual excitations are basically the same, but just differ in the gain and slope depth parameters. The reason is empirically, after comparing the harmonic and residual spectral shape of several SMS analysis of singer recordings. The next figure shows these differences.

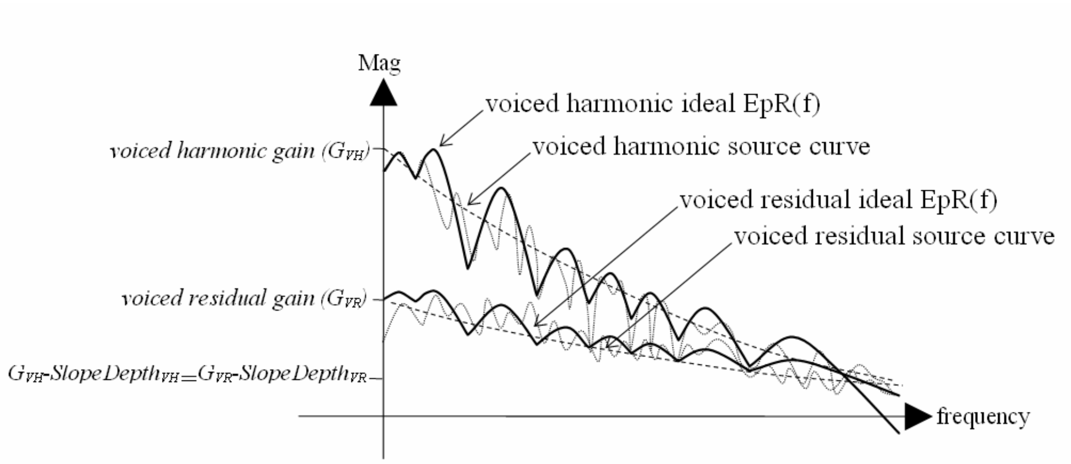

Mag
voiced harmonic ideal EpR(f)
voiced harmonic source curve
voiced harmonic gain (G_{VH})
voiced residual ideal EpR(f)
voiced residual source curve
voiced residual gain (G_{VR})
G_{VH}-SlopeDepth$_{VH}$=G_{VR}-SlopeDepth$_{VR}$
frequency

Figure 59: Differences between harmonic and residual EpR filters

The differential spectral shape envelope actually stores the differences (in dB) between the ideal EpR model (*iEpR*) and the real harmonic spectral shape (*HSS*) of a singer's performance. We calculate it as an equidistant envelope with 30 Hz steps.

$$DSS(f) = envelope_{i=0..}\left[30i, HSS_{dB}(30i) - iEpR_{dB}(30i)\right] \qquad (24)$$

2.4.2.5.3 EpR phase alignment

The phase alignment of the harmonics at the beginning of each period is obtained from the EpR spectral phase envelope. A time shift is applied just before the synthesis, in order to get the actual phase envelope at the synthesis time (usually it will not match the beginning of the period). This phase alignment is then added to the voiced harmonic excitation spectrum phase envelope.

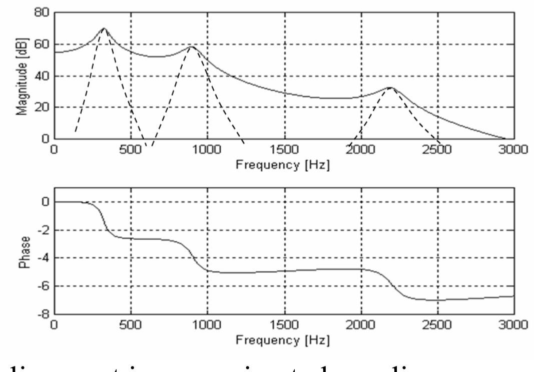


Figure 60: The phase alignment is approximated as a linear segment, with a phase shift for each resonance

The EpR spectral phase model assumes that each filter resonance (not the source resonance) produces a linear shift of π on the flat phase envelope with a bandwidth depending on the estimated resonance bandwidth. Although this value has been obtained empirically, the symmetric second order resonance model itself has a phase shift under the resonance peak. The value of this phase shift depends on the center frequency and the bandwidth of the resonance. This phase model is especially important in order to get more intelligible sounds and more natural low pitch male voices.

2.4.2.5.4 Implementation of the EpR filter

The EpR filter is implemented in the frequency domain. The input is the spectrum that results out from the voiced harmonic excitation or from the voiced residual excitation. Both inputs are supposed to be approximately flat spectrums, so we just need to add the EpR resonances and source curve to the amplitude spectrum. In the case of the voiced harmonic excitation we also need to add the EpR phase alignment to the phase spectrum.

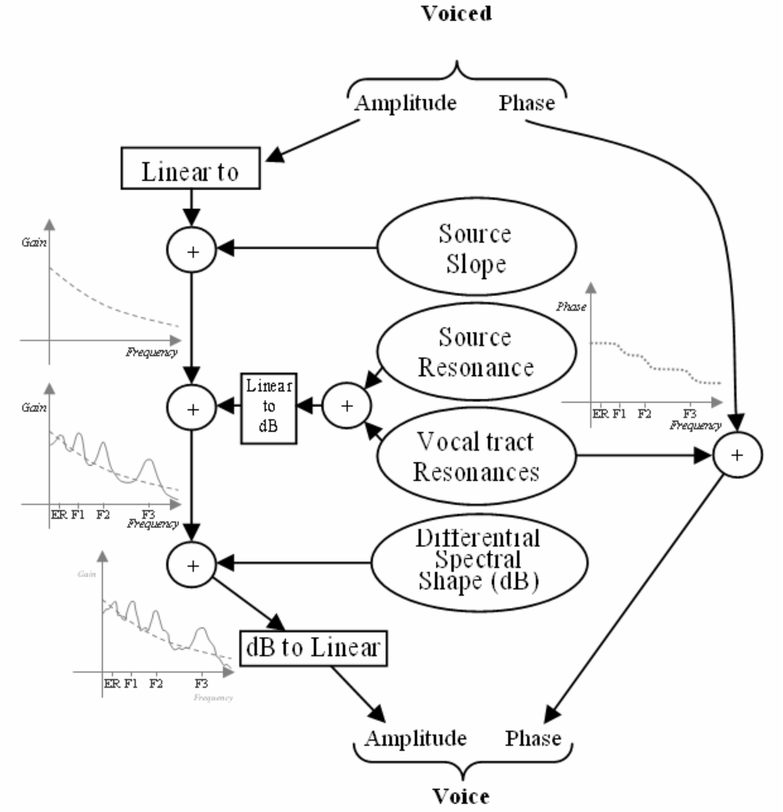


Figure 61: Frequency domain implementation of the EpR model

For each frequency bin we have to compute the value of the EpR filter. This implies a considerable computational cost, because we have to calculate the value of all the resonances. However, we can optimize this process by assuming that the value of the sum of all the resonances is equal to the maximum amplitude (dB) of all the filter and excitation resonances (over the source curve). Then we can even do better by only using the two neighbor’s resonances for each frequency bin. This is not a low-quality approximation of the original method because the differential spectral shape envelope takes care of all the differences between the model and the real spectrum.

If we want to avoid the time domain voiced excitation, especially because of the computational cost of the fractional delay and the FFT, we can change it to be directly generated in the frequency domain. From the pitch and gain input we can generate a train of deltas in frequency domain (sinusoids) that will be convolved with the transform of the synthesis window and then synthesized with the standard frame based SMS synthesis, using the IFFT and overlap-add method. However, the voice quality is a little distorted due to the fact that the sinusoids are assumed to have constant amplitude and frequency along the frame duration.

2.4.3 Pressure waves technologies

2.4.3.1 Glottal pulse model

As seen in the previous section, voice can be modeled as a source, compound of glottal pulse input and aspiration noise, cascaded with an all pole vocal tract filter and a derivative lip radiation filter. The model assumes there is no source and tract interaction or any form of nonlinearity and can be simplified by folding the lip radiation filter into the glottal pulse waveform to give its derivative as shown in Figure 61.

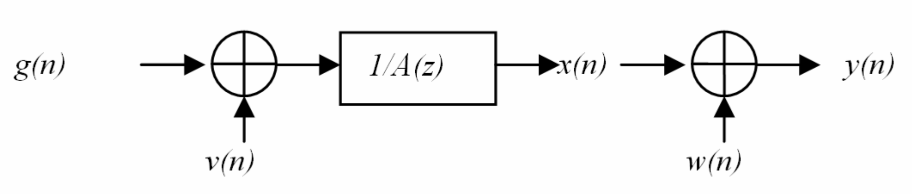


Figure 62: Source filter model representation

Different models exist for the derivative glottal waveform being the most relevant the Rosenberg-Klatt (RK) model, see Klatt and Klatt (1990), and the Liljencrants-Fant (LF) model, see Fant et al. (1985).

The RK glottal pulse model is a simplified derivative version of the more general and popular LF model. The RK model is based on a two-piece polynomial representation proposed in Rosenberg (1971), and according to Jinachitra and Smith (2005) can be described as:

$$g(n)=\begin{cases} 2a_g n/f_s-2b_g(n/f_s)^2, & 0\le n\le T_0\cdot OQ\cdot f_s \\ 0, & T_0\cdot OQ\cdot f_s\le n\le T_0\cdot f_s \end{cases} \qquad (25)$$

with

$$a_g=\frac{27\cdot AV}{4\cdot\left(OQ^2\cdot T_0\right)} \quad b_g=\frac{27\cdot AV}{4\cdot\left(OQ^3\cdot T_0^{\,2}\right)} \qquad (26)$$

where T_0 is the fundamental period, f_s is the sampling frequency, AV is the amplitude parameter, and OQ is the open quotient of the glottal source.

The LF model improves the accuracy of the RK glottal derivative model while keeping the number of parameters small and fairly easy to estimate. The model fits well into the source filter formulation, improves the sound quality of the analysis/synthesis procedure and can capture the characteristics of breathy, normal and pressed voice modes. However, its nonlinear elements (exponentials and sinusoids) do not allow using the joint source-filter parameter estimation procedure proposed in Jinachitra and Smith (2005).

The LF derivative of volume velocity model is defined as:

$$U_g^{'}(t) = E_0 \cdot e^{\alpha t} \cdot \sin\left(\omega_g \cdot t\right) \qquad \mathit{for} \qquad t_1 \le t \le t_3 \qquad (27)$$

$$U_g^{'}(t) = -\frac{EE}{e \cdot TA} \cdot \left[e^{-\alpha \cdot (t - t_3)} - e^{-\alpha \cdot (T_0 - t_3)} \right] \qquad \mathit{for} \qquad t_3 \le t \le T_0 \qquad (28)$$

where ω_g is the pulsation defined by the duration of the opening branch of the glottal pulse:

$$\omega_g = \pi / t_2 \qquad (29)$$

EE is the excitation strength as denoted in Ananthapadmanabha and Fant (1982), "*the derivative of the glottal flow signal is the effective excitation of the vocal tract (EE)*", *TA* is the time constant of the exponential curve and describes the "*rounding of the corner*" of the waveform between *t*4 and *t*3, E_0 is the scale factor, and *α* is the coefficient of the exponentially growing sinusoid.

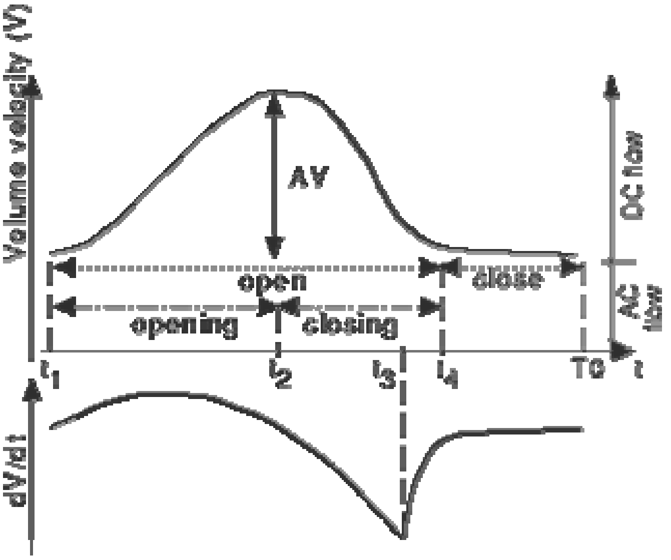


Figure 63: Hand drawn representation of a LF glottal waveform and its time derivative taken from Hanson (1995) where T_0 is the duration of the period, t_1 is the beginning of the separation of the vocal folds and onset of the airflow, t_2 is the instant of the maximum glottal flow through the glottis with maximum amplitude *AV*, t_3 is the moment of glottal closure, and *t*4 is the instant of complete glottal closure and when no airflow occurs. Finally, it is assumed that the area below the modeled curve for the [t_1,t_2] segment is equal to the area of the [t_2,T_0] part of the waveform.

Important parameterization of the glottal flow is given by the Open Quotient and the Speed Quotient. The values of these parameters are highly correlated to different types of phonation.

The Open Quotient (OQ) indicates the duty ratio of the glottal airflow and is defined as the ratio of the time in which the vocal folds are open and the whole pitch period duration ($(t_4$-$t_1)/T_0$ in Figure 63). Substantial changes in the spectrum of the voice excitation imply notable ratio variations.

The Speed Quotient or skewness is an indicator of the glottal pulse asymmetry and is defined as the ratio of rise and fall time of the glottal flow ((t_2-t_1)/(t_4-t_2) in Figure 63). The glottal airflow is usually skewed to the right, which means that the decrease of the airflow is faster than its increase as reported in Rothenberg (1981), Ananthapadmanabha (1984) and Titze (1988).

It is worth stressing the relevance of the Glottal Closure Instants (GCI), noted in Figure 63 as t_3. The estimation of the instants of glottal closure in voiced speech enables the use of larynx synchronous processing techniques where the characteristics of the glottal excitation waveform are separated from those of the vocal tract filter and treated independently in subsequent processing.

2.4.3.2 Overlap and add

Overlap and Add (OLA) is a technique based on windowing consecutive and overlapped segments of the real sound waveform. Using OLA, most frequent transformations are achieved by scaling and repositioning along time each of the individual grains excised from the waveform, see Roads (1978). When the grains are reallocated, the output signal is obtained by interpolating the segments, deriving into the destruction of the original phase relationships, pitch period discontinuities, and distortions.

2.4.3.2.1 Synchronous overlap and add

The Synchronized Overlap-add (SOLA) was originally presented in Roucos and Wilgus (1986). The technique used correlation to reallocate the grains. SOLA computes the cross-correlation of consecutive overlapping grains inside the overlapping area to find the maximum similarity instant. At such instant grains are concatenated by the interpolation of their samples. SOLA improved significantly aforementioned OLA drawbacks and proved to be robust in the presence of correlated or uncorrelated noise, see Wayman et al. (1992).

Evolutions of SOLA include Waveform Similarity Overlap-add (WSOLA), see Verhelst and Roelands (1993), a technique that uses waveform similarity, for example cross-correlation or cross-average magnitude difference function, as synchronicity agent; or Pitch Synchronous Overlap-add (PSOLA), a technique that use pitch as synchronicity agent.

2.4.3.2.2 Pitch synchronous overlap and add

Pitch Synchronous Overlap-add (PSOLA is a technique that operates the grains pitch-synchronously, see Charpentier and Stella (19686). One of the challenges of this technique is the estimation of the pitch period of the signal, even in those cases where a fundamental frequency is missing. There is a popular evolution of PSOLA called Pitch Synchronous Overlap-add (TD-PSOLA), see Moulines et al. (1989), Moulines and Charopentier (1990), and Dutilleux et al. (2002).. For the case of voice, TD-PSOLA estimates those instants corresponding to the maximum amplitude or glottal pulses to synchronize during the voiced fragments of the sound, and performs asynchronously during the unvoiced regions.

2.4.3.3 Formant synthesis

FOF stands for *"Fonctions d'onde formantiques"*, formant wave functions in English, a voice synthesis proposal by Xavier Rodet, see Rodet (1984) and Rodet et al. (1984), consisting in the overlap of time domain functions that emulate individual vocal formants behavior. FOF uses five formant wave functions to mimic the resonance spectrum of the first five formants. Each formant wave function is repeated at every fundament voice period and is parameterized by center frequency, amplitude, bandwidth, decay, and initial phase.

Although initial proposal was meant to synthesize only voiced utterances, later work, see Richard et al. (1992) extended the use of FOFs to unvoiced phonemes.

In 2004, Jordi Bonada proposed in Bonada (2004) an analysis synthesis technique somehow evolved from the FOF concept. His approach was to estimate the radiated voice pulses from spectral analysis instead of generating them as the summation of five formant wave functions. Once estimated the voice pulse, an OLA procedure generates the periodic train of pulses at the desired pitch.

2.4.3.4 Concatenative

Concatenative refer to those synthesis technologies based on concatenating over time elemental units of sound. The elemental units that chain one after the other to generate the synthesis are snippets of real recordings that most of times require from prior transposition and timescaling to adjust to the synthesis score. Although conceptually close to the OLA family, concatenative elemental units enclose higher-level concepts and embrace wider bits; rather than a frame or a grain, they map to notes, note to note transitions, stationeries, attacks, releases, and also diphones, triphones, or syllables for the case of voice.

2.4.3.4.1 Concatenation of sound pieces

Raw concatenation always causes spectral shape and phase discontinuities, consecutive units come from different contexts and their timbre, amplitude and phase properties at the boundaries are almost always substantially different. In order to avoid (at least perceptually) such discontinuities, spectral shape and phase corrections are applied progressively to those frames that surround the juncture so that they reach coherent values at their borders.

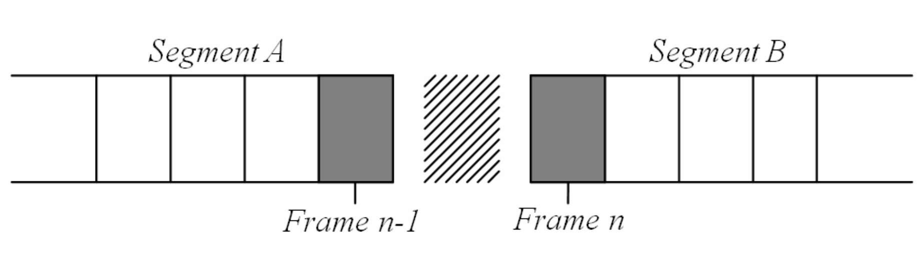


Figure 64: Segments concatenation illustration: last frame at segment A and first frame at segment B have to reach consecutive frames coherence.

2.4.3.4.2 Phase concatenation

In order to avoid phase discontinuities at the segment boundaries, we have come out with a phase continuity condition that takes care of the boundary phases and the possibly different transposition factors applied to each segment. In Figure 65 we can see a representation of the two frames around the boundary; where *frame n-1* is the last frame of the left segment and *frame n* is the first frame of the right segment f_{n-1} and f_n refer to the frequencies of the i^{th} harmonic.

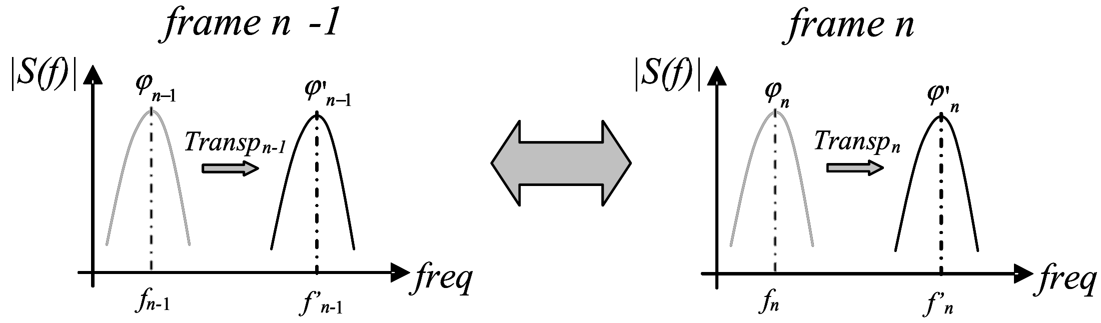


Figure 65: Concatenation boundary

The basic condition for phase continuity comes from the assumption that the frequency of each harmonic varies linearly between this two consecutive frames. In that case the phase relation between both frames should be

$$\varphi_n^i = \varphi_{n-1}^i + 2\pi \frac{f_{n-1}^i + f_n^i}{2} \Delta t \tag{30}$$

where φ_n^i and φ_{n-1}^i are the phases of the i^{th} harmonic at the right and left frame respectively. Thus, the desired phase for the left frame ϕ_{n-1}^i should be

$$\phi_{n-1}^i = \varphi_n^i - 2\pi \frac{f_{n-1}^i + f_{n-1}^i}{2} \Delta t \tag{31}$$

But in fact we are not just concatenating two segments, but also transposing them with different transposition factors (*transp*$_{n\text{-}1}$ and *transp*$_n$). We should distinct then between the original frequency of each segment ($f_{n\text{-}1}$ and f_n) and the transposed ones ($f''_{n\text{-}1}$ and f''_n), where $f''_{n\text{-}1}=f_{n\text{-}1}\cdot transp_{n\text{-}1}$ and $f''_n=f_n\cdot transp_n$. The basic condition should be applied to the transposed frequencies and phases. This can be expressed as

$$\begin{aligned}\phi_{n-1}^{i} = \varphi_{n}^{i} - 2\pi \cdot \frac{f_{n-1}^{i} \cdot transp_{n-1} + f_{n}^{i} \cdot transp_{n}}{2} \cdot \Delta t \\ + 2\pi\left(i+1\right)\Delta t \cdot pitch_{n} \cdot \left(transp_{n} - 1\right)\end{aligned} \qquad (32)$$

This correction is applied either to the left or to the right segment around the boundary, and spread along several frames in order to get a smooth transition. We can rewrite the previous equation as

$$\phi_{n-1}^{i} = \varphi_{n}^{i} - \Delta\varphi_{c} \qquad (33)$$

where $\Delta\varphi_c$ is the phase correction that guarantees the phase continuation of the i^{th} harmonic. In Figure 66 we can see an example where this phase correction is spread along 5 frames on the left part of the boundary.

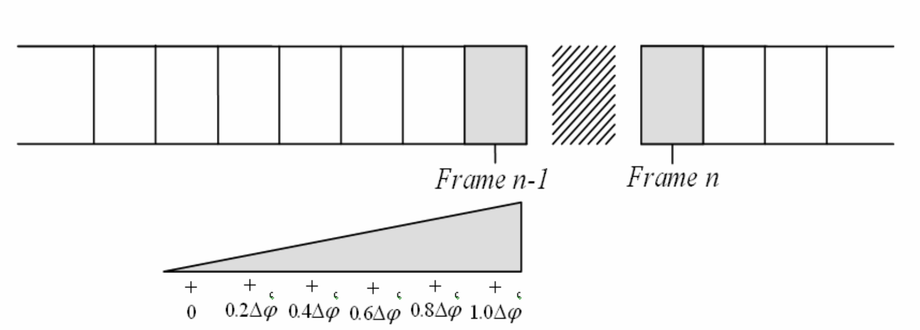


Figure 66**:** Spreading the phase difference in concatenation

Since the impulsive periods of left ant right segments are not aligned, we will often have big phase correction values. Therefore it's better if we calculate how much we should move the right segment (Δt_{sync}) in order to align the beginning of both periods, so that the phase correction to be applied is minimized. We could approximate this time shifting by assuming that the beginning of the periods will be aligned if the phase of the fundamental is continued. This can be expressed as

$$\Delta t_{sync} = \frac{-\Delta\varphi_{c}}{2\pi pitch_{n} \cdot transp_{n}} \qquad (34)$$

where $\Delta\varphi$ is calculated as in equation 34 for the particular case of i=1 (the fundamental). Finally, if we combine both equations (phase continuity plus period alignment) we obtain

$$\begin{aligned}\phi_{n-1}^{i} &= \varphi_{n}^{i} + 2\pi\left(i+1\right) pitch_{n} \cdot transp_{n} \cdot \Delta t_{sync} \\ &-2\pi \cdot \frac{f_{n-1}^{i} \cdot transp_{n-1} + f_{n}^{i} \cdot transp_{n}}{2} \cdot \Delta t \\ &+2\pi\left(i+1\right)\Delta t \cdot pitch_{n} \cdot \left(transp_{n}-1\right)\end{aligned} \tag{35}$$

2.4.3.4.3 Spectral shape concatenation

In order to avoid spectral shape discontinuities at the segment boundaries, we can take advantage of the EpR interpolation which can handle resonances (i.e. formant-like functions) in a nice way. Therefore, it is needed to have an EpR estimation for both frames at the boundary (frame n-1 and frame n). Once the model is estimated for these frames, an intrinsic formant mapping is applied, formant 0 to formant 0, formant 1 to formant 1, and so on. The interpolation can be decomposed into a spectral stretching plus a differential envelope. This differential envelope is obtained subtracting the stretched left EpR (using the formant mapping between EpR$_{n-1}$ to EpR$_n$) to the right EpR.

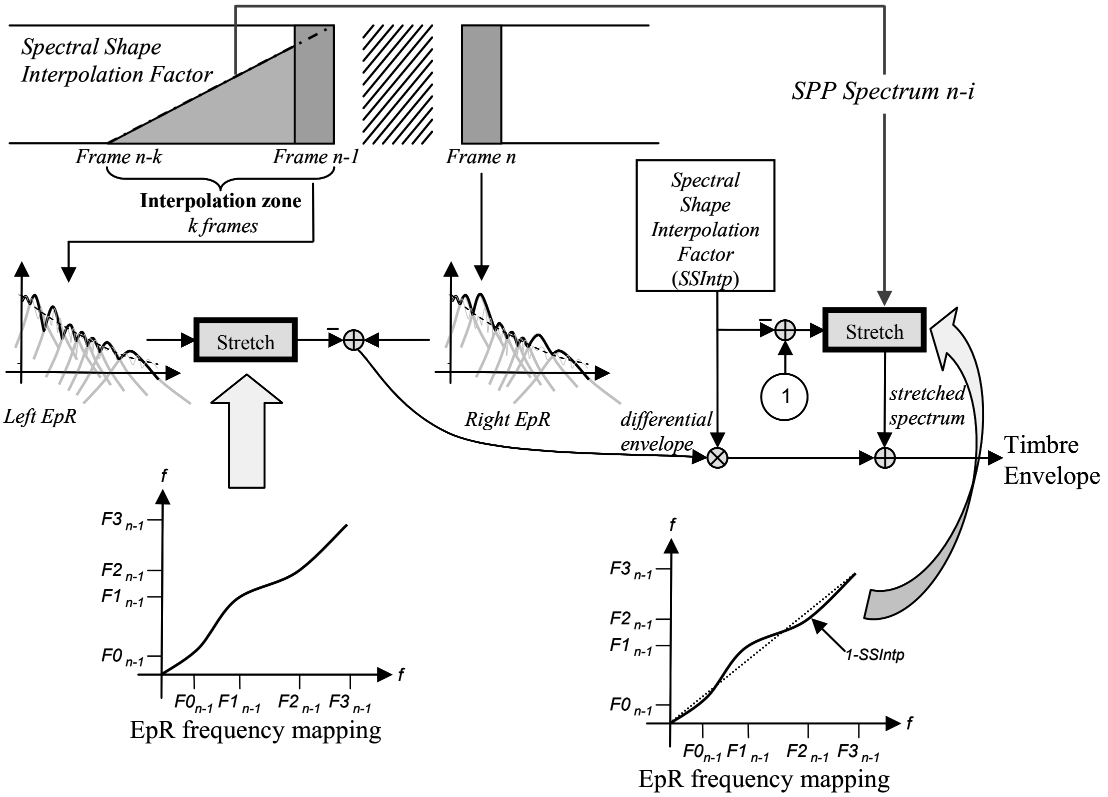


Figure 67: The sample transformation

In Figure 67 we can see how we have implemented this algorithm. Notice that the spectral shape interpolation is spread along several frames in a similar way to the phase concatenation, but with the addition of the spectrum stretching. The frames in the interpolation zone are stretched using the interpolation between the EpR_{n-1} to EpR_n mapping and the identity mapping (y=x), with a mapping interpolation factor equal to 1 minus the spectral shape interpolation factor (1-*SSIntp*). This spectral shape interpolation factor (*SSIntp*) goes from 0 (begin of the interpolation zone, $frame_{n-k}$) to 1 (end of the interpolation zone, $frame_{n-1}$).

Chapter 3

Spectral based description of the singing voice

3.1 Introduction

A whole bunch of general audio descriptors exist and are used in many different audio processing applications. A reasonable compilation of them can be found in Peeters (2004). Although the majority of descriptors presented in this chapter are exclusively voice and singing voice specific, some of them require, or are based on, generic audio descriptors. This does not pretend to be an exhaustive compilation on voice descriptors but rather a general overview of the different voice specific attributes that can be automatically extracted from human voice signals.

The chapter expands from low level to high level descriptors and focuses on presenting the notions and algorithms of those attributes mostly related with the author's contributions.

3.2 Voice excitation descriptors

Voice excitation descriptors refer to those attributes that describe the voice excitation under the source-filter formulation.

3.2.1 Amplitude

Amplitude measures the sound pressure displacement in between the equilibrium atmospheric level. Although voice presents low pressure values (in normal conversation amplitude a particle may be displaced only about one millionth of an inch), its dynamic range (from softest to loudest) is large.

Sound amplitude does not relate directly to the sensation of loudness produced on the listener. Sound loudness is a subjective term describing the strength of the ear's perception of a sound. It is intimately related to sound intensity, defined as the sound power per unit area. Sound intensity must be factored by the ear's sensitivity to the particular frequencies contained in the sound.

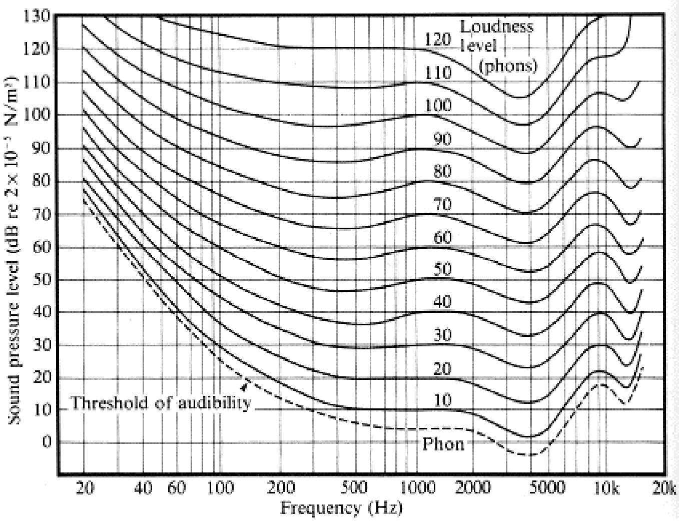


Figure 68: Curves of equal loudness named isophones determined experimentally by Robinson & Dadson in 1956, following the original work of Fletcher and Munson (1993). Although isophones have only been proven valid for pure stable sinusoids, they can be used as well for approximate loudness estimations of complex mixtures, see Pfeiffer (1999).

Amplitude is directly related to the acoustic energy or intensity. Several different measures exist for estimating the amplitude or energy time envelope of a sound as an indication of its level. Amplitude is commonly derived from rectified wave, whether by summing its samples, computing its mean or picking the peak value. Same for energy but using squared wave.

Moreover amplitude and energy envelopes are usually smoothed along time using low pass filters. Thus, even though everyone uses their own formulation, [SB] define 'total amplitude' of a sound clip as the sum of its absolute values

$$At = \sum_{i=1}^{N} |x_i| \qquad (36)$$

and [P] recommends the Root Mean Square (RMS),

$$RMS = \sqrt{\frac{\sum_{i=1}^{N} x_i^2}{N}} \qquad (37)$$

Most commonly, amplitude envelope *A(n)* and energy envelope *E(n)* are defined as:

$$A(n)=\frac{1}{N}\cdot\sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1}\left|x(n+m)\right|\cdot w(m) \tag{38}$$

$$E(n)=\frac{1}{N}\cdot\sum_{m=-\frac{N}{2}}^{\frac{N}{2}-1}\left[x(n+m)\right]^2\cdot w(m) \tag{39}$$

where *w(m)* is an N-point window.

Blaauw proposes in Blaauw (2005) a voice specific amplitude envelope detector with adaptative smoothing criteria depending on the windowed audio frequency content. The detector computes the zero crossing rate and the derivative of a normalized coarse (~11.6 ms window) peak envelope. These two values, together with pitch estimation when available, decide the most adequate smooth function to apply to a fine (~3.85 ms window) peak envelope. With this, fine amplitude envelope is computed in such a way transients in plosives are preserved while sustained vowels and fricatives present smooth wrappers.

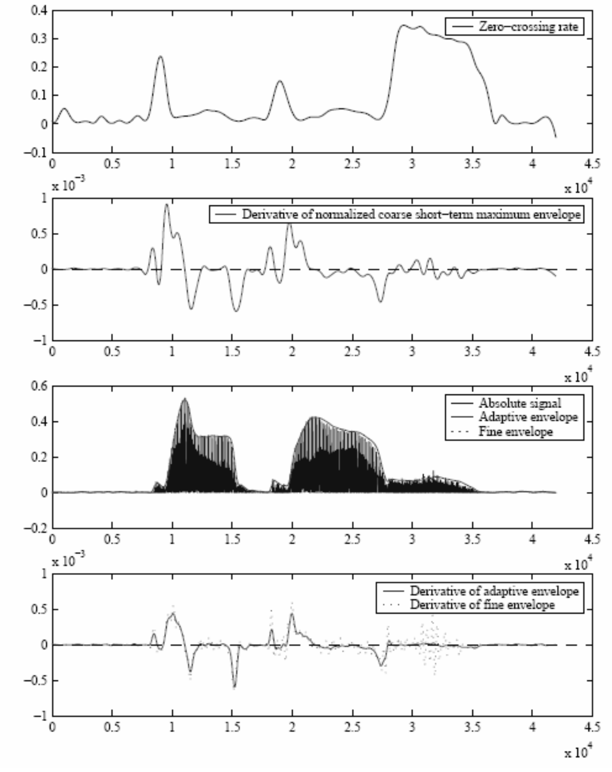


Figure 69: Illustration of Blaaw's voice amplitude envelope estimation. Taken from Blaauw (2005) with permission of author.

Alternative amplitude measures for modulated speech signals components are based on the Analytic Signal and the Teager Energy Operator, see Vakman (1996). Being *x(t)* the original signal and *z(t)* its analytic:

$$x(t) = a(t) \cdot \cos[\Phi(t)] \tag{40}$$

$$z(t) = x(t) + j \cdot \hat{x}(t) = r(t) \cdot \exp[j \cdot \Phi(t)] \tag{41}$$

and its Hilbert transformed $\hat{x}(t) = x(t) * \left(\frac{1}{\pi \cdot t} \right)$
the Hilbert Transform separation algorithm is given by:

$$r(t) = \sqrt{x^2(t) + \hat{x}^2(t)} \approx |a(t)| \tag{42}$$

And defining *x(t)* as the original signal and Ψ_c[*x(t)*] as its Teager operated.

$$x(t) = a(t) \cdot \cos\left[\int_0^t \omega_i(\tau) \cdot d\tau \right] \tag{43}$$

$$\Psi_c[x(t)] = \dot{x}(t) - x(t) \cdot \ddot{x}(t) \approx [a(t) \cdot \omega_i(t)]^2 \tag{44}$$

it is proved in Maragos et al. (1993) that the Energy Operator Separation Algorithm is given by:

$$\sqrt{\frac{\Psi_c[\dot{x}(t)]}{\Psi_c[x(t)]}} \approx |a(t)| \tag{45}$$

3.2.2 Pitch

Pitch estimation is a very old and complex subject that has been addressed countless times with all kind of different approaches in different fields of the digital signal processing discipline. Several articles report evaluation of most popular implementations and present comparisons between their performances, Rabiner et al. (1976), Fonollosa (1996), Cuadra et al. (2001), and Gerhard (2003), .

The implementations that address specifically pitch analysis for human voice signals can be fairly classified into those operating mainly in time and those operating mainly in frequency. Time domain techniques for voice fundamental frequency estimation rely on different techniques of period detection to find the time elapsed between repetitions of a slow-time-varying pattern, see Schaffer et al. (1974), Rabiner (1977), Sukkar et al. (1988), Sood and Krishnamurthy (1992), Terez (2002), and Cheveigné and Kawahra (2002). Frequency domain techniques for voice fundamental frequency estimation rely on harmonic peaks detection to obtain the harmonic structure that explains better the input voice spectra (magnitude and / or phases), see Schroeder (1968), Hess (1992), Maher and Beauchamp (1994), Kim and Kim (1996) and Lu (1999). There are as well hybrid approaches that rely precisely on the combination of operations in both domains, see Boyanov et al. (1993).

In general, basic time domain approaches tend to perform more efficiently in terms of computational cost while basic spectral domain approaches tend to perform more accurately. However, although pitch estimation is considered a solved topic, there is not a single pitch estimator with perfect performance in the analysis of the singing voice. Inputs such as high pitched singing, growls, pressed phonation uttering, fried attacks, voiced fricatives and plosives, unvoiced transitions and others usually result into misleading estimations. According to Sood and Krishnamurthy (2004), best pitch detectors rarely achieve even 97% accuracy for human voice.

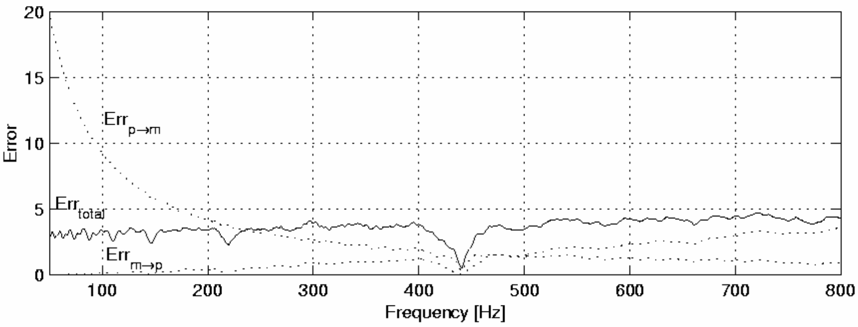


Figure 70: Error functions for the Two Way Mismatch procedure pitch detection procedure proposed in Maher and Beauchamp (1994)

3.2.3 Vibrato and tremolo

Vibrato can be defined as a continuous sinusoid-like modulation in the fundamental frequency of an instrument. Although some researchers believe vibrato in voice appears as a muscular fatigue symptom, see Sundberg (1987), the fact is singers use it as an expressive resort to enhance a melody and to make long sustained notes be perceived more alive. Most performers attach their vibrato to a voice amplitude modulation named tremolo that resonates at the same vibrato frequency.

The analysis of the vibrato and tremolo and the parameterization of their respective depths and their common rate has been a frequent subject of study. Most related literature dig for implications in physiology and musicology disciplines at the time they present digital signal processing algorithms for detection and estimation of the vibrato and tremolo pulsations. Most relevant literature in this field include Sundberg (1978, 1979), Castallengo et al. (1989), Horii (1989), Shipp et al. (1990), Herrera and Bonada (1998), Leydon et al. (2003), and Jeon and Driessen (2005).

3.2.4 Rough and growl

Standing on the voice-source model point of view, the vocal disorders that are being considered here come basically from the aperiodicities of the voiced excitation, that is, from the periodic train of pulse-like waveforms that corresponds to the voiced glottal excitation.

Roughness in voice can come from different pathologies such as biphonia, or diplophonia, and can combine with many other voice tags such as hoarse or creaky, see

Titze (1994). In this paper we will not stick to the rigorous rough voice definition but we will refer to rough voice as the one due to cycle to cycle variations of the fundamental frequency (jitter), and the period amplitude (shimmer).

Singers in jazz, blues, pop and other music styles often use the growl phonation as an expressive accent. Perceptually, growl voices are close to other dysphonic voices such as hoarse or creaky, however, unlike these others, growl is always a vocal effect and not a permanent vocal disorder.

According to Sakakibara et al. (2004) growl comes from simultaneous vibrations of the vocal folds and supra glottal structures of the larynx. The vocals folds vibrate half periodically to the aryepiglottic fold vibration generating sub-harmonics.

3.2.4.1 The growl observation

The behavior of the growl sub-harmonics in terms of magnitude and phase vary quite a lot from one voice to another, from one pitch to another, from one phrase to another, etcetera. However certain patterns appear quite frequently. These patterns, which are explained next, are the ones that the growl effect applies.

If a growl utterance is observed in time domain, it is most of the times easy to recognize which is the real period of the signal and which is the macro period due to growling as it is in Figure 70. In the observations made growl phonation appeared to have from two to five sub-harmonics. In Figure 71 example, the spectrum presents three sub-harmonics placed at $F_0 \cdot (m+k/4)$ (for *m*=0..number of harmonics, and *k*=1..3). Thus, three inner periods can be distinguished in between a growl macro period.

Regarding the magnitudes, in the band that goes from the fundamental frequency up to approximately 1500 Hz, the sub-harmonic peaks are commonly located below the spectral envelope (defined by the harmonic peaks). In this band, the closer the sub-harmonic is to the nearest harmonic, the higher its magnitude is. In the upper band, from approximately 1500 Hz to half the sampling rate, sub-harmonics go along with the harmonic spectral shape.

Regarding the phase, for a growl utterance with *N* sub-harmonics, the typical behavior of the phases of the sub-harmonics is to get approximately aligned with the phase of the left harmonic peak every *N*+1 periods as illustrated in figure 6.

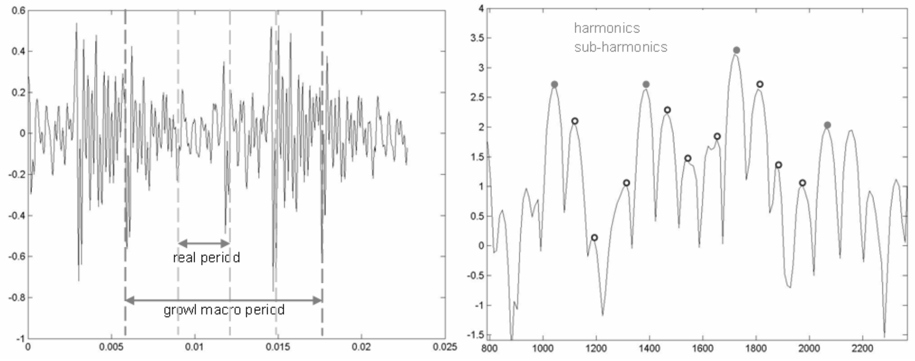


Figure 71: Partial waveform (upper) and spectrum representation of a growl utterance

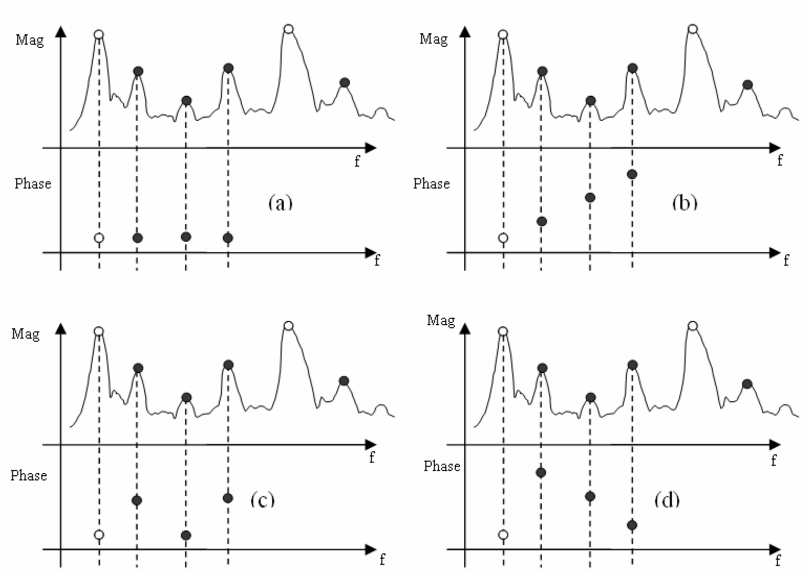


Figure 72: Representation of the spectrum phase behavior of a growl voice in the beginning of four consecutive periods (a,b,c,d) for N=3 subharmonics

For the harmonic peaks, harmonic *i* is always 1 cycle below harmonic *i*+1. In between them, for the sub-harmonics peaks, the peak phase can be generally expressed as:

$$\varphi^{sh}_{p,k} = \varphi^{h}_{i} + \frac{2\pi}{N+1}\cdot(k+1)\cdot p \quad \text{for } k\text{=0,1,2 and } p\text{=0,1,2,3} \tag{46}$$

being *p* the inner period index (*p*=0 for Figure 71.a and *p*=3 for Figure 71.d), *k* the sub-harmonic peak index in between consecutive harmonic peaks, and *N* the number of sub-harmonics.

3.2.5 Pulse regularity

Voice pulses regularity is commonly given by two parameters: jitter and shimmer. Jitter describes the non-uniformities of the voice pulses along time while shimmer describes the voice pulses variations along amplitude. Human voice excitation, because of its inherent nature, does never produce a train of pulses equally spaced in time and equally leveled in amplitude from one period to the next.

Although different formulations exist for the jitter and shimmer, they are all distortion measures that describe the same effect. Spectral based method exist for jitter and shimmer estimations based on the sinusoidal of the peaks; however most efficient and consistent methods give results based on time analysis. One of these different measures is given by the ratio between the summation of the absolute deviations from a regression line approximation of the magnitudes and the summation of magnitudes:

$$\mathrm{Re}\,gularity = \frac{\sum \left| PulseMagnitude[i] - LinRgr[i] \right|}{\sum PulseMagnitude[i]} \tag{47}$$

Where magnitude can refer whether to pitch or amplitude and *LinRgr*[i] is a first order polynomial approximation of the magnitudes neighboring a center pulse. Both pitch and amplitude magnitudes can be derived from standard estimation algorithms. The regression approximation is estimated by simply tracing a line from the first neighboring magnitude taken into consideration to the last neighboring magnitude taken into consideration.

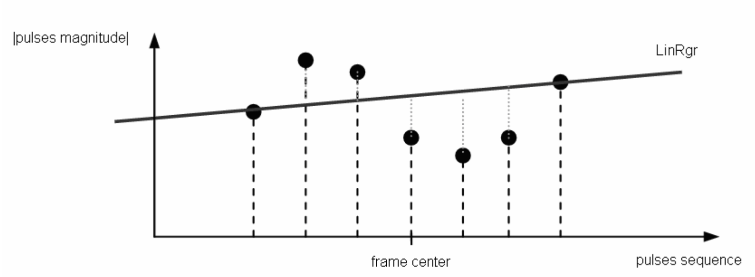


Figure 73: Illustration of a possible data set of pulses magnitudes and subsequent regression line. Seven pulses are taken into consideration in this example.

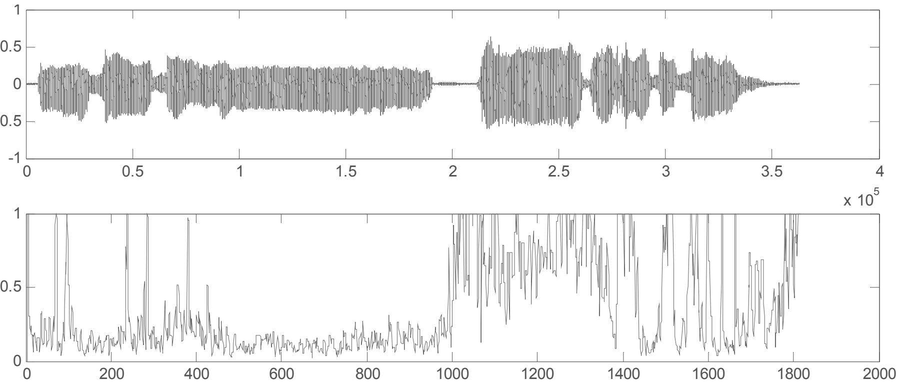


Figure 74: Pulse amplitude regularity descriptor for a couple of sung musical phrases. First phrase (on the left) has normal phonation. Second phrase (on the right) has an intentional rough phonation. As aforementioned in previous section, vocal disorders present significant deviations of amplitude or period in the train of glottal pulses.

3.2.6 Breathiness

The breathy voice is caused by a particular mode of phonation. According to Sundberg (1987), there is a phonatory dimension ranging from pressed to breathy phonation. When a high sub-glottal pressure is combined with a high degree of adduction, a pressed phonation occurs. If the sub-glottal pressure decreases, jointly with a lower adduction, it is called a *flow phonation*. Then, if the adduction force is still reduced, the vocal folds fail to make contact, producing the known breathy phonation. Most relevant studies in this specific topic have been carried out by Rahul Shrivastav, see Shrivastav and Sapienza (2003) and Shrivastav (2003).

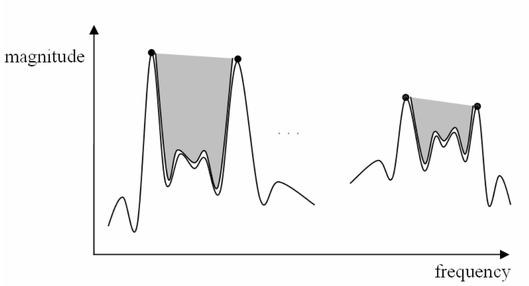


Figure 75: Hand drawn illustration of two different inter-harmonic peak areas. Harmonic peaks are marked with a black dot. Areas are grey shadowed and are their top boundary is defined by the linear interpolation between consecutive harmonic peaks.

Perceptually, a breathy phonation results into a harmonic spectrum mixed with high-frequency noise, due to the air turbulence. Thus, the standard approach for breathiness characterization is to examine the harmonic content at high frequencies. Next, an algorithm for breathiness estimation based on harmonic spectral peaks detection is presented. The estimator computes the area between the harmonic peaks envelope and the actual spectrum, as illustrated in Figure 75. Since breathiness nature is mostly perceived when singing long sustained notes, algorithms do not take into account the non-stationeries.

The equation 9 gives us the final formula for the *Breathiness Factor*. In this case, experimental results showed that the best frequency boundaries were *4kHz* and *9kHz*.

$$Breathiness = 1 - \sum_{k=4000\cdot\frac{N_s}{F_s}}^{9000\cdot\frac{N_s}{F_s}} \frac{envelope[k] - spec[k]}{N_b} \tag{48}$$

where N_s is the Fast Fourier Transform size, F_s is the sampling rate and N_b the number of bins used considered in the summation.

A more computational efficient approximation can be used instead for the estimation of the breathiness. The approximation consists on averaging for all peaks in the range [1, 5] kHz the difference between the averaged magnitude between a peak and its consecutive and the magnitude value and the frequency bin placed in the middle of them:

$$Breathiness = \frac{1}{N}\sum_{n} \frac{PeakMag[n] + PeakMag[n+1]}{2} - PeakMag\left[\frac{2n+1}{2}\right] \tag{49}$$

Moreover, a normalization correction factor is added to the obtained in order to make the algorithm as pitch uncorrelated as possible. This correlation with the pitch is due to the fact that an individual voice singing at different fundamental frequencies will present different spacing between harmonic peaks. This factor is:

$$Breathiness+=Correction=\begin{cases}-9.8-35(pitch-150)/100 & pitch<110\\ -12(pitch-150)/100 & 110<pitch<290\\ -14-2(pitch-150)/100 & pitch>290\end{cases} \tag{50}$$

where threshold values are indicated in Hz

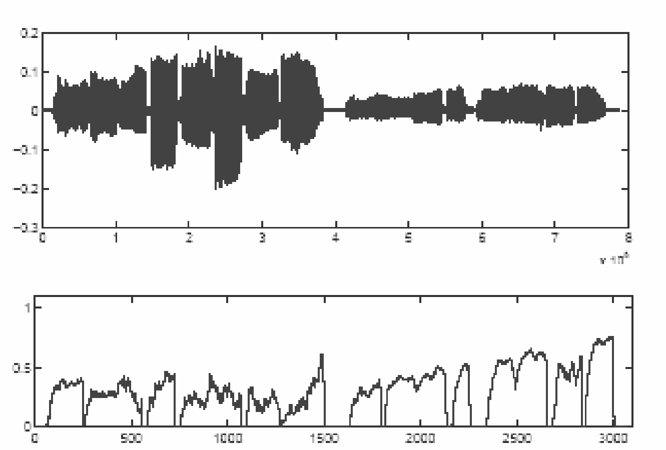


Figure 76: Breathiness factor computed as the spectral peak envelope differential using formula F. The sound example is a sung musical scale with modal phonation (first) and breathy phonation (second).

3.3 Timbre descriptors

Timbre can be defined as the combination of qualities of a sound that distinguishes it from others of the same pitch and volume. Timbre is one of the main descriptors in voice, holding language and speaker information: different phonemes have different timbres making language understandable and different speakers possess different timbres for each phoneme making individuals recognizable, see Zhang (2003) and Bartsch and Wakefield (2004).

Timbre is tied to the notion of sound perception and thus, because of the close relationship between spectral domain and human sound perception means, to the distribution of sonic energies along the spectrum. For such reason timbre is also very commonly referred to as spectral envelope.

Voice timbre can be parameterized using all kind of different models and algorithms such as Filter coefficients, Cepstrum coefficients, Spectral Envelope, Formant-based representation, Geometric representation, or Wavelets-based representation.

Filter coefficients refer to representations such as the LPC filter coefficients representation.

Cepstrum is defined as the Fourier Transform of the log (with unwrapped phase) of the Fourier Transform of a signal and can be intuitively interpreted as the magnitude spectrum of a signal, low pass filtered to get rid of the rapid fluctuations.

Spectral envelopes refer to an amplitude curve in the frequency domain defined by approximately equidistant or logarithmically spaced frequency points, being one of its most typical representatives the sinusoidal harmonic envelope.

Formant-based representation require from formant detection, a challenging research topic, especially in cases such as high pitched singing or real-time applications. Formants are defined by a set of three dimensions: frequency, amplitude and bandwith; and most of the times include a resonant filter to model that can emulate formant's behavior.

Geometric representations describe the amplitude curve of the spectral envelope using break-point functions or splines.

Wavelets refer to a representation of a signal based on scaled and translated copies of a fast decaying oscillating waveform and can be interpreted as a multi-resolution Shot-time Fourier Transform; see Chan (1995) for an introduction.

Only some of these can be considered voice-specific (LPC and formant-based) while the rest can describe any pseudo-harmonic instrument timbre. All sort of different approaches are overviewed and compiled in Shwarz (1998), Shwarz and Rodet (1999), Kleijn (2000,2003), Röbel and Rodet (2005), and Ekman et al, (2006).

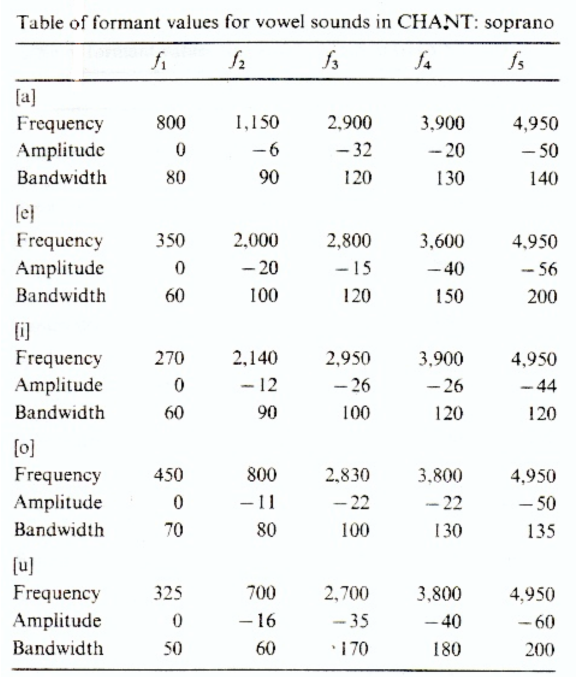


Table of formant values for vowel sounds in CHANT: soprano

	f_1	f_2	f_3	f_4	f_5
[a]					
Frequency	800	1,150	2,900	3,900	4,950
Amplitude	0	−6	−32	−20	−50
Bandwidth	80	90	120	130	140
[e]					
Frequency	350	2,000	2,800	3,600	4,950
Amplitude	0	−20	−15	−40	−56
Bandwidth	60	100	120	150	200
[i]					
Frequency	270	2,140	2,950	3,900	4,950
Amplitude	0	−12	−26	−26	−44
Bandwidth	60	90	100	120	120
[o]					
Frequency	450	800	2,830	3,800	4,950
Amplitude	0	−11	−22	−22	−50
Bandwidth	70	80	100	130	135
[u]					
Frequency	325	700	2,700	3,800	4,950
Amplitude	0	−16	−35	−40	−60
Bandwidth	50	60	170	180	200

Figure 77: Table containing the frequency, magnitude, and bandwidth of the first five vowel formants of a soprano singer. Taken from Rodet and Bennett (1989) with author's permission

There are specific cases in which partially meaningful descriptions of voice timbre fulfil the requirements. So is the use-case we present next where a timbre-related estimation is obtained in real time. The scenario will be used to present some of the most popular voice timbre parameterizations as well as some straight-forward alternatives.

3.3.1 Whawactor: real time timbre variation estimation

Although the wah-wah effect was initially developed by trumpet players using mutes in the early days of jazz, it has became known as a guitar effect ever since Jimi Hendrix popularized Vox Cry-baby pedal in the late 60's. A wah-wah guitar pedal contains a resonant bandpass filter with a variable center frequency that is changed by moving the pedal back and forth with your foot. Usually, a knob controls the mix between original and filtered guitar signals as represented in Figure 76.

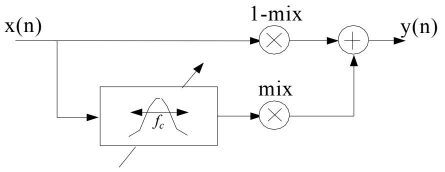


Figure 78: general wah-wah effect block diagram

The explanation of the why wah-wah effect resembles human [wa-wa] utterance is found on the voice spectral characteristics of the vocalic phonemes [u] and [a], in particular, on the first formant location. Considering the [u] vowel first formant is around 350 Hz, and the [a] vowel first formant is around 700 Hz, the [u] to [a] articulation produces a modulated sound due to the trajectory of the first formant that is perceived as the effect of a resonant filter moving upwards in frequency.

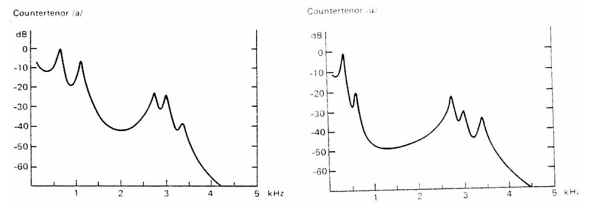


Figure 79: Spectral envelopes of vowels [a] (left) and [u] (right) for a countertenor, scanned from Rodet and Bennett (1989) with permission of authors; the formants different locations are clearly distinguishable.

There is already a musical instrument named talk-box and interface developed by ATR Media Integration & Communication Research Labs that profits from the link between the wah-wah effect and the [wa-wa] utterance. The talk-box redirects the sound of an amplifier to the performer's mouth where lips and vocal cavities modulate the sound; a small microphone collects the resulting shaped sound and sends it to the amplifier. The interface, called Mouthesizer in Lyons et al. (2001), uses a video camera to measure the opening of the performer's mouth and changes the wah-wah filter centre frequency according to this measure. It was in fact the multifaceted requirements of such a system what made us think about an alternative straightforward solution.

The Wahwactor is a two-input and one-output system. Out of the two input tracks, one of the tracks may be considered a control rather than a proper input since it is the one in charge of driving the transformations to be applied to the audio signal. In the context of the Wahwactor, the audio signal is typically a guitar signal and the control signal is a voice [wa-wa] utterance signal.

First, the voice signal is analyzed to pick up a meaningful descriptor. Then a simple conversion (shift, scale and smooth), is used as the centre frequency of the wah-wah filter, through which the guitar signal is sent to be mixed with the original.

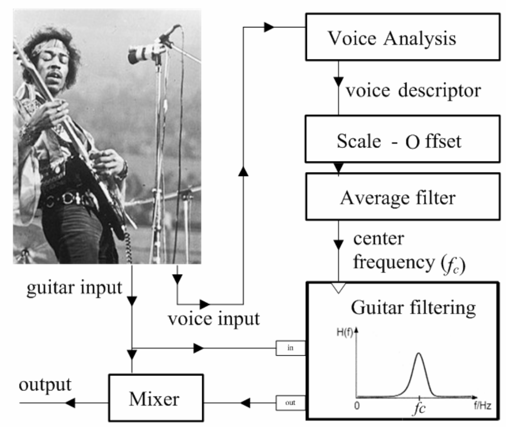


Figure 80: The Wahwactor block diagram.

To work in real-time, the Wahwactor uses a frame-by-frame algorithm described by the diagram illustrated in figure 3. The voice signal is sampled at 44100 Hz and analyzed using a 2100 sample Hamming window and a 2048 point Fast Fourier Transform. The guitar signal is sampled at 44100 Hz and filtered using 2100 sample length buffers. The algorithm uses a 1050 sample hop size so that we have a 50% overlap in synthesis. This overlap is necessary to smooth the filter phase frame to frame variations.

The voice analysis step performs the extraction of the voice descriptor that is mapped to the control of the resonance filter frequency. Next, five different voice descriptors are presented and evaluated as candidates. Being aware of the fact that lower formants contain most of the phonetics (intelligibility) whether higher formants relate to personality, the proposed descriptors focus their analysis on the low/mid-band spectra.

3.3.1.1 'Cepstrum': MFCC's variation

Mel-Frequency Cepstral Coefficients (MFCC's) are considered to be a very useful feature vector for representing the timbral characteristics of the human voice. Here we propose the '*Cepstrum*' descriptor to be the sum of the variations of all MFCC's but the first, which is the energy coefficient. This is:

$$Cepstrum = \sum_{i=1}^{N-1} \Delta MFCC_i \tag{51}$$

where N is the number of cepstral coefficients, and delta MFFC's are computed as the maximum variation between current frame and previous ones:

$$\Delta MFCC_i = \max_m \left(MFCC_i^k - MFCC_i^{k-m} \right) \tag{52}$$

where i is the cepstral coefficient index, k is the current frame index and m=1,2,3.

The computation of the MFCC's has been implemented using Malcolm Slaney's Auditory Toolbox, see Slaney (1998), taking N=13, and using 40 filters inside the (0.7, 6) KHz band.

3.3.1.2 'LPC': LPC roots

As already explained in previous chapters, Linear Predictive Coding (LPC) models the human vocal tract as an infinite impulse response filter. With such model, the formant frequencies can be estimated from the roots of this vocal tract filter. LPC algorithm gives the coefficients of an all pole filter in the form:

$$H(z) = \frac{b_0}{a_0 + a_1 z^{-1} + a_2 z^{-2} + \ldots + a_N z^{-N}} \tag{53}$$

This filter can be obtained from an estimation $x_p(n)$ of current sample $x(n)$ using last N samples, minimizing the prediction error $err(n) = x(n) - x_p(n)$ in the least square sense and using the Levinson-Durbin autocorrelation method to solve.

LPC filter can be also expressed as:

$$H(z) = \frac{b_0}{\left(1 - p_1 z^{-1}\right)\left(1 - p_2 z^{-1}\right)\ldots\left(1 - p_N z^{-1}\right)} \tag{54}$$

whrere we can translate the phases of each of the poles θ_i into frequencies (as shown in Figure 81):

$$f_i = \frac{\theta_i}{2\pi} \cdot F_S \tag{55}$$

where F_S is the sampling frequency.

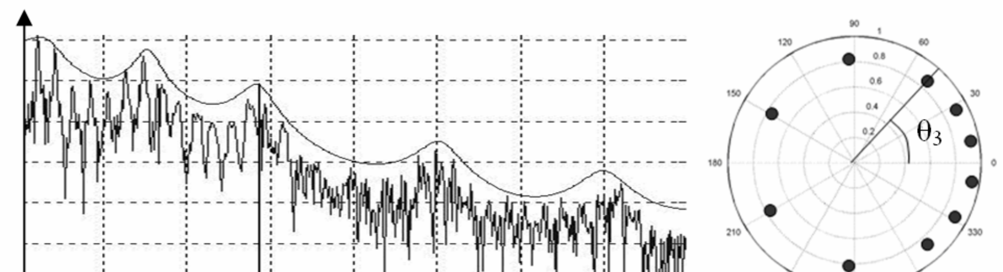


Figure 81: On the left, log magnitude spectrum and its LPC representation. On the right, third formant pole (f3) polar representation (θ3).

The proposed *'LPC'* descriptor is the angle of the root of the LPC filter that has the maximum amplitude inside the [0, π) phase segment. For the LPC analysis, the voice signal is down-sampled to approximately 4 KHz and the number of LPC coefficients is set to 4.

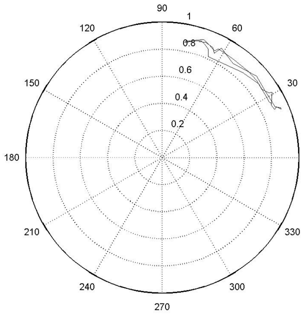


Figure 82: Polar coordinate plot of the 'LPC' root trajectory for a [wa-wa] utterance.

3.3.1.3 'Slope': Low-band Harmonics Slope

The *'Slope'* descriptor is defined as the slope of the harmonic peaks of the spectrum in the [500, 1500] Hz band. The computation of this descriptor employs a pitch detection algorithm based on the Two-way Mismatch Procedure, see Maher and Beauchamp (1994) and uses peak detection and peak continuation algorithms from Zolzer (2002). The slope of the harmonic peaks is obtained using a least-squares regression line.

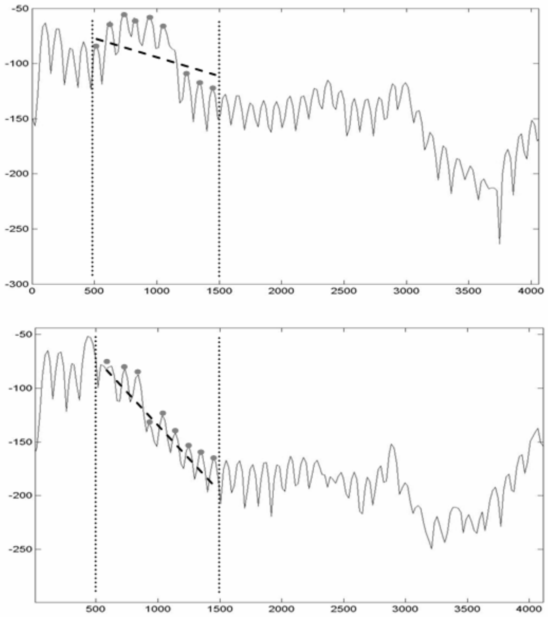


Figure 83: Log magnitude spectra representation of vowels [a] (upper) and [u] (lower) showing considered harmonic peaks and an approximation of their slope.

Notice that such descriptor presents inconsistency with very high pitched voices: a voice whose pitch is close to 800 Hz will only have one harmonic peak in the analysis band. Although such high pitch values are not usual, we have to bear in mind a couple of cases. First, the guitar player frequently utters the [wa-wa] at the pitch of the guitar notes that are being performed, singing in falsetto if necessary. Second, the pitch detection may sometimes give one octave high errors.

3.3.1.4 'Centroid': Low-band Spectral Centroid

The spectral centroid is the barycentre of the magnitude spectrum, see Peeters (2004) and it is usually defined as in Serra and Bonada (1998):

$$Centroid = \sum_{k=0}^{N-1} \frac{k \cdot f_s}{N} \cdot \frac{|X(k)|}{\sum_{k=0}^{N-1} |X(k)|} \tag{56}$$

where k is the spectral bin index, N the number of points of the FFT, f_s the sampling rate frequency, and *X(k)* the sound spectrum. Our *'Centroid'* descriptor is a particularization of the definition above that only takes into account those frequency bins k that fulfill:

$$k_F = round\left(500 \cdot \frac{N}{f_s}\right) < k < round\left(1500 \cdot \frac{N}{f_s}\right) = k_L \tag{57}$$

Notice here that the descriptor suffers from harmonic peaks that move around the boundaries of the computation frequency band along time, getting in and out from frame to frame. This effect becomes problematic when these swerving peaks are the prominent peaks of a formant.

3.3.1.5 'Area': Low-band Spectral Weighted Area

The *'Area'* descriptor is defined as

$$Area = \sum_{k=k_F}^{k_L} \frac{k \cdot f_s}{N} \cdot |X(k)| \tag{58}$$

where k_F and k_L take the values defined in equation 4. Since f_s/N is the inter-spectral bins frequency step, the descriptor can be understood as the local low-band linearly-weighted spectrum area.

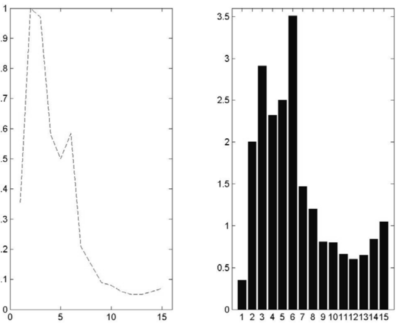


Figure 84: Low-band linear magnitude spectrum (left) and its 'Area' descriptor representation (right).

This descriptor is somewhat related to the weighted additive difference between consecutive spectral shapes used to get the onset detection feature of high frequency content in Masri and Batterman (1996).

In the spectral domain, energy increases linked to transients tend to appear as a broadband event. Since the energy of the signal is usually concentrated at low frequencies, changes due to transients are more noticeable at high frequencies, see Rodet and Jaillet (2001). To emphasize this, the spectrum can be weighted preferentially toward high frequencies before summing to obtain a weighted energy measure. In Masri (1996), Masri proposes a high frequency content (HFC) linear weight function where each bin's contribution is in proportion to its frequency. The HFC function produces sharp peaks during attack transients and is notably successful when faced with percussive onsets, where transients are well modeled as bursts of white noise.

$$\widetilde{E}(n) = \frac{1}{N} \cdot \sum_{k=-\frac{N}{2}}^{\frac{N}{2}-1} |k| \cdot |X_k(n)|^2 \tag{59}$$

3.3.1.6 Evaluation of the descriptors

As pointed out in Koestoer and Paliwal (2001), the greatest disadvantage of using LPC with pseudo-harmonic sounds is that the algorithm tends to envelope the spectrum as tightly as possible. This might cause, for high pitched sounds and high order LPC's, the algorithm to descend down to the level of residual noise in between two harmonic peaks.

In terms of robustness, the '*Slope*' descriptor cannot be considered a good candidate because of its high pitch inconsistency. Nor does the *'Centroid'* seem to be the perfect choice because of the swerving peaks problem. Although it may give the impression that the '*Area*' descriptor should also suffer from this problem, the linear weighting attenuates its effects to the point that it is unnoticeable. In fact, taking a look at *'Centroid'* and '*Area*' descriptors in Figure 86, we observe that noisy secondary peaks appear in *'Centroid'* valleys whereas valleys in the *'Area'* descriptor are smoothly shaped. Finally, because the '*LPC*' descriptor is extremely dependent on the sub-sampling frequency and number of coefficients parameters, it can not be considered a robust descriptor.

In terms of computational cost, we have estimated the seconds it takes to compute each of the descriptors for a specific [wa-wa] utterance. Results are shown in Figure 85. From these results, computational cost considerations force us to discard '*Cepstrum*' and '*Slope*' descriptors.

In terms of reliability, by taking a look at Figure 86, we can state that all descriptors but the '*Cepstrum*', which lacks smoothness, are valid choices. However, if we pay attention to the descriptors behaviour over the fifth [wa] utterance of the sample (which has a long linear progressive transition from [u] to [a]), we can consider '*LPC*' and '*Area*' descriptors to be the ones that are better linked to the phonetic evolution. At least, since it is difficult to measure such a concept, they are the ones that are better linked to the user's intentions.

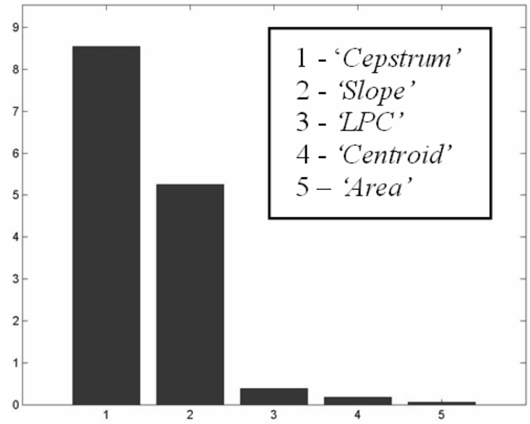


Figure 85*:* Computational cost in seconds of all proposed descriptors for the utterance used in Figure 86.

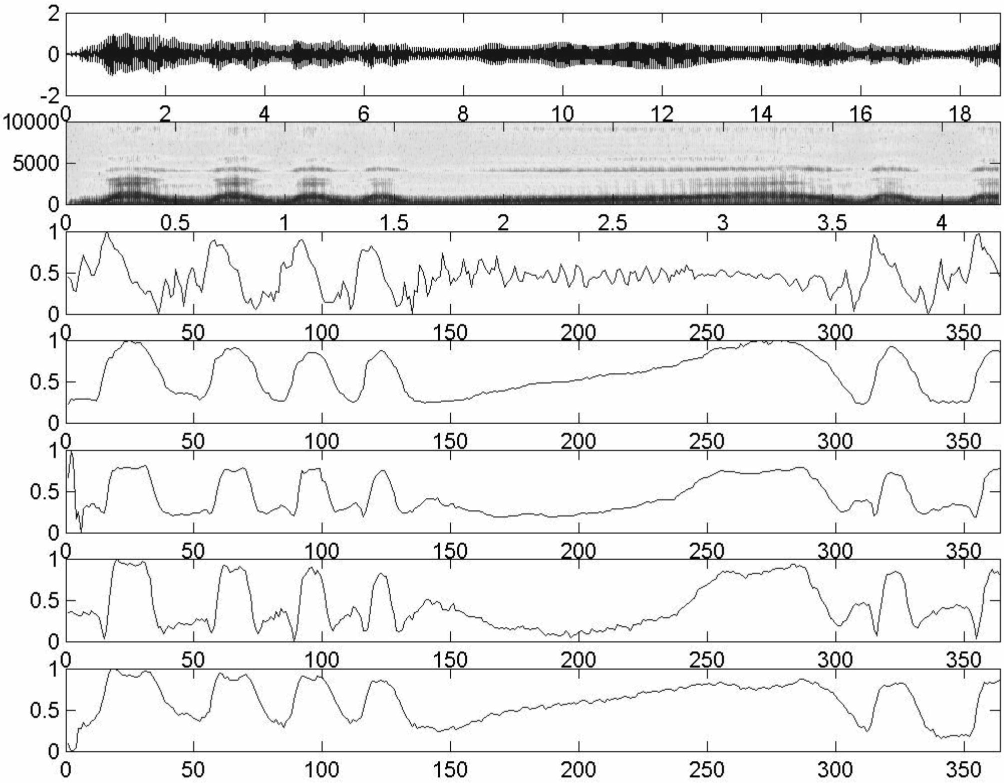


Figure 86*:* In descending order: sound waveform, spectrogram, and normalized 'Cepstrum', 'LPC', 'Slope', 'Centroid', and 'Area' descriptors envelopes of a [wa-wa-wa-wa-wa-wa-wa] utterance.

As a conclusion, '*LPC*' and '*Area*' would be the best choices for the control parameter. However, although both are relatively cheap in terms of computation cost, the *'Area'* descriptor is much better in terms of robustness. Thus, the current Wahwactor implementation uses the '*Area*' descriptor.

3.3.2 Nasality

Nasal consonants are produced when the nasal and oropharyngeal resonators couple. The closed oral cavity and the sinus structure are joined to form cavities to the main airflow course through pharynx and nasal tract. Nasalized vowels occur when the oral cavity is not completely closed by tongue or lips, air flows through both nasal and oral cavities. These two cavities affect each other and when they resonate together it results into a loss of amplitude or antiresonance at certain frequencies. They can even cancel out each other if resonance frequencies are close enough.

According to Fujimura (1962) and Lieberman and Blumstein (1988) nasal sounds present their main characteristic spectral properties in the 200 to 2500 Hz range. Nasals consonants usually present their first formant at around 300 Hz and their antiformant at around 600 Hz. Thus they concentrate energies in the lower frequencies region and present little energy in the antiformant surroundings. Nasals also present murmur resonant peaks, being the most significant at around 250 Hz with a secondary minor peak at around 700 Hz.

When in regular speech palate and pharynx tissues do not close together, air escapes constantly through the nose as a symptom of what is know as hypernsaslity. Hypernasal speech is in fact an indicative of an anatomical, neurological or peripheral nervous systems. Cairns and his colleagues presented in 1996 a technique for hypernasality detection using Teager energy operator non linearity, see Cairns et al. (1996). The algorithm assumed normal speech could be expressed as the summation of formants at various frequencies:

$$S_{normal} = \sum_{i=1}^{I} F_i(\omega) \tag{60}$$

while in the case of nasal speech there would also be antiformants (*AF*) and nasal formants (*NF*) involved:

$$S_{nasal} = \sum_{i=1}^{I} F_i(\omega) - \sum_{k=1}^{K} AF_k(\omega) + \sum_{m=1}^{M} NF_m(\omega) \tag{61}$$

And since a primary cue for nasality is the intensity reduction of their formant, we can simplify previous equation with a low-pass filter:

$$S_{normal-LPF} = F_1(\omega) \tag{62}$$

$$S_{nasal-LPF} = F_1(\omega) - \sum_{k=1}^{K} AF_k(\omega) + \sum_{m=1}^{M} NF_m(\omega) \tag{63}$$

The multicomponent nature of the low pass filtered nasalized speech can be exploited through the non linearity of the Teager Energy Operator (TEO), see Teager (1989) and Kaiser (1990), defined over *x(n)* as:

$$\Psi_d[x(n)] = x^2(n) - x(n-1) \cdot x(n+1) \tag{64}$$

The result of applying the TEO over low pass filtered version of normal and nasalized speech would be:

$$\Psi_d\left[S_{normal-LPF}(n)\right]=\Psi_d\left[f_1(n)\right] \tag{65}$$

$$\Psi_d\left[S_{nasal-LPF}(n)\right]=\Psi_d\left[f_1(n)\right]-\sum_{k=1}^{K}\Psi_d\left[a\cdot f_k(n)\right]+\sum_{m=1}^{M}\Psi_d\left[n\cdot f_m(n)\right]+\sum_{j=1}^{K+M+1}\Psi_{cross}[\ ] \tag{66}$$

while the result of applying the TEO over normal and nasalized utterances after applying a band pass filtered centered around the first formant would be:

$$\Psi_d\left[S_{normal-BPF}(n)\right]=\Psi_d\left[f_1(n)\right] \tag{67}$$

$$\Psi_d\left[S_{nasal-BPF}(n)\right]=\Psi_d\left[f_1(n)\right] \tag{68}$$

The comparison between the outputs of the TEO after a low pass filter and after a band pass filter centered on the first formant form the basis of the hypernasality detection. Taking profit of this algorithm, a simplified real time adaptation for standard nasality estimation can be proposed.

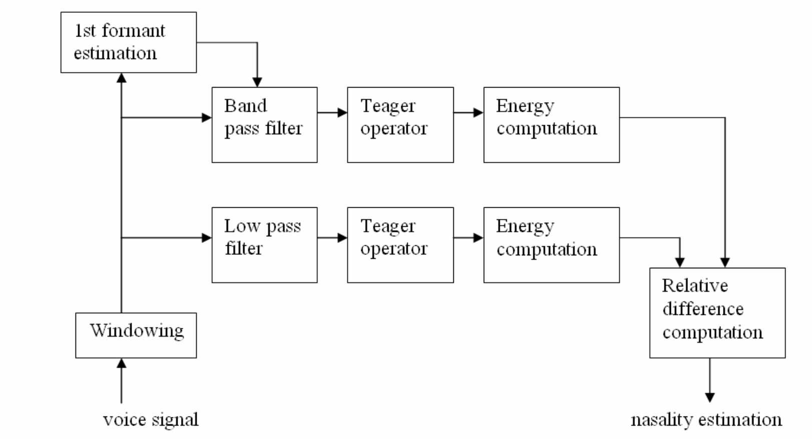


Figure 87: Block diagram of the implementation of an algorithm for nasality detection
A rectangular 30 ms window length window with a 50% overlap is applied.

The first formant is roughly estimated as the centroid of the magnitude spectra power in the]300,1000[Hz band:

$$F_1=\frac{\sum_i f_i\cdot a_i}{\sum_i a_i} \quad \textit{for}\text{ all } i \text{ such that } f_i \in]300,1000[\tag{69}$$

The low pass filter is defined as a 41 coefficient FIR filter with its cut off frequency set to 1000 Hz and its reject gain set to -50 dB.

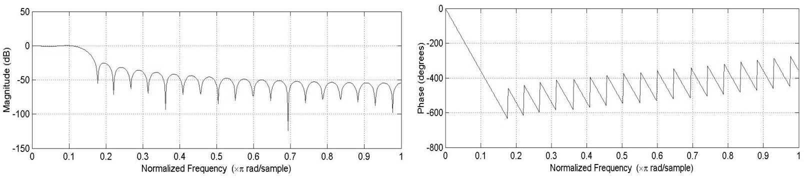


Figure 88: magnitude and phase spectra of the low pass filter.

The band pass filter is a Gabor filter with its center frequency being defined at each analysis step according to the outcome of the first formant estimator. The width of the band pass filter is set to 500 Hz.

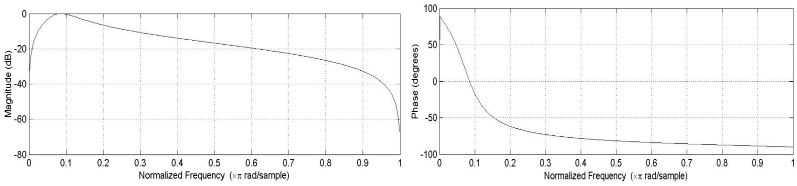


Figure 89: Magnitude and phase spectra of the Gabor band pass filter at 700 Hz.

Energy is computed as the squared sum of all samples contained in a frame:

$$E_s = \sum_n |s(n)|^2 \tag{70}$$

And the relative difference computation compares the energy estimations of the band pass filtered (E_{BPF}) signal and the low pass filtered signal (E_{LPF}):

$$r = \frac{|E_{BPF} - E_{LPF}|}{\max(E_{BPF}, E_{LPF})} \tag{71}$$

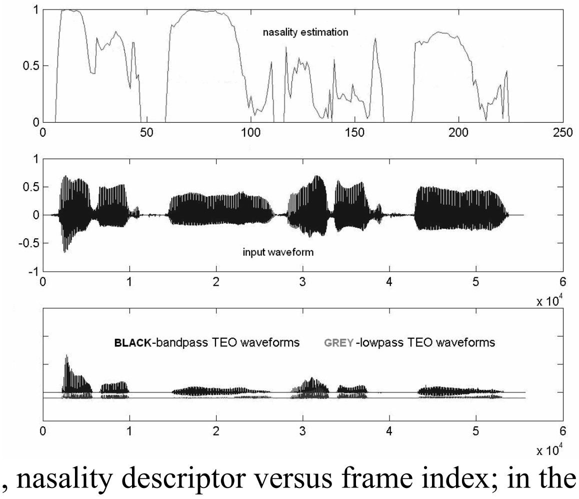


Figure 90: On top, nasality descriptor versus frame index; in the middle waveform (normalized amplitude versus sample index); on the bottom, band pass and low pass waveforms along sample index. The sound represented is a phrase repeated twice, first nasal, and then normal.

3.4 Musical meaningful parameters

This section presents a couple of descriptors that tackle directly concepts that any musician can understand regardless of their technical knowledge. First the melody, second the onset. Onsets detection is in fact the first step for tempo and beat detection.

3.4.1 Melody

Extracting melody from singing voice is hot topic in research, especially in the field of Music Information Retrieval (MIR), see McNab et al. (1996), Radova and Silhan (2002), and Gómez et al. (2003), where melody abstraction is believed to be one of the requirements for the deployment of smart human adapted retrieval engines. It has been in fact one of MIR standard challenges, namely Query by Humming, the one that rocketed in a certain way the number of teams working around melody extraction, see McNab et al. (2000), Yang (2001), Hu and Dannenberg (2002), Zhu and Shasha (2003), Viitaniemi (2003), and Lau et al. (2005). Nearly all applications available with humming search features use relative measures of pitch (i.e. intervals) and duration. Others, such as Watzazong[43] or Midomi[44] are based on comparing the input against a database of other users humming or whistle performances; in such case no melody extraction is necessarily required.

Melody extraction in Query by Humming systems runs as an offline process, handling at any time of the performance analysis, all past and future information. But tools for real time melody extraction exist for other type of applications.

Voice to MIDI (Musical Instrument Digital Interface) format is an example of application performing real time melody estimation. There exist several hardware and software commercial voice to MIDI converters. Examples of software applications are MidiVoicer[45] and Digital Ear[46]. Regarding hardware converters, there are two specially remarkable: Vocal to MIDI[47], and MidiVox[48].

Figure 91: MidiVox general view

Some of these converters might be useful and give reasonable results in some cases but they generally lack of robustness. The problem comes from the singing voice real-time note onset detection. Such critical process is in charge of deciding at each analysis frame time if the current voice data belongs to a new note or if it has to be considered part of the

[43] http://www.watzatsong.com

[44] http://www.midomi.com

[45] http://www.e-xpressor.com/m_voicer.html

[46] http://www.digital-ear.com/digital-ear/info.htm

[47] http://www.geocities.com/vocal2midi

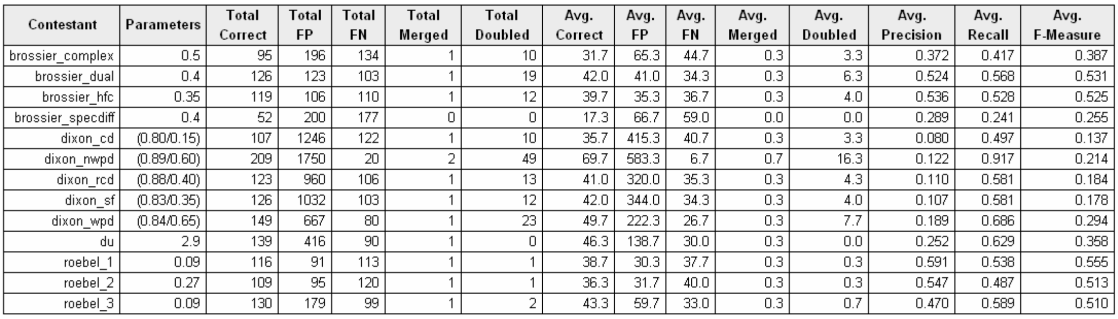
[48] http://www.healingmusic.net

preceding note. This decision has to be taken with no frame delay and with no prior knowledge on the final melody (not even key or/and scale). The considerable complexity of this problem makes it nearly impossible to avoid the converters outcome false notes.

3.4.2 Singing voice onset

As of today, there is no onset detection algorithm with acceptable results when dealing with solo singing voice. In Figure 91, showing MIREX2006 results for singing voice onset detection, we can observe approximately half of the algorithms fail completely while the other half reach around a 50% Average F-measure. Top result is 55% and can be achieved using an algorithm based on transient / non-transient spectral peaks classification originally proposed in Robel (2005). Usually, a beat is perceived whenever occur any of the following cases:

1. Sudden energy modulation. Most of beat detection algorithms work with energy related attributes, and some of them do it exclusively.
2. Sudden frequency modulation. Since Duxbury et al. (2003) frequency modulation information is also taken into account in beat detection algorithms. This information is usually computed by means of deviation from the stable phase behavior or by peak classification Robel (2005)
3. Sudden timbre modulation. Klapuri (1999) and all other algorithms that stand on top of that consider energies in different frequency bands.



Contestant	Parameters	Total Correct	Total FP	Total FN	Total Merged	Total Doubled	Avg. Correct	Avg. FP	Avg. FN	Avg. Merged	Avg. Doubled	Avg. Precision	Avg. Recall	Avg. F-Measure
brossier_complex	0.5	95	196	134	1	10	31.7	65.3	44.7	0.3	3.3	0.372	0.417	0.387
brossier_dual	0.4	126	123	103	1	19	42.0	41.0	34.3	0.3	6.3	0.524	0.568	0.531
brossier_hfc	0.35	119	106	110	1	12	39.7	35.3	36.7	0.3	4.0	0.536	0.528	0.525
brossier_specdiff	0.4	52	200	177	0	0	17.3	66.7	59.0	0.0	0.0	0.289	0.241	0.255
dixon_cd	(0.80/0.15)	107	1246	122	1	10	35.7	415.3	40.7	0.3	3.3	0.080	0.497	0.137
dixon_nwpd	(0.89/0.60)	209	1750	20	2	49	69.7	583.3	6.7	0.7	16.3	0.122	0.917	0.214
dixon_rcd	(0.88/0.40)	123	960	106	1	13	41.0	320.0	35.3	0.3	4.3	0.110	0.581	0.184
dixon_sf	(0.83/0.35)	126	1032	103	1	12	42.0	344.0	34.3	0.3	4.0	0.107	0.581	0.178
dixon_wpd	(0.84/0.65)	149	667	80	1	23	49.7	222.3	26.7	0.3	7.7	0.189	0.686	0.294
du	2.9	139	416	90	1	0	46.3	138.7	30.0	0.3	0.0	0.252	0.629	0.358
roebel_1	0.09	116	91	113	1	1	38.7	30.3	37.7	0.3	0.3	0.591	0.538	0.555
roebel_2	0.27	109	95	120	1	1	36.3	31.7	40.0	0.3	0.3	0.547	0.487	0.513
roebel_3	0.09	130	179	99	1	2	43.3	59.7	33.0	0.3	0.7	0.470	0.589	0.510

Figure 92: MIREX2006 table of results for the beat onset in singing voice signals. Taken from the contest public site[49]

From recent comparative study, see Bello et al. (2005) of onset detection algorithms, the spectral complex difference stands out as a good candidate for beat onset function. The ability of this function relies on the assumption spectral properties at onset events suffer unpredictable rapid changes. By working in the complex domain note onsets come from significant energy variations in the magnitude spectrum, and/or from significant deviations phase values in the phase spectrum, presumably caused by pitch variations. The detection function Γ*(m)*, at frame *m*, is defined as the summation of the Euclidean distance between observed spectrum $S_k(m)$ and a prediction of it $\hat{S}_k(m)$, for all frequency bins *k*:

[49] http:// www.music-ir.org/mirex2006/index.php/Audio_Onset_Detection_Results:_Solo_Singing_Voice

$$\Gamma(m)=\sum_{k=1}^{N}\sqrt{\left|\hat{X}_k(m)\right|^2+\left|X_k(m)\right|^2-2\cdot\left|\hat{X}_k(m)\right|\cdot\left|X_k(m)\right|\cdot\cos(\Delta\varphi(m))} \tag{72}$$

For a reasonable hop size, locally stationary sinusoids are assumed to have approximately constant amplitude and instantaneous frequency along adjacent frames. Constant instantaneous frequency means constant phase increment from frame to frame,

$$\varphi_k(m)-\varphi_k(m-1)\approx\varphi_k(m-1)-\varphi_k(m-2) \tag{73}$$

thus we can define estimated phase spectra as previous phase plus previous phase variation

$$\hat{\varphi}_k(m)=\varphi_k(m-1)+\left[\varphi_k(m-1)-\varphi_k(m-2)\right] \tag{74}$$

and we can define our estimated magnitude spectra as being the same as previous frame,

$$\left|\hat{X}_k(m)\right|=\left|X_k(m-1)\right| \tag{75}$$

resulting into

$$\Gamma(m)=\sum_{k=1}^{N}\sqrt{\left|X_k(m)\right|^2+\left|X_k(m-1)\right|^2-2\cdot\left|X_k(m)\right|\cdot\left|X_k(m-1)\right|\cdot\cos(\varphi_k(m)-2\cdot\varphi_k(m-1)+\varphi_k(m-2))} \tag{76}$$

An alternative to this beat onset function exists based on aforementioned Teager Energy Operator and the Energy Separation Algorithm

Originally introduced by Teager when demonstrating the non-linear behavior of the vocal tract production mechanisms, see Teager 1989, and first documented by Kaiser in Kaiser (1990), the Teager Energy Operator (TEO) is a non-linear operator defined over a discrete signal *x(n)* as:

$$\Psi_d\left[x(n)\right]=x^2(n)-x(n+1)\cdot x(n-1) \tag{77}$$

The TEO has proven to be sensitive to multicomponent signals, see Maragos (1991) and has been used to separate the FM and AM voice components. Given an AM-FM model for speech, in Maragos et al. (1993), it was developed the Energy Separation Algorithm (ESA) to isolate the AM and the FM components. Being *f*(*n*) the FM contribution, and *a*(*n*) the AM contribution at sample *n*, the separation equations are:

$$f(n)\approx\frac{1}{2\pi T}\cdot\arccos\left(1-\frac{\Psi_d\left[y(n)\right]+\Psi_d\left[y(n+1)\right]}{4\cdot\Psi_d\left[x(n)\right]}\right) \tag{78}$$

$$|a(n)| = \sqrt{\frac{\Psi_d[x(n)]}{1 - \left(\frac{\Psi_d[y(n)] + \Psi_d[y(n+1)]}{4 \cdot \Psi_d[x(n)]}\right)^2}} \tag{79}$$

where *T* is the input signal period, and *y*(*n*) is the first derivative of *x*(*n*). It is shown in Maragos et al. (1993) *y*(*n*) can be approximated as $y(n) = x(n) - x(n-1)$

Such equations provide at each sample an estimation of the envelope and instantaneous frequency at a very small computational complexity and with an instantaneous-adapting nature. Based on these, an algorithm has been implemented and tuned for beat detection in solo singing voice signals.

The algorithm, based on frame analysis, resamples first the input signal at 8 kHz and applies a 1 kHz bandwidth Gabor filter centered at 900 Hz. Next, it computes *f*(*n*) and *a*(*n*), and averages them along each frame to obtain a frame mean value.

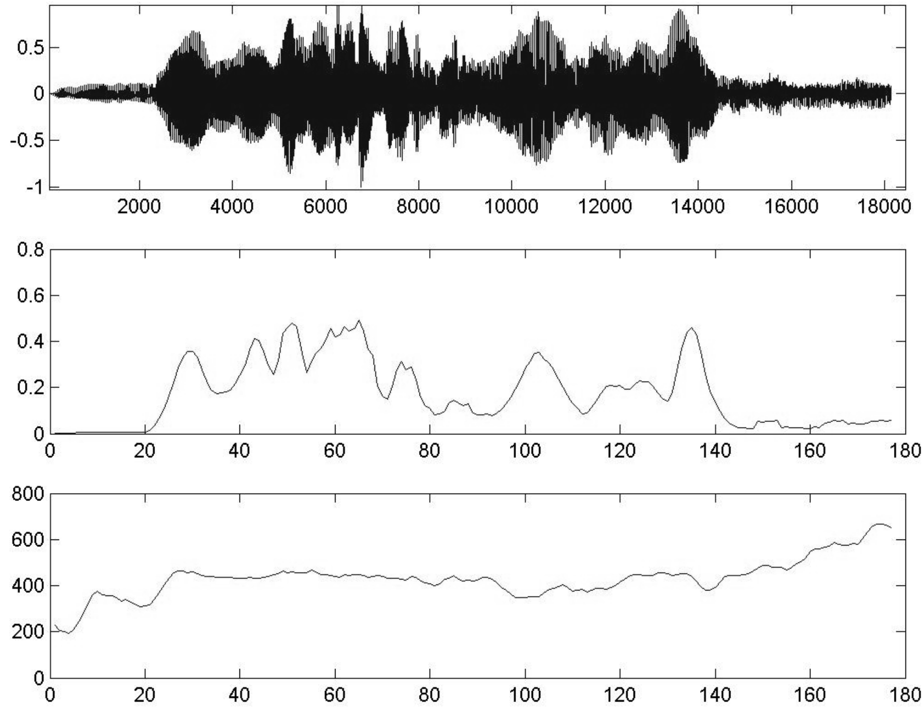


Figure 93: Obtained *a*(*n*) (middle) and *f*(*n*) (lower) mean frames contributions from the implemented algorithm applied over a singing voice signal (upper) with vibrato modulations.

Next step is to apply a procedure by which significant variations of mean frame *f*(*n*) and *a*(*n*) are detected. The current implementation of the algorithm runs a valley detection algorithm only over *a*(*n*) mean frame values to detect sudden modulation variations. Thus, current implementation fails to detect those onsets due exclusively to a variation in the fundamental frequency.

In order to evaluate the algorithm, tests were performed using MIREX2006 singing voice sample database. In this process, it was found the three different manual annotations of the beats used as ground truth had larger deviations between them than the tolerance used to perform evaluations. For such reason tolerance value was set to the maximum deviation found in the annotated files, i.e. 150 milliseconds. The results obtained are:

Average F-Measure: 0.7782
Average Precision: 0.7649
Average Recall: 0.799
Total number of ground truth onsets: 74
Total Correct: 56.6667
Total False Positives: 18.3333
Total False Negatives: 17.3333
Total Merged: 0.3333
Total Doubled: 0
Average absolute distance: 0.0612

Even without exploiting both modulation contributions, results show the suitability of using FM and AM descriptors for the estimation of voice onset, and consequently also for the estimation of solo singing voice tempo and beat attributes.

3.5 Conclusions

A variety of algorithms for the automatic description of the voice and singing voice have been introduced in this chapter. Although attributes such as melody are referred to as high level attributes, none of the descriptors presented in this chapter get to bridge the semantic gap. In this sense some experiments were carried out in order to find the perfect linear combination of descriptors such as the ones presented in this chapter that could correlate the most with user subjective ratings of singing voice attributes such as “crystal clean”. Surprisingly enough it was found no ground truth data could be user generated without excessively lack of consistency, suggesting highest levels of representation for the singing voice are far too subjective dependent to be handled in such way.

From all algorithms presented, two of them make use of the Teager non linear operator: nasality and onsets. In both cases, the algorithms using the operator have demonstrated to outcome reasonable results, and, furthermore, computationally efficiency. Such discovery reveals many other non linear operators may exist with high correlations with some voice specific attributes.

Chapter 4

Techniques for the transformation of the singing voice

4.1 Introduction

In this chapter different algorithms for the transformation of the singing voice are presented. These transformations are achieved by the manipulation of objects and parameterizations that belong to signal models described in Chapter 2 or rather by the modification of concepts and descriptors presented in Chapter 3. Since phase locked vocoder like techniques have been proven along last few years to be the spectral based technologies with better sound quality in voice transposition, most of the algorithms are based on them.

The first part of the chapter stands on the source filter representation of the voice, presenting those transformations that belong to the timbre first and to the excitation next. For those that tackle the voice excitation, different sets of transformations are described for emulating vocal disorders, modifying the dynamics and transposing the sound. Those that tackle the timbre are basically an introduction grouped in two categories, one working with the spectral shape as a whole, and the other working with harmonics spectral peaks as elements.

The second part of the chapter raises the grade of abstraction and addresses what has been called as high-level transformations. These transformations deal with more universally meaningful controls such as singer age or number of unison singers, and most of them are achieved by the combinations and linkage of lowest level transformations.

A special section about singing voice driven synthesis closes the chapter. This section presents two main different real time morph algorithms for morphing singing voice whether with another voice or whether with an instruments.

4.2 Vocal tract transformations

This section presents the very basic notion of an equalization filter; i.e. the modification of the distribution of energies along the spectrum. First filtering operational procedure is presented for the cases of dealing with the sinusoidal model and dealing with the phase locked vocoder harmonic peak regions. Next and last, a timbre transformation control named timbre mapping is introduced. Other timbre related transformations will appear in the higher level transformations section.

4.2.1 Equalization filters

Many different implementations of numerous algorithms exist for equalizing a signal. This section introduces two different approaches, standing each of them on top of a model; one on the sinusoidal plus residual and the other on the phase locked vocoder.

In the sinusoidal plus residual model, the equalization algorithm can take advantage of the signal decomposition to modify the amplitude of any arbitrary partial present in the sinusoidal component. Thus, the filter does not need to be characterized by a traditional transfer function, and a more complex function can be defined by summing delta-functions:

$$H(f) = \sum \delta(f_i) \cdot g_i \tag{80}$$

where g_i is the gain applied to the i^{th} partial of frequency f_i. Using this filter representation we can think of an equalization that would filter out even harmonic energies turning the voice sound into a clarinet-like.

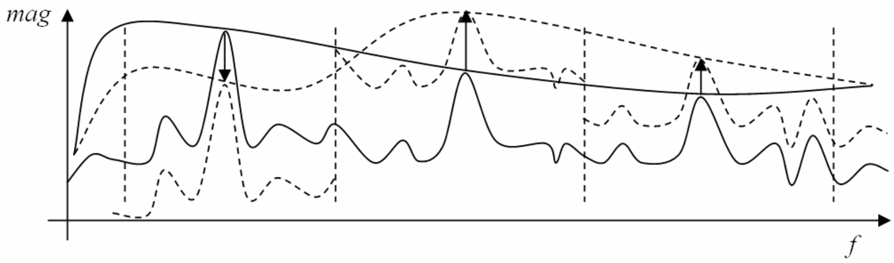


Figure 94: Representation of an equalization based on the phase locked vocoder model. The harmonic peaks define the original spectral envelope which is shown with a continuous line in the Figure 2.17. On the other hand, we have an envelope drawn as a dashed line in same figure that defines the desired timbre

In the phase locked vocoder model, equalization is not only applied to the sinusoids or harmonic peaks but to the whole harmonic peak region. In such model the local behaviour of the peak region has to be preserved both in amplitude and phase. Thus, once the algorithm computes the necessary gain to raise or drop the harmonic peak so that it follows the new timbre envelope, that gain is added to all the bins that belong to that peak region (amplitude linear addition).

4.2.2 Timbre mapping

Singing voice timbre modification allows the creation of clone performances with completely different voice personality. Timbre transformation can be achieved by modifying the input spectral shape by shifting and scaling processes. Timbre mapping gives name to a control interface for timbre mapping that allows applying both shift and scale over whatever spectral band. Timbre mapping is performed by assigning to a certain frequency f_y, the amplitude of another frequency f_x being both frequencies mapped by a function $g(f)$ so that $g(f_x) = f_y$ and thus:

$$TibreMapping\left[\left|SpectralShape\left(f_x\right)\right|\right] = \left|SpectralShape\left(g\left(f_x\right)\right)\right| = \left|SpectralShape\left(f_y\right)\right| \quad (81)$$

Scaling would be a particular case of timbre mapping in which the mapping function would be one of the family $g(x) = k \cdot x$ with k being the scale factor, with values usually between 0.5 and 2. In general timbre scaling effect resembles modifications over the size and shape of the instrument. In the case of the voice, timbre scaling is coupled to the modification of the vocal tract length, mostly of the infra laryngeal area. Whenever k takes values underneath 1, the spectral shape will expand, thus the emulated vocal tract will compress and voice will turn more kid alike. Whenever k takes values over 1, the spectral shape will compress and thus the emulated vocal tract will expand, resulting into a voice that sounds like coming from a big fat man.

The shifting transformation would be a particular case of the timbre mapping as well in which $g(x)$ can be expressed as $g(x) = x + c$ with c being the offset factor. This shift is performed in such a way that no new harmonic content is generated; just the amplitude envelope of the spectrum is modified.

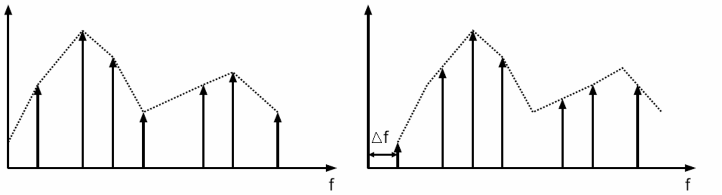


Figure 95: Representation of a Spectral shape shift transformation of value Δf

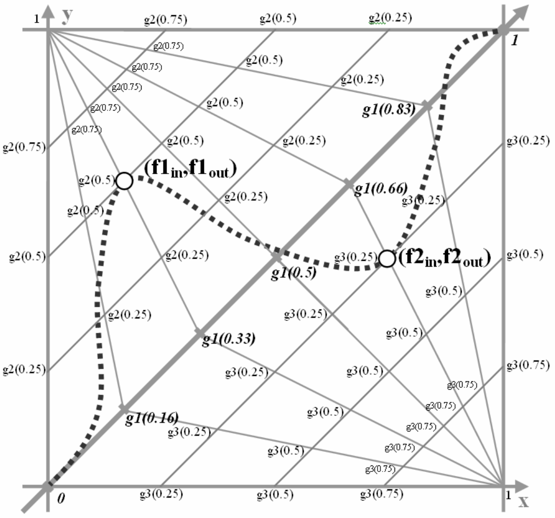


Figure 96: Sketch of the timbre mapping control with normalized frequency axis. Two sample points are defined: $f1_{in}$=$g1$(0.33), $f1_{out}$= $f1_{in}$+$g2$(0.5)·(1- $f1_{in}$), and $f2_{in}$=$g1$(0.66), $f2_{out}$= $f2_{in}$-$g3$(0.25)·$f2_{in}$. The $g1$, $g2$ and $g3$ functions define the type of scale used, linear, logarithmic, or other. The normalized input frequencies, f_{in}, are defined in the diagonal

axis trough the scale function *g1*. The normalized output frequencies f_{out} are defined in the axis that go from the diagonal to the vertex points x=1 or y=1, trough the scale functions *g2* and *g3* respectively. The sensitivity of the control interface has to be tuned to take profit of all the drawing area.

The timbre mapping graphical interface control allows the user to draw all different kinds of functions over a perceptual significant scaled map. This functions rule the mapping between the input voice spectral amplitudes and the output synthesis voice spectral amplitudes. This vocal effect can be seen as a local spectral shape stretch-compress or also as a formant mapping since with prior knowledge of the formant positions the algorithm could reallocate them with the adequate mapping function.

4.3 Voice excitation transformations

This section presents different voice excitation transformations. A first group of them relate to the emulations of vocal disorders such as roughness or breathiness. Vocal disorders have been largely studied as pathology in the field of phoniatry, however, in the context of popular singing, vocal disorders not always come from pathologies but sometimes healthy voices use them as an expressive recourse.

The goal of the algorithms presented here is to achieve natural modifications of the voice excitation in order to enhance singing voice in music productions. Unlike many emulators of vocal disorders, the algorithms presented in this section arise from spectral models and work with frequency domain techniques.

4.3.1 Roughness

Using the nomenclature presented in Chapter 3, we can say the most common techniques used to synthesize rough voices work with the source - filter model and reproduce the jitter and shimmer aperiodicities in time domain, see Titze (1994). These aperiodicities can be applied to the voiced pulse-train excitation by taking real patterns that have been extracted from rough voices recordings or by using statistical models, see Childers (1990).

The main idea behind the algorithm for turning a normal phonation voice into a rough voice is to take the original input signal, transpose it down a certain number of octaves, and then take that transposed signal and shift it and overlap it with a certain amount of randomness (Figure 97) to re-synthesize the original voice with its new rough character.

The shift applied to each of the N shifted versions of the transposed signal is:

$$Shift_i = i \cdot T_0 + \mathrm{X}_i \quad for\ i=0,1,..,N\text{-}1 \tag{82}$$

where T_0 is the original period and X_i is a zero mean random variable. These differently shifted N versions of the transposed signal are then scaled by a unity mean random variable Y_i and finally overlapped.

In order to take in what the outcome of such system is, figure 3 shows a figurative time domain representation of all steps for N=2.

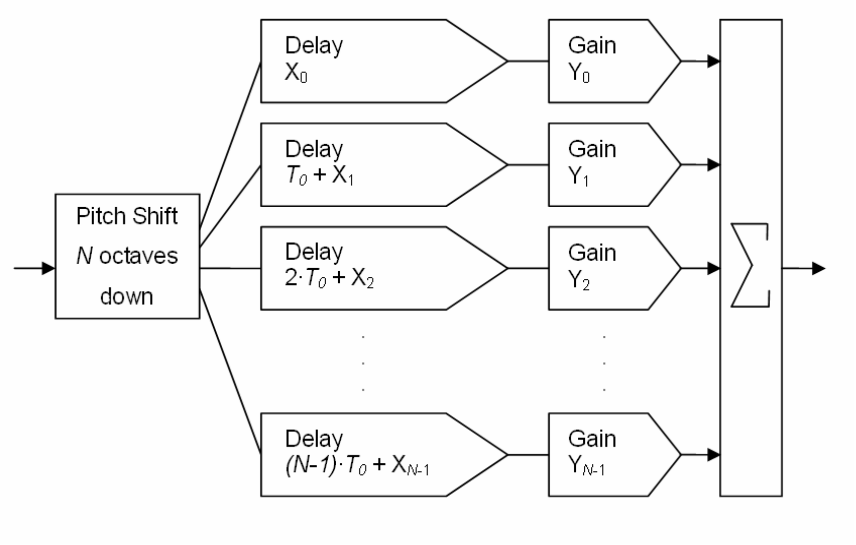
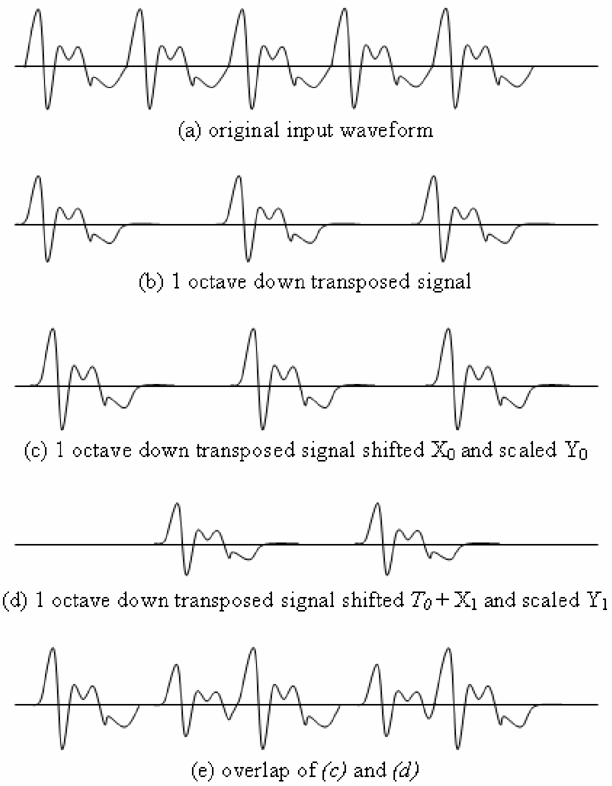


Figure 97: Block diagram of the rough emulator.



(a) original input waveform

(b) 1 octave down transposed signal

(c) 1 octave down transposed signal shifted X_0 and scaled Y_0

(d) 1 octave down transposed signal shifted $T_0 + X_1$ and scaled Y_1

(e) overlap of *(c)* and *(d)*

Figure 98: Figurative representation of the waveforms at different points of the algorithm for N=2

In order to allow the effect to fit in a real time environment, final implementation of the roughness transformation is a simplified version of the previously described algorithm.

Thus, Figure 97 does not illustrate real results since the implemented frame-based analysis does not take into account the relationship between the frame rate and the input period, nor the analyzed frame history. Also, the real implementation changes X and Y stochastic variables values at every frame time. This scenario does not allow generating patterns such as the ones represented in the figure. This is also the reason why even though theoretically, with such algorithm, the higher *N* is the more control over isolated periods of the signal, the implemented system does not fulfil this rule. All performed simplifications are described next.

The one octave down transposition is accomplished by adding pure sinusoids to the spectrum in the sub-harmonic frequencies. More precisely, adding the main lobe bins of the analysis window and taking the phase from the closest harmonic peak and shifting it with the corresponding offset as in Schoentgen (2001):

$$\Delta\varphi = 2\pi\cdot f_h\cdot\left(\frac{f_{sh}}{f_h}-1\right)\cdot\Delta t \tag{83}$$

Where f_{sh} is the sub-harmonic frequency to fill, f_h is the frequency of the closest harmonic peak, and Δt is the frame time. Since the greater *N* is, the more computationally expensive the effect is, we have taken the minimum *N* value, *N*=2.

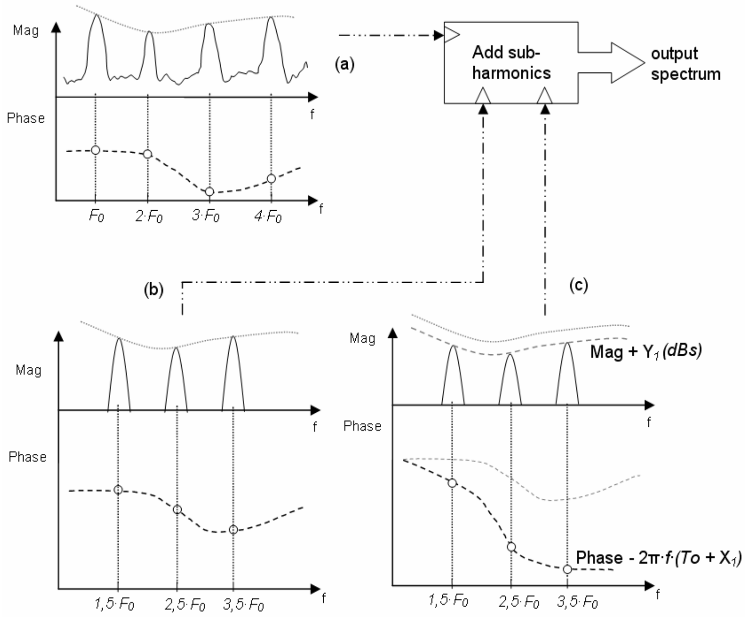


Figure 99: Representation of the roughness straight forward implemented system where (a) is the original input signal spectrum (b) are the sub-harmonics that generate the transposed version of the signal, and (c) are the sub-harmonics that generate figure 3.d signal

The jitter and shimmer stochastic variables of the first channel are set to its mean value X_0=0 and Y_0=1. Thus, the output of this first channel will be a one octave down transposition of the original input. This is a not very risky simplification for N=2 since it can be seen as moving X_0 and Y_0 randomness to X_1 and Y_1.

The stochastic variables X_1 and Y_1 are defined to have a normal distribution with variances 3% of the input signal period, and 3 dBs respectively.

The random scaling due to Y_1, as well as the random time shift due to $T_0 + X_1$ are only applied to the sub-harmonics. The only reason for doing such oversimplification is to reduce the computational cost of the algorithm since with this only half of the peaks to which the random variables should be computed and applied are actually processed. The time shift is applied in frequency domain by adding the corresponding constant slope phase offset to the phase of the sub-harmonics spectrum as represented in the spectrum of Figure 99.c.

Only sub-harmonics inside the F_0-8000 Hz band are added to the spectrum. Upper sub-harmonics are not significantly relevant in terms of acoustic perception to reproduce the rough effect, and the first sub-harmonic (placed at $0.5F_0$) is assumed to be, based on the observations, always masked by the amplitude of the fundamental peak.

4.3.2 Growl

Based on the most frequently observed growl spectral symptoms described in chapter 3, the implemented algorithm is based on filling the original spectrum with sub-harmonics.

Since growl is not a permanent disorder, transformation can not be applied all along the performance. Therefore, the implementation includes an automatic growl control (as shown Figure 100) that is in charge of deciding how much of the effect has to be applied at each time depending of the input singing voice. This control is mainly based on the first derivatives of the fundamental frequency and energy and has control on how many sub-harmonics have to be added, and their phase and magnitude patterns (including the gain of the sub-harmonics).

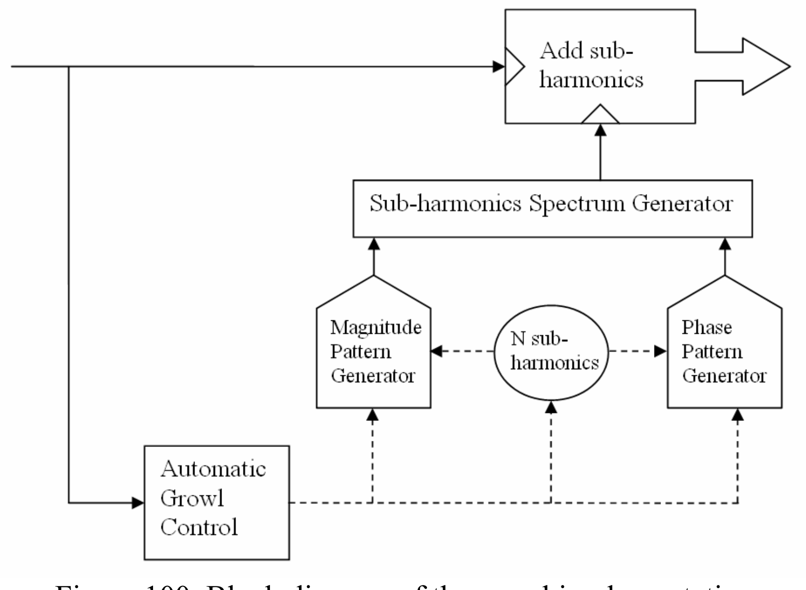


Figure 100: Block diagram of the growl implementation

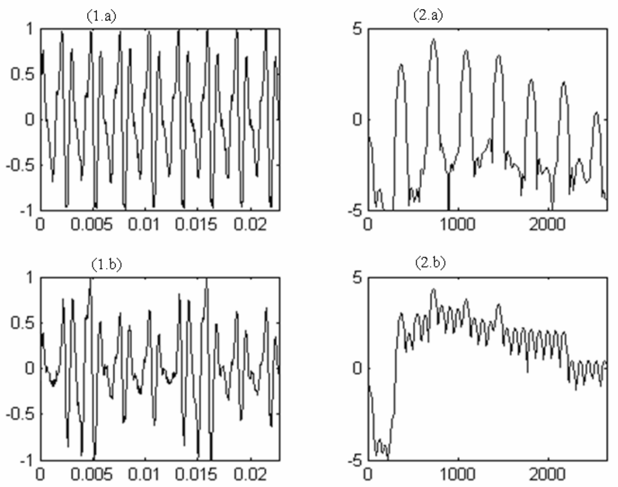


Figure 101: Partial views of waveform (1) in seconds and magnitude spectra (2) in Hz from both original (a) and transformed (b)

With such implementation, the algorithm is able to reproduce growl sub-period amplitude patterns as the one shown in Figure 101. In the waveform view of the transformed voice we can observe the different amplitudes of each of the four periods of the growl macro period. This amplitude modification is achieved by applying phase alignment patterns extracted from real growl analysis to the sub-harmonics.

4.3.3 Breathiness

Breathiness is a very different pathology to hoarseness. The main character of a breathy voice is not the roughness but the breeziness, where harmonic components hide under the turbulent noise.

Breath transformation can be achieved by applying at the same time two different algorithms: a mid-high boost frequency filter to the input spectrum, and a shape distortion over to the main love harmonic peak magnitude spectrum.

The frequency filter is applied to the harmonic peaks with a gain given by the breathiness transformation intensity control. Gain values range from 0 to 12 dB. Filtering can be applied as in section 4.2.1. with a slight modification. Those bins inside the harmonic peak region but not bodying the harmonic peak itself are modified in a more moderate scale using a step-like offset function. Moreover, the filter includes a frequency gain compressor that adjusts those harmonic peaks placed over 1,500 Hz with a given gain above *5* dBs. Moreover the bins considered to be not part of the harmonic peak are not raised up to the same amount.

On top of this filter, based on the assumption each harmonic peak region will include the peak and its surrounding noise, the algorithm flattens down the main lobe of the peak. After several tests it was found more realistic breathiness synthesis were obtained by distorting the harmonic magnitude shape of the peak instead of applying an offset gain. Final implementation of the algorithm applies a -30 dB gain to the three main bins around the estimated peak frequency for all peaks but the first four.

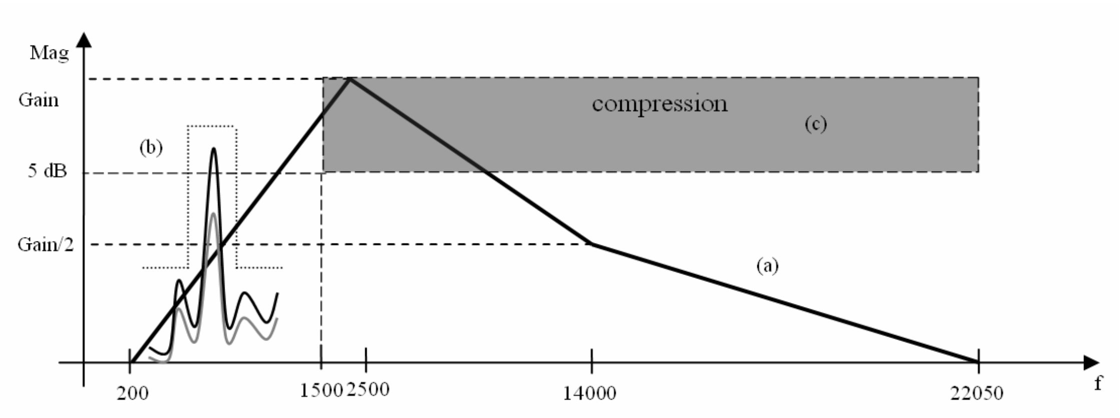


Figure 102: Representation of the three different considerations inside the breathiness filter: (a) the filter itself, (b) the harmonic peak offset function, and (c) the gain compression. Representation uses a unscaled magnitude versus frequency plot where Sr means sampling frequency and assuming an input voice sampled at 44,100 Hz..

An alternative to this algorithm exists in which breathiness effect is achieved by replacing the original harmonic peaks by peaks taken from a real breath recording analysis. This approach uses the phase locked vocoder model for the replacement and preserves the original harmonic distribution as well as the original spectral shape.

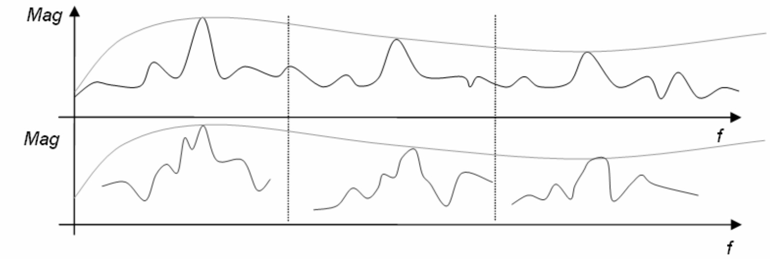


Figure 103: Representation of the peak replacement approach for the breathiness effect

4.3.4 Whisper

Whisper transformation can be achieved by mixing a whisper template with the input singing voice.

Whisper template is obtained from recording a real whisper utterance, editing it to remove non-stationary boundaries, inverse filtering it using a manual estimation of the vocal tract function, and storing it as an infinite loop. Later, in synthesis, the algorithm picks one template frame at each step and equalizes it using a smoothed spectral shape estimation of the input voice frame. Next in synthesis, both the original input and the equalized template are filtered to smooth the mix and improve the naturalness of the blending. These two filters define their envelope as a function of the synthesis voice fundamental frequency.

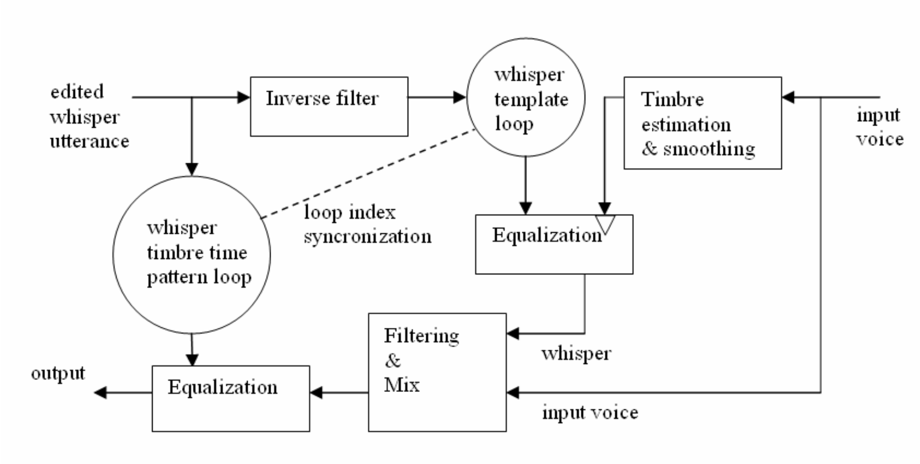


Figure 104: General block diagram of the whisper transformation

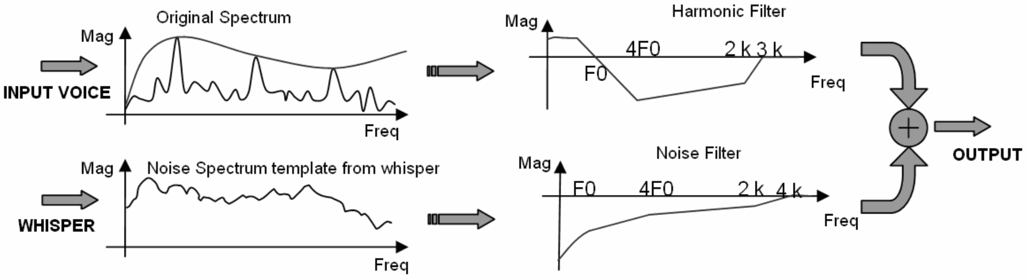


Figure 105: Representation of the whisper transformation filtering operations where F0 means synthesis fundamental frequency.

As a final step, the output in Figure 105 is equalized to preserve the time behavior of the whisper utterance. The equalization applied at each synthesis step is obtained as the difference between the spectral shape of the whisper utterance current frame and an average spectral shape of the whisper utterance.

4.3.5 Dynamics

The most commonly known effects that are considered to transform loudness are the ones that modify the sound intensity level: the volume change, the tremolo, the compressor, the expander, the noise gate and the limiter, see Verfaille et al. (2005). But loudness is also used in a musical sense in order to represent the sound level of an acoustical instrument. The mechanisms that relate the actions of a player with the sound level produced by a given instrument are usually so complex that seldom this feature can be uncorrelated from others such as timbre. Thus, the difference in sound between playing a soft and a loud note in an instrument is not only its sound level.

In the case of a piano, for example, in Solà (1997) a transformation was implemented in order to obtain all possible musical loudness (dynamics of a note) out of a single previously analyzed note. It was concluded, that the most feasible implementation was based on taking the highest possible dynamic as a starting point.

Then, it is just a matter of subtracting the spectral information that is not needed to obtain the notes that have lower dynamic values.

In the case of the singing voice, some studies have been carried out, see Fant (1960) and Sundberg (1973). According to Sundberg (1987) and using his nomenclature, it is possible, under certain conditions, to infer the source spectrum modifications from uttering the same vowel at different loudness of phonation. From this assumption, in Fabig and Janer (2004) we can find a method for modifying the loudness of the singing voice based on the automatic estimation and latter modification of EpR's excitation parameters. The modification is performed by equalizing the harmonic peaks as in section 4.2.1 with a filter obtained as the difference between the automatic estimation of EpR's voice excitation envelope and a desired voice excitation envelope derived from the user's control.

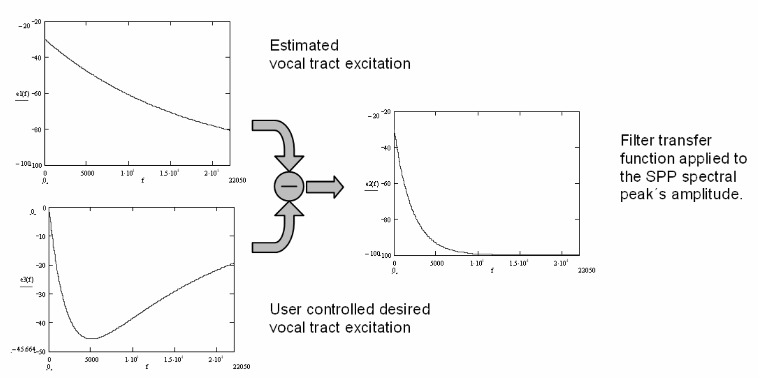


Figure 106: Dynamic modification filter. Taken from Fabig and Janer (2004) with permission of authors.

4.3.6 Pitch transposition

Transposition in the spectral domain is achieved by reallocating harmonic content to a new set of harmonic frequencies that are obtained by applying a constant scale factor to the originals. Standard pitch transposition of singing voice is always defined with timbre preservation, this means, leaving the original spectral shape unmodified. If we define timbre as the envelope described by the amplitudes and frequencies of the harmonics, or its approximation,

$$Sshape = \{(f_1, a_1)(f_2, a_2)...(f_N, a_N)\} \tag{84}$$

the resulting sinusoidal spectrum $X_{transp}(f)$ after a transposition of value *k* will be of the form:

$$X_{transp}(f) = \sum \delta(k \cdot f_i) \cdot Sshape(k \cdot f_i) \tag{85}$$

Using the phase locked vocoder model, transposition is carried out by shifting harmonic peak regions in frequency. The amount of frequency shift is calculated for each harmonic peak and is applied as a constant to the whole peak region, i.e. the linear frequency displacement for all the bins of a region will be the same. Therefore, the local amplitude spectrum of each region will be kept as it is, thus preserving the window convolution with each harmonic peak. In most cases, the harmonic peak region frequency displacement will be a non integer value and thus the spectrum will have to be interpolated. The algorithm we present here uses a 3rd order spline interpolation as a good compromise of quality versus computational cost.

For the i^{th} harmonic peak, the new frequency value is

$$f_i^{new} = transp \cdot f_i \tag{86}$$

and the shift to be applied

$$\Delta f_{Transp} = \left| f_i^{new} - f_i \right| = f_i \cdot \left| transp - 1 \right| \tag{87}$$

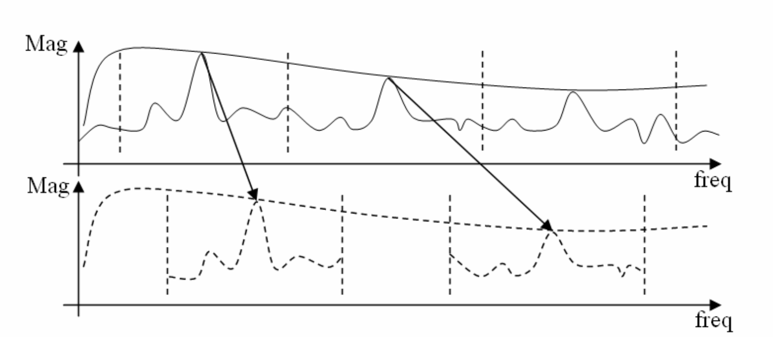


Figure 107: example of an SPP transposition with transposition factor > 1 preserving the original spectral shape. The original SPP spectrum is shown on the top, the arrows show the displacement in frequency that is applied to each harmonic, and at the bottom we can see the resulting spectrum. The gaps between SPP regions are filled with -200 dB constant spectral amplitude.

In Figure 107 we can see an example of upwards transposition. For downwards transpositions, the harmonic peak regions overlap in the resulting spectrum. Overlapping is achieved by adding the complex values at each spectral bin.

4.3.6.1 Phase correction

When one harmonic peak region is shifted in frequency in a frame by frame process, the phase needs to be corrected in order to continue the harmonics quasi-sinusoidal waves. For the harmonic i^{th}, the ideal phase increment between two consecutive frames is:

$$\Delta\varphi_i = 2\pi f_i \Delta t \tag{88}$$

where the time increment between frames is Δt. If we transpose a sample by a factor *transp*, then the ideal phase increment should be

$$\Delta\varphi_i = 2\pi f_i \cdot transp \cdot \Delta t \qquad (89)$$

Therefore, the amount of phase that should be added to the spectrum phase in order to continue the harmonic ith is

$$\Delta\varphi_i = 2\pi f_i \cdot transp \cdot \Delta t - 2\pi f_i \Delta t = 2\pi f_i \left(transp - 1\right)\Delta t \qquad (90)$$

This phase increment is added to all the bins that belong to the ith region, as we can see in Figure 108. This way we can preserve the phase consistency across bins, i.e. the vertical phase coherence, see Laroche (1999)

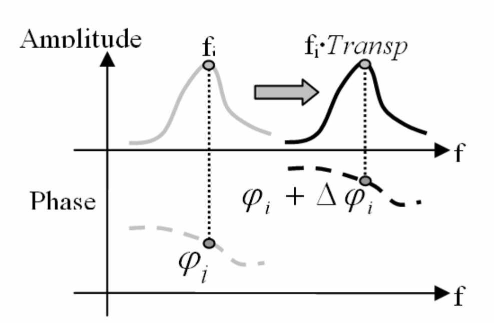


Figure 108: Phase shift in SPP transposition

The phase increment applied to each harmonic is added to an accumulated phase, frame by frame. In fact, this accumulated phase is the phase finally added to each harmonic peak region. However, this implementation results in the lost of the harmonic phase alignment after several frames because for each harmonic it uses the frequency calculated from the spectrum, and these frequencies don't follow a perfect harmonic scale. We can see this effect in Figure 109, where the dashed line is the ideal accumulated phase (perfect harmonic relation) and the solid line is the actual accumulated phase.

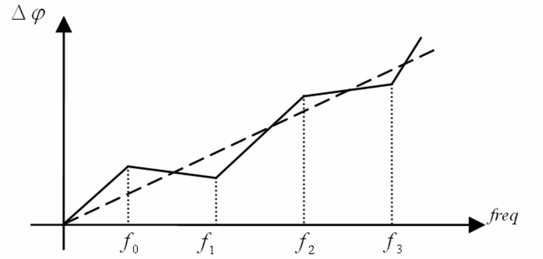


Figure 109: Accumulated phase after several frames

This loss of phase alignment produces some fuzziness and loss of presence in the synthesized voice, especially for low pitches. In order to avoid such problem the algorithm considers voice has a perfect harmonic distribution of the harmonics, in which case, the equation of the phase increment is:

$$\Delta\varphi_i = 2\pi f_0 \left(i+1\right)\left(transp-1\right)\Delta t \tag{91}$$

where f_0 is the fundamental frequency (*pitch*) and i is the index of the harmonic (i equals 0 for the fundamental).

4.3.6.2 Peak picking

Peak picking refers to the set of rules that decide, when transposing and equalizing, which of the original harmonic peak regions have to be used to fill the transformed spectrum. Standard peak picking rule is based on using the one closer in frequency. With this the algorithm tries to preserve the color of the noise attached to each harmonic peak region.

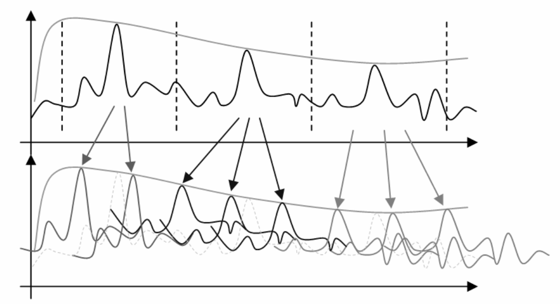


Figure 110: Peak picking and locating in a downwards transposition

However, sometimes, especially when transposing down, peaks partially masked in the original spectrum are raised by the equalization and become clearly perceptible. When the original peak has a noisy or unstable character, since gain offset is applied to all the bins inside the region, including the ones surrounding the peak, the instability or noisiness of the peak becomes noticeable. Even more, there might be cases in which the algorithm uses and raises a noisy peak twice or more causing its instability to spread along the region in which it is used.

When any of these occur, distortions and artifacts may arise. To avoid so, the algorithm forces all peaks under 900 Hz and over -30 dB to have perfect harmonic frequencies and forces all peaks under 1500 Hz and over -30 dB to be perfect harmonic peaks. With this, the algorithm cleans most significant peaks and avoids aforementioned problems.

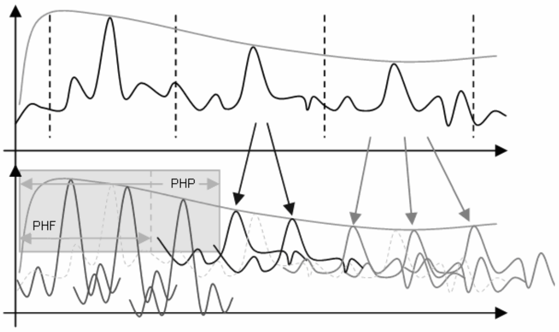


Figure 111: Peak cleaning in a downwards transposition. PHF stands for Perfet Harmonic Frequencies and PHP stands for Perfect Harmonic Peaks

However, peak cleaning introduces a problem for originally non-modal voices such as rough, husky or breathy voices to some degree: cleaning the low frequency most prominent peaks causes unnatural modifications in the original character of the voice. In order to skip such undesired effect, the peak picking strategy was redefined in order to take into account the closest distances in both domains: frequency and magnitude.

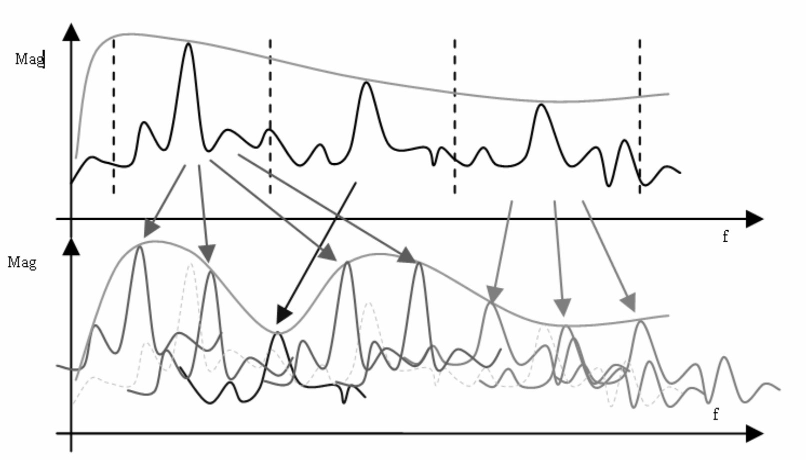


Figure 112: Peak peaking double-sided (magnitude and frequency) strategy example

With such double-sided criteria no artificial noisiness appear in synthesis at the same time transposition transformations preserve the original nature of the input singing voice.

4.3.6.3 Phase alignment

Phase alignment is applied in synthesis as follows. First, the algorithm applies the time shift (corresponding to the estimated maximally flat phase alignment value for the fundamental φ'$_0$) to move the phases of all harmonics of the analysis frame spectrum so that the phase of the fundamental φ_0 is set to φ'$_0$. This time shift for certain frame is computed as:

$$TimeShift1 = GetPhaseDiff\left(\varphi_0{}', \varphi_0\right) / \left(2\pi \cdot f_0\right) \qquad (92)$$

where *GetPhaseDiff* computes the unwrapped phase difference and f_0 is the estimated pitch. Once the time shift has been applied, the phases of the frame spectrum will be on its maximally flat phase alignment position. That will be the phase alignment to preserve.

The variable *BestPeak* stores which of the original harmonic peak region index is the best candidate as a source to fill each synthesis harmonic region. As pointed out in section 4.3.6.2, best candidates are chosen taking into consideration not only transposition but also, if present, timbre mapping transformation. There is no restriction on the fundamental, so *BestPeak*[0] is not forced any more to be 0. Now, using the *BestPeak* we define the new phase alignment for the transformed frame.

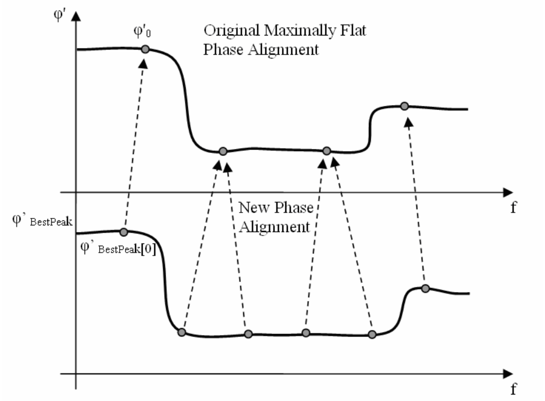


Figure 113: Hand drawn representation of the estimation of the synthesis phase alignment in downwards transposition. Dots represent peak locations.

Next, the algorithm applies the corresponding time shift in order to move the phase of the fundamental in the new phase alignment φ'$_{BestPeak[0]}$ to synthesis phase, taking into account the phase offset due to transposition. This time shift is computed as:

$$TimeShift2 = GetPhaseDiff\left(\varphi_{BestPeak[0]} + PhOffset[0], \varphi'_{BestPeak[0]}\right) \Big/ \left(2\pi \cdot f_0 \cdot transp\right) \quad (93)$$

where $\varphi_{BestPeak[0]}$ is the phase of the peak *BestPeak*[0] in the analysis spectrum, and *PhOffset*[*k*] is the phase that has to be added due to transposition to each harmonic peak region, obtained for the fundamental in frame *i* as:

$$PhOffset_i[0] = PhOffset_{i-1}[0] + 2\pi \cdot f_0 \cdot \left(iBestPeak[0]+1\right) \cdot \left(transp_0 - 1\right) \cdot T_{fr} \quad (94)$$

where *transp*$_0$ is the necessary transposition value to fill the synthesis fundamental harmonic peak region and T_{fr} is the frame time.

When *BestPeak*[*k*] changes for a certain peak *k* from one frame to the next, the algorithm applies a gradual correction to the phase of that harmonic peak region in order to solve the phase alignment envelope discontinuity. The difference between the new phase value and the previous phase value is gradually applied in 0.1π steps along consecutive frames, as shown in Figure 114.

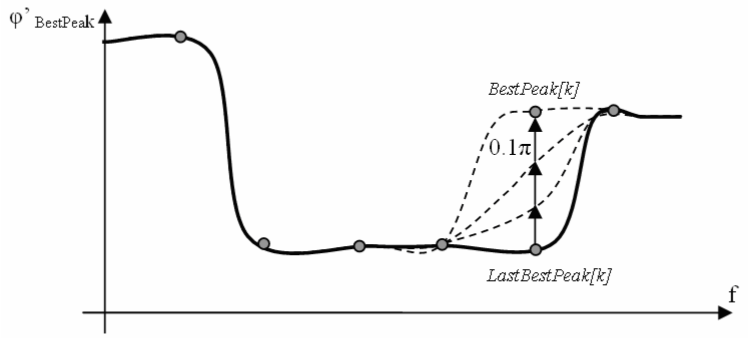


Figure 114: Phase alignment correction for consecutive frame *BestPeak*[*k*] changes

Finally, special care has to be taken every time *BestPeak*[*k*] changes occur in the fundamental. In these cases, a time shift offset has to be computed and accumulated just like phase offset is. This time shift offset is computed for frame *i* as:

$$\begin{aligned} &TimeShiftOffset_i = TimeShiftOffset_{i-1} - TimeShift2 + \\ &GetPhaseDiff\left(\varphi_{LastBestPeak[0]} + PhOffset_{i-1}[0] + PhOffset_{LastBestPeak[0]}[0], \varphi'_{LastBestPeak[0]}\right) \big/ \left(2\pi \cdot f_0 \cdot transp\right) \end{aligned} \tag{95}$$

where *LastBestPeak[k]* stores lastest valid peak indexes and $PhOffset\varphi_{LastBestPeak[0]}[0]$ is defined as:

$$PhOffset_{LastBestPeak[0]}[0] = 2\pi \cdot f_0 \cdot \left(LastBestPeak[0]+1\right) \cdot \left(transp_{LastBestPeak[0]} - 1\right) \cdot T_{fr} \tag{96}$$

where $transp_{LastBestPeak[0]}$ is the transposition factor assigned to *LastBestPeak*[0] harmonic peak region.

4.3.6.4 Partial dependent frequency scaling

Other transformations based on harmonic transposition exist in which different transposition factors can be applied to each harmonic peak. Note that, regardless of the model behind the harmonic peak transposition (sinusoidal or harmonic peak region), the resulting sound will be inharmonic. An example of such kind of transformations can be obtained by applying a frequency shift factor to all the partials of our sound:

$$f_i = f_i + k$$

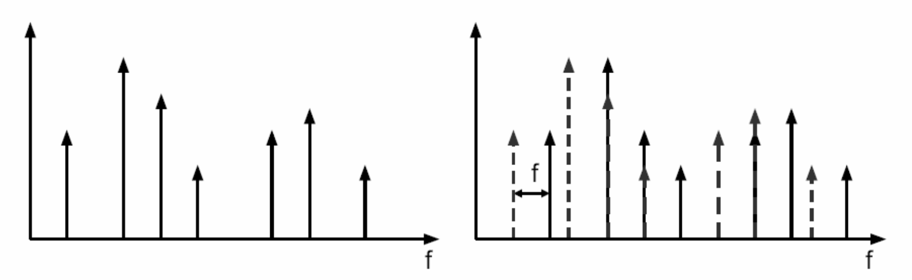


Figure 115: Representation of a frequency shift transformation. Deltas represent harmonic peaks

In the same way, all harmonic peaks can be scaled multiplying them by a given scaling factor. The relative shift of every partial could be computed, for exaple, depending on its original partial index, ruled by a stretching factor following the formula:

$$f_i = f_i \cdot fstretch^{\;(i-1)} \tag{97}$$

This kind of stretching can be observed in a real piano sound. Thus, frequency stretching could be used for example, whenever morphing voice with a piano

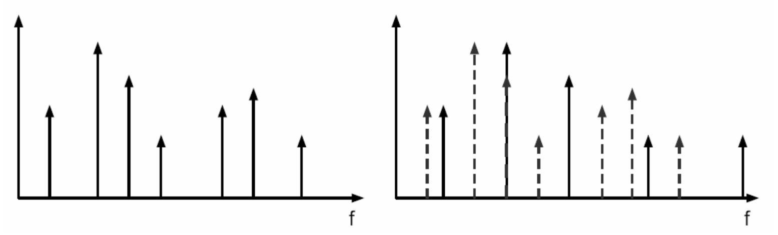


Figure 116: Representation of a frequency stretch transformation. Deltas represent harmonic peaks

4.4 High level transformations

In the context of this work, the term higher level is usually used to refer to those transformations that rather modify more abstract and perceptual attributes of the voice such as (melody, personality or density). Sometimes the term also refers to those algorithms that are applied as a function of the properties obtained in a prior analysis of the input voice, see Verfaille et al. (2005) and Amatriain et al. (2003) or even as a complex combination of lower level transformations, see Serra and Bonada (1998). The boundary to differentiate higher level transformations does not draw a clear line and is in fact an appreciation that is susceptible of discussion.

4.4.1 Pitch based high level transformations

Transformations such as intonation, pitch discretization or vibrato apply direct modifications to the fundamental frequency envelope. Although these transformations could be considered low level transformations they deal with melodic concepts, some depend on the analyzed attributes, and they change the way the melody is perceived by the listener.

4.4.1.1 Intonation

Intonation transformation is achieved by stretching or compressing the difference between the analysis pitch envelope and a low pass filtered version of it. The goal of the transformation is to sharpen or flatten the non-stationeries of a vocal performance such as attacks, transitions or releases.

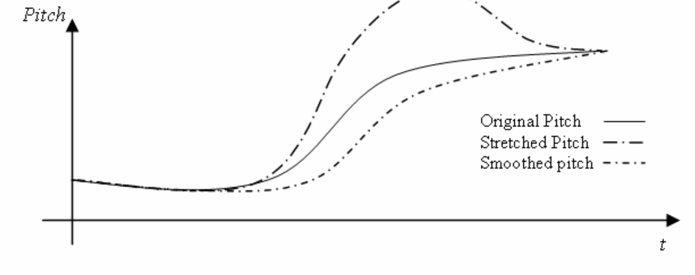


Figure 117: Representation of the effects produced to the pitch envelope due to the two different intonation transformation configurations

The effect of the intonation transformation is a subtle modification of the phrasing of the melodic lines.

4.4.1.2 Pitch discretization to the temperate scale

Pitch discretization to the temperate scale is indeed a very particular case of pitch transposition where the pitch is quantified to one of the 12 semitones in which an octave is divided. The transformation can be accomplished by forcing the pitch to take the nearest frequency value of the equal temperate scale. This effect is widely used on vocal sounds, see sections 2.2.1.4.1 and 2.2.2.

Using the sinusoidal model for a formulation of the transformation, a perfect harmonic sound sinusoids fulfil

$$f_i = f_0 \cdot i \tag{98}$$

where f_0 is the frequency of the fundamental and $i=1..N$ where N is the number of sinusoids. The algorithm computes the new fundamental frequency applying the following formula:

$$f_0' = 55 \cdot \left(\left(2^{\left(1/12 \right)} \right)^{\left(round \left(12 \cdot \log\left(f_0/55 \right) / \log(2) \right) \right)} \right) \tag{99}$$

where it is assumed 55 is the frequency in *Hz* that corresponds to an A0.

From this new fundamental f_0', we can compute the transposition factor, defined as:

$$k = \frac{f_0'}{f_0} \tag{100}$$

and apply it using a pitch transposition algorithm

4.4.1.3 Vibrato

The emulation of vibrato and tremolo can be achieved in real-time by applying low frequency modulations to the frequency (vibrato) and to the amplitude (tremolo) of the partials. As already mentioned in section 3.2.3, vibrato and tremolo modulations share the resonating frequency f_m. In order to apply vibrato the algorithm modulates the fundamental frequency f_0 as:

$$f_0' = f_0 \cdot (1 + c \cdot \sin(2\pi \cdot f_m)) \tag{101}$$

where c is the vibrato depth, usually around 75 cent, and f_m ranges from 4 to 7 Hz. With this the algorithm obtains a transposition factor and applies the corresponding transposition with timbre preservation.

For the tremolo, the modulation over the harmonic peaks amplitude a_i can be applied as:

$$a_i' = a_i + t(i) \cdot \frac{\sin(2\pi \cdot f_m) - 1}{2} \text{ (dB)} \qquad (102)$$

where the modulation depth *t(i)* would apply different depths at different frequencies emulating the spectral tilt variations suffered in a real tremolo sound. This curve could be a sampled version of the curve shown in Figure 118.

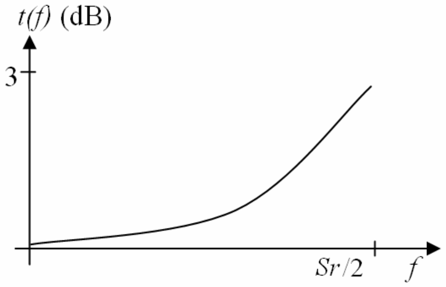


Figure 118: Exemplification of a possible *t(f)* curve, where *Sr* is the sampling rate

When real time is not necessary, more realistic vibrato transformation can be achieved using real vibrato recording. The recordings are analyzed and their patterns of modulation over the fundamental frequency and over the EpR's resonances are used as template and applied to the input voice. For the modulation of the EpR attributes, an estimation of the EpR model is required for the input voice as well.

4.4.2 Identity transformations

There are transformations whose main goal is to modify the person behind a singing performance. These transformations do not target a specific singer sound alike (see section 4.5.1) but they rather pursue a generic resemblance to communities such as men, woman, kids, aged, or even robot or beasts.

Three different examples of such kind of transformation are given next in this section.

4.4.2.1 Gender change

Gender transformation is performed by means of a combination of pitch transposition, timbre mapping, and spectral shape shift.

Pitch transposition is applied using the simplest rule, for woman to man transpose downwards and for man to woman transpose upwards. This basic rule is in fact a direct implication of the result of many years of human evolution: statistically, and because of their different physiology, men speech and sing at lower pitches than women.

But men do not only sing at lower pitches but they also have, in general, bigger vocal tract organs. Therefore, timbre mapping is applied to compensate such differences. For the man to woman transformation the spectral shape is compressed and vice versa for the woman to man.

The theoretical explanation for the spectral shift to be used in this effect is that women change their formant frequencies in accordance to the pitch. That is, when a female singer rises up the pitch, the formants move along with the fundamental. Thus, to convert a male into a female a spectral shift is applied as a function of the new synthesis pitch. Also, to convert a female into a male voice the algorithm also applies a shift in the spectral shape. This shifting has to be applied in a way the formants of the female voice remain stable along different pitches

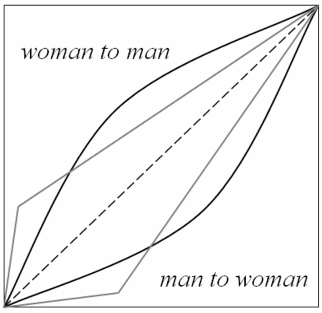


Figure 119: Illustration of four possible mapping functions of the timbre mapping. Those on the upper triangle belong to woman to man transformation and those on the lower to the man to woman. All sort of different compress / stretch functions can be useful for the gender change. Examples based on quadratic functions are plotted in black, mapping functions compound of lines are plotted in grey.

4.4.2.2 Age transformations

Voice transformations exist for converting a middle age man into a kid or into an elder person.

Kid transformation is basically based on applying transposition upwards (of approximately an octave), timbre mapping, and intonation. Regardless of the gender of the input singer, timbre mapping function expands the spectral shape in a similar way it is done for male to female transformation. Intonation is applied to expand original pitch variations to emulate the kid's reduced control over pitch.

Aged transformation uses vibrato, breathiness, roughness and timescaling. Vibrato is applied with exceptionally fast modulation rate and depth values in order to resemble false notes in their performance. Breathiness and roughness are applied in order to emulate slight vocal disorders, and timescaling can be applied, when no background music, to slow down the uttering rate.

Anyway, in both cases phonetic transformation is required in order to achieve perfectly natural sounding. Both kids and aged people present characteristic uttering peculiarities that, if automatically and synthetically emulated, could improve the quality of such age transformations.

4.4.2.3 Exotic alienations

Exotic alienations refer to transformations such as comic, ogre or alien. These effects combine standard natural sounding transformations tuned with extreme parameters settings together with artificial sounding transformations. Such is the case of ogre transformation, where extreme timbre mapping and transposition is applied together with some frequency stretching; or the case of alien, where a combination of vibrato and frequency stretching is used to achieve an outer space effect. However robot transformation is an exception to this sort of combinations. Robot effect is obtained by simply setting, for each synthesis frame, the spectral phase envelope to constant zero.

4.4.3 Multiplication

Multiplication section embraces those singing voice transformations that turn a single voice input into a group of voices.

4.4.3.1 Harmonizer

Harmonizing can be defined as "*mixing a sound with several pitch-shifted versions of it*" [Gus]. Most of the times, the pitch shifted versions are forced to be in tune with the original. The harmonizing effect requires from two parameters: the number of harmonies and the pitch for each of these. The pitch of each of the harmonies is typically specified by the key and chord of harmonization. Sometimes, the key and chord is estimated from the analysis of the input pitch and the melodic context, renaming the effect as smart or intelligent harmonizing, see Pachet and Roy (1998), Tokumaru et al. (1998), Abrams et al. (1999).

The formulation of the harmonizer for a number H of harmonies can be:

$$X'(f)=X(f)+\sum_{h=1}^{H} X_{transp}(f,h) \tag{103}$$

Where $X_{transp}(f,h)$ is the original sinusoidal spectrum transposed (with timbre preservation) by a factor that depends on h.

The actual implementation of the algorithm includes additional processing in order to minimize the correlation between all different voices in the harmony and thus improve the reality of the result. On one hand, the algorithm applies time-varying delays in the unvoiced parts of each generated voice track. These delays can follow a given table or can be calculated following a random Gaussian distribution for which the user can specify its mean and deviation. On the other hand, the algorithm applies time-varying dynamic transformation to the different voices.

This harmonizing algorithm has been ported to a real time MIDI keyboard harmonizer, with which singers can harmonize live their performance by playing a keyboard. In this environment, the implementation includes a manager for the note to note transitions and a manager for the chord to chord transitions.

Note to note transitions are handled as represented in Figure 120. When the player releases a note, the note is hold during a certain stand by time (150 milliseconds). If no notes are played during this time, release-like envelopes are used to stop the synthesis of that voice track once the stand by time has elapsed. However, if a new note is played during stand by, a smooth note to note fundamental frequency transition is applied to transit to next note.

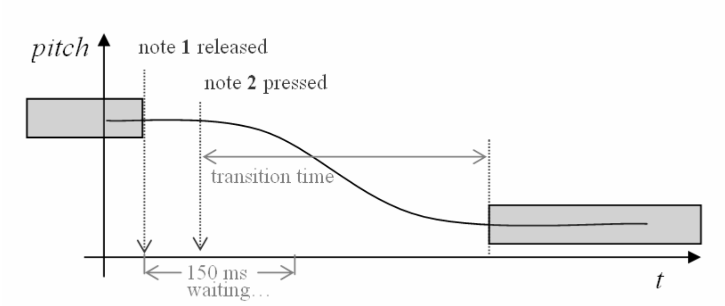


Figure 120: Harmonizer note to note transition

Chord to chord transitions manager tries to match each key note of the previous chord with each key note of the actual chord with a strategy ruled by the key number. Each chord is represented as a vector whose elements are the key numbers of each note sorted from lowest to highest. The match is performed by linking notes that share the vector index in consecutive chords. Thus, the lowest will be linked with the lowest and so on.

4.4.3.2 Choir

Choir transformation converts the sound of a single voice input into a unison choir of singers. Choir effect has been typically approached generating multiple clones of the input voice, where each clone had slight modifications on the pitch, time and timbre envelopes emulating different singers in a choir, see Schnell et al. (2006). This is the implementation available in most audio processing software under the name of chorus. The results of such approach are far from being natural and they are usually perceived as a phaser-flanger like transformation rather than a chorus. Alternatively, Kahlin and Ternström (1999) tried to achieve ensemble perception by amplitude modulations in the spectral harmonic peaks of the signal with no definitive results. More recently, Bonada proposed a choir transformation based on the morph (see section 4.5) between the input voice and a real choir recording as shown in Figure 121, see Bonada (2005).

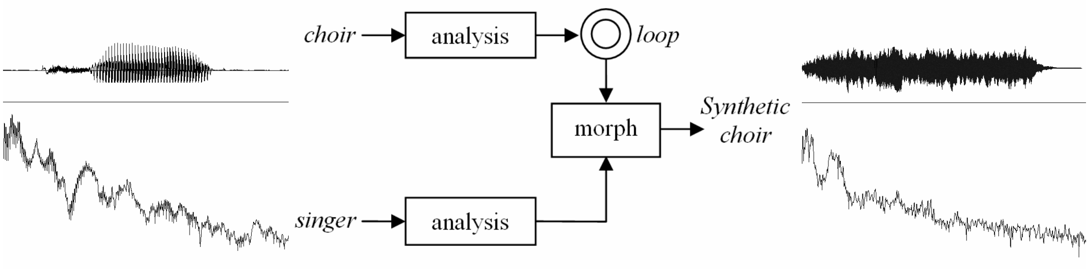


Figure 121: System overview (middle), and visual comparison of the spectrums of a voice solo (left) and a unison choir (right). Taken from Bonada (2005) with permission of the author.

The choir recording is edited to remove the non stationeries, is analyzed and stored in a loop buffer. The morph module combines the pitch and timbre from the input singer with the local spectrum of the choir.

4.4.4 Voice enhancement

This section presents a new method for esophageal voice enhancement using speech digital signal processing techniques based on modeling radiated voice pulses in frequency domain. Though this might seem far away form the main subject of this work it is specifically included due to two main reasons. First, the proposed algorithm comprises very interesting and novel approaches with clear applications on the digital processing of the singing voice. The second, singing is also a discipline for the laryngectomized community. Prove of it is the laryngectomized choir of León, in Spain, where more than twenty post neck surgery patients sing together, record albums, and perform live.

Figure 122: The laryngectomized choir of Leon[50]

Most laryngectomized patients suffer from voice blackout after neck surgery, having mainly two ways to recover voice. One consists in using an Artificial Larynx, a hardware device which held against the neck or cheek produces sound vibrations in the throat, while the speaker articulates with the tongue, palate, throat and lips as he does for the usual vocalization. Those devices can be easily mastered, but the sound quality is rather electronic and artificial. Besides, one hand is employed to hold the device during speech, disturbing the gestural communication.

The alternative consists in training esophageal speech, a way of speech production based on the technique in which the patient transports a small amount of air into the esophagus. Probably due to an increased thoracic pressure, the air is forced back past the pharyngo-esophageal segment to induce resonance and allow speech. Rapid repetition of such air transport can produce understandable speech. However, on average, esophageal voice results in low-pitched (~50Hz), low intensity speech, and with a poor degree of intelligibility. On the other side, this technique is able to preserve the speaker's individuality, since the speech is generated by his own vocal organs, and the speaker is able to use his hands and facial expression freely and actively for a more natural communication.

Several researches which pretend to enhance and clarify the esophageal speech have been reported so far in Nakamura and Shikano (1996) and Sawada et al. (2004), as well as studies which account for its lack of phonetic clarity, see Robins et al. (1994) and Bellandese et al. (2001). Besides, an electronic larynx device[51] with the aim of improving

[50] http://www.foros.com/alle/coro.asp

[51] http://www.larynxlink.com/suppliers/RometBrochure.pdf

the esophageal voice is commercially available since few years ago. This device consists of a small circuit board, a compact microphone and a speaker, and uses formant synthesis techniques to produce the synthetic voice. One of its main disadvantages is that it fails to keep the speaker's individuality.

A new algorithm has been designed to: improve intelligibly of esophageal speech, allow laryngectomized patients have intuitive control on the resynthesized prosody, minimize traces of artificialness, and resemble pre-surgery patient healthy voice. This algorithm uses straightforward digital signal processing system based on analysis, transformation and resynthesis to enhancement the esophageal voice.

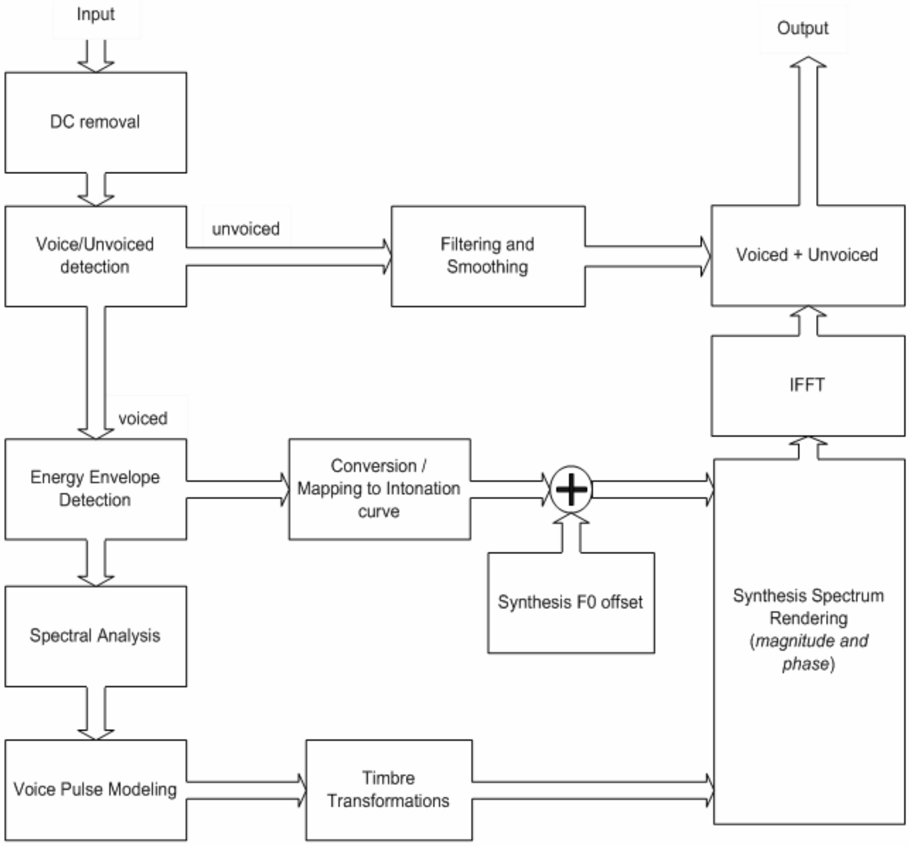


Figure 123: General block diagram of the Voice Enhancer

The proposed voice enhancer is based on the Voice Pulse Modeling (VPM) algorithm, see Bonada (2004) which intends to combine the waveform preservation ability of time-domain techniques with the flexible and wide transformations of frequency-domain techniques, while at the same time avoiding the complexity and the contextual problems of them.

The implementation of the system is based on frame analysis. After filtering, the input voice frame is classified as voiced or unvoiced. For the voiced frames spectral phase and magnitude (timbre and pitch) information is made up to feed the VPM. The voiced spectrum is then converted to time domain by the IFFT module and added to the processed unvoiced utterances.

4.4.4.1 The preprocess and the unvoiced cycle

The preprocess block for the first prototype implementation consists on a simple high pass filter (DC removal in figure 1). The filter is applied to get rid of the lowest frequencies (from 10 Hz to 30 Hz approximately) typically present in esophageal voice recordings due to powerful respiration.

Once filtered, there is a voiced / unvoiced detection which does not rely on fundamental frequency analysis as it does traditionally, but it bases on a combination of signal dynamics and spectral centroid. For those frames that are considered unvoiced, second step equalization is applied in order to reduce unvoiced consonant perceptual forcefulness.

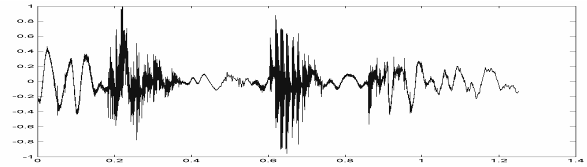


Figure 124: Waveform (normalized magnitude versus sample index) of the recording of the Spanish word '*tomate*' uttered by a laryngectomized patient

4.4.4.2 Feeding the synthesis engine

If voiced, the utterance is analyzed and resynthesized with new timbre, phase and prosody. Because of the rough nature of esophageal speech, no pitch analysis is performed. Thus, spectral timbre envelope and frequency phase alignment are not computed using harmonic peak information as usually done in the VPM technique (figure 3).

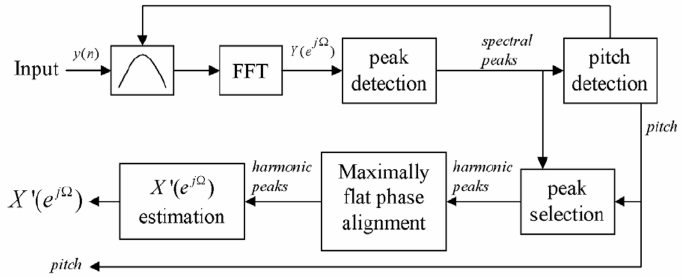


Figure 125: Block diagram of the VPM analysis process taken from Bonada (2004) with permission of author

For the case of the timbre, we track spectral envelope using a bank of non-overlapped constant bandwidth (~175 Hz) filters equally spaced in frequency. Frequency phase alignment envelope is derived directly from the resulting dB spectral magnitude envelope by low pass filtering, offsetting, shifting and scaling its y-axis to fulfill phase 0 at frequency 0 and approximately π phase drops under most prominent formants. A smoothing function ensures no phase alignment discontinues at consecutive frames.

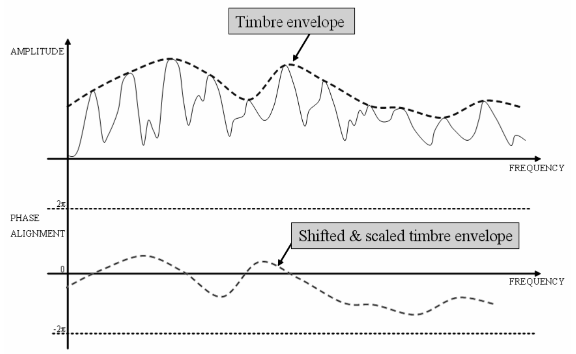


Figure 126: Hand drawn representation of the phase alignment envelope generation process from the shift and scale of the estimated timbre envelope

These frequency magnitude and phase envelopes are then used by the VPM to model the frequency response of a single voice pulse centered in the analysis window. Next, together with the pitch the rendering module generates the synthesis frequency domain spectrum out of the transformed voice pulse models.

Synthesis pitch envelope *$f_0(t)$* is obtained by means of filtering, scaling and offsetting the energy envelope (in dB scale) detected in the analysis:

$$f_0(t) = filter\left[10 \cdot \log\left(s^2(t)\right)\right] \cdot \alpha + \beta \tag{104}$$

Where α=12.5 and β=-2400 have been found to be appropriate values for the male voice used as an example. However, values should be tuned for each specific case in order to fit as much as possible original patient's average speech fundamental frequency and prosody intonation.

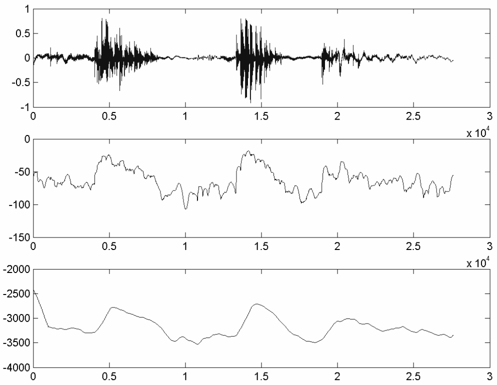


Figure 127: Along sample index, upper plot shows the original input signal after preprocess, mid plot shows its energy envelope in dBs, and lower plot shows the computed healthy synthesis pitch in cents of a tone

4.4.4.3 Healthy pitch and timbre

Although alaryngeal speakers suffer mainly from changes in those characteristics related to the voice source, relevant changes occur as well in the vocal cavity transmission characteristics.

Recent studies in which acoustical analysis of vowels was carried out show alaryngeal speakers place their formants in higher frequencies. According to Cervera et al. (2001) the explanation of this symptom seems to be that total laryngectomy results in shortened vocal tract relative to normal subjects. In fact, in Diedrich and Youngstrom (1966) it is demonstrated using data of a patient captured before and after laryngectomy that effective vocal tract length is reduced after neck surgery.

In order to compensate the aforementioned formant shift, the resulting bank magnitude envelope is frequency-varying scaled using a timbre mapping function such as:

$$|X_s(f)| = |X_a(f^{\alpha})| \tag{105}$$

where $|X_s(f)|$ is the synthesis spectral envelope and $|X_a(f)|$ the spectral envelope obtained in the analysis.

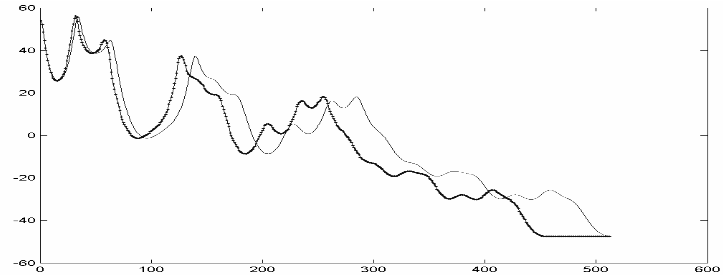


Figure 128: Log spectral magnitude versus frequency index plot of analysis (lighter) and synthesis (darker) timbre envelopes for an [o] utterance

For values of α around 1.02 the frequency down shift achieved is around 60 Hz for the first formant and 160 Hz for the second formant, which are the values spotted in [7] as average formant rise values.

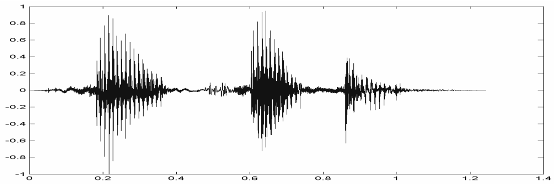


Figure 129: Waveform (normalized magnitude versus sample index) of the resynthesized '*tomate'* utterance represented in figure 2

4.5 Morph and control, voice driven synthesis

Morphing is a technique with which, out of two or more elements, we can generate new ones with hybrid properties. With different names, and using different signal processing techniques, the idea of audio morphing is well known in the Computer Music community, see Serra (1994), Tellman and Haken (1995), Osaka (1995), Slaney et al. (1996) and Settel and Lippe (1996). In most of these techniques, the morph is based on the interpolation of sound parameterizations resulting from analysis/synthesis techniques, such as the short-time Fourier Transform, LPC or sinusoidal models.

4.5.1 Singing voice conversion

Singing voice conversion is a very particular case of audio morphing. The goal is to be able to morph two singing voice signals in such a way we can control the individuality of the resulting synthetic voice. Whenever this control is performed by means of modifying a reference voice signal matching its individuality parameters to another, we can refer to it as voice conversion, see Abe (1992).

Specific automatic conversion for the singing did not appear since the end of the 1990's. Yoram Meron proposed using Hidden Markov Models (HMMs) to map an individual singer sinusoidal phoneme models with a target sound, see Meron (1992). At the same time, people at the Music Technology Group developed a real time singing voice automatic impersonation transformation for a karaoke environment; see Cano and Loscos (1999), deBoer et al. (2000), and Cano et al. (2000). The system used the sinusoidal plus residual model to morph two performances, seed's and target's, time synchronized with a real time audio alignment based on a HMM's. Some details are given next for the latter approach.

A karaoke-type application for PC was implemented in which the user could sing like his/her favorite singers. That is, an automatic impersonating system with which the user could morph his/her voice attributes, such as pitch, timbre, vibrato and articulations with the ones from a prerecorded singer we call target. In this particular implementation the target's performance of the song to be morphed is recorded and analyzed beforehand. In order to incorporate to the user's voice the corresponding characteristics of the target's, the system first recognizes what the user is singing (phonemes and notes), looks for the same sounds in the target performance (i.e. synchronizes the sounds), interpolates the selected voice attributes, and synthesizes the output morphed voice. All this is accomplished in real-time.

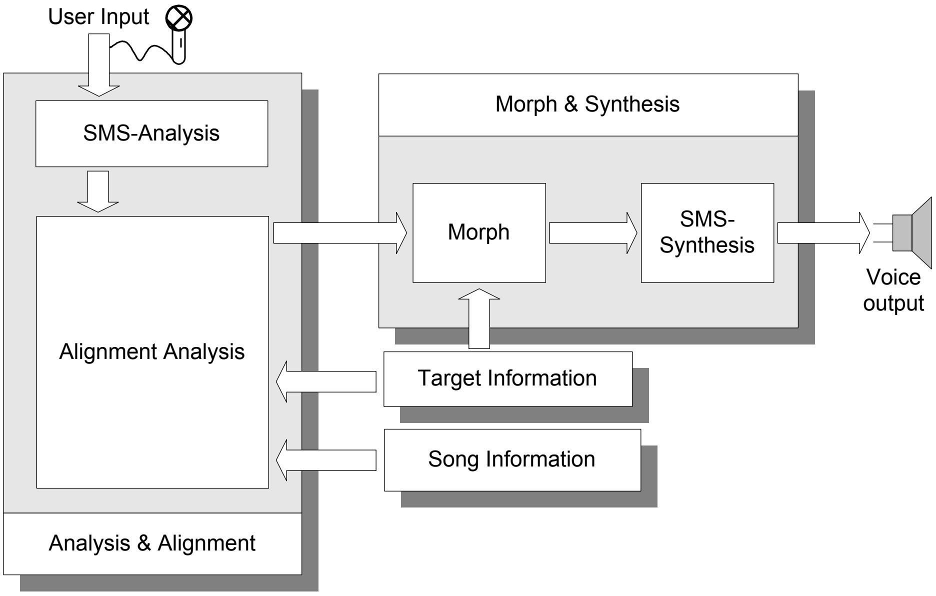


Figure 130: System block diagram with an overview of the whole process.

Figure 130 shows the general block diagram of the voice impersonator system. The system relies on two main algorithms that define and constrict the architecture: the SMS analysis/synthesis technique, see section 2.4.2.4.2, and a Hidden Markov Model based Automatic Speech Recognizer (ASR). The SMS technique is the one responsible of providing a suitable parameterization of the singing voice in order to perform the morph in a flexible and musical-meaningful way and the ASR is the one responsible of aligning user's with target's singing performances.

Before the morph takes place, it is necessary to supply information about the song to be morphed and the song recording itself (Target Information and Song Information). The system requires the phonetic transcription of the lyrics, the melody, and the actual recording to be used as the target audio data. Thus, a good impersonator of the singer that originally sang the song has to be recorded. This recording has to be analyzed with SMS, segmented into morphing units, and each unit labeled with the appropriate note and phonetic information of the song. This preparation stage is done semi-automatically, using a non-real time application developed for this task.

The first module of the running system includes the real-time analysis and the recognition/alignment steps. Each analysis frame, with the appropriate parameterization, is associated with the phoneme of a specific moment of the song and thus with a target frame. Once a user frame is matched with a target frame, the algorithm morphs those interpolating data from both frames and synthesizes the output sound. Only voiced phonemes are morphed and the user has control over which and by how much each parameter is interpolated. The frames belonging to unvoiced phonemes are left unprocessed. Thus unvoiced synthesis always takes user's consonants.

4.5.1.1 Front end parameterization of the singing voice

The speech assumption that the signal can be regarded as stationary over an interval of a few milliseconds is kept true for the singing voice. Thus, the prime function of the front-end parameterization stage is to divide the input speech into blocks and from each block extract the features. The spacing between blocks is 5.8 ms and the blocks are overlapped with an analysis window of approximately 20ms. Various possible choices of vectors together with their impact on recognition performance are discussed in Haeb-Umbach (1993). The sound features extracted in the system front-end are: Mel Cepstrum, delta Mel Cepstrum, energy, delta energy and voiceness.

To compute the Mel-Frequency Cepstral Coefficients (MFCCs) coefficients, the Fourier spectrum is smoothed by integrating the spectral coefficients within triangular frequency bins arranged on a non-linear scale called the Mel-scale. The system uses 24 of these triangular frequency bins (from 40 to 5000 Hz). The Mel-scale is designed to approximate the frequency resolution of the human ear being linear up to 1000 Hz and logarithmic thereafter. In order to make statistics of the estimated speech power spectrum approximately Gaussian, log compression is applied to the filter-bank output. The final processing stage is to apply the Discrete Cosine Transform to the log filter-bank coefficients. This has the effect of compressing the spectral information into the lower order coefficients and it also decorrelates them.

The voiceness vector consists of a Pitch Error measure and Zero Crossing Rate. The pitch error component of the voiceness vector is extracted from the fundamental frequency analysis. To obtain the zero crossing rate measure we use the formula:

$$Z_s = \frac{1}{N} \sum_{n=m-N+1}^{m} \left(\frac{\left|\mathrm{sgn}\{s(n)\} - \mathrm{sgn}\{s(n-1)\}\right|}{2} w(m-n) \right) \qquad (106)$$

where *sgn*{*s*(n)} = +1, if *s*(n)>=0; -1 if *s*(n) <0, *N* is the number of frame samples, *w* is the frame window and *s* is the input signal.

The acoustic modeling assumes that each acoustic vector is uncorrelated with its neighbors. This is a rather poor assumption since the physical constraints of the human vocal apparatus ensure that there is continuity between successive spectral estimates. However, considering differentials to the basic static coefficients greatly reduces the problem.

All these extracted features are quantized using a Lindo Buzo and Gray (LBG), see Lindo et al. (1980) algorithm approach.

4.5.1.2 Alignment

An aligner is a system that automatically time aligns speech signals with the corresponding text. This application emerges from the necessity of building large time-aligned and phonetically labeled speech databases for ASR systems. The most extended and successful way to do this alignment is by creating a phonetic transcription of the word sequence comprising the text and aligning the phone sequence with the speech using a HMM speech recognizer, see Waibel and Lee (1990).

The phoneme alignment can be considered speech recognition without a large portion of the search problem. Since the string of spoken words is known the possible paths are restricted to just one string of phonemes. This leaves time as the only degree of freedom and the only thing of interest then is to place the start and end points of each phoneme to be aligned. For the case of aligning singing voice to the text of a song, more data is available out of the musical information of which the algorithm can take profit: the time at which the phoneme is supposed to be sung, its approximate duration, and its associated pitch.

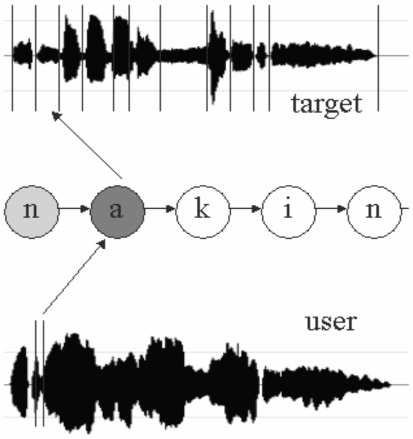


Figure 131: Recognition and matching of morphable units.

The aligner presented is this section is a phoneme-based system that can align the singing voice signal to the lyrics both real and non-real time. In no real time the aligner is used in the preparation of the target audio data, to fragment the recording into morphable units (phonemes) and to label them with the phonetic transcription and the musical context. This is done out of real-time for a better performance. And when real time, as the singer performs, the voice can be analyzed and morphed with the corresponding piece of the target depending on which phoneme of the lyrics is currently being sung.

In this type of systems, contextual effects cause large variations in the way that different sounds are produced. Although training different phoneme HMMs for different phoneme contexts (i.e. triphonemes) present better phonetic discrimination, it is not recommended in the case no large database is available, which is the case for the singing voice. HMMs can have different types of distribution functions: discrete, continuous, and semi continuous. Discrete distribution HMMs match better with small train database and are more efficient computationally [You96]. Because of all this considerations, the nature of the elements is chosen to be discrete.

The most popular way in which speech is modeled is as a left-to-right HMM with 3 states. The implementation of the system also fits 3 states to most of the phonemes (except for the plosives) as an approach to mimic the attack, steady state and release stages of a note. The plosives are modeled with 2 states to take into consideration somehow their intrinsic briefness. The silence is modeled with 1 state as it is in speech since silence is generally stationary and has no temporal structure to exploit.

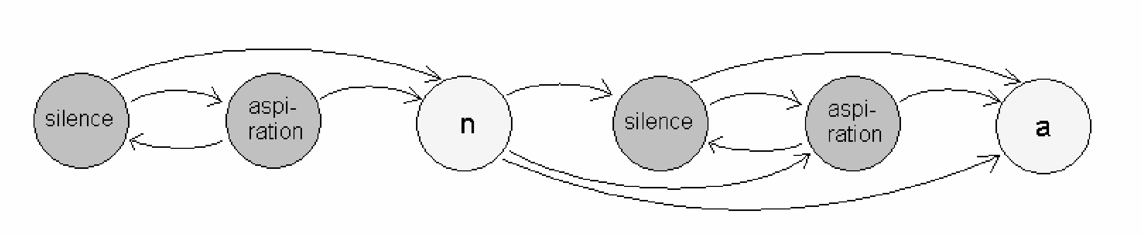


Figure 132: Concatenation of silences and aspirations in the composite song network

The alignment process starts with the generation of a phonetic transcription out of the lyrics text. This phonetic transcription is used to build the composite song network concatenating the models of the phonemes transcribed.

The phonetic transcription previous to the alignment process has to be flexible and general enough to account for all the possible realizations of the singer. It is very important to bear in mind the non-linguistic units silence and aspiration as they appearance cannot be predicted. Different singers place silences and aspirations in different places. This is why while building the composite song network, between each pair of phoneme models, we insert both silence and aspiration models.

Viterbi algorithm is the usual decoding choice in the text to speech alignment problem, see Rabiner and Juang (1993). This algorithm gives as result the most probable path through the models, giving the points in time for every transition from one phoneme model to the following. The alignment resultant from the Viterbi decoding will follow the most likely path, so it will decide which the most probable phoneme sequence is.

To achieve a low-delay Viterbi, some modifications have been introduced in the standard decoding, see Loscos and Cano (1999). To compensate a possible loss of robustness, some strategies on discarding phony candidates are introduced to preserve a good accuracy. During the low-delay alignment we have several hypotheses on our location in the song with similar probability. We use heuristic rules as well as musical information from the score to discard candidates.

4.5.1.3 Morph and synthesis

Depending on the phoneme the user is singing, a unit from the target is selected. Once the target unit has been chosen, each frame from the user is morphed with a different frame from the target's unit, advancing sequentially in time. The system features controls on the interpolation of the different parameters extracted at the analysis stage, such as amplitude, fundamental frequency, spectral shape, residual signal, etc. However, by default, the amplitude is not interpolated in order to give the user the feeling of being in control.

In most cases the durations of user and target phonemes are different. If a given user's phoneme is shorter than the one from the target the system simply skips the remaining part of the target phoneme and goes directly to the articulation portion. In the case when the user sings a longer phoneme than the one present in the target data, the system enters in the loop mode. Each voiced phoneme of the target has a loop point frame, marked in the preprocessing, non-real time analysis. The system uses this frame to loop-synthesis in case the user sings beyond that point in the phoneme. Once this last frame in the target is reached, the rest of the user's frames are interpolated with it as shown in Figure 3.

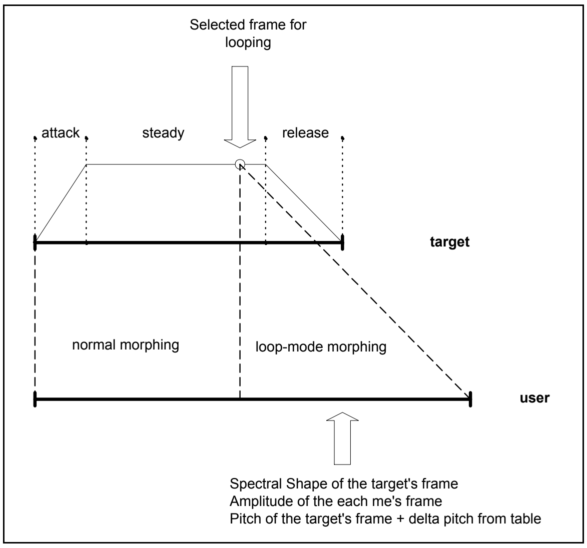

Selected frame for looping
attack
steady
release
target
normal morphing
loop-mode morphing
user
Spectral Shape of the target's frame
Amplitude of the each me's frame
Pitch of the target's frame + delta pitch from table

Figure 133: Loop synthesis diagram.

The frame used as a loop frame requires a good spectral shape and a pitch as close as possible to the note that corresponds to that phoneme. In order to avoid static artificial sounding due to the repetition of a single frame, templates obtained from the analysis of a longer phoneme are used to generate synthetic variations of the loop frame. Once all the chosen parameters have been interpolated in a given frame they are added back to the basic SMS frame of the user. The synthesis is done with the standard synthesis procedures of SMS.

4.5.2 Voice orchestration

The term voice orchestration refers to the creation of musical arrangements for the different instruments in an orchestra by means of singing voice melody phrasing and vocal instrument mimicking.

This section presents a prototype with limited features, named Larynxophone, for the voice orchestration of a trumpet. Similar to the singing voice impersonator, the Larynxophone processes can be decomposed in non real time processes, which are the ones that take place as a prelude, before the application runs, and the processes that take place in real time, which are the ones that occur while the user is performing.

The analysis used for both the trumpet samples and the voice signal captured by the microphone is frame-based, and uses spectral domain techniques that stand on the Spectral Peak Processing, see section 2.4.2.2.

4.5.2.1 Non-real time processes: instrument database creation

The non-real time processes focus on the wind instrument database creation. That is, on recording real performances, editing and cutting them into notes, analyzing and labelling them, and storing them as an instrument database.

In the current implementation, the database contains only three trumpet notes at A3, A#5, and C5. For each sample the database contains a binary file in which the necessary data resulting from the analysis is stored.

4.5.2.2 Real time processes: voice analysis, morph and instrument synthesis

The real-time processes start with a frame based spectral analysis of the input voice signal out of which the system extracts a set of voice features. This voice feature vector decides which sample has to be fetched from the database, and controls the cross-synthesis between the instrument sample and the voice signal.

Because of the nature of the prototype database, the criterion to decide the trumpet sample to pick at each frame time is “take the nearest sample in pitch”, which is a particularization of the more general criterion “take the sample with most similar expression”. From that sample, the trumpet frame is chosen sequentially, taking into account loops. The frame is transformed to fit the user’s energy and tuning note specification, for which energy correction and transposition with spectral shape preservation is applied, with similar techniques to those described in [4]. Finally, the synthesis is in charge of concatenating the synthesis frames by inverse frequency transformation and the necessary window overlap-add related processes.

Regardless the modest suitability of using a voice to MIDI converter, it is important and probably a key factor to decide how to map voice attributes to MIDI messages.

Basically, driving a synthesis with voice utterances requires from a mapping between pitch / dynamic related voice attributes and MIDI messages: note on / note off (key number and key velocity), poly-phonic key pressure (aftertouch), and pitch bend change. Of course neither these voice attributes fulfil vocal expressivity space nor the

MIDI messages are able to reproduce all possible wind instrument expressive nuances; however, when adequately mapped will allow a useful basic control.

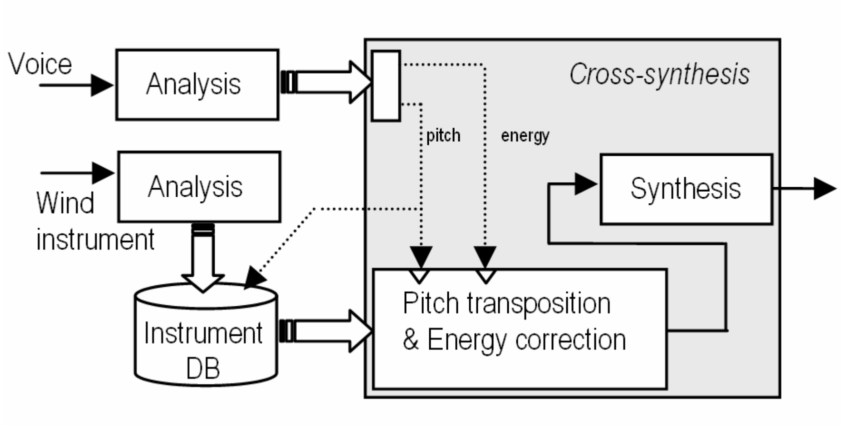


Figure 134: Larynxophone block diagram

The proposal is to use the pitch envelope obtained in the fundamental frequency analysis to fill the key number and its associated modulation along the note, the pitch bend; and to use the EpR excitation parameter, see section 2.4.2.5.1, to fill the key velocity and its associated modulation, the aftertouch.

Obviously though, this process has to run on the fly and this means once a frame has been detected as the onset of a note, the converter takes the current pitch and dynamic values (possibly averaged along some short past history) as the mean over the note. Thus, all follow-ing frames that are considered part of that note define aftertouch and pitch bend messages from the difference between its current values and the onset frame values.

The cross-synthesis between wind instrument samples and voice utterances is a shortcut technique that avoids intermediate MIDI conversions. Taking profit of morph algorithms, we can extend the voice control further than pitch, dynamics and their associated modulations and set off continuous control over, for example, the sharpness of the attack or the instrument timbre modulations.

4.6 Conclusions

A significant group of transformations for the sensing voice based on spectral analysis and modelling have been introduced in this chapter.

Regarding the excitation based transformations, although algorithms have proven to be suitable in changing the voice character, the naturalness of the effect is highly dependent on input voice. For different types of voice, different tessitura, different expressions, etcetera, different values of transformation parameters are required. In that sense, a dynamic automatic control over transformation parameters has to be found. Concretely, in the growl effect patterns extracted from real growl recordings are roughly reproduced in synthesis. This means the period to period amplitude envelope inside a growl macro-period is not only included in the phase alignment of the sub-harmonics but also in the sub-harmonics amplitudes. However, it is a tedious job to find the sub-harmonic amplitudes and phase alignment required for a certain made-up amplitude envelope. It is also remarkable no control over the jitter is available with the current growl implementation.

Dynamic transformation is probably the one lacking the most of naturalness. Although relatively modest values of the transformation parameter can result into natural sounding, more significant changes are not perceived as natural. Probably, the solution can be found by stating voice at different dynamics (from whisper to shout) can be considered different instruments and thus morphing techniques are required as in Kawahara (2003).

As a general comment, it is a requirement in any research discipline to establish a criteria by which technologies and algorithms can be assessed and compared between them. However, when it comes to evaluate the quality of a singing voice transformation, regardless if quality refers to naturalness, intelligibility, any other perceptual attributes, two options exist. The first, which is probably the most commonly accepted by the scientific community, is to organize and perform human perceptual tests. However, an alternative validation procedure exists based on the deployment of the algorithm. In such procedure, technologies are judged by integrating them into an application and evaluating the results from the user feedback perspective. In this sense, it can be said, that the fact that most of the transformations presented in this chapter are integrated into commercial products, such as Yamaha's Vocaloid or Pinnacle's Studio10, is itself, in a way, a validation of the algorithms.

Chapter 5

Conclusions and future work

5.1 Author's contribution

This work is a consequence of over nine years of research in the topic of digital signal processing of the singing voice. Over all these years the author has been involved in the creation of new technologies and the implementation of new singing related applications. However, above all, the author credits the team he has worked with inside the Music Technology Group. Mostly all the contributions gathered in this work were born out of a team work and emerge from the constant interaction between researcher colleagues.

The author's main contributions relate to the following voice technologies: sinusoidal plus residual model, more concretely, the Spectral Modelling Synthesis (SMS), the Spectral Peak Processing (SPP), the Excitation plus Residual (EpR) model of the voice, and the Concatenative models. These are the technologies on top of which most new algorithms have been built. The SMS has been extensively used by the author for the analysis of the signal. SMS sinusoidal plus residual formulation allows a very interesting point of view for the study of voice signal properties. Basic contributions were done by the author in the SMS implementation for the improvement of pitch analysis and residual manipulation in pitch transposition. Spectral Peak Processing (SPP) was jointly developed with Jordi Bonada at the Music Technology Group (MTG) at the same time Laroche worked with the Rigid Phase Locked Vocoder (RPLV). At the time we found out about the RPLV, it was decided not to go public without offering improvements to this technique family. Our most significant new contributions were the strategy which decided the source peak harmonic region to use in synthesis and the maximally flat phase alignment preservation. EpR voice model was initially conceived for a project that addressed the singing voice synthesis. The scope of the model, also developed by the singing voice team in the MTG, was to get the main attributes of the voice at the time it could avoid the unsolved problem of formant estimation. For such purpose we created the concept of resonances, excitation envelope, and residual envelope, which was added to make the model lossless. EpR demonstrated to be a very robust and useful tool for modelling the voice. Based on real voice EpR analysis, patters were build for vibrato, note to note transitions, phonetic transitions, nasalization, and others expressive nuances. The author worked in applying afterwards those patterns to synthetic voices obtaining satisfactory results and in the automatic estimation of EpR's resonances. Also related with the singing synthesizer project, the author was involved in the creation of a new sound concatenation technology that could create a smooth synthesis out of the collage of real recording snippets taken from different contexts. In this topic, author's main contribution relate to the phase and timbre continuation algorithms.

On top of these technologies, the author has worked in many different descriptors and transformations, namely in the rough and growl analysis and transformation, the breathiness estimation and emulation, the pitch detection and modification, nasality identification, voice to melody conversion, beat onset detection for the solo singing voice, real time singing voice morphing, and voice to instrument transformation. Moreover, together with the co-writers of the chapter "*Spectral Processing*" from Zölzer (2004), the author contributed significantly to the matlab source coding included the chapter.

5.2 Conclusions and future work

Singing voice processing itself is a field which does not lay anymore in the scientific community but has mostly all migrated to the industry. According to the author opinion, future research guidelines related to the singing voice topic are not defined anymore inside the overviews of voice musical production but inside the world of human computer interaction, where the source material is not solo voice but music.

From this point of view, the technologies that related with the singing voice should be able to recognize, describe and understand lead voices and vocal harmonies inside a complex sound mix. Moreover, new technologies should also be able to mold and adapt the sung content to whatever new shape and support is required. The first part of this discourse is closely related with music data mining or music information retrieval (MIR); from song analysis, we should be able to recognize a singer, query songs by type of singing such as folk or enka, find songs from similar singers, etcetera. But the second part of the discourse does not belong to MIR but to technologies that will probably emerge from those presented in this work; this is, from song analysis, we should be able to remix the lead voice inside the song, change the melody or / and the lyrics of a song, or even, change the singer we want to perform that song.

Said so, the AudioScanner project presented in section 1.2.1.8 is a first step towards singing voice manipulation inside complex sound mixtures. It already allows, under certain constraints, to harmonize Norah Jones or isolate Mick Jagger lead vocals from their albums CD tracks. Moreover, source separation technologies can take current standards of signal analysis a bit further by discriminating different sound objects inside a song; i.e. the analysis of song tempo detection would be significantly improved when drum tracks could be isolated from the rest of non percussive instruments.

References

Abe, M. (1992). *A Study on Speaker Individuality Control*. PhD thesis, NTT, Human Interface Laboratories, Japan

Abrams, S., Oppenheim, D. V., Pazel, D. and Wright, J. (1999). Higher-level composition control in lusic sketcher: Modifiers and smart harmony. *Proceedings of the International Computer Music Conferences*, Beijing, China.

Amatriain, X., Bonada, J., Loscos, A., and Serra, X. (2001). Spectral Modeling for Higher-level Sound Transformation. *Proceedings of MOSART Workshop on Current Research Directions in Computer Music*, Barcelona, Spain.

Amatriain, X., Bonada, J., Loscos, A., and Serra, X. (2002). Spectral Processing. Udo Zölzer, editor, *DAFX: Digital Audio Effects*, page 554, John Wiley & Sons Publishers.

Amatriain, X., Bonada, J., Loscos, A., Arcos, J., and Verfaille, V. (2003). Content-based Transformations. *Journal of New Music Research*, volume 32.

Ananthapadmanabha T.V., Fant G. (1982). Calculation of true glottal flow and its components. *Speech Communication*, volume 1, pages 167-184.

Ananthapadmanabha, T. V. (1984). *Acoustic analysis of voice source dynamics*. Progress and Status Report, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden.

Arroabaren, I. (2004). *Un modelo matemático para el canto lírico*. PhD thesis, Universidad de Navarra, Spain.

Atal, B. S. and Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *Journal of the Acoustical Society of America*, volume 50, pages 637-655.

Atal, B. S. and Remde, J. R. (1982). A new model of LPC excitation for producing natural-sounding speech at low bit rates. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 614-617, Paris, France.

Bartsch, M. A. and Wakefield, G. H. (2004). Singing voice identification using spectral envelope estimation. *IEEE Transactions on Speech and Audio Processing*.

Bellandese, M., Lerman, J. and Gilbert, J. (2001). An Acoustic Analysis of Excellent Female Oesophageal, Tracheoesophageal and Laryngeal Speakers. *Journal of Speech, Language and Hearing Research*, volume 44, pages 1315-1320.

Bello, J. P., Daudet, L., Abdallah, S., Duxbury, C., Davies, M. and Sandler, M. B. (2005). A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing*.

Blaauw, M. (2005). *A Method of Adaptive Amplitude Envelope Estimation Optimized for Singing Voice*. Internal report, Music Techonlogy Group, Universitat Pompeu Fabra.

Boer, M., Bonada, J., Cano, P., Loscos, A. and Serra, X. (2000). Singing Voice Impersonator Application for PC. *Proceedings of International Computer Music Conference*, Berlin, Germany.

Bonada J. (2005). Voice Solo to Unison Choir Transformation. *Proceedings of the 118th American Engineering Society*, Barcelona, Spain.

Bonada, J. (2004). High Quality Voice Transformations Based On Modeling Radiated Voice Pulses In Frequency Domain. *Proceedings of 7th International Conference on Digital Audio Effects*; Naples, Italy.

Bonada, J. and Loscos, A. (2003). Sample-based singing voice synthesizer by spectral concatenation. *Proceedings of Stockholm Music Acoustics Conference*, Stockholm, Sweden.

Bonada, J., Celma, O., Loscos, A., Ortolà, J., and Serra, X. (2001). Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models. *Proceedings of International Computer Music Conference*, Havana, Cuba.

Bonada, J., Loscos, A., Cano, P., and Serra, X. (2001). Spectral Approach to the Modeling of the Singing Voice. *Proceedings of 111th AES Convention*, New York, USA.

Bonada, J., Loscos, A., Mayor, O., and Kenmochi, H. (2003). Sample-based singing voice synthesizer using spectral models and source-filter decomposition. *Proceedings of 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications*, Firenze, Italy.

Boyanov, B., Ivanov, T., Hadjitodorov, S. and Choolet, G. (1993). Robust Hybrid Pitch etector. *Electronic Letters*, volume 29.

Cairns, D. A., Hansen, J. H. L. and Kaiser, J. F. (1996). Recent advances in hypernasal speech detection using the nonlinear Teager energy operator. *Proceedings of the Fourth International Conference on Spoken Language*, volume: 2, pages 780-783.

Cairns, D. A., Hansen, J. H. L. and Riski, J. E. (1996). A noninvasive technique for detecting hypernasal speech using a nonlinear operator. *IEEE Transactions on Biomedical Engineering*, volume 43, pages 35

Cano, P. and Loscos, A. (1999). *Singing Voice Morphing System based on SMS*. Graduate Thesis. Polytechnic University of Catalonia, Spain.

Cano, P., Loscos, A., and Bonada, J. (1999). Score-Performance Matching using HMMs. *Proceedings of International Computer Music Conference*, Beijing, China

Cano, P., Loscos, A., Bonada, J., Boer, M. and Serra, X. (2000). Voice Morphing System for Impersonating in Karaoke Applications. *Proceedings of International Computer Music Conference*, Berlin, Germany

Cano, P., Loscos, A., Bonada, J., de Boer, M., and Serra, X. (2000). Voice Morphing System for Impersonating in Karaoke Applications. *Proceedings of International Computer Music Conference*, Berlin, Germany.

Carre. M. (2002). *Systémes de recherche de documents musicaux par chantonnement*. PhD thesis, Ecole Nationale Supérieure des Télécommunications, Paris, France.

Castallengo, M., Richard, G. and d'Alessandro, C. (1989). Study of vocal pitch vibrato perception using synthesis. *Proceedings of the 13th International Congress on Acoustics*. Yugoslavia.

Cervera, T., Miralles, J. L. and González, J. (2001) Acoustical Analysis of Spanish Vowels Produced by Laringectomized Subjects. *Journal of Speech, Language, and Hearing Research*, volume 44, pages 988-996.

Chan, Y. T. (1995). *Wavelet Basics*. Kluwer Academic Publications.

Charpentier, F. J. and Stella, M. G. (1986). Diphone synthesis using an overlap-add technique for speech waveforms concatenation. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

Cheveigné, A., Kawahra, K. (2002). YIN, a fundamental frequency estimator for speech and music. *Journal of the Acoustical Society of America*, volume 111, pages 1917-1930.

Childers, D. G. (1990). Speech Processing and Synthesis for Assessing Vocal Disorders. *IEEE Engineering in Medicine and Biology Magazine*, volume 9, pages 69-71.

Childers, D. G. (1994). Measuring and Modeling Vocal Source-Tract Interaction. *IEEE Transactions on Biomedical Engineering*.

Childers, D. G. and Hu, H. T. (1994). Speech synthesis by glottal excited linear prediction. *Journal of the Acoustical Society of America*, volume 96, pages 2026-2036.

Concatenation-based midi-to-singing voice synthesis. *Proceedings of Audio Engineering Society 103rd Convention*, New York, USA.

Cook, P. (1990). *Identification of control parameters in an articulatory vocal tract model with applications to the synthesis of singing*. PhD thesis, Stanford University, CCRMA, USA.

Cook, P. (1992). SPASM: a Real-Time Vocal Tract Physical Model Editor/Controller and Singer: the Companion Software Synthesis System. *Computer Music Journal*, volume 17, pages 30-44.

Cook, P. R. and Leider C. (2000). Squeeze Vox: ANew Controller for Vocal Synthesis Models. *Proceedings of International Computer Music Conference*, Berlin, Germany.

Cuadra, P., Master, A. and Sapp C. (2001). Efficient Pitch Detection Techniques for Interactive Music. *Proceedings of the International Computer Music Conference*, Havana, Cuba

Diedrich, W. M. and Youngstrom, K. A. (1996). *Alaryngeal Speech*. Springfield, Charles C. Thomas.

Ding, Y. and Qian X. (1997). Processing of musical tones using a combined quadratic polynomial-phase sinusoid and residual (quasar) signal model. *Journal of the Audio Engineering Society*.

Doi, T., Nakamura, S., Lu, J., Shikano, K. (1996). Improvement in esophageal speech by replacing excitation components in cepstrum domain. *Proceedings of the Acoustical Society of Japan*, pages 253-254.

Dubnov, S. (2006). YASAS - Yet Another Sound Analysis - Synthesis Method. *Proceedings of the International Computer Music Conference*, New Orleans, USA.

Dudley, H. (1936). *Synthesizing speech*. Bell Laboratories Record.

Dudley, H. (1939). Remaking speech. *Journal of the Acoustical Society of America*, volume 11, pages 169-177.

Dutilleux, P., De Poli, G. and Zölzer, U. (2002). *Time-segment Processing*. John Wiley & Sons.

Dutoit, T. and Leich H. (1993). MBR-PSOLA : Text-toSpeech Synthesis Based on an MBE Resynthesis of the Segments Database. *Speech Communication*, volume 13, pages 435-440.

Duxbury, C. Bello, J. P., Davies, M., and Sandler, M. (2003). A combined phase and amplitude based approach to onset detection for audio segmentation. *Proceedings of the 4th European Workshop on Image Analysis for Multimedia Interactive Services*, London, UK.

Ekman, L. A., Kleijn, W., Murthi, M. N. (2006). Spectral Envelope Estimation and Regularization. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Fabig, L. and Janer, J. (2004). Transforming Singing Voice Expression - The Sweetness Effect. *Proceedings of 7th International Conference on Digital Audio Effects*; Naples, Italy.

Fant, G. (1960). *Acoustic Theory of Speech Production*. The Hague: Mouton.

Fant, G. (1982). *Preliminaries to analysis of the human voice.* Progress and Status Report, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden.

Fant, G., Liljencrants, J., and Lin, Q. (1985). *A four parameter model of vocal flow*. Progress and Status Report, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden.

Fitz, K., and Haken, L. (1996). Sinusoidal modeling and manipulation using lemur. *Computer Music Journal*, volume 20, pages 44-59.

Fitz, K., Haken, L. and Christensen P. (2000). A New Algorithm for Bandwidth Association in Bandwidth-Enhanced Additive Sound Modeling. *Proceedings of the International Computer Music Conferences*, Berlin, Germany.

Flanagan, J. (1972). *Speech Analysis, Synthesis*, and Perception, Springer-Verlag, Berlin-Heidelberg-New York.

Flanagan, J. L. and Golden, R. M. (1966). Phase Vocoder, *Bell System Technical Journal*, pages 1493-1509.

Fletcher, H. and Munson, W. A. (1933). Loudness, its definition,mesurement and calculation. *Journal of the Acoustical Society of America*, volume 5, pages 82-108.

Fonollosa, J. A. (1996). A Comparison of Several Recent Methods of Fundamental Frequency and Voicing Decision Estimation. *Proceedings of the 4th International Conference on Spoken Language Processing.*

Fujimura, O. (1962). Analysis of nasal consonants. *The Journal Of The Acoustical Society Of America*, volume 34, pages 1865-1875.

George, E. B. and Smith, M. J. (1992). An analysis by synthesis approach to sinusoidal modeling applied to the analysis and synthesis of musical tones. *Journal of the Audio Engineering Society*, volume 40, pages 497-516.

Gerhard, D. (2003). *Computationally measurable differences between speech and song*. PhD Thesis, Simon Fraser University, Department of Computer Science, Vancouver, Canada.

Gerhard, D. (2003). *Pitch Extraction and Fundamental Frequency: History and Current Techniques*. Technical Report TR-CS 2003-06.

Gold, B. and Rader, C. M. (1967). The channel vocoder. *IEEE Transactions on Audio and Electroacoustics*, volume 15, pages 148–161.

Gómez, E., Klapuri, A. and Meudic, B. (2003). Melody Description and Extraction in the Context of Music Content Processing. *Journal of New Music Research*, volume 32.

Goodwin, M. M. (1997). *Adaptive Signal Models: Theory, Algorithms, and Audio Applications*. PhD thesis, University of California, Berkeley, USA.

Griffith, D. W. and Lim, J. S. (1988). Multiband excitation vocoder. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, volume 36, pages 236-243.

Haeb-Umbach, R., Geller, D. and Ney, H. Improvements in connected digit recognition using linear discriminant analysis and mixture densities. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

Hämäläinen, P., Mäki-Patola, T., Pulkki, V. and Airas, M. (2004). Musical Computer Games Played by Singing. *Proceedings of the 7th International Conference on Digital Audio Effects*, Naples, Italy.

Hanson, H. M. (1995). *Glottal Characteristics of Female Speakers*. PhD thesis, Harward University, Cambridge, USA.

Henrich, N. (2001). *Etude de la source glottique en voix parlée et chantée : modélisation et estimation, mesures acoustiques et électroglottographiques, perception*. PhD thesis, University of Paris, France.

Herrera, P. and Bonada, J. (1998). Vibrato Extraction and Parameterization in the Spectral Modeling Synthesis Framework. *Proceedings of the Digital Audio Effects Workshop*, Barcelona, Spain.

Hess, W. (1992). *Pitch and voicing determination. Advances in speech signal processing*. M. M. Sondhi and S. Furui, Marcel Dekker, editor, New York.

Horii, Y. (1989). Acoustic analysis of vocal vibrato: A theoretical interpretation of data. *Journal of Voice*, volume 3, pages 36-43.

Horii, Y. (1989). Frequency modulation characteristics of sustained /a/ sung in vocal vibrato. *Journal of Speech and Hearing Research*, volume 32, pages 829-836.

Hu, N. and Dannenberg, R. B. (2002). A Comparison of Melodic Database Retrieval Techniques using Sung Queries. *Proceedings of the Joint Conference on Digital Libraries*, pages 301-307, Oregon, USA.

Hu, W. T. and Wu H. T. (2000). A Glottal-Excited Linear Prediction (GELP) Model for Low-Bit-Rate Speech Coding. *Proceedings of the National Science Council*, volume 24, pages 134-142, China.

Jeon, C. and Driessen, P.F. (2005). Intelligent Artificial Vocal Vibrato Effecter Using Pitch Detection and Delay-Line. *Proceedings of the Pro Audio Expo and Convention*, Barcelona, Spain.

Jinachitra, P., Smith, J. O. (2005). Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm. *Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, New Paltz, New York, USA.

Kahlin, D. and Ternström, S. (1999). The chorus effect revisited: Experiments in frequency domain analysis and simulation of ensemble sounds. *Proceedings of IEEE Euromicro*, Milan, Italy.

Kaiser, J. F. (1990). On a simple algorithm to calculate the `energy' of a signal. *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

Kawahara, H. (2003). Exemplar-based Voice Quality Analysis and Control using a High Quality Auditory Morphing Procedure based on STRAIGHT. Voice Quality: Functions, Analysis and Synthesis, *Proceedings of the ISCA Tutorial and Research Workshop*, Geneva, Switzerland

Kelly, J. L. and Lochbaum, C. C. (1962). Speech synthesis. *Proceedings of the 4th International Congress on Acoustics*, pages 1-4.

Kim, H. L., Kim, D. H., Ryu, Y. S., Kim, Y. K. (1996). A study on pitch detection using the local peak and valley for Korean speech recognition. *Proceedings of the IEEE TENCON.*

Kim, Y. E. (2003). *Singing Voice Analysis / Synthesis*. PhD thesis, Massachusetts Institute of Technology, USA.

Klapuri, A. (1999). Sound onset detection by applying psychoacoustic knowledge. *Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, Arizona, USA.

Klatt, D. (1980). Software for a Cascade/Parallel Formant Synthesizer. *Journal of the Acoustical Society of America*, volume 67, pages 971-995.

Klatt, D. H. and Klatt, L. C. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of Acoustical Society of America*, volume 87, pages 820-857.

Klatt, D. H., and Klatt, L. C. (1990). Analysis, synthesis and perception of voice quality variations among female and male talkers. *Journal of Acoustical Society of America*, volume 87, pages 820-857.

Kleijn, W. B. (2000). Representing Speech. *Proceedings of the European Signal Processing Conference.*

Kleijn, W. B. (2003). Signal Processing Representations of Speech. *Journal of the ACM Transactions on Information and Systems.*

Kob, M. (2002). *Physical modeling of the singing voice*. PhD thesis, Instute für Technische Akustik, RWTH, Logos Verlag Berlin, Germany.

Koestoer, N. and Paliwal, K.K. (2001). Robust spectrum analysis for applications in speech processing. *Proceedings of the Microelectronic Engineering Research Conference*, Brisbane, Australia.

Kroon, K., Deprettere, E. and Sluyter, R. (1986). Regular-pulse excitation: A novel approach to effective and efficient multi-pulse coding of speech. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 34, pages 1054-1063.

Laroche, J. and Dolson, M. (1999). Improved phase vocoder time-scale modification of audio. *IEEE Transactions on Acoustics, Speech, and Signal Processing.*

Laroche, J. and Dolson, M. (1999). New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects. *Proceedings of IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

Larsson, B. (1977). *Music and singing synthesis equipment (MUSSE)*. Quarterly Progress and Status Report, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden.

Lau, E., Ding, A. and Calvin J. (2005). *MusicDB: A Query by Humming System*. Final Project Report, Massachusetts Institute of Technology, USA.

Lee, M. E. (2005). *Acoustic Models for the Analysis and Synthesis of the Singing Voice*. PhD thesis, School of Electrical and Computer Engineering, Georgia Institute of Technology, USA.

Levine, S., Smith III, J. O. (1998). A Sines+Transients+Noise Audio Representation for Data Compression and Time/PitchScale Modifications. *Proceedings of the 105th Audio Engineering Society Convention*, San Francisco, USA.

Leydon, C., Bauer, J. J. and Larson, C. R. (2003). The role of auditory feedback in sustaining vocal vibrato. *Journal of Acoustical Society of America*, volume 114, pages 1571-1581.

Lieberman, P and Blumstein, S.(1988). *Speech physiology, speech perception, and acoustic phonetics*. Cambridge University Press.

Lieberman, P. and Blumestein S. (1988). *Speech Physiology, speech perception and acoustics phonetics*. Cambridge University Press.

Linde Y., Buzo A. and Gray R. M. (1980). An Algorithm for vector quantizer design. *IEEE Transactions on Communication*, volume 28, pages 84-95.

Lomax, K. (1997). *The Analysis and Synthesis of the Singing Voice*. PhD thesis, Oxford University, Engalnd.

Loscos, A. and Aussenac, T. (2005). The Wahwactor: a voice controlled wah-wah pedal. *Proceedings of International Conference on New Interfaces for Musical Expression*, Vancouver, Canada

Loscos, A. and Resina, E. (1998). SMSPerformer: A real-time synthesis interface for SMS. *Proceedings of COST G6 Conference on Digital Audio Effects*, Barcelona, Spain.

Loscos, A., Cano, P. and Bonada, J. (1999). Singing Voice Alignment to text. *Proceedings of International Computer Music Conference*, Beijing, China.

Loscos, A., Cano, P., and Bonada, J. (2002). Low-Delay Singing Voice Alignment to Text. *Proceedings of International Computer Music Conference*, Beijing, China.

Lu, H. L. (1999). *A Hybrid Fundamental Frequency Estimator for Singing Voice*. Internal report, CCRMA, Standford, 1999

Lu, H. L. (2002). *Toward a high-quality singing synthesizer with vocal texture control.* PhD thesis, Stanford University, CCRMA, USA.

Lyons, M. J., Haehnel M. and Tetsutani, N. (2001). The Mouthesizer: A Facial Gesture Musical Interface. *Conference Abstracts*, Siggraph 2001, Los Angeles, page 230.

Macon, M. W., Jensen-Link, L., Oliverio, J., Clements, M. and George, E. B. (1997). Concatenation-based MIDI-to-Singing Voice Synthesis, *Proceedings of Audio Engineering Society 103 International Conference*, New York, USA.

Maher R., Beauchamp J. (1994). Fundamental frequency estimation of musical signals using a Two-Way Mismatch procedure. *Journal of Acoustical Society of America*, volume 95, pages 2254-2263.

Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE Acoustics, Speech, and Signal Processing*, volume 63, pages 1973-1986.

Makhoul, J., Roucos, S. and Gish, H. (1985). Vector quantization in speech coding. *Proceedings of the IEEE*, volume 73, pages 1551-1588.

Maragos, P., Kaiser, J. F. and Quatieri T. F. (1993). On Amplitude and Frequency Demodulations using Energy Operators. *IEEE Transactions of Signal Processing*, volume 41, pages 1532-1550.

Maragos, P., Kaiser, J. F. and Quatieri, T. F. (1993). On amplitude and frequency demodulation using energy operators. *IEEE Transactions on Signal Processing*, volume 41, pages 1532-1550.

Maragos, P., Quatieri, TH. F. and Kaiser, J.F. (1991). Speech Nonlinearities, Modulations, and Energy Operators. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada.

Masri, P. (1996). *Computer Modeling of Sound for Transformation and Synthesis of Musical Signal*. PhD thesis, University of Bristol, Bristol, U.K.

Masri, P. and Batterman, A. (1996). Improved model of attack transients in music analysis-resynthesis. *Proceedings of the International Computer Music Conferences*, Hong Kong, China.

McAulay, R. J. and Quatieri, T. F. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Transactions on Acoustics, Speech, and Signal* Processing.

McNab R. J., Smith L. A., Witten I. H. and Henderson C. L. (2000). Tune Retrieval in the Multimedia Library. *Multimedia Tools and Applications*, pages 113-132.

McNab, R. J., Smith, L. A. and Witten, I. A. (1996). *Signal Processing for Melody Transcriptio*n. Senior Income Guarantee working paper, volume 95.

Mellody, M. (2001). *Signal Analysis of the Female Singing Voice: Features for Perceptual Singer Identity*. PhD thesis, University of Michigan, USA.

Meron, Y. (1999). *High Quality Singing Synthesis using the Selection-based Synthesis*. PhD thesis, University of Tokyo.

Miller, N. J. (1973). *Filtering of Singing Voice Signal from Noise by Synthesis*. PhD thesis, University of Utah, USA.

Miller, N. J. (1975). Pitch detection by data reduction. *IEEE Transactions on Acoustics, Speech and Signal Processing, Special Issue on IEEE Symposium on Speech Recognition*, volume 23, pages 72-79.

Moorer, J. A. (1978). The use of the phase vocoder in computer music applications. *Journal of the Audio Engineering Society*, volume 26, pages 42-45.

Moulines, E. and Charpentier, F. (1990). Pitch synchronous waveform processing techniques for text to speech synthesis using diphones. *Speech Communication*, volume 9, pages 453-467.

Moulines, E., Charpentier, F., and Hamon, C. (1989). A diphone synthesis system based on time-domain prosodic modifications of speech. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.*

Oppenheim, A. V. and Schafer, R. W. (1989). *Discrete-Time Signal Processing*. Prentice-Hall, Englewood Cliffs, NJ.

Osaka, N. (1995). Timbre Interpolation of sounds using a sinusoidal model. *Proceedings of International Computer Music Conference*, Banff, Canada.

Pachet, F. and Roy, P. (1998). Reifying chords in automatic harmonization. *Proceedings of the Workshop on Constraints for Artistic Applications.*

Peeters, G. (2004). *A large set of audio features for sound description (similarity and classification) in the CUIDADO project.* CUIDADO I.S.T. Project Report.

Peeters, G. and Rodet, X. (1999). Sinola: A new analysis/synthesis method using spectrum peaks shape distortion, phase and reassigned spectrum. *Proceedings of the International Computer Music Conferences*, Beijing, China.

Pfeiffer, S. (1999). The Importance of Perceptive Adaptation of Sound Features in Audio Content Processing. *SPIE Storage and Retrieval for Image and Video Databases VII*, pages 328-337, San Jose, California, USA.

Prame, E. (2000). *Vibrato and intonation in classical singing*. PhD thesis, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden.

Puckette, M. S. (1995). Phase-locked vocoder. *Proceedings of IEEE Conference on Applications of Signal Processing to Audio and Acoustics*.

Rabiner, L. (1977). On the use of Autocorrelation Analysis for Pitch Detection. IEEE *Transactions on Acoustics, Speech, and Signal Processing*, volume 25.

Rabiner, L. and Juang, B. H. (1993). *Fundamentals of Speech Recognition*. Prentice Hall.

Rabiner, L. R. and Schafer, R.W. (1978). *Digital Processing of Speech Signals*. Prentice-Hall, Englewood Cliffs, NJ.

Rabiner, L., Cheng, M., Rosenberg, A., McGonegal, C. (1976). A comparative Performance Study of Several Pitch Detection Algorithms. *IEEE Transactions on Acoustics, Speech, And Signal Processing*, volume 24.

Radova, V., Silhan, O. (2002). Method for the Segmentation of Voiced Speech Signals into Pitch Period. *Proceedings of the 5th Nordic Signal Processing Symposium*, Norway.

Richard, G., d'Alessandro, C. and Grau, S. (1992). Unvoiced speech synthesis using poissonian random formant wave functions. *Signal Processing VI: European Signal Processing Conference*, pages 347–350.

Roads, C. (1978). Granular Synthesis of Sound. *Computer Music Journal*, volume 2, pages 61-62.

Robbins, J., Fisher, H., Blom, E. and Singer, M. A comparative acoustic study of normal, oesophageal and tracheoesophageal speech production. *Journal of Speech and Hearing Disorders*, volume 49, pages 202-210.

Robel, A. (2005). Onset detection in polyphonic signals by means of transient peak classification. *MIREX Online Proceedings, International Symposium on Music Information Retrieval*, London, U.K.

Röbel, A. and Rodet, X. (2005). Efficient Spectral Envelope Estimation and its application to pitch shifting and envelope preservation. *Proceedings of the 8th International Conference on Digital Audio Effects*, Madrid, Spain.

Rodet, X. (1984). Time-domain formant-wave-function synthesis. *Computer Music Journal*, volume 8, pages 9-14.

Rodet, X. (1984). Time-domain formant-wave-function synthesis. *Computer Music Journal*, 8(3):9–14, 1984.

Rodet, X. (1997). Musical Sound Signal Analysis/Synthesis Sinusoidal+Residual and Elementary Waveform Models. *Proceedings of the IEEE Time-Frequency and Time Scale Workshop*, Coventry, Grande Bretagne.

Rodet, X. and Bennett, G. (1989). Synthesis of the Singing Voice. M. Mathews and J. Pierce, editors, *Current Directions in Computer Music Research*, Cambridge, MIT Press, pages 19-44.

Rodet, X. and Depalle, P. (1992). A new additive synthesis method using inverse Fourier transform and spectral envelopes. . *Proceedings of International Computer Music Conference*, San Jose, California, USA.

Rodet, X. and Jaillet, F. (2001). Detection and modeling of fast attack transients. *Proceedings of the International Computer Music Conferences*, Havana, Cuba.

Rodet, X., Potard, Y. and Barrière, J. B. (1985). *CHANT - de la synthèse de la voix chantée à la synthèse en général*. Report, Institut de Recherche et Coordination Acoustique/Musique, Paris, France.

Rodet, X., Potard, Y. and Barriere, J.-B. (1984). The CHANT project: From the synthesis of the singing voice to synthesis in general. *Computer Music Journal*, volume 8, pages 15-31.

Rosenberg, A. (1971). Effect of glottal pulse shape on the quality of natural vowels. *Journal of the Acoustical Society of America*, volume 49, pages 583–590.

Rosenberg, S. (1970). Glottal pulse shape and vowel quality. *Journal of the Acoustical Society of America*, volume 49, pages 583-590.

Ross, M., Shaffer, H., Cohen, A., Freudberg, R. and Manley, H. (1974). Average Magnitude Difference Function Pitch Extractor. *IEEE Transactions on Acoustics, Speech, And Signal Processing*, volume 22.

Rothenberg, M. (1981). Some relations between glottal air flow and vocal fold contact area. *American Speech Language Hearing Association Reports*, volume 11, pages 88-96.

Roucos, S. and Wilgus, A. M. (1986). High Quality Time-Scale Modification for Speech. *Proceedings of the IEEE Proceedings on Acoustics, Speech, and Signal Processing*, Tokyo, Japan, Sakakibara, K. I., Fuks, L., Imagawa H. and Tayama, N. (2004). Growl voice in ethnic and pop styles. *Proceedings of the International Symposium on Musical Acoustics*, Nara, Japan.

Sawada, H. and Nakamura, M. (2004). Mechanical Voice System and its Singing Performance. *Proceedings of the IEEE International Conference on Intelligent Robots and Systems*, pages 1920-1925.

Sawada, H., Takeuchi, N. and Hisada, A. (2004). A Real-time Clarification Filter of a Dysphonic Speech and Its Evaluation by Listening Experiments, *International. Proceedings of the International Conference on Disability, Virtual Reality and Associated Technologies*, pages 239-246.

Scheirer, E. D. and Kim, Y. E. (1999). Generalized audio coding with MPEG-4 structured audio. *Proceedings of Audio Engineering Society 17th International Conference*, Florence, Italy.

Schnell, N., Peeters, G., Lemouton, S., Manoury, P. and Rodet, X. (2002). Synthesizing a choir in real-time using Pitch Synchronous Overlap Add (PSOLA). *Proceedings of the IEEE 1st Benelux Workshop on Model based Processing and Coding of Audio*.

Schoentgen, J. (2001). Stochastic models of jitter. *Journal of the Acoustical Society of America*, volume 109, pages 1631-1650.

Schroeder, M. and Atal, B. (1985). Code excited linear prediction (CELP): high quality speech at low bit rate. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages. 937-940, Tampa, FL, USA

Schroeder, M. R. (1968). Period histogram and product spectrum: New methods for fundamental frequency detection. *Journal of the Acoustical Society of America*, volume 43, pages 829-834.

Schroeder, M. R. (1999). *Computer Speech: Recognition, Compression, and Synthesis*. Springer Verlag, Berlin, Germany.

Schroeder, M. R. and Atal, B. S. (1984). Code-excited linear prediction (CELP): High-quality speech at very low bit rates. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 937-940.

Schwarz, D. (1998). *Spectral Envelopes in Sound Analysis and Synthesis*. PhD thesis, Institute of Computer Science of the University of Stuttgart, Germany.

Schwarz, D. and Rodet, X. (1999) Spectral Envelope Estimation and Representation for Sound Analysis-Synthesis. *Proceedings of the International Computer Music Conference*, Beijing, China.

Serra, X. (1994). Sound hybridization techniques based on a deterministic plus stochastic decomposition model. *Proceedings of International Computer Music Conference*, San Francisco, USA.

Serra, X. (1997). *Sound Modelling with Sinusoids plus Noise*. Swets & Zeitlinger Publishers.

Serra, X. and Bonada, J. (1998). Sound Transformations Based on theSMS High Level Attributes. *Proceedings of the Conference on Digital Audio Effects*, Barcelona, Spain.

Serra, X. and Smith, J. O. (1990). Spectral modeling synthesis: A sound analysis/synthesis system based on a deterministic plus stochastic decomposition. *Computer Music Journal*, volume 14, pages 12–24.

Settel, Z. and Lippe, C. (1996). Real-Time Audio Morphing. *Proceedings of the 7th International Symposium on Electronic Art*.

Settle, J. and Lippe, C. (1994). Real-time musical applications using the fft-based resynthesis," *Proceedings of the International Computer Music Conference*, Aarhus, Denmark.

Shipp, T., Doherty, T. and Haglund, S. (1990). Physiologic factors in vocal vibrato production. *Journal of Voice*, volume 4, pages 300-304.

Shrivastav, R. (2003). The use of an auditory model in predicting perceptual ratings of breathy voice quality. *Journal of Voice*, volume 17, pages 502-512.

Shrivastav, R. and Sapienza, C. M. (2003). Objective measures of breathy voice quality obtained using an auditory model. *Journal of the Acoustical Society of America*, volume 114, pages 2217-2224.

Singhal, S. and Atal, B. S. (1989). Amplitude optimization and pitch prediction in multipulse coders. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, volume 37, pages 317-327.

Slaney, M. (1998). *Audiotory Toolbox: A Matlab toolbox for Auditory Modeling Work, version 2*. Technical Report, Interval Research Corporation.

Slaney, M., Covell, M., Lassiter, B. (1996). Automatic audio morphing. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1001-1004.

Solà, J.M. (1997). *Disseny i Implementació d'un Sintetitzadorde Piano*. Graduate Thesis. Polytechnic University of Catalonia, Spain.

Sood S. and Krishnamurthy, A. (2004). A robust on-the-fly pitch (OTFP) estimation algorithm. *Proceedings of the Annual ACM international Conference on Multimedia*, pages 280-283

Sood, S., Krishnamurthy, A. (1992). A robust on-the-fly pitch (OTFP) estimation algorithm. *Proceedings of the 12th Annual ACM international Conference on Multimedia*, New York.

Sukkar, R. A., LoCicero, J. L., Picone, J. W. (1988). Design and implementation of a robust pitch detector based on a parallel processing technique. *Proceedings of the IEEE Journal on Selected Areas in Communications*, volume 6, pages 441-451.

Sundberg, J. (1973). The source spectrum in professional singing. *Folia Phoniatrica*, volume 25, pages 71-90.

Sundberg, J. (1978). Effects of the vibrato and the 'singing formant' on pitch. *Journal of Research in Singing*, volume 5, pages 5-17.

Sundberg, J. (1979). Maximum speed of pitch changes in singers and untrained subjects. *Journal of Phonetics*, volume 7, pages 71-79

Sundberg, J. (1987). *The Science of Singing Voice*. Illinois Universitary Press.

Teager, H. M. and Teager, S. M. (1989). Evidence for Nonlinear Sound Production Mechanisms in the Vocal Tract. *Speech Production and Speech Modelling*, W. J. Hardcastle and A. Marchal, editors, France.

Tellman, E., Haken, L. and Holloway, B. (1995). Timbre Morphing of Sounds with Unequal Number of Features. *Journal of Audio Engineering Society.*

Terez, D. (2002) Robust pitch determination using nonlinear state-space embedding. *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, volume 1, pages 345-348.

Ternström, S. (1989). *Acoustical Aspects of Choir Singing*. PhD thesis, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden.

Thibault, T. (2004). *High-level Control of Singing Voice Timbre Transformations, Sound Processing and Control*. Master's thesis, McGill University, Faculty of Music, Montreal, Canada.

Thomasson, M. (2003). *From Air to Aria. Relevance of Respiratory Behaviour to Voice Function in Classical Western Vocal Art*. PhD thesis, Royal Institute of Technology, Department of Speech, Music & Hearing, Stockholm, Sweden.

Titze, I. R. (1988). The physics of small-amplitude oscillation of the vocal folds. *Journal of the Acoustical Society of America*, volume 83, pages 1536-1552.

Titze, I. R. (1994). Summary Statement. *Proceedings of the Workshop on Acoustic Voice Analysis, Summary Statement*. National Center for Voice and Speech, Denver, Colorado.

Tokumaru, M., Yamashita, K., Muranaka, N. and Imanishi, S. (1998). Membership functions in automatic harmonization system. *Proceedings of the IEEE 28th International Symposium on Multiple-Valued Logic*, pages 350-355

Uneson, M. (2003). *Burcas - a simple concatenation-based midi-to-singing voice synthesis system for Swedish*. PhD thesis, Lund University, Sweeden.

Vakman, D. (1996). On the analytic signal, the Teager-Kaiser energy algorithm, and other methods for defining amplitude and frequency. *IEEE Transactions on Signal Processing*, volume 44, pages 791-797.

Valimaki, V. and Karjalainen, M. (1994). Improving the Kelly-Lochbaum vocal tract model using conical tube sections and fractional delay filtering techniques. *Proceedings of the International Conference on Spoken Language Processing*.

Vercoe, B. L., Gardner W. G. and E. D. Scheirer. (1998). Structured audio: Creation, transmission, and rendering of parametric sound representations. *Proceedings of the IEEE*, volume 86, pages 922–940.

Verfaille, V., Zölzer, U. and Arfib, D. (2005). Adaptive Digital Audio Effects (A-DAFx):A New Class Of Sound Transformations. *IEEE Transactions on Acoustics, Speech, and Signal Processing*.

Verhelst, W. and Roelands, M. (1993). An overlap-add technique based on waveform similiarity (WSOLA) for high-quality time-scale modifications of speech. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*.

Verma, T. and Meg, T. H. (1998). Time scale modifications using a sines + transients + noise signal model. *Proceedings of the Digital Audio Effects Workshop*, November 1998.

Verma, T. and Meng, T. (1998). An analysis/synthesis tool for transient signals that allows a flexible sines+transients+noise model for audio. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

Verma, T. and Meng., T. (2000). Extending spectral modeling synthesis with transient modeling synthesis. *Computer Music Journal*, volume 24, pages 47-59.

Verma, T., Levine, S. and Meng, T. (1997). Transient modeling synthesis: a flexible transient analysis/synthesis tool for transient signals. *Proceedings of the International Computer Music Conferences*, Thessoloniki, Greece.

Viitaniemi, T. (2003). *Probabilistic models for the transcription of single-voice melodies*. Master's thesis, Tampere University of Technology, Finland.

Viitaniemi, T., Klapuri A. and Eronen, A. (2003). *A probabilistic model for the transcription of single-voice melodies*. Finnish Signal Processing Symposium, Tampere University of Technology.

Vinyes, M. Bonada, J., and Loscos, A. (2006). Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking. *Proceedings of Audio Engineering Society 120th Convention*, Paris, France

Waibel, A. and Lee, K. F. (1993). *Readings in Speech Recognition*. Morgan Kaufmann.

Wayman, J. L., Reinke, R. E. and Wilson, D. L. (1989). High quality speech expansion, compression, and noise filtering using the SOLA method of time scale modification. *Proceedings of the 23d Asilomar Conference on Signals, Systems and Computers*, volume 2, pages 714-717.

Yang, C. (2001). *Music Database Retrieval Based on Spectral Similarity*. Stanford University Database Group Technical Report.

Zhang, T. (2003). Automatic singer identification. *Proceedings of the IEEE Conference on Multimedia and Expo*.

Zhu, Y. and Shasha, D. (2003). Query by humming: a time series database approach. *Proceedings of the International Conference on Management of Data / Principles of Database Systems*.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge, Massachusetts.

Zolzer, U. (2002). *DAFX - Digital Audio Effects*. Wiley, John & Sons.

Annex A

Related publications by the author

A.1 Journal articles

Amatriain, X. Bonada, J. Loscos, A. Arcos, J. Verfaille, V. 2003.
'Content-based Transformations'
Journal of New Music Research Vol.32 .1

Abstract: Content processing is a vast and growing field that integrates different approaches borrowed from the signal processing, information retrieval and machine learning disciplines. In this article we deal with a particular type of content processing: the so-called content-based transformations. We will not focus on any particular application but rather try to give an overview of different techniques and conceptual implications. We first describe the transformation process itself, including the main model schemes that are commonly used, which lead to the establishment of the formal basis for a definition of content-based transformations. Then we take a quick look at a general spectral based analysis/synthesis approach to process audio signals and how to extract features that can be used in the content-based transformation context. Using this analysis/synthesis approach we give some examples on how content-based transformations can be applied to modify the basic perceptual axis of a sound and how we can even combine different basic effects in order to perform more meaningful transformations. We finish by going a step further in the abstraction ladder and present transformations that are related to musical (and thus symbolic) properties rather than to those of the sound or the signal itself.

A.2 Book chapters

Amatriain, X. Bonada, J. Loscos, A. Serra, X. 2002.
'Spectral Processing'
Udo Zölzer Ed., DAFX: Digital Audio Effects, p.554 John Wiley & Sons Publishers.

Description: Digital Audio Effects (DAFX) is the name chosen for the European Research Project COST G6. DAFX investigates the use of digital signal processing, its application to sounds, and its musical use designed to put effects on a sound. The aim of the project and this book is to present the main fields of digital audio effects. It systematically introduces the reader to digital signal processing concepts as well as software implementations using MATLAB. Highly acclaimed contributors analyze the latest findings and developments in filters, delays, modulators, and time-frequency processing of sound. Features include chapters on time-domain, non-linear, time-segment, time-frequency, source-filter, spectral, bit stream signal processing; spatial effects, time and frequency warping and control of DAFX. Also include MATLAB implementations throughout the book illustrate essential DSP algorithms for sound processing, and accompanying website with sound examples available. The approach of applying digital signal processing to sound will appeal to sound engineers as well as to researchers and engineers in the field of signal processing.

A.3 Conferences

Loscos, A. Resina, E. 1998.
'SMSPerformer: A real-time synthesis interface for SMS'
Proceedings of COST G6 Conference on Digital Audio Effects 1998. Barcelona

Abstract: SmsPerformer is a graphical interface for the real-time SMS synthesis engine. The application works from analyzed sounds and it has been designed to be used both as a composition and a performance tool. The program includes programmable time-varying transformations, MIDI control for the synthesis parameters, and performance loading and saving options.

Loscos, A. Cano, P. Bonada, J. 1999.
'Low-Delay Singing Voice Alignment to Text'
Proceedings of International Computer Music Conference 1999. Beijing, China

Abstract: In this paper we present some ideas and preliminary results on how to move phoneme recognition techniques from speech to the singing voice to solve the low-delay alignment problem. The work focus mainly on searching the most appropriate Hidden Markov Model (HMM) architecture and suitable input features for the singing voice, and reducing the delay of the phonetic aligner without reducing its accuracy.

Cano, P. Loscos, A. Bonada, J. 1999.
'Score-Performance Matching using HMMs'
Proceedings of International Computer Music Conference 1999. Beijing, China

Abstract: In this paper we will describe an implementation of a score-performance matching, capable of score following, based on a stochastic approach using Hidden Markov Models.

Cano, P. Loscos, A. Bonada, J. de Boer, M. Serra, X. 2000.
'Voice Morphing System for Impersonating in Karaoke Applications'
Proceedings of International Computer Music Conference 2000. Berlin, Germany

Abstract: In this paper we present a real-time system for morphing two voices in the context of a karaoke application. As the user sings a pre-established song, his pitch, timbre, vibrato and articulation can be modified to resemble those of a pre-recorded and pre-analyzed recording of the same melody sang by another person. The underlying analysis/synthesis technique is based on SMS, to which many changes have been done to better adapt it to the singing voice and the real-time constrains of the system. Also a recognition and alignment module has been added for the needed synchronization of the user's voice with the target's voice before the morph is done. There is room for improvements in every single module of the system, but the techniques presented have proved to be valid and capable of musically useful results.

de Boer, M. Bonada, J. Cano, P. Loscos, A. Serra, X. 2000.
'Singing Voice Impersonator Application for PC'
Proceedings of International Computer Music Conference 2000. Berlin, Germany

Abstract: This paper presents the implementation aspects of a real-time system for morphing two voices in the context of a karaoke application. It describes the software design and implementation under a PC platform and it discusses platform specific issues to attain the minimum system delay.

Bonada, J. Loscos, A. Cano, P. Serra, X. 2001.
'Spectral Approach to the Modeling of the Singing Voice'
Proceedings of 111th AES Convention. New York, USA

Abstract: In this paper we will present an adaptation of the SMS (Spectral Modeling Synthesis) model for the case of the singing voice. SMS is a synthesis by analysis technique based on the decomposition of the sound into sinusoidal and residual components from which high-level spectral features can be extracted. We will detail how the original SMS model has been expanded due to the requirements of an impersonating applications and a voice synthesizer. The impersonating application can be described as a real-time system for morphing two voices in the context of a karaoke application. The singing synthesis application we have developed generates a performance of an artificial singer out of the musical score and the phonetic transcription of a song. These two applications have been implemented as software to run on the PC platform and can be used to illustrate the results of all the modifications done to the initial SMS spectral model for the singing voice case.

Bonada, J. Celma, O. Loscos, A. Ortolà, J. Serra, X. 2001.
'Singing Voice Synthesis Combining Excitation plus Resonance and Sinusoidal plus Residual Models'
Proceedings of International Computer Music Conference 2001. Havana, Cuba

Abstract: This paper presents an approach to the modeling of the singing voice with a particular emphasis on the naturalness of the resulting synthetic voice. The underlying analysis/synthesis technique is based on the Spectral Modeling Synthesis (SMS) and a newly developed Excitation plus Resonance (EpR) model. With this approach a complete singing voice synthesizer is developed that generates a vocal melody out of the score and the phonetic transcription of a song.

Amatriain, X. Bonada, J. Loscos, A. Serra, X. 2001.
'Spectral Modeling for Higher-level Sound Transformation'
Proceedings of MOSART Workshop on Current Research Directions in Computer Music. Barcelona

Abstract: When designing audio effects for music processing, we are always aiming at providing higher-level representations that may somehow fill in the gap between the signal processing world and the end-user. Spectral models in general, and the Sinusoidal plus Residual model in particular, can sometimes offer ways to implement such schemes.

Bonada, J. Loscos, A. Mayor, O. Kenmochi, H. 2003.
'Sample-based singing voice synthesizer using spectral models and source-filter decomposition'
Proceedings of 3rd International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications. Firenze, Italy

Abstract: This paper is a review of the work contained in the insides of a sample-based virtual singing synthesizer. Starting with a narrative of the evolution of the techniques involved in it, the paper focuses mainly on the description of its current components and processes and its most relevant features: from the singer databases creation to the final synthesis concatenation step.

Bonada, J. Loscos, A. 2003.
'Sample-based singing voice synthesizer by spectral concatenation'
Proceedings of Stockholm Music Acoustics Conference 2003. Stockholm, Sweden

Abstract: The singing synthesis system we present generates a performance of an artificial singer out of the musical score and the phonetic transcription of a song using a frame-based frequency domain technique. This performance mimics the real singing of a singer that has been previously recorded, analyzed and stored in a database, in which we store his voice characteristics (phonetics) and his low-level expressivity (attacks, releases, note transitions and vibratos). To synthesize such performance the systems concatenates a set of elemental synthesis units (phonetic articulations and stationeries). These units are obtained by transposing and time-scaling the database samples. The *concatenation of these transformed samples is performed by spreading out the spectral shape and phase discontinuities of the boundaries along a set of transition frames that surround the joint frames. The expression of the singing is applied through a Voice Model built up on top of a Spectral Peak Processing (SPP) technique.*

Loscos, A. Bonada, J. 2004.
'Emulating Rough and Growl Voice In Spectral Domain'
Proceedings of 7th International Conference on Digital Audio Effects; Naples, Italy

Abstract: This paper presents a new approach on transforming a modal voice into a rough or growl voice. The goal of such transformations is to be able to enhance voice expressiveness in singing voice productions. Both techniques work with spectral models and are based on adding sub-harmonics in frequency domain to the original input voice spectrum.

Cano, P. Fabig, L. Gouyon, F. Koppenberger, M. Loscos, A. Barbosa, A. 2004.
'Semi-Automatic Ambiance Generation'
Proceedings of 7th International Conference on Digital Audio Effects; Naples, Italy

Abstract: Ambiances are background recordings used in audiovisual productions to make listeners feel they are in places like a pub or a farm. Accessing to commercially available atmosphere libraries is a convenient alternative to sending teams to record ambiances yet they limit the creation in different ways. First, they are already mixed, which reduces the flexibility to add, remove individual sounds or change its panning. Secondly, the number of ambient libraries is limited. We propose a semi-automatic system for ambiance generation. The system creates ambiances on demand given text queries by fetching relevant sounds from a large sound effect database and importing them into a sequencer multitrack project. Ambiances of diverse nature can be created easily. Several controls are provided to the users to refine the type of samples and the sound arrangement.

Loscos, A. Celma, O. 2005.
'Larynxophone: Using Voice as A Wind Controller'
Proceedings of International Computer Music Conference 2005; Barcelona

Abstract: In the context of music composition and production using MIDI sequencers, wind instrument tracks are built on the synthesis of music scores that have been written using whether MIDI keyboards or mouse clicks. Such modus operandi clearly handicaps the musician when it comes to shape the resulting audio with the desired expression.
This paper presents a straightforward method to create convincing wind instrument audio tracks avoiding intermediate MIDI layers and easing expression control. The method stands on the musician ability to mimic, by singing or humming, the desired wind instrument performance. From this vocal performance, a set of voice features are extracted and used to drive a real-time cross-synthesis between samples of a wind instrument database and the musician's voice signal.

Janer, J. Loscos, A. 2005.
'Morphing techniques for enhanced scat singing'
Proceedings of 8th Intl. Conference on Digital Audio Effects; Madrid, Spain

Abstract: In jazz, scat singing is a phonetic improvisation that imitates instrumental sounds. In this paper, we propose a system that aims to transform singing voice into real instrument sounds, extending the possibilities for scat singers. Analysis algorithms in the spectral domain extract voice parameters, which drive the resulting instrument sound.
A small database contains real instrument samples that have been spectrally analyzed offline. Two different prototypes are introduced, reproducing a trumpet and a bass guitar respectively.

Loscos, A. Aussenac, T. 2005.
'The Wahwactor: a voice controlled wah-wah pedal'.
Proceedings of 2005 International Conference on New Interfaces for Musical Expression; Vancouver, Canada

Abstract: Using a wah-wah pedal guitar is something guitar players have to learn. Recently, more intuitive ways to control such effect have been proposed. In this direction, the Wahwactor system controls a wah-wah transformation in real-time using the guitar player's voice, more precisely, using the performer [wa-wa] utterances. To come up with this system, different vocal features derived from spectral analysis have been studied as candidates for being used as control parameters. This paper details the results of the study and presents the implementation of the whole system.

Mayor, O. Bonada, J. Loscos, A. 2006.
'The Singing Tutor: Expression Categorization and Segmentation of the Singing Voice'
Proceedings of 121st Convention of the Audio Engineering Society; San Francisco, CA, USA

Abstract: Computer evaluation of singing interpretation has traditionally been based exclusively on tuning and tempo. This article presents a tool for the automatic evaluation of singing voice performances that regards on tuning and tempo but also on the expression of the voice. For such purpose, the system performs analysis at note and intra-note levels. Note level analysis outputs traditional note pitch, note onset and note duration information while Intra-note level analysis is in charge of the location and the expression categorization of note's attacks, sustains, transitions, releases and vibratos. Segmentation is done using an algorithm based on untrained HMMs with probabilistic models built out of a set of heuristic rules. A graphical tool for the evaluation and fine-tuning of the system will be presented. The interface gives feedback about analysis descriptors and rule probabilities.

Loscos, A. Wang, Y. Boo, J. 2006.
'Low Level Descriptors for Automatic Violin Transcription'
Proceedings of 7th Intl. Conference on Music Information Retrieval; Victoria, Canada

Abstract: On top of previous work in automatic violin transcription we present a set of straight forward low level descriptors for assisting the transcription techniques and saving computational cost. Proposed descriptors have been tested against a database of 1500 violin notes and double stops.

Loscos, A. Bonada, J. 2006.
'Esophageal Voice Enhancement by Modeling Radiated Pulses in Frequency Domain'
Proceedings of 121st Convention of the Audio Engineering Society; San Francisco, CA, USA

Abstract: Although esophageal speech has demonstrated to be the most popular voice recovering method after laryngectomy surgery, it is difficult to master and shows a poor degree of intelligibility. This article proposes a new method for esophageal voice enhancement using speech digital signal processing techniques based on modeling radiated voice pulses in frequency domain. The analysis-transformation-synthesis technique creates a non-pathological spectrum for those utterances featured as voiced and filters those unvoiced. Healthy spectrum generation implies transforming the original timbre, modeling harmonic phase coupling from the spectral shape envelope, and deriving pitch from frame energy analysis. Resynthesized speech aims to improve intelligibility, minimize artificial artifacts, and acquire resemblance to patient's pre-surgery original voice.

Vinyes, M. Bonada, J. Loscos, A. 2006.
'Demixing Commercial Music Productions via Human-Assisted Time-Frequency Masking'
Proceedings of AES 120th Convention; Paris, France

Abstract: Audio Blind Separation in real commercial music recordings is still an open problem. In the last few years some techniques have provided interesting results. This article presents a human-assisted selection of the DFT coefficients for the Time-Frequency Masking demixing technique. The DFT coefficients are grouped by adjacent pan, inter-channel phase difference, magnitude and magnitude-variance with a real-time interactive graphical interface. Results prove an implementation of such technique can be used to demix tracks from nowadays commercial songs. Sample sounds can be found at http://www.iua.upf.es/~mvinyes/abs/demos.

Bonada, J. Blaauw, M. Loscos, A. 2006.
'Improvements to a Sample-Concatenation Based Singing Voice Synthesizer'
Proceedings of 121st Convention of the Audio Engineering Society; San Francisco, CA, USA

Abstract: This paper describes recent improvements to our singing voice synthesizer based on concatenation and transformation of audio samples using spectral models.

Improvements include firstly robust automation of previous singer database creation process, a lengthy and tedious task which involved recording scripts generation, studio sessions, audio editing, spectral analysis, and phonetic based segmentation; and secondly synthesis technique enhancement, improving the quality of sample transformations and concatenations, and discriminating between phonetic intonation and musical articulation.

Bonada, J. Blaauw, M. Loscos, A. Kenmochi, H. 2006.
'Unisong: A Choir Singing Synthesizer'
Proceedings of 121st Convention of the Audio Engineering Society; San Francisco, CA, USA

Abstract: Computer generated singing choir synthesis can be achieved by two means: clone transformation of a single voice or concatenation of real choir recording snippets. As of today, the synthesis quality for these two methods lack of naturalness and intelligibility respectively. Unisong is a new concatenation based choir singing synthesizer able to generate a high quality synthetic performance out of the score and lyrics specified by the user. This article describes all actions and techniques that take place in the process of virtual synthesis generation: choir recording scripts design and realization, human supervised automatic segmentation of the recordings, creation of samples database, and sample acquiring, transformation and concatenation. The synthesizer will be demonstrated with a song sample.

Annex B

Related patents by the author

PITCH CONVERSION DEVICE AND PROGRAM

Publication number: JP2006064799
Publication date: 2006-03-09
Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10L21/04; G10H1/00; G10L11/00; G10L21/00; G10H1/00; G10L11/00;
- European:
Application number: JP20040244693 20040825
Priority number(s): JP20040244693 20040825

Abstract of JP2006064799
PROBLEM TO BE SOLVED: To obtain an output sound of natural sound quality by a pitch conversion device which uses phase vocoder technology.
SOLUTION: An amplitude spectrum (A) is obtained by analyzing the frequency of an input speech waveform through FFT (Fast Fourier Transform) analytic processing. A plurality of local peaks P<SB>0</SB>to P<SB>2</SB>etc., of spectrum intensity are detected on the amplitude spectrum (A) and spectrum distribution regions R<SB>0</SB>etc., are designated by the local peaks. As shown in (B), the spectrum distribution regions R<SB>0</SB>etc., are moved on a frequency axis according to an input pitch to vary the pitch. At this time, the difference [Delta]f<SB>1</SB>between a harmonic frequency f<SB>1</SB>and a complete harmonic frequency 2f<SB>0</SB>is held and when a harmonic frequency f<SB>11</SB>after pitch variation is determined, the frequency obtained by shifting the complete harmonic frequency 2f<SB>01</SB>after the pitch variation corresponding to the difference [Delta]f<SB>1</SB>is regarded as the harmonic frequency f<SB>11</SB>. Other harmonic frequencies such as f<SB>2</SB>are similarly obtained.

SPEECH PROCESSING APPARATUS AND PROGRAM

Publication number: JP2006017946
Publication date: 2006-01-19
Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10L13/00; G10L13/02; G10L21/04; G10L21/00; G10L13/00; G10L21/00;
- European: G10L13/02E
Application number: JP20040194800 20040630
Priority number(s): JP20040194800 20040630

Abstract of JP2006017946
PROBLEM TO BE SOLVED: To generate natural output speech from input speech as regards technology to change the characteristics of the speech.
SOLUTION: An envelope specification section 23 generates input envelope data DEVin indicating the spectrum envelope EVin of an input speech signal Sin. A template acquisition section 33 reads spectrum data DSPt for conversion indicating the frequency spectrum SPt of speech for conversion out of a memory section 51. A data generation section 3a specifies a frequency spectrum SPnew which is the frequency spectrum of a shape corresponding to the frequency spectrum SPt of the speech for conversion and has the spectrum envelope which nearly coinciding with the spectrum envelope EVin of the input speech on the basis of the input envelope data DEVin and the spectrum data DSPt for conversion and generates new spectrum data DSPnew indicating the frequency spectrum SPnew. A reverse FFT section 15 and an output processing section 16 generate an output speech signal Snew on the basis of the new spectrum data DSPnew.

DEVICE AND PROGRAM FOR IMPARTING SOUND EFFECT

Publication number: JP2006010908
Publication date: 2006-01-12
Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10L13/00; G10L21/04; G10L13/00; G10L21/00;
- European:
Application number: JP20040186014 20040624
Priority number(s): JP20040186014 20040624

Abstract of JP2006010908
PROBLEM TO BE SOLVED: To provide a sound effect imparting device and a sound effect imparting program capable of performing various sound conversion.
SOLUTION: A FFT part 2 analyzes the frequency of an input sound to determine spectrum distribution area including local peaks respectively. A spectrum deformation

part 4 varies spectrum intensities of the respective spectrum distribution areas according to spectrum intensity variation tables stored in various tables 5. At this time, the degree of effect can be adjusted with parameters from a parameter specification part 6. The spectrum deformed by a spectrum deformation part 4 is converted by an IFFT part 7 into time areas, which are put together again and output. Values of the parameters may be varied according to the frequency of a spectrum distribution area.

DEVICE AND PROGRAM FOR IMPARTING SOUND EFFECT

Publication number: JP2006010906
Publication date: 2006-01-12
Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10L13/00; G10H1/00; G10H1/16; G10H1/00; G10H1/06; G10L13/00;
- European: G10H1/00S; G10H1/10
Application number: JP20040186012 20040624
Priority number(s): JP20040186012 20040624

Abstract of JP2006010906
PROBLEM TO BE SOLVED: To impart real distortion effect to an input sound.
SOLUTION: An FFT part 2 analyzes the frequency of the input sound to detect a local peak and a pitch. A sub-harmonics imparting means 5 adds a spectrum between local peaks of the input spectrum according to a parameter from a parameter specification part 8. The frequency spectrum to which sub-harmonics are added is converted by an IFFT part 9 into time regions, which are put together again and output. A 1st sub-harmonics addition part 6 imparts distortion effect like a creak by adding a spectrum irregularly varying in gain between local peaks of the input spectrum and a 2nd sub-harmonics addition part 7 imparts distortion effect like a growl by adding a plurality of spectra differing in frequency between the local peaks of the input spectrum.

MUSICAL SOUND PROCESSING EQUIPMENT, MUSICAL SOUND PROCESSING METHOD, AND MUSICAL SOUND PROCESSING PROGRAM

Publication number: JP2005121742
Publication date: 2005-05-12
Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10H1/10; G10H1/06; (IPC1-7): G10H1/10
- European:
Application number: JP20030354089 20031014
Priority number(s): JP20030354089 20031014

Abstract of JP2005121742

PROBLEM TO BE SOLVED: To provide a musical sound processing equipment which makes the impartation of natural ensemble effects etc., possible, a musical sound processing method, and a musical sound processing control program.
SOLUTION: The musical sound processing equipment is provided with an ensemble effect imparting section which imparts ensemble effects to inputted single sound speech. The ensemble effect imparting section is equipped with a wobble signal generating unit and a spectrum changing section. The spectrum changing section extracts the respective harmonic overtone regions from the inputted single sound speech and divides the respective harmonic overtone regions to a plurality of regions. The spectrum changing section applies respectively different modulations to each of a plurality of the spectrum components existing in the respective regions based on the number-of-persons feeling assignment information supplied from a number-of-persons feeling parameter input section and the wobble signals supplied from the respective wobble signal generating sections.

APPARATUS, METHOD, AND PROGRAM FOR MUSICAL SOUND PROCESSING

Publication number: JP2005107315
Publication date: 2005-04-21
Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10H1/10; G10H1/06; (IPC1-7): G10H1/10
- European:
Application number: JP20030342254 20030930
Priority number(s): JP20030342254 20030930

Abstract of JP2005107315
PROBLEM TO BE SOLVED: To provide an apparatus, a method, and a program for musical sound processing that imparts natural unison singing effect.
SOLUTION: A pseudo-random signal generation unit 500 comprises a plurality of pseudo-random signal generation parts 510. The respective pseudo-random signal generation parts 510 generate pseudo-random signals differing in way of varying etc., and supply those pseudo-random signals to corresponding clone signal generation parts 410. A pseudo-random signal generated by each pseudo-random signal generation part 510 has its white noise processed by an LPF. The cutoff frequency of the LPF is set to a value matching natural ways of variation and fluctuation when a person sings most (in concrete, about 2 Hz). Each clone signal generation part 410 uses the pseudo-random signal as a modulation signal to perform voice conversion.

VOICE PROCESSING DEVICE, VOICE PROCESSING METHOD, AND VOICE PROCESSING PROGRAM

Publication number: JP2005099509
Publication date: 2005-04-14

Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10L21/04; G10H1/043; G10L21/00; G10H1/04; (IPC1-7): G10H1/043; G10L21/04
- European:
Application number: JP20030334130 20030925
Priority number(s): JP20030334130 20030925

Abstract of JP2005099509

PROBLEM TO BE SOLVED: To provide a voice processing device, a voice processing method, and a voice processing program, which give vibrato accompanied with natural tone variation.
SOLUTION: A template creation part 900 creates a template set TS concerned with a target voice which is supplied from a target input part 950 and to which vibrato is applied. A pitch, a gain, and an inclination of a spectrum (slope) of each frame of the target voice are recorded in the template set TS created by the template creation part 900. An vibrato applying part 400 changes pitches, gains, and slopes of a singing voice (input voice) of an amateur singer or the like under the control of a vibrato application control part 700 and performs inverse FFT of spectrums obtained by the changed pitches, gains, and slopes and supplies a voice obtained by inverse FFT (namely, a voice having the changed pitches, gains, and slopes) to a voice output part 500.

VOICE PROCESSING APARATUS AND PROGRAM

Publication number: US2006004569
Publication date: 2006-01-05
Inventor: YOSHIOKA YASUO (JP); LOSCOS ALEX (ES)
Applicant: YAMAHA CORP (JP)
Classification:
- international: G10L19/14; G10L13/02; G10L21/00; G10L19/00; G10L13/00; G10L21/00
- European: G10L13/02E
Application number: US20050165695 20050624
Priority number(s): JP20040194800 20040630

Abstract of US2006004569

Envelope identification section generates input envelope data (DEVin) indicative of a spectral envelope (EVin) of an input voice. Template acquisition section reads out, from a storage section, converting spectrum data (DSPt) indicative of a frequency spectrum (SPt) of a converting voice. On the basis of the input envelope data (DEVin) and the converting spectrum data (DSPt), a data generation section specifies a frequency spectrum (SPnew) corresponding in shape to the frequency spectrum (SPt) of the converting voice and having a substantially same spectral envelope as the spectral envelope (EVin) of the input voice, and the data generation section generates new spectrum data (DSPnew) indicative of the frequency spectrum (SPnew). Reverse FFT

section and output processing section generates an output voice signal (Snew) on the basis of the new spectrum data (DSPnew).

SOUND EFFECT APPLYING APPARATUS AND SOUND EFFECT APPLYING PROGRAM

Publication number: US2005288921
Publication date: 2005-12-29
Inventor: YOSHIOKA YASUO (JP); LOSCOS ALEX (ES)
Applicant: YAMAHA CORP (JP)
Classification:
- international: G10H1/00; G10H1/00; (IPC1-7): G10L21/00
- European: G10H1/00S; G10H1/10
Application number: US20050159032 20050622
Priority number(s): JP20040186012 20040624

Abstract of US2005288921
In a sound effect applying apparatus, an input part frequency-analyzes an input signal of sound or voice for detecting a plurality of local peaks of harmonics contained in the input signal. A subharmonics provision part adds a spectrum component of subharmonics between the detected local peaks so as to provide the input signal with a sound effect. An output part converts the input signal of a frequency domain containing the added spectrum component into an output signal of a time domain for generating the sound or voice provided with the sound effect.

VOICE SYNTHESIZING APPARATUS CAPABLE OF ADDING VIBRATO EFFECT TO SYNTHESIZED VOICE

Publication number: EP1291846
Publication date: 2003-03-12
Inventor: YOSHIOKA YASUO (JP); LOSCOS ALEX (ES)
Applicant: YAMAHA CORP (JP)
Classification:
- international: G10L13/00; G10K15/04; G10L11/00; G10L13/02; G10L13/04; G10L13/08; G10L21/04; G10K15/04; G10L11/00; G10L13/00; G10L21/00; (IPC1-7): G10L13/08; G10L13/06
- European: G10L13/08P
Application number: EP20020019741 20020903
Priority number(s): JP20010265489 20010903

Abstract of EP1291846
A voice synthesizing apparatus comprises: storage means for storing a first database storing a first parameter obtained by analyzing a voice and a second database storing a second parameter obtained by analyzing a voice with vibrato; inputting means for inputting information for a voice to be synthesized; generating means for generating a third parameter based on the first parameter read from the first database and the second

parameter read from the second database in accordance with the input information; and synthesizing means for synthesizing the voice in accordance with the third parameter. A very real vibrato effect can be added to a synthesized voice.

SINGING SYNTHESIZING METHOD, DEVICE, AND RECORDING MEDIUM

Publication number: JP2003255998
Publication date: 2003-09-10
Inventor: KENMOCHI HIDENORI; JORDI BONADA; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10L21/04; G10H7/00; G10L13/00; G10L13/02; G10L13/06; G10L21/00; G10H7/00; G10L13/00; (IPC1-7): G10L21/04; G10L13/00; G10L13/06
- European: G10H7/00C; G10L13/02
Application number: JP20020052006 20020227
Priority number(s): JP20020052006 20020227

Abstract of JP2003255998
PROBLEM TO BE SOLVED: To synthesize a natural singing voice or singing voice of high quality.

METHOD AND DEVICE FOR VOICE SYNTHESIS AND PROGRAM

Publication number: JP2003076387
Publication date: 2003-03-14
Inventor: YOSHIOKA YASUO; ALEX ROSUKOSU
Applicant: YAMAHA CORP
Classification:
- international: G10L13/00; G10K15/04; G10L11/00; G10L13/02; G10L13/04; G10L13/08; G10L21/04; G10K15/04; G10L11/00; G10L13/00; G10L21/00; (IPC1-7): G10L13/00; G10K15/04; G10L13/04; G10L21/04
- European: G10L13/08P
Application number: JP20010265489 20010903
Priority number(s): JP20010265489 20010903

Abstract of JP2003076387
PROBLEM TO BE SOLVED: To provide a voice synthesizer which can give very real vibratos. SOLUTION: The voice synthesizer is provided with a storage means for storage of a first database for storage of first parameters obtained by voice analysis and a second database for storage of second parameters obtained by vibrato voice analysis, an input means for input of information of a voice to be synthesized, an addition means which adds the first and second parameters read out from the storage means on the basis of inputted information and generates a third parameter, and a voice synthesizing means which synthesizes the voice on the basis of the third parameter.

SINGING VOICE SYNTHESIZING METHOD

Publication number: US2003221542
Publication date: 2003-12-04
Inventor: KENMOCHI HIDEKI (JP); LOSCOS ALEX (ES); BONADA JORDI (SE)
Applicant:
Classification:
- international: G10L21/04; G10H7/00; G10L13/00; G10L13/02; G10L13/06; G10L21/00; G10H7/00; G10L13/00; (IPC1-7): G10H7/00
- European: G10H7/00C; G10L13/02
Application number: US20030375420 20030227
Priority number(s): JP20020052006 20020227

Abstract of US2003221542

A frequency spectrum is detected by analyzing a frequency of a voice waveform corresponding to a voice synthesis unit formed of a phoneme or a phonemic chain. Local peaks are detected on the frequency spectrum, and spectrum distribution regions including the local peaks are designated. For each spectrum distribution region, amplitude spectrum data representing an amplitude spectrum distribution depending on a frequency axis and phase spectrum data representing a phase spectrum distribution depending on the frequency axis are generated. The amplitude spectrum data is adjusted to move the amplitude spectrum distribution represented by the amplitude spectrum data along the frequency axis based on an input note pitch, and the phase spectrum data is adjusted corresponding to the adjustment. Spectrum intensities are adjusted to be along with a spectrum envelope corresponding to a desired tone color. The adjusted amplitude and phase spectrum data are converted into a synthesized voice signal.

SINGING VOICE SYNTHESIZING METHOD

Publication number: EP1505570
Publication date: 2005-02-09
Inventor: KENMOCHI HIDEKI (JP); BONADA JORDI (ES); LOSCOS ALEX (ES)
Applicant: YAMAHA CORP (JP)
Classification:
- international: G10L13/02; G10L21/02; G10L13/00; G10L21/00; (IPC1-7): G10L13/02
- European: G10L13/02E
Application number: EP20030017548 20030806
Priority number(s): EP20030017548 20030806

Abstract of EP1505570

A frequency spectrum is detected by analyzing a frequency of a voice waveform corresponding to a voice synthesis unit formed of a phoneme or a phonemic chain. Local peaks are detected on the frequency spectrum, and spectrum distribution regions including the local peaks are designated. For each spectrum distribution region, amplitude spectrum data representing an amplitude spectrum distribution depending on a frequency axis and phase spectrum data representing a phase spectrum distribution depending on the frequency axis are generated. The amplitude spectrum data is adjusted to move the amplitude spectrum distribution represented by the amplitude spectrum data along the frequency axis based on an input note pitch, and the phase spectrum data is adjusted corresponding to the adjustment. Spectrum intensities are adjusted to be along with a spectrum envelope corresponding to a desired tone color. The adjusted amplitude and phase spectrum data are converted into a synthesized voice signal.

VOICE CONVERTER FOR ASSIMILATION BY FRAME SYNTHESIS WITH TEMPORAL ALIGNMENT

Publication number: US2005049875
Publication date: 2005-03-03
Inventor: KAWASHIMA TAKAHIRO (JP); YOSHIOKA YASUO (JP); CANO PEDRO (ES); LOSCOS ALEX (ES); SERRA XAVIER (ES); SCHIEMENTZ MARK (ES); BONADA JORDI (ES)
Applicant: YAMAHA CORP (US)
Classification:
- international: G10L13/02; G10L21/00; G10L13/00; G10L21/00; (IPC1-7): G10L13/00
- European: G10L13/02E
Application number: US20040951328 20040927
Priority number(s): US20040951328 20040927; JP19990300268 19991021; JP19990300276 19991021; US20000693144 20001020

Abstract of US2005049875
A voice converting apparatus is constructed for converting an input voice into an output voice according to a target voice. In the apparatus, a storage section provisionally stores source data, which is associated to and extracted from the target voice. An analyzing section analyzes the input voice to extract therefrom a series of input data frames representing the input voice. A producing section produces a series of target data frames representing the target voice based on the source data, while aligning the target data frames with the input data frames to secure synchronization between the target data frames and the input data frames. A synthesizing section synthesizes the output voice according to the target data frames and the input data frames. In the recognizing feature analysis, a characteristic analyzer extracts from the input voice a characteristic vector. A memory memorizes target behavior data representing a behavior of the target voice. An alignment processor determines a temporal relation between the input data frames and the target data frames according to the characteristic vector and the target behavior data so as to output alignment data. A target decoder produces the target data frames according to the alignment data, the input data frames and the source data containing phoneme of the target voice.

SOUND SIGNAL PROCESSING APPARATUS, SOUND SIGNAL PROCESSING METHOD, AND SOUND SIGNAL PROCESSING PROGRAM

Publication number: US2006272488
Publication date: 2006-12-07
Inventor: FUJISHIMA TAKUYA (JP); LOSCOS ALEX (ES); BONADA JORDI (ES); MAYOR OSCAR (ES)
Applicant: YAMAHA CORP (JP)
Classification:
- international: G10H1/36; G10H1/36;
- European:
Application number: US20060439818 20060524
Priority number(s): JP20050154738 20050526

Abstract of US2006272488
A sound signal processing apparatus which is capable of correctly detecting expression modes and expression transitions of a song or performance from an input sound signal. A sound signal produced by performance or singing of musical tones is input and divided into frames of predetermined time periods. Characteristic parameters of the input sound signal are detected on a frame-by-frame basis. An expression determining process is carried out in which a plurality of expression modes of a performance or song are modeled as respective states, the probability that a section including a frame or a plurality of continuous frames lies in a specific state is calculated with respect to a predetermined observed section based on the characteristic parameters, and the optimum route of state transition in the predetermined observed section is determined based on the calculated probabilities so as to determine expression modes of the sound signal and lengths thereof.

DEVICE, METHOD, AND PROGRAM FOR DECIDING VOICE QUALITY

Publication number: JP2006195449
Publication date: 2006-07-27
Inventor: FUJISHIMA TAKUYA; SEKINE SATOSHI; KAMIYA SHINGO; BONADA JORDI; LARS FABIG; OSCAR MAYOR; ALEX LOSCOS
Applicant: YAMAHA CORP
Classification:
- international: G10L11/00; G10K15/04; G10L11/00; G10K15/04;
- European:
Application number: JP20050361612 20051215
Priority number(s): JP20040363730 20041215; JP20050361612 20051215

Abstract of JP2006195449
PROBLEM TO BE SOLVED: To provide a device for deciding voice quality which can objectively decide the quality (voice quality) of the voice signal of the utterance made by man.
SOLUTION: A device for deciding voice quality is equipped with a physical parameter analyzing section which finds the physical parameters as the physical features of an

inputted voice signal from the voice signal and a voice quality decision section which decides the voice quality of the speech signal, based on the physical parameters. The physical parameters are a spectral tilt, a spectrum excitation, a formant sharpness, harmonic stability, a valley depth attenuation, waveform peak level stability, energy, pitch, etc., and scales for deciding the voice quality are the echo degree, transparency, stability, high-tone tolerance, low-tone tolerance, etc.