Universitat Ramon Llull

TESI DOCTORAL

Títol	Conveying expressivity and vocal effort transformation in synthetic speech with Harmonic plus Noise Models
Realitzada per	Àngel Calzada Defez
en el Centre	Escola d'Enginyeria i Arquitectura la Salle
i en el	Departament de Comunicacions i Teoria de senyal
Dirigida per	Joan Claudi Socoró Carrié

Acknowledgements

The following text expresses my gratitude to those who somehow helped me along this thesis, and I dedicate this work to them in Catalan. The rest of this thesis is written in English.

Aquest treball m'agradaria dedicar-lo especialment a mons pares, Àngel Calzada i Lluïsa Defez, ja que sense el seu suport mai hagués pogut, ni tant-sols imaginar, poder arribar fins aquí. També a la resta de la família, Albert, Lluï i Berna, que m'han perdonat les nombroses hores d'absència dedicades a la tesi.

En segon lloc m'agradaria dedicar-lo als meus amics Ivan Blanco, Laia Albaladejo, Mar Moral, i tots aquells que m'heu fet costat en aquest llarg viatge. També agrair a totes aquelles persones com els nomenats anteriorment i, la colla de Vilassar, companys de l'EUPMt entre altres que m'han ajudat realitzant de forma desinteressada les avaluacions del treball presentat en aquesta tesi i qui mai oblidaran el TA-MU-MA.

També voldria agrair a Joan Claudi Socoró pel seu ajut al llarg d'aquesta tesi i les hores dedicades a orientar-me i animar-me a presentar aquest treball. També voldria agrair als meus companys d'usMIMA, en Marc Benet, Immaculada Herrero, i Markus Wilhelms, pels seus ànims i per deixar-me dedicar hores de dedicació a la nostre startup a finalitzar la tesi.

Finalment voldria agrair al Comissionat per a Universitats i Recerca del *DIUE* de la Generalitat de Catalunya i del Fons Social Europeu qui ha dipositat en mi el seu suport econòmic a través de la beca *FI* (2010FI B 01083) i dels dos ajuts per a la mobilitat *BE* (2010 BE1 00503 i 2011 BE1 01084) que em van permetre realitzar les estades als centres DFKI-LT (Saarbrücken, Alemanya) i CSTR (Edimburg, Regne Unit).

A tots vosaltres moltíssimes gràcies.

Abstract

This thesis was conducted in the Grup en Tecnologies Mèdia (GTM) from *Escola d'Enginyeria i Arquitectura la Salle*. The group has a long trajectory in the speech synthesis field and has developed their own Unit-Selection Text-To-Speech (US-TTS) which is able to convey multiple expressive styles using multiple expressive corpora, one for each expressive style. Thus, in order to convey aggressive speech, the US-TTS uses an aggressive corpus, whereas for a sensual speech style, the system uses a sensual corpus. Unlike that approach, this dissertation aims to present a new schema for enhancing the flexibility of the US-TTS system for performing multiple expressive styles using a single neutral corpus. The approach followed in this dissertation is based on applying Digital Signal Processing (DSP) techniques for carrying out speech modifications in order to synthesize the desired expressive style. For conducting the speech modifications the Harmonics plus Noise Model (HNM) was chosen for its flexibility in conducting signal modifications.

Voice Quality (VoQ) has been proven to play an important role in different expressive styles. Thus, low-level VoQ acoustic parameters were explored for conveying multiple emotions. This raised several problems setting new objectives for the rest of the thesis, among them finding a single parameter with strong impact on the expressive style conveyed. Vocal Effort (VE) was selected for conducting expressive speech style modifications due to its salient role in expressive speech. The first approach working with VE was based on transferring VE between two parallel utterances based on the Adaptive Pre-emphasis Linear Prediction (APLP) technique. This approach allowed transferring VE but the model presented certain restrictions regarding its flexibility for generating new intermediate VE levels. Aiming to improve the flexibility and control of the conveyed VE, a new approach using polynomial model for modelling VE was presented. This model not only allowed transferring VE levels between two different utterances, but also allowed to generate other VE levels than those present in the speech corpus. This is aligned with the general goal of this thesis, allowing US-TTS systems to convey multiple expressive styles with a single neutral corpus. Moreover, the proposed methodology introduces a parameter for controlling the degree of VE in the synthesized speech signal. This opens new possibilities for controlling the synthesis process such as the one in the CreaVeu project using a simple and intuitive graphical interfaces, also conducted in the GTM group. The dissertation concludes with a review of the conducted work and a proposal for schema modifications within a US-TTS system for introducing the VE modification blocks designed in this dissertation.

Keywords: Voice Quality, Vocal Effort, Expressive Speech Synthesis, Unit Selection, Text-to-Speech, Harmonic-plus-Noise-Model

Resumen

Esta tesis se llevó a cabo en el Grup en Tecnologies Mèdia (GTM) de la *Escuela de Ingeniería y Arquitectura la Salle*. El grupo lleva una larga trayectoria dentro del campo de la síntesis de voz y cuenta con su propio sistema de síntesis por concatenación de unidades (US-TTS). El sistema permite sintetizar múltiples estilos expresivos mediante el uso de corpus específicos para cada estilo expresivo. De este modo, para realizar una síntesis agresiva, el sistema usa el corpus de este estilo, y para un estilo sensual, usa otro corpus específico para ese estilo. La presente tesis aborda el problema con un enfoque distinto proponiendo cambios en el esquema del sistema con el fin de mejorar la flexibilidad para sintetizar múltiples estilos expresivos a partir de un único corpus de estilo de habla neutro. El planteamiento seguido en esta tesis esta basado en el uso de técnicas de procesamiento de señales (DSP) para llevar a cabo modificaciones del señal de voz para que este exprese el estilo de habla deseado. Para llevar acabo las modificaciones de la señal de voz se han usado los modelos harmónico más ruido (HNM) por su flexibilidad para efectuar modificaciones de señales.

La cualidad de la voz (VoQ) juega un papel importante en diferentes estilos expresivos. Por ello se exploró la síntesis expresiva basada en modificaciones de parámetros de bajo nivel de la VoQ. Durante este estudio se detectaron diferentes problemas que dieron pié a los objetivos planteados en esta tesis, entre ellos el encontrar un único parámetro con fuerte influencia en la expresividad. El parámetro seleccionado fue el esfuerzo vocal (VE) por su importante papel a la hora de expresar diferentes emociones. Las primeras pruebas se realizaron con el fin de transferir el VE entre dos realizaciones con diferente grado de VE de la misma palabra usando una metodología basada en un proceso filtrado de pre-émfasis adaptativo con coeficientes de predicción lineales (APLP). Esta primera aproximación logró transferir el nivel de VE entre dos realizaciones de la misma palabra, sin embargo el proceso presentaba limitaciones para generar niveles de esfuerzo vocal intermedios. A fin de mejorar la flexibilidad y el control del sistema para expresar diferentes niveles de VE, se planteó un nuevo modelo de VE basado en polinomios lineales. Este modelo permitió transferir el VE entre dos palabras diferentes e incluso generar nuevos niveles no presentes en el corpus usado para la síntesis. Esta flexibilidad está alineada con el objetivo general de esta tesis de permitir a un sistema US-TTS expresar múltiples estilos de habla expresivos a partir de un único corpus de estilo neutro. Además, la metodología propuesta incorpora un parámetro que permite de forma sencilla controlar el nivel de VE expresado en la voz sintetizada. Esto abre la posibilidad de controlar fácilmente el proceso de síntesis tal y como se hizo en el proyecto CreaVeu usando interfaces simples e intuitivas, también realizado dentro del grupo GTM. Esta memoria concluye con una revisión del trabajo realizado en esta tesis y con una propuesta de modificación de un esquema de US-TTS para expresar diferentes niveles de VE a partir de un único corpus neutro.

Palabras clave: Cualidad de la voz, Esfuerzo Vocal, Síntesis expresiva, Selección de unidades,Conversión texto-habla, Modelos harmónicos más ruido

Resum

Aquesta tesi s'ha dut a terme dins del Grup en Tecnologies Mèdia (GTM) de l'*Escola d'Enginyeria i Arquitectura la Salle*. El grup te una llarga trajectòria dins del cap de la síntesi de veu i fins i tot disposa d'un sistema propi de síntesi per concatenació d'unitats (US-TTS) que permet sintetitzar diferents estils expressius usant múltiples corpus. De forma que per a realitzar una síntesi agressiva, el sistema usa el corpus de l'estil agressiu, i per a realitzar una síntesi sensual, usa el corpus de l'estil corresponent. Aquesta tesi pretén proposar modificacions del esquema del US-TTS que permetin millorar la flexibilitat del sistema per sintetitzar múltiples expressivitats usant només un únic corpus d'estil neutre. L'enfoc seguit en aquesta tesi es basa en l'ús de tècniques de processament digital del senyal (DSP) per aplicar modificacions de senyal a la veu sintetitzada per tal que aquesta expressi l'estil de parla desitjat. Per tal de dur a terme aquestes modificacions de senyal s'han usat els models harmònic més soroll per la seva flexibilitat a l'hora de realitzar modificacions de senyal.

La qualitat de la veu (VoQ) juga un paper important en els diferents estils expressius. És per això que es va estudiar la síntesi de diferents emocions mitjançant la modificació de paràmetres de VoQ de baix nivell. D'aquest estudi es van identificar un conjunt de limitacions que van donar lloc als objectius d'aquesta tesi, entre ells el trobar un paràmetre amb gran impacte sobre els estils expressius. Per aquest fet l'esforç vocal (VE) es va escollir per el seu paper important en la parla expressiva. Primer es va estudiar la possibilitat de transferir l'VE entre dues realitzacions amb diferent VE de la mateixa paraula basant-se en la tècnica de predicció lineal adaptativa del filtre de pre-èmfasi (APLP). La proposta va permetre transferir l'VE correctament però presentava limitacions per a poder generar nivells intermitjos d'VE. Amb la finalitat de millorar la flexibilitat i control de l'VE expressat a la veu sintetitzada, es va proposar un nou model d'VE basat en polinomis lineals. Aquesta proposta va permetre transferir l'VE entre dues paraules qualsevols i sintetitzar nous nivells d'VE diferents dels disponibles al corpus. Aquesta flexibilitat esta alineada amb l'objectiu general d'aquesta tesi, permetre als sistemes US-TTS sintetitzar diferents estils expressius a partir d'un únic corpus d'estil neutre. La proposta realitzada també inclou un paràmetre que permet controlar fàcilment el nivell d'VE sintetitzat. Això obre moltes possibilitats per controlar fàcilment el procés de síntesi tal i com es va fer al projecte CreaVeu usant interfícies gràfiques simples i intuïtives, també realitzat dins del grup GTM. Aquesta memòria conclou presentant el treball realitzat en aquesta tesi i amb una proposta de modificació de l'esquema d'un sistema US-TTS per incloure els blocs de DSP desenvolupats en aquesta tesi que permetin al sistema sintetitzar múltiple nivells d'VE a partir d'un corpus d'estil neutre.

Paraules clau: Qualitat de la veu, Esforç Vocal, Síntesi expressiva, Selecció d'unitats, Conversió Text-Parla, Models Harmònic més Soroll.

Contents

Acknowledgements **i**

Abstract **vii**

Index **ix**

List of Figures **xiii**

List of Tables **xv**

Acronyms **xvii**

1 Introduction **1**

1.1 Research group 1

1.2 Motivation 3

1.3 Objectives 4

1.4 Contents 8

2 Background **11**

2.1 Synthesis techniques 12

2.1.1 Formant synthesis 12

2.1.2 Diphone concatenation 14

2.1.3 Unit-selection synthesis 16

2.1.4 Statistical synthesis 21

2.2 Expressive speech synthesis . . . 24

2.2.1 Corpus-based approaches . . . 25

2.2.2 Mapping codebooks . . . 26

2.2.3 Frequency warpping functions . . . 27

2.2.4 Artificial Neural Networks . . . 28

2.2.5 Probabilistic Linear transformations . . . 28

2.2.6 Rule based conversion . . . 29

2.2.7 Speaker interpolation . . . 29

2.2.8 Hidden Markov Models . . . 31

2.3 Vocal effort . . . 31

2.3.1 Definition . . . 32

2.3.2 Spectral behavior and Voice Quality (VoQ) . . . 33

2.4 Adaptive Pre-Emphasis Linear Prediction for VE modification . . . 35

2.5 Sinusoidal and harmonics models overview . . . 37

2.6 GTM's current US-TTS implementation and scope of this dissertation . . . 43

3 Expressivity modification based on prosody and VoQ low-level parameters using HNM 47

3.1 VoQ measurements from HNM parameters . . . 49

3.2 Expressive transformation based on VoQ modifications . . . 52

3.2.1 Expressive speech corpus . . . 52

3.2.2 HNM parameter modifications . . . 55

3.3 Evaluation and results . . . 56

3.4 Conclusions . . . 58

4 Transferring vocal effort with LPC 61

4.1 NECA database and HNM parameterization . . . 62

4.2 Spectral emphasis modification with HNM . . . 65

4.3 Evaluation and results . . . 67

4.4 Discussion 73

4.5 Conclusions 76

5 Vocal effort interpolation 79

5.1 Parametric spectral model and codebooks 81

5.2 Vocal effort modification with the polynomial model 83

5.3 Evaluation and results 89

5.4 Discussion 92

5.5 Conclusions 95

6 Conclusions 97

6.1 Thesis review 97

6.2 Future work 104

6.3 Research contributions 109

Appendices 113

A Harmonics plus Noise Model (HNM) implementation 113

A.1 Harmonic part modelling 116

A.1.1 Forcing harmonic frequencies constraint with Lagrange multiplyiers 122

A.1.2 Summary of the iterative algorithm for harmonic part computation 124

A.1.3 Harmonic part signal reconstruction 124

A.2 Stochastic part modelling 126

A.3 Time modifications 130

A.4 Pitch modifications 134

A.5 Window characteristics 137

A.6 Linear prediction analysis with HNM 139

Bibliography 143

List of Figures

1.1 Speech synthesis approaches comparing task difficulty and synthesis quality 5

2.1 Klatt's formant synthesis software schema 13

2.2 Visual example of unit type impact on word coverage. 17

2.3 Block diagram of an HMM synthesis system. 21

2.4 HMM Model junctions examples of (a) discontinuities and (b) oversmoothing 24

2.5 Speech spectra for male speakers at different VE levels. 33

2.6 Speech spectra for female speakers at different VE levels. 34

2.7 Formant variation in shouted speech. 34

2.8 Adaptive Pre-emphasis Linear Prediction block diagrams. 37

2.9 Block diagram for transferring vocal effort with APLP. 38

2.10 Graphical representation of the time evolution for basis function in the: (a) Fourier, (b) Chirplet, (c) fractional Fourier and (d) fan-Chirp transforms. 41

2.11 Narrowband spectrogram of the same large-bandwidth speech segment, with substantial frequency modulation, using the discrete Fourier and fan-chirp transforms. 42

2.12 General pipeline of a US-TTS system and components to be modified. 44

2.13 Proposed general scheme modifications of the current US-TTS system to include vocal effort. 45

3.1 Block diagram of the proposed VoQ transformation system. 54

3.2 CMOS evaluation for using VoQ with prosody or only prosody. 57

3.3 MOS signal quality evaluation for emotion conversion with HNM using: a) Prosody only, b) prosody, Jitter and Shimmer. 58

4.1 Time mapping function construction from phoneme boundary marks for the logatom */O/-/t/-/a:/-/p/-/o/*. 63

4.2 Reference analysis time instants mapped to high or low VE version. 64

4.3 Block diagram of the proposed vocal effort transformation methodology. 68

4.4 Detail of the Spectral Emphasis (SE) transformation block depicted in figure 4.3. . . . 68

4.5 Boxplots for (a) vocal effort approximation to the reference audio by conversion, (b) averaged value of the proximity to the reference and (c) the signal's quality preference by conversion type, (d) average of the signal's quality preference. 70

4.6 Evaluation of vocal effort proximity to target by the expert evaluators subgroup (26.83% of all evaluators). 71

4.7 Boxplots for (a) perceived vocal effort and (b) subjective quality in MOS for each vocal effort conversion. 72

5.1 Codebooks generation from the NECA coprus. This composition is replicated for each VE level and gender. 82

5.2 Schematic diagram of the proposed vocal effort interpolation method. 84

5.3 Temporal linear interpolation of model coefficients. 85

5.4 VE model interpolation for different interpolation factor (γ) values at multiples of Fundamental Frequency or Pitch (f_0). 86

5.5 Scaled interpolated model. 87

5.6 Multiplication factors ($\boldsymbol{\alpha}$) before and after being raised to the power of β 88

6.1 Example of a possible VE model interpolations. 105

6.2 Proposed US-TTS speech synthesis block to allow the VE modification. 106

A.1 Overview of the HNM schema implemented. 115

A.2 Graphical representation of the analysis windows used for frame k at analysis time t_a^k and frame $k-1$ centred at t_a^{k-1}. 116

A.3 Frame synthesis Overlap-and-Add for the harmonic part reconstruction (A.27). 125

A.4 Example of subtracting the synthesized harmonic part from the original signal. 127

A.5 Amplified time scale of the example shown in figure A.4. 128

A.6 Spectrum using 4096-point Discrete Fourier Transform (DFT) of the signal presented in figure A.4. 129

A.7 Time modifications. (a) original and (b) modified time instants and signal OLA. . . . 131

A.8 Effects of performing time modifications related with linear phase term. 132

A.9 Vocal tract phase envelope obtained by linear interpolation of the complex amplitudes. 134

A.10 Pitch modification effects on concatenation boundaries. 136

A.11 Effects of a wide main lobe in the window spectrum. 138

A.12 Verification of sidelobe presence in the analysis window for multiple combinations of a,b pairs. 139

A.13 Verification of main lobe bandwidth for multiple combinations of a,b pairs. 140

A.14 Window used for signal analysis. 141

List of Tables

2.1 Recognition rates in percentage and Average Success ratings 26

2.2 Examples of successful prosody rules for emotion expression. 30

3.1 VoQ parameters and modification factors for neutral-to-expressive conversion 54

4.1 Used logatoms transcription. 68

5.1 Excerpt of the codebook corresponding to *High* vocal effort level for the female speaker. 83

5.2 Transcription of the stimuli used in the first evaluation using SAMPA notation. 90

5.3 Preference of the proposed method with the reference (APLP with HNM, presented in Chapter 4) according to vocal effort performance in the first evaluation. 90

5.4 Transcriptions of the stimuli used for the second evaluation using the SAMPA notation. 91

5.5 Perception of the extreme vocal effort levels synthesized against neutral level. 92

5.6 Perception of the interpolated vocal effort levels synthesized against high, neutral and low VE levels. 92

Acronyms

aDFT	Adaptive Discrete Fourier Transform
AGAUR	Agency for Administration of University and Research Grants
AG	Accent Group
aHM	Adaptive Harmonic Model
aHM-AIR	Adaptive Harmonic Model with Adaptive Iterative Refinement
AIR	Adaptive Iterative Refinement
ANN	Artificial Neural Networks
APLP	Adaptive Pre-emphasis Linear Prediction
aQHNM	Adaptive Quasi Harmonic plus Noise Model
AR	AutoRegressive
ASR	Automatic Speech Recognition
BWE	Bandwidth Expansion
BMM	buried Markov models
CBR	Case Based Reasoning
CMOS	Comparative Mean Opinion Scale
CSTR	Centre for Speech Technology Research
DFKI-LT	German Research center for Artificial Intelligence - Language Technology Lab
DFT	Discrete Fourier Transform

DFW	Dynamic Frequency Warping
DNN	Deep Neural Network
DSP	Digital Signal Processing
DTW	Dynamic Time Warping
ESS	Expressive Speech Synthesis
f_0	Fundamental Frequency or Pitch
FChT	Fan-Chirp Transform
FFT	Fast Fourier Transform
FT	Fourier Transform
GMBM	Gaussian Mixture Bigram Models
GMM	Gaussian Mixture Models
GPMM	Grup de Processament MultiModal
GTAM	Grup de Tecnologies Audiovisuals i Multimèdia
GTM	Grup en Tecnologies Mèdia
GUI	Graphical User Interface
HCI	Human Computer Interaction
HM	Harmonic Models
hammI	Hammarberg Index
HNM	Harmonics plus Noise Model
HNR	Harmonic-to-Noise Ratio
HMM	Hidden Markov Models
HMM-TTS	Hidden Markov Model based Text-To-Speech
HSMM	Hidden Semicontinuous Markov Models
IIR	Infinite Impulse Response

LP	Linear Prediction
LPC	Linear Prediction Coefficient
LSF	Linear Spectral Frequencies
LSP	Line Spectrum Pair
MAP	Maximum A Posteriori
MFCC	Mel Frequency Cepstral Coefficients
ML	Maximum Likelihood
ML	Machine Learning
MLLR	Maximum Likelihood Linear Regresion
MOS	Mean Opinion Score
MSc	Master of Sciences
NLP	Natural Language Processing
OLA	Overlap-and-Add
pe1000	Relative amount of energy above 1000 Hz
Ph.D.	Philosophiae Doctorate
PP	Peak Picking
PSOLA	Pitch Synchronous OverLap-and-Add
QHM	Quasi-Harmonic Model
aQHM	Adaptive Quasi-Harmonic Model
eaQHM	Extended Adaptive Quasi-Harmonic Model
SAMPA	Speech Assessment Methods Phonetic Alphabet
SE	Spectral Emphasis
SPL	Sound Pressure Level
STASC	Speaker Transformation Algorith using Segmental Codebooks

STFT	Short Time Fourier Transform
TD-PSOLA	Time Domain Pitch Synchronous Overlap-and-Add
TRUE	Testing platfoRm for mUltimedia Evaluation
TTS	Text-To-Speech
UAB	Universitat Autonoma de Barcelona
US	Unit Selection
US-TTS	Unit-Selection Text-To-Speech
VE	Vocal Effort
VoQ	Voice Quality
VT	Voice Transformation
WGN	White Gaussian Noise

CHAPTER 1

Introduction

Contents

1.1 Research group 1
1.2 Motivation 3
1.3 Objectives 4
1.4 Contents 8

The research presented in this dissertation has been conducted as a prerequisite for the completion of the Philosophiae Doctorate (Ph.D.) program *"IT and its management* from *Universitat d'enginyeria i arquitectura La Salle* which belongs to *Universitat Ramon Llull* in Barcelona, Spain. The research has been conducted inside the research group Grup en Tecnologies Mèdia (GTM), which belongs to the Department of Communications and Signal Theory of La Salle Campus Barcelona (Ramon Llull University), with official reference *2009-SGR-293*.

1.1 Research group

The author enrolled in the GTM, from the Engineering Department in La Salle from Universitat Ramon Llull, which comes from the fusion of two prior recognized research groups: Grup de Processament MultiModal (GPMM), with official reference *2005-SGR-00806*, and the Grup de Tecnologies Audiovisuals i Multimèdia (GTAM), with official reference *2005-SGR-00682*. The research group involves different multimedia technology fields such as signal processing (in speech, audio, and image), the transmission of multimedia data, synthesis with graphics and 3D animation, and the evaluation of the final user experience on using the multimedia products.

The author has developed his research inside the speech processing area in the Human Computer Interaction (HCI) research line inside the department of Comunicacions i Teoria del senyal. The group began the study of speech synthesis in the mid 80's with the work of Josep Martí [Martí, 1985]. Different synthesis techniques such as articulatory, formant and linear prediction based synthesis were studied. Following the appearance of Pitch Synchronous OverLap-and-Add (PSOLA) [Moulines and Charpentier, 1990], a few years later, the group changed to concatenative synthesis implementing a Catalan synthesizer [Camps et al., 1992, Guaus et al., 1996] based on the PSOLA technique. With the correct level of intelligibility of the synthesis systems, a parallel research line was opened. It was focused on conveying emotions in the speech synthesized. In collaboration with the Publicity and Audiovisual department of the Universitat Autonoma de Barcelona (UAB) a new research line was established in the study of acoustic modeling of expressions [Rodríguez et al., 1999] using a text-to-speech conversion system [Rodríguez et al., 1999, Iriondo et al., 2000]. Next a hybrid system with Harmonics plus Noise Model (HNM) and PSOLA for improving quality was proposed [Iriondo et al., 2002]. The author has been working in the development of the digital signal processing block based on HNM synthesis which was first presented in his master thesis [Calzada, 2008]. The HNM library was implemented with the Matlab® framework. The system has evolved since then with several improvements which solve different problems such as acoustic quality and flexibility for prosodic and, most recently, Voice Quality (VoQ) modifications. This system has been used for several research projects such as:

- **CreaVeu:** The aim was to create a speech synthesis system capable of producing customized voices with a Graphical User Interface (GUI). The project code was funded by Agency for Administration of University and Research Grants (AGAUR) under the code *2010-VALOR-00164* [Creaveu, 2013].

- **Reune-T:** The main objective of Reune-T is the creation of a collaborative virtual environment. This application allows hosting virtual meetings with a great sense of reality. Reune-T interface in several platforms, such as PC and mobile. The project was funded by Ministry of Science and Innovation from Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica, 2008-2011 with the code *PPT-430000-2008-36* [Alías, 2011].

- **CuentaCuentos 2.0:** CuentaCuentos 2.0 is a customized multi-platform system (computer, mobile phone, and interactive TV), which allows sharing and publishing tales. The project was funded by Spanish Industry, Tourism and Trade Council Plan Avanza I+D, subprogram Avanza Contenidos with code *TSI-070100-2008-19* [Alías, 2008].

For more detailed information about the evolution of the GTM's Text-To-Speech (TTS) can be found in [Alías and Iriondo, 2002, Formiga et al., 2010].

During his Ph.D. the author of this dissertation realized two stays abroad collaborating with two European research institutions: German Research center for Artificial Intelligence - Language Technology Lab (DFKI-LT) Saarbrücken, Germany, in 2011 and Centre for Speech Technology Research (CSTR) University of Edinburgh in 2013. Both stays were funded by AGAUR with the grants *2010 BE1 00503* and *2011 BE1 01084*. Likewise, AGAUR also has funded the research carried on during the three Ph.D. years with FI grants from *Programa d'ajuds per a la contractació de personal investigator novell (2010 FI_B 01083, 2011 FI_B1 00023* and *2012 FI_B2 00152).*

1.2 Motivation

Nowadays technology constitutes an important part of our lives, we constantly take advantage of technology in all of its fields, from small digital cameras to huge communication networks. Technology, usually, makes our life easier, but sometimes the interaction with it might not be comfortable or the user experience is not natural, thus preventing the technology expansion. Many efforts have been made to improve the technology interfaces in order to make the interaction with devices, or automated services, more comfortable and similar to human-to-human interaction. The main communication method used in this human-to-human interaction is speech. Thus, many efforts have been made in the field of speech processing during the past years. Many improvements have been made from the first speech synthesis system, whose main purpose was to be able to get a machine talking in a intelligible way, to the ultimate cutting-edge speech technologies. Moreover, there is recognised demand for speech technologies in the health environment [Ebert, 2011, Rupal, 2013] where these techniques present a strong potential for quality of life improvement for speech impaired persons, the reason why some projects have been started on this topic [Jreige et al., 2009, Yamagishi et al., 2012, Matsui et al., 2013]. But whereas the intelligibility of the speech synthesis systems has currently reached an excellent level, the resulting synthesis is still far from natural and lacking of "human" component [Campbell, 2005].

In order to synthesize more expressive speech, current technology uses either huge databases with a specific corpus for each expressive style, or application of signal processing techniques to modify the signal characteristics based on established rules. The former approach is only used in Unit-Selection Text-To-Speech (US-TTS) synthesis systems, which is the framework developed in the research group, whereas the latter can be applied in many synthesis techniques. The quality obtained with US-TTS systems depends directly on the database signal quality and diversity, thus large databases for each style are needed in order to assure high quality synthetic speech. This requirement is not a feasible option when working with limited resources, and that brings the motivation to explore signal modification techniques to be able to synthesize multiple speech styles from one unique database. But as pointed out in [Arslan, 1999, Taylor, 2007, Huang et al., 2001, Toda, 2003], these modifications on the signal usually degrade the overall quality. The author believes that with a parametric representation of the speech signal, such as Harmonics plus Noise Models (e.g. [Stylianou et al., 1997, Stylianou, 2001, Banos et al., 2008]), and an accurate understanding of the model and its correct use, the problem can be efficiently solved.

While many studies focus on prosody [Murray et al., 2000, Stallo, 2000], others rely on the importance of other features such as voice quality (VoQ) [Heuft et al., 1996, Rank and Pirker, 1998, Drioli et al., 2003]. The author participated in [Monzo et al., 2010] where HNM was used to modify not only prosody but also VoQ in expressive speech for US-TTS. This first attempt encouraged the author to continue working with HNM as a speech signal representation for enhancing the quality of Expressive Speech Synthesis (ESS) systems using *VoQ* parameters following the current trend [Drioli et al., 2003, Kim et al., 2006, Banos et al., 2008, Erro et al., 2009, Türk and Schröder, 2008, Degottex and Stylianou, 2012, Kafentzis et al., 2014b].

1.3 Objectives

Figure 1.1 shows the TTS research approaches for enhancing synthetic speech signal quality and system flexibility in order to achieve a perfect unconstrained synthesis. The axis *Task difficulty* can be understood as the adaptation to multiple domains, such as in [Alías et al., 2008], or to other forms of expressiveness other than neutral expressive style, as used in this dissertation. Some approaches focus on the speech corpus construction such [Alías et al., 2008] where the flexibility of the system was improved by means of incorporating new domain-specific sub-corpora to the system, or in [Golipour et al., 2013] where the prosody realisation of the units in the corpus database is extended by modifying the units with signal processing techniques and then incorporating the processed units back into the corpus. The research presented in this dissertation also contributes in the pathway of improving the flexibility of future ESS. However, instead of improving the flexibility by increasing the corpus size, the proposed approach is based on applying signal processing-based schemes for conducting vocal effort (VE) modifications. Vocal Effort (VE) is defined in section 2.3, and in this work it has been represented by means of slow variations of the spectral envelope, which has a very close relation with voice quality characteristics of speech. Moreover, VE is closely related to activation in the emotion dimension [Schröder and Grice, 2003].

The quality and flexibility of US-TTS systems depends largely on the speech corpus size (see section

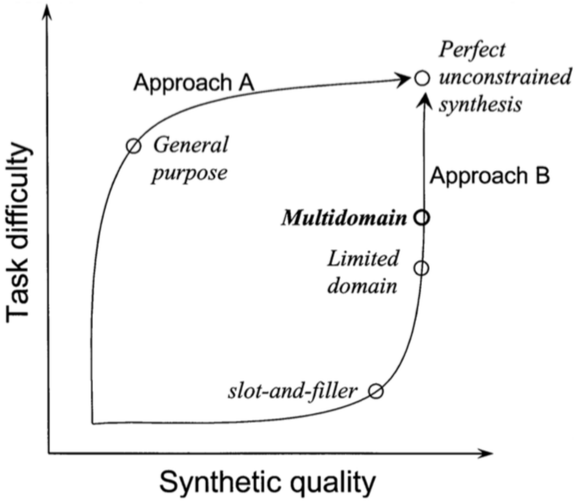


Figure 1.1 — Graphical representation of the different approaches to TTS research towards perfect unconstrained speech synthesis, as a function of the task difficulty and the obtained synthetic speech quality. From [Alías et al., 2008]

2.1.3 for deeper explanation). This means that in order to improve the output synthetic speech quality or to enhance its expressiveness (e.g. increasing the number of expressive styles to synthesize), the corpus sizes can be increased in order to incorporate new realizations of specific units that were not previously well represented (reducing the presence of acoustic artefacts in the synthetic signal) and also new templates of the new required expressiveness (e.g. new speaking styles that the TTS can imitate) [Alías et al., 2008]. Instead, the use of signal processing techniques to modify both the prosody and specific spectral properties of an original neutral speech signal can be an alternative proposal to increase the expressiveness. Hence, this work is focused on proposing new schemes based on signal processing in conjunction with new models that allow the modification of the essential properties of speech that convey expressiveness.

The goal of this dissertation is to contribute to the flexibility of future ESS systems based on Unit Selection (US) by proposing a new architecture of a Text-To-Speech (TTS) and the main signal processing procedures for conducting the speech modifications that will finally extend the system's flexibility. After some preliminary studies that have been also incorporated in this work, this thesis is focused on developing a signal processing technique that permits controlling the vocal effort (VE) level in the synthesized speech signal. In order to maximize the flexibility and control of these speech signal modification techniques, it is desirable to work with a parametric representation instead of directly using the speech signal. The use of a parametric model will enable developing TTS user interfaces that easily allow fine control of the expressiveness degree in the synthesized speech by controlling multiple speech parameters (for instance VE or aspiration noise) via sliders as was performed in [Creaveu, 2013]. To that end, in this work the first step was to validate a model which was appropriate for conducting speech signal modifications without significant quality degradations.

Then, the main goal of this dissertation is: to **propose an architecture introducing signal processing procedures that provides new expressive capabilities for a US-TTS that uses a single speech corpus with neutral expressivity**. From this, the following research questions arise:

Q1. Is the Harmonics plus Noise Model (HNM) appropriate in the aim of conducting the necessary speech signal processing-based modifications to convey expressiveness in the output signal when the source signal has a neutral expressiveness? Moreover, can the inclusion of VoQ in addition to prosodic attributes in the pipeline for transforming neutral speech into more expressive signals improve the perception of the conveyed expressiveness?

Q2. Is it possible to transfer VE levels between two parallel signals using a spectral envelope representation based on HNM?

Q3. Is it possible to synthesize VE levels different from the ones available in the speech recordings used for modelling purposes and controlling the VE level using a simple parameter?

Q4. Which are the main modifications of a US-TTS block diagram that can lead to obtain a more flexible system, able to control the degree of delivered expressiveness using a neutral speech units database?

As explained in the previous section 1.1, the author developed an HNM library in his Master of Sciences (MSc) studies following the trends in the speech synthesis field. This fact induced the author to test if the Harmonics plus Noise Models were suitable for modifying other speech characteristics than prosody, for instance voice quality. First, some preliminary experiments permitted us to prove that HNM was an appropriate mathematical speech model to convey expressiveness, and that the modification of VoQ can produce better results than modifying only prosodic attributes (see Chapter 3). Secondly, procedures based on HNMs for transferring and interpolating VE were proposed (see chapters 4 and 5, respectively).

In order to answer the proposed research questions (*Q1*-*Q4*) aiming to design a new signal processing scheme that improves the flexibility of US-TTS systems, this dissertation was based on the following hypotheses which were validated with the proper experiments presented:

H1. Harmonics plus Noise Model (HNM) is a suitable speech representation for conducting VoQ modifications for expressive speech synthesis, and VoQ can be added as an extra acoustic feature to improve the expressiveness in synthetic speech.

H2. HNM can be used for transferring VE between two original recordings with different VE levels.

H3. A parametric VE model based on the HNMs could be used for synthesizing other VE levels than the ones present in the utterances available in the speech corpus.

H4. The flexibility of a US-TTS that uses a neutral speech corpus can be increased providing an output speech with an expressiveness that can be controlled using the HNMs and the required signal processing for modifying VE.

For validating each one of the previous hypothesis (*H1-H4*) the following objectives were propounded:

O1. Conduct several experiments for transforming neutral speech into different expressive speech styles using the HNMs. In these experiments compare transformations produced modifying only prosodic attributes against transformations where a preselected set of low-level VoQ parameters are added to the transformation pipeline (in addition to prosodic parameters).

O2. Represent VE with a low-order AutoRegressive (AR) filter applied to the harmonic component and use this representation for exchanging VE between two parallel utterances using HNMs. Verify that the VE transfer can be conducted for all the VE levels available in the corpus.

O3. Propose a parametric model with easy and stable interpolation properties for modelling and modifying the spectral envelope. Use the parametric model for modifying VE, transferring VE levels between different speech utterances, and evaluate the obtained results. Finally, generate intermediate VE models by means of using the utterances with the available VE levels in the speech database interpolating the model's parameters and use them for synthesis, and perform the required assessments.

O4. Obtain a more flexible US-TTS system that, using a neutral corpus of speech units, allows the user to easily control the VE level conveyed in the synthesized signal.

1.4 Contents

This dissertation is structured as follows:

- In this first chapter the enrolled Ph.D. program and the joined research group have been briefly described. Next, motivations for the chosen topic have been exposed and the chapter ends listing the main objectives to accomplish and the hypotheses assessed in the following chapters.
- **Chapter 2** introduces the basic concepts and the state of the art of the field explored in this work with a brief explanation of the main speech synthesis techniques, the approaches for synthesizing expressive speech styles and the structure of a US-TTS system. Next, the concept of vocal effort is introduced and a review of the state of the art regarding vocal effort modifications and harmonic models is presented. The chapter concludes presenting the Harmonics plus Noise Model implementation used in this dissertation.
- **Chapter 3** develops objective *O1*, explaining a first attempt of using HNMs for conducting prosody and Voice Quality modifications for transforming from neutral to expressive speech, validating hypothesis *H1*. The obtained results demonstrate the capabilities of HNM for conducting VoQ modifications and the improvement in the perceived expressiveness when combining VoQ with prosody modifications in synthesised speech, answering the research question *Q1*.
- **Chapter 4** develops objective *O2*, presenting a procedure for transferring vocal effort between two parallel utterances with different VE levels using HNMs and a copy-synthesis approach, thus answering the research question *Q2*. The proper procedure is evaluated with subjective listening tests. The obtained results are exposed and analysed, validating the hypothesis *H2*. The chapter concludes with a discussion of the proposed methodology and its limitations.
- **Chapter 5** develops objective *O3*, presenting a new methodology, which can be applied for transferring VE between non-parallel utterances, and a parametric model for modifying vocal effort which extends the flexibility of the proposed methodology explained in chapter 4. The performance of the proposed methodology in this chapter is compared against the methodology introduced in the previous chapter, and also its capability for interpolating vocal effort levels is evaluated, which validates the hypothesis *H3* and thus answers research question *Q3*.
- **Chapter 6** discusses the results obtained in this dissertation and concludes, addressing objective *O4* by proposing a US-TTS system design incorporating the modular solutions presented in the previous chapter. Finally, further research to be conducted is proposed as a means to extend the insights reached in this Ph.D. work.

- **Appendix A** explains the Harmonics plus Noise Model (HNM) implementation used in the experiments presented in this dissertation. The notation used along all the chapters is based on the definitions present in this appendix.

CHAPTER 2

Background

Contents

2.1 **Synthesis techniques** 12

2.1.1 Formant synthesis 12

2.1.2 Diphone concatenation 14

2.1.3 Unit-selection synthesis 16

2.1.4 Statistical synthesis 21

2.2 **Expressive speech synthesis** 24

2.2.1 Corpus-based approaches 25

2.2.2 Mapping codebooks 26

2.2.3 Frequency warpping functions 27

2.2.4 Artificial Neural Networks 28

2.2.5 Probabilistic Linear transformations 28

2.2.6 Rule based conversion 29

2.2.7 Speaker interpolation 29

2.2.8 Hidden Markov Models 31

2.3 **Vocal effort** 31

2.3.1 Definition 32

2.3.2 Spectral behavior and Voice Quality (VoQ) 33

2.4 **Adaptive Pre-Emphasis Linear Prediction for VE modification** 35

2.5 **Sinusoidal and harmonics models overview** 37

2.6 GTM's current US-TTS implementation and scope of this dissertation . 43

This section presents the basic concepts and background knowledge relevant for the different areas related with the presented dissertation.

Section 2.1 lists different techniques to produce speech signals. Next, in section 2.2 the different approaches to produce expressive speech are explained with their benefits and drawbacks. In section 2.3 the concept of vocal effort (VE) is introduced, defining it and explaining its relation with this dissertation. Section 2.5 reviews the evolution of the Harmonic Models (HM) up to their current state of the art situation. Finally A explains the implementation of HM used in the experiments presented in this thesis.

2.1 Synthesis techniques

Many studies have listed the importance of speech parameters such as *Fundamental Frequency or Pitch (f_0)* mean and range, Voice Quality (henceforth VoQ) or articulatory precision, among others, that allow us to model specific emotions in speech [Burkhardt and Sendlmeier, 2000]. Next, the most used techniques for speech synthesis will be reviewed, paying attention to the synthesized signal quality and system requirements. The following section focuses in more detail on the different approaches for conveying emotion and expressiveness in the synthesized speech. Several techniques available in the literature are reviewed, from specific emotion corpus unit selection techniques to the most cutting-edge voice conversion methods, which have already been used for converting not only identity between speakers but also spectral characteristics correlated with emotions for a given speaker.

2.1.1 Formant synthesis

Formant synthesis is also known as **synthesis-by-rule** [Schröder, 2001, Taylor, 2007]. This term is used to emphasize that at run time during the process of synthesizing the speech signal no human recorded waveform signal is involved. In this technique, the synthesis process is entirely done with parametric data and generating synthetic excitation signals [Schröder, 2004]. Formant synthesis is based on a source-filter model. The parameters are usually formant frequencies and their bandwidths which specify the filter characteristics along the signal generation process, thus defining the time evolution of the filter response, and pitch for the excitation signal, which is a periodic pulse for voiced sounds and white noise for obstruent and unvoiced sounds [Taylor, 2007]. The excitation signal is modified via mean pitch value and pitch range. Formant synthesis usually models the oral and nasal cavities separately, thus the excitation signal passes through vocal tract filters and if needed also across filters modelling the nasal cavity, for example for nasal sounds [Taylor, 2007]. These filters may be arranged together either in cascade or in parallel mode. In figure 2.1 Klatt's formant synthesis software schema is presented. Klatt's software implemented both architectures, serial (or cascade) and parallel. Finally the outputs are combined and passed through a radiation component which simulates the load and propagation characteristics of the lips and nose.

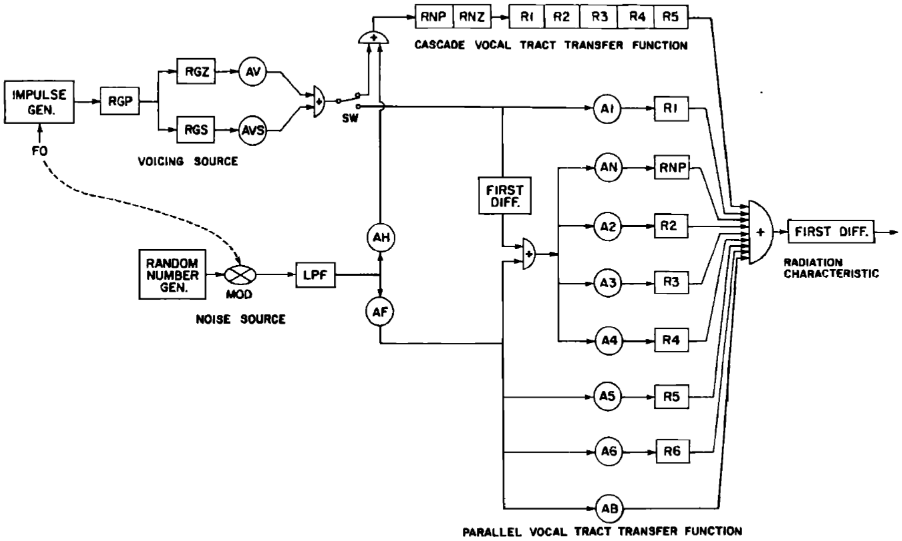


Figure 2.1 — Klatt's formant synthesis software schema from [Klatt, 1980].

The first formant synthesizers were developed by Walter Lawrence, with the parallel formant synthesizer named *Parametric Artificial Talker (PAT)* in 1953, and the second one developed by Gunnar Fant from KTH with the cascade formant synthesizer called *OVE II*. But the extended use of formant synthesis arrived in the early eighties with Dennis Klatt. In 1979, with Jonathan Allen and Sheri Hunnicut, he developed the MIT system named *MITalk*. Two years later they introduced the famous *Klattalk* which improved the voicing source. This formed the basis for the most commercial system used in the twentieth century, *DECtalk*. An example of DECtalk can be found in Stephen Hawking's speech synthesis system. Although the system that Hawking uses is a deprecated version with good intelligibility but low quality synthesis, he keeps that version because the sound of that voice has been extensively associated with him thus giving to that synthesized signal sound a specific identity.

Formant synthesis has been recognized to allow high degree control. The intelligibility of the synthetic signal is more than acceptable but the overall audio quality does not sound natural. Formant synthesis systems usually sound "robot-like" [Schröder, 2004, Huang et al., 2001, Taylor, 2007]. The main reasons for this lacking of naturalness are primarily attributed to a simplistic source model and missing speech dynamics information [Taylor, 2007]. Although it has been proven that formant synthesis systems may reach extremely high quality, even to be impossible to distinguish from the original speech [Klatt, 1987], this requires great effort of manually tuning the system parameters [Huang et al., 2001, Taylor, 2007]. Despite its “robot like” quality, formant synthesis techniques are still used due to their flexibility [Burkhardt, 2009]. Some implemented systems based on formant synthesis techniques are: *EmoSyn* [Burkhardt, 2015b], *AffectEditor* [Cahn, 1990], HAMLET [Murray, 2008] and AmhTTS [Anberbir and Takara, 2009]. KTH has a free downloadable multiplatform vowel formant synthesizer available at their website [Beskow, 2015].

Formant synthesis was relegated to a second place over 1985 by diphone concatenation techniques, which will be discussed in the following section. However some work has been done to fuse format synthesis and diphone concatenation [Carlson, 2002, Öhlin and Carlson, 2004].

2.1.2 Diphone concatenation

In 1968 Rex Dixon and David Maxey created the first approach in concatenative synthesis with diphones parametrized with formants. In 1977 Joe Olive and his colleagues at Bell Labs used linear prediction diphones. A few years later, in 1985, with the development of the Pitch Synchronous OverLap and Add (PSOLA) prosody modification technique, by France Telecom’s Charpentier and Moulines, concatenative systems begun to increase their presence in scientific community [Huang et al., 2001, Hamon et al., 1989]. Nowadays it remains one of the most used techniques.

In the previous section we mentioned the unnatural sounding of formant synthesis pointing to a simplistic source model and the complexity for managing the dynamics of the filter parameters as the main causes [Taylor, 2007]. In diphone concatenation synthesis, instead of using a model for both, source and filter, the original waveform is used. The synthesis is done by concatenating original human speaker recordings. Thus, using original speech signals instead of a model, more high quality speech and naturalness can be reached at synthesis. The synthesised signal’s quality directly depends on the quality of recorded speech signal stretches [Zen et al., 2009]. This signal stretches are called units. In diphone concatenation the basic unit is the diphone. A diphone comprises two half-phones. Usually a diphone is understood to be the portion of signal between the centres of two correlative phones [Schröder, 2004]. Therefore a database to store the diphone units is needed. This database must contain enough units to meet all feasible combinations [Mori et al., 2006]. Nevertheless, not all possible combinations must be considered but only those that may occur in spoken language. Obviously the final quality will be reduced in case no units matching the requirements are found. This requirements might be:

- ▷ contain the specific phonemes to be synthesized.
- ▷ reduce concatenation mismatches at unit boundaries, minimizing both spectral (i.e., obtaining smooth and realistic formant trajectories) and prosodic (i.e., smooth and realistic energy and f_0 contours) discontinuities.

The required unit database can be further reduced when working in limited domains, where diphone diversity requirements are relaxed, while maintaining the overall high quality [Huang et al., 2001, Zen et al., 2009, Black and Lenzo, 2000]. However, in order to improve the synthesized speech quality it is mandatory to minimize spectral and prosodic discontinuities.

Two main approaches for minimizing these discontinuities are: i) keep multiple instances for specific diphones, in order to meet spectral and prosodic possible occurrences, or ii) contain units with average parameters as the most representative units, and then apply signal processing techniques to finally reach the spectral and prosodic target requirements. The modification of units with signal processing techniques usually degrades the signal quality [Huang et al., 2001, Taylor, 2007] although it is still higher than formant synthesis [Schröder, 2004]. Diphone concatenation systems usually incorporate a signal processing module that tunes few prosodic parameters (i.e., pitch and duration of units, and sometimes also intensity). Voice quality has been disregarded in the development of this kind of systems [Schröder, 2001]. Some studies mark the importance of voice quality parameters to improve the naturalness of the synthesized signal [Heuft et al., 1996, Rank and Pirker, 1998, Drioli et al., 2003, Monzo et al., 2010] whereas others assure that with prosody modifications are enough [Murray et al., 2000, Stallo, 2000]. The reason for this discrepancy may be because the lack of exhaustive and conclusive studies in emotion and expressiveness of human speech. [Montero et al., 1999] concluded that each emotion is better characterized by a different set of parameters. Thus while Montero considers anger and happiness as segmental emotions (which are related to VoQ), surprise and sadness are classified as prosodic emotions. But prosody and voice quality relevance depends not only on the specific emotion but also on the speaker conveying that emotion. [Schröder, 1999] concludes that each person has its own strategies for conveying specific emotions which may involve duration and intonation.

One of the main problems in formant synthesis systems commented in section 2.1.1 was the lack of accurate time dynamics information. In diphone concatenation systems, parameter dynamics are held in the diphone itself since each unit composes a specific coarticulation [Sproat, 1998]. Coarticulation is the change of acoustic characteristics for a given phone influenced by surrounding phones [Huang et al., 2001]. Thus the unit database must contain at least one record of each feasible coarticulation available. This is one of the reasons for working with diphones instead of working directly with phones. Another advantage of working with diphones is that concatenation discontinuities or artefacts are reduced. If phones are considered as the basic units; then, in the junction of two units, the spectral discontinuity would be considerable since the spectrum of both units would rarely coincide. On the other hand, if we work with diphones, the junction will be formed by two units sharing the same phone, thus being more likely to have a similar spectrum and reducing audible artefacts due to discontinuities. For instance, consider the word *man* with phonetic transcription /m { n/[1]. Each phone contains its own formants. Thus when concatenating the phones corresponding to phonemes /m/ with /{/ and /n/ we will face spectral discontinuities in each boundary. Instead, working with diphones, for synthesizing the word *man* we would have to concatenate /#-m/,/m{/,/{n/ and /n-#/, where # stands for silence. We can see that in diphone synthesis the spectral formants at the boundaries will be more similar due to concatenating the same half phone. But recall that, in this first concatenative approach, similar does not mean equal, because there is only one occurrence of each unit. For example, imagine we want to synthesize the word *casper* and we have units /"k{/ from the word *Canada* and /{s/ from *Fast*. In the junction of both units the phoneme is conceptually the same, /{/, but audio segments do not come from the same signal so, at least, there will be some discontinuities.

In conclusion, the diphone approach offers higher quality than formant synthesis, though control is reduced to prosody modifications such pitch and duration and sometimes intensity. The unit databases can be quite small since only feasible diphone occurrences in spoken language are needed, hence the corpus size may be reduced without compromising synthesis quality (this is especially true when working in restricted domains). Coarticulation dynamics are recorded intrinsically around the center of diphone units. This properties made diphone concatenation one of the most used techniques [Montero et al., 1999, Schröder, 1999, Murray et al., 2000, Türk et al., 2005]. These are some diphone synthesis implementation examples: EmoFilt [Burkhardt, 2015a, Burkhardt, 2005], MARY [DFKI, 2015, Schröder and Trouvain, 2003], CereVoice [Aylett and Pidcock, 2007].

2.1.3 Unit-selection synthesis

As already commented in the previous section, when concatenating stretches of original speech signals, we may encounter mismatches at the unit boundaries, which degrade the overall quality. Using units larger than diphones allow covering longer speech segments with lower unit concatenations. But working with larger units implies a database increase to cover all possible combinations. On the other hand, in diphone concatenation we only kept one instance for a specific diphone and this unit was the best representative unit so a minor modification is required to get the target requirements. But this signal manipulation may degrade the overall quality of the synthesized signal as well [Taylor,

[1] Phonetic descriptions correspond to Speech Assessment Methods Phonetic Alphabet (SAMPA) notation.

2007, Huang et al., 2001, Toda, 2003]. To avoid signal modifications, instead of holding only one realization of each unit, multiple instances of the same unit in different contexts must be kept. For example, consider having multiple recordings of the diphone /"k/, where one would be stressed, the other is located at the beginning of a word, another is placed at the end of a question sentence (where pitch usually rises), and so on. Storing multiple instances of the units, or using larger units than diphones, entails an increase of the database. Therefore two questions arise: which unit size must be chosen? And, storing multiple instances for each unit in our database, which criteria is used for choosing the best unit among them? Implications (and their magnitude) related to modifying unit length and storing multiple instances for each unit in the database will be explained next.

Unit length

As already mentioned, to reduce the number of concatenations in a synthetic speech, a larger unit length is needed. However, this requires an increase not only in unit length but also in the amount of units recorded in order to cover the same number of combinations than using a smaller unit length. Figure 2.2 depicts, for different unit types and lengths, the increase of the number of units required to cover a certain number of surnames. For the specific case depicted in 2.2, the number of units required is proportional to the unit length, thus the minimum requirements are met for diphones while the maximum are for 2-syllable.

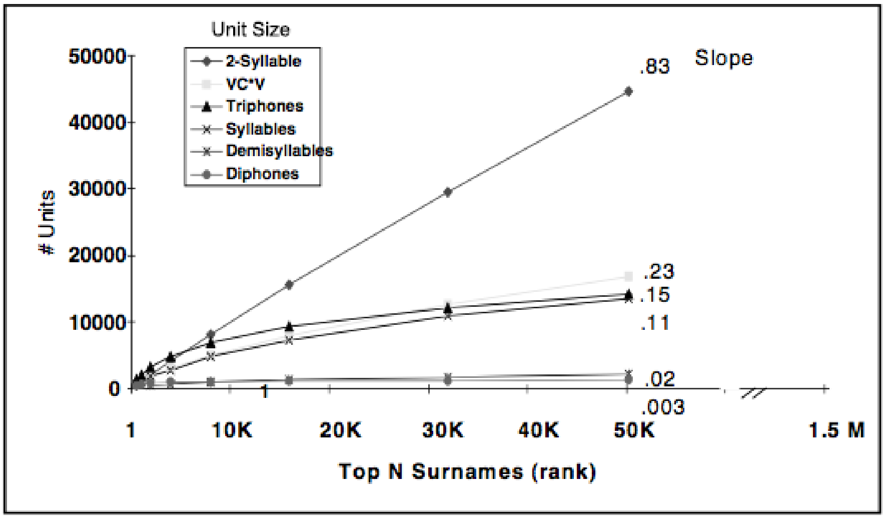


Figure 2.2 — Visual example of unit type impact on word coverage. The figure shows the number of units required for covering a certain number of surnames based on different unit types. Taken from [Macchi, 1998].

Considering a pronunciation system composed by N phones and M syllables, we proceed to list[2] some of the most commonly used unit types, the minimum number of units needed (if only one instance is recorded) and some reference using that unit type :

- ▷ **Diphones:** The unit, formed by two half-phones, begins at the centre of one phone and extends to the centre of the following phone. There are less than N^2 since not all combinations are feasible in the language.
- ▷ **Phones:** There are N units or phones [Hunt and Black, 1996, Saito et al., 1996].
- ▷ **Demi-syllables:** Are the syllable equivalent of half-phones, units either extend from a syllable start boundary to the mid-point of a syllable (the middle of the vowel) or extend from this mid-point to the end of the syllable. There are 2M demi-syllables [Pearson et al., 1998] .
- ▷ **Di-syllables:** The equivalent to di-phones, units that extend from the middle of one syllable to the middle of the next. There are M^2 units [Law and Lee, 2000].
- ▷ **Syllables:** Syllable defined as a vowel sound, a diphthong, or a syllabic consonant, with or without preceding or following consonant sounds. Is a uninterrupted segment of speech. The length and rules depend on the language of speech recordings. The minimum requirement is M units [Matoušek et al., 2005].
- ▷ **Words :** Words defined as a single distinct meaningful element of speech [Vosnidis and Digalakis, 2001].

Unit selection costs

On the other side, a possible solution for reducing, or even avoiding, signal modifications in order to synthesize the output signal from a selected sequence of units is to keep multiple instances for a given unit. However, the next question arises: “How do we choose the better unit among all available instances in the database?”. It can be approached specifying some kind of parameters and measures to be able to rank all the options available. Andrew Hunt and Alan Black proposed a method to evaluate the suitability of a specific unit for a given target [Hunt and Black, 1996]. This method is based in the concept of cost. The proposed method evaluates each unit considering the two major problems in unit selection synthesis. The suitability of the unit for a given target is called **target cost**, $C^t(t_i, u_i)$, and expresses the distance between the unit (u_i) and the given target (t_i). Also, the mismatch produced when concatenating the unit (u_i) with its precedent unit (u_{i-1}) is evaluated with **concatenation**

[2] The complete list with more references can be found in [Taylor, 2007].

cost, $C^c(u_{i-1}, u_i)$, sometimes referred to as joint cost. Usually, each cost is computed as the weighted sum of specific target and concatenation costs measured over a set of p and q parameters, respectively as

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \tag{2.1}$$

$$C^c(u_{i-1}, u_i) = \sum_{j=2}^{q} w_j^c C_j^c(u_{i-1}, u_i) \tag{2.2}$$

, where w_j^t and w_j^c are the cost weights for the p and q parameters involved in target and concatenation costs, respectively. With these weights an adjustment of the relevance of each parameter among the others involved in the computation can be done. The tuning of these weights is one of the crucial points in unit selection systems [Alías et al., 2006, Formiga and Alías, 2007], however the weight tuning is out of the scope of this work. Hunt and Black's most revolutionary contribution is not the calculation of this weights but the cost function concept for selecting the best unit. The total cost for a given sequence of n target-unit (t_i^n, u_i^n) pairs is

$$\begin{aligned} C(t_1^n, u_1^n) &= \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) + \sum_{i=2}^{n}\sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \end{aligned} \tag{2.3}$$

, where S denotes silence and $C^c(S, u_1)$ and $C^c(u_n, S)$ define the start and end conditions given by the concatenation of the first and last units with silence.

So the unit selection task is defined as the selection of a set of units $(\bar{u}_1^n)$ which minimize the total cost defined in (2.3) described

$$\bar{u}_1^n = \min_{u_1,\ldots,u_n} C(t_1^n, u_1^n) \tag{2.4}$$

Unit selection may be considered a general technique and diphone concatenation one of its branches since both share many similarities. Unit selection has become a solution for some of the main problems encountered in diphone concatenation. But it is clear that for either reducing the number of concatenations by enlarging the basic unit length, or minimizing the use of signal processing techniques to reach the target requirements by storing multiple instances for each unit, in unit selection synthesis the database increases considerably compared to diphone concatenation systems, which constitutes one of its main drawbacks [Barra-Chicote et al., 2010]. Also, the quality of unit selection relies on being able to find the suitable unit in the database, and when this does not happen the overall quality decreases considerably. This entails that the quality of unit selection synthesis depends principally on the database, the bigger the database, the better the quality achieved [Banos et al., 2008, Black et al., 2007, Schröder, 2001, Zen et al., 2009, Golipour et al., 2013]. Thus, building the database requires special effort to design it carefully in order to keep it as small as possible without losing generalisation [Schröder, 2004]. Some people even minimize the database size on purpose losing generalisation but not quality when working in reduced scopes or limited domains [Alías et al., 2005, Black and Lenzo, 2000]. Campbell, on the other hand, considers that the aspects which make speech natural for an ordinary conversation are not totally fulfilled when acted speech is recorded in the corpus. He points out that research studies concentrate on specific emotions but those emotions do not convey the way people communicate with each other the major part of the day. He considers that more subtle expressions are used than those typically studied, the "big six" [Cornelius, 2000]. Also, he takes into consideration the "observer's paradox" which states that people behave differently when faced with a microphone and thus he has built a specific database with everyday conversational recordings in which the subject is given a microphone which records everything the subject says, more or less like the famous television show Big Brother. [Campbell, 2004, Campbell, 2005]

The unit size has been changing depending on the purpose. For synthesis, many people uses diphone units [Monzo et al., 2008, DFKI, 2015], but in other scenarios such as mapping sounds between two languages for speech-to-speech translation, working with phonemes is not suitable since depending on the languages, there might not be a clear relation. Otherwise with words, there would be many mismatches, thus Accent Group[3] (AG) may be a more suitable type of unit to be considered for speech synthesis [Agüero et al., 2006].

Some of the most well known systems which implement unit selection strategies are Loquendo [Loquendo, 2015], EmoFilt [Burkhardt, 2015a, Burkhardt, 2005], WebSphere from IBM [Websphere,], MARY TTS [DFKI, 2015], and CereVoice [Aylett and Pidcock, 2007].

[3] In English, AGs are defined as the stressed syllable and all preceding non-stressed syllables. In Spanish and Catalan AG is defined as the content word and all the preceding function words. [Agüero et al., 2006]

2.1.4 Statistical synthesis

In concatenative speech synthesis systems, huge databases are usually built to store the units which later will be concatenated in the synthesis process. So this approach can be considered a *memorizing* and unit sequence *ordering*. Instead of memorising data, statistical synthesis methods try to generalize certain knowledge changing the paradigm of *memorising* to *learning*. This *learning* is done through statistical machine learning techniques.

Hidden Markov Models (henceforth HMM) are by far the most widely used statistical modelling framework in speech synthesis. Their use in the speech processing field began in the mid 80's and was focused on the recognition paradigm [Schwartz et al., 1984]. It was first introduced in speech synthesis by Tokuda [Tokuda et al., 1995]. This approach resulted in unacceptable quality but the flexibility of the model made it worth investigating. Since that first approach many studies have been carried out improving the different parts of the system which is depicted in figure 2.3.

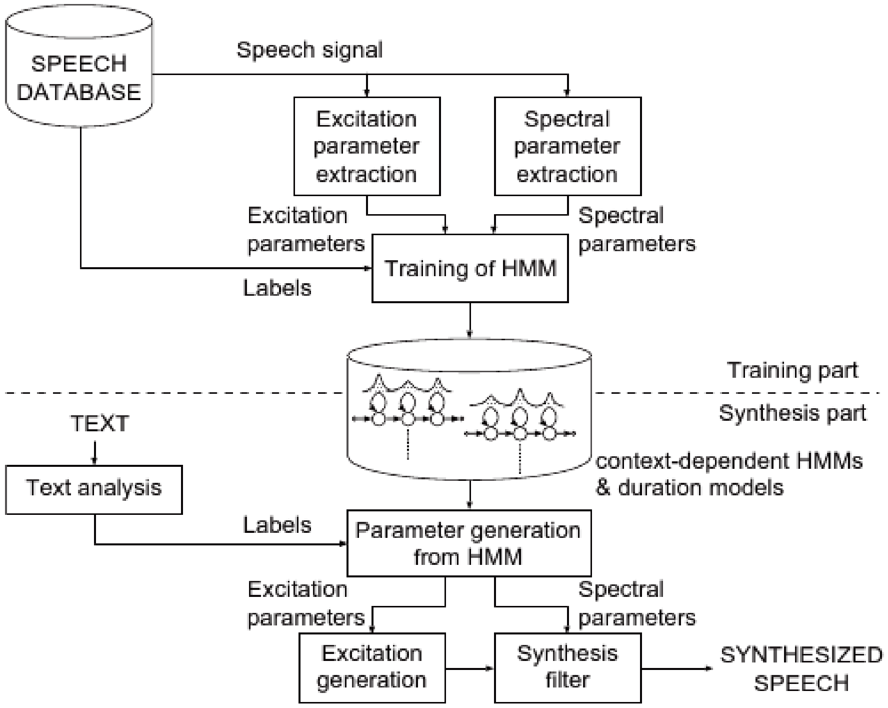


Figure 2.3 — Block diagram of an HMM synthesis system from [Zen et al., 2009]

Statistical synthesis systems, as can be seen in figure 2.3, can be divided into two main stages, the training and synthesis. In the training stage, the first step is to extract parameters from the signal using any parametric representation (MFCC [Black et al., 2007],HNM [Banos et al., 2008], Line Spectrum Pair (LSP) [Kim et al., 2006], f_0 [Chen et al., 2014]). These parameters are used to extract models by use of generative techniques such as HMM [Black et al., 2007], Hidden Semi-continuous Markov Models (acshsmm) [Yamagishi and Kobayashi, 2007] or buried Markov models (acsbmm) [Bulyko et al., 2002] among others. Usually a maximum likelihood (ML) criterion is used to estimate the model parameters as:

$$\hat{\lambda} = arg \max_{\lambda} \{p\,(O \mid \mathcal{W}, \lambda)\} \tag{2.5}$$

where λ is a set of model parameters, O is a set of training data (referring to the aforementioned speech signal parameters) and $\mathcal{W}$ is a set of word sequences corresponding to O. Then, at synthesis time, the generation of the speech parameters ($\hat{o}$) for a given word sequence to be synthesized (w) from a set of estimated model parameters ($\hat{\lambda}$) is produced with

$$\hat{o} = arg \max_{o} \left\{p\left(o \mid w, \hat{\lambda}\right)\right\} \tag{2.6}$$

And finally, the speech waveform is generated from the estimated parameters with the suitable technique, usually a vocoder-based reconstruction using the excitation-filter approach of speech production.

The main advantages of HMM synthesis systems are:

▷ **Flexibility:** Voice characteristics can be easily modified by appropriately changing the parameters of the model. This has been used for speaker adaptation [Masuko et al., 1997, Tamura et al., 2001], speaker interpolation [Yoshimura et al., 2000] or eigenvoice technique [Shichiri et al., 2002].

▷ **Multiple language support:** Using the core system for synthesizing in multiple languages is easily achieved because only contextual factors to be used depend on each language and can be learned during the training phase. Up to now, many languages have been used in HMM-based speech synthesis (HMM-TTS); some of them are: Arabic [Abdel-Hamid et al., 2006], Catalan [Bonafonte et al., 2008], or Spanish [Gonzalvo et al., 2006].

▷ **Multiple expressive style:** This can be done by re-estimating the existing average voice models with only a few utterances of another, more expressive voice, using adaptation techniques [Tachibana et al., 2008]. The most used are Maximum a Posteriori (MAP) [Gauvain and Lee, 1994] and Maximum Likelihood Linear Regresion (MLLR) [Leggetter and Woodland, 1995].

- ▷ **Take advantage of ASR improvements:** Also, many improvements in Automatic Speech Recognition (ASR) can be applied in synthesis. This includes the use of eigenvoices [Kuhn et al., 2000].

- ▷ **Small footprint:** Statistical synthesis systems require by far fewer resources than unit selection because the acoustic model statistics are stored, rather than multiple templates of speech units. For instance, a system requiring 2 Megabytes without compression was demonstrated in Blizzard Challenge 2005 [Zen et al., 2007].

On the other hand, widely accepted HMM drawbacks are:

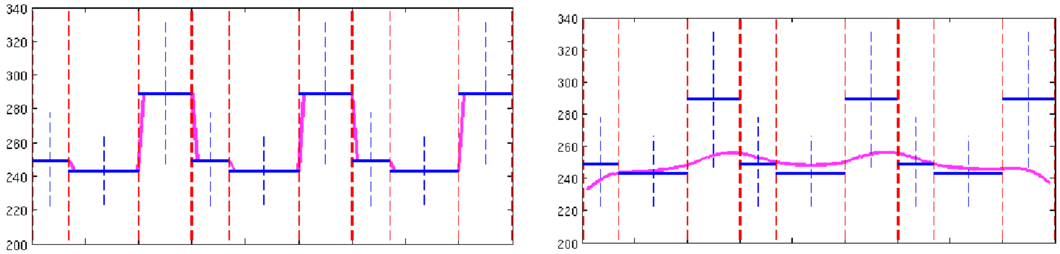
- ▷ **Vocoder sound:** HMM speech synthesis uses a vocoder to generate the final signal waveform. This audio is reported to sound *buzzy*. But recent studies have been done to solve this problem by improving the excitation source modelling [Kim et al., 2006,Banos et al., 2008]. Results show an improvement in the overall synthesized speech quality.

- ▷ **Modelling accuracy:** Although several of the models have been proposed (HSMM, BMM), HMM is still the most widely used because other approaches require an increasing number of model parameters, and also various essential algorithms such as phonetic decision-tree-based context clustering need to be modified to cope with this dynamical approach [Zen et al., 2009].

- ▷ **Oversmoothing:** The speech parameter generation algorithm is used to generate spectral and excitation parameters from the HMM by applying constraints between static and dynamic features. The averaging done in the modelling process increases the robustness of the system against data sparseness. Also, the use of dynamic feature constraints enhances the system by enabling it to generate smooth trajectories in the model junction boundaries as can be seen in figure 2.4. The oversmoothing problem is directly related to a typical muffled and flat sound. Many efforts have been done (i.e., Global Variance [GV], Minimum Generation Error [MGE]) to improve the quality due to the oversmoothing effect.

Some new approaches try to combine the best of both methods, unit-selection and HMM techniques, the so-called hybrid speech synthesis systems [Aylett and Yamagishi, 2008, Banos et al., 2008,Gonzalvo Fructuoso, 2010,Tiomkin et al., 2011,Guner and Demiroglu, 2012,Phung et al., 2013].

Finally, with the recent improvements achieved in the speech recognition field using neural networks, during the last years some systems have tried to incorporate them also for speech synthesis [Zen et al., 2013,Zen and Senior, 2014,Tokuda and Zen, 2015]. [Zen et al., 2013] compared the performance of a speech synthesis system based on Deep Neural Network (DNN) with a HMM-based system with



(a) Discontinuity in model boundaries

(b) Oversmoothed signal

Figure 2.4 — HMM Model junctions examples of (a) discontinuities and (b) oversmoothing

several hidden layer configurations. In the subjective test performed, the subjects were asked for their preferred speech signal being the DNN-based system the preferred one for all the configurations implemented. [Wu et al., 2015] also compared a DNN-based synthesis system with the HMM approach. Wu exposes two main weaknesses of HMM-based speech synthesis systems being: the density function over the acoustic features and the decision-tree driven parameterisation of the model, in which parameters are shared across groups of linguistic contexts. Moreover to comparing DNN with HMM approaches, Wu proposes the use of multi-task learning for a DNN-based speech synthesis system, for additional supervision during the training stage of the neural network, and the use of stacked bottleneck features as additional context features for improving the synthesized speech parameter dynamics. In the conducted test, Wu obtained an objective improvement over baseline DNN and the benchmark HMM system, based on a Mel Cepstral Distortion and V/UV error rate criteria, using multi-task training, but it was not reflected in the subjective measures. On the other hand, using stacked bottleneck features the speech synthesis was perceived as more natural than regular DNN approach.

2.2 Expressive speech synthesis

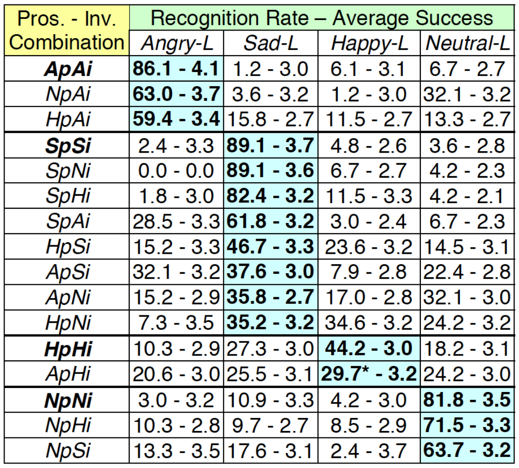
As explained in section 1.2, expressive speech synthesis has been one of the hot topics in speech processing over the last decade [Bailly et al., 2003, Danieli, 2006, Sharma et al., 2013]. The main focus is to make synthesis sound natural. To achieve this purpose, some studies focus on introducing emotion into the final synthesis [Bulut et al., 2002, Eide et al., 2004, Erro et al., 2009, Barra-Chicote et al., 2010]; on the other hand, other studies disagree with the idea of focusing uniquely on emotions [Campbell, 2005], thus they pay more attention to conversational events instead of emotions [Campbell, 2007, Adell, 2009].

From now on we will focus on the emotion paradigm, without underestimating the relevance of conversational events for achieving naturalness. According to the parameters that best convey emotions, the two main trends are VoQ and prosody modifications. Many discussions can be found about whether prosody is able to convey different emotions by itself, or VoQ is really needed for some specific emotions. The basic prosody parameters are pitch, duration, and energy. VoQ on the other hand focuses more on spectral characteristics such as the harmonic-to-noise ratio (HNR), Hammarberg Index or Drop-off of spectral energy above 1 Khz. These VoQ characteristics are briefly explained in section 3.1 together with the methodology developed by [Monzo, 2010] for its computation from an HNM parametrisation. While some studies consider prosody sufficient for emotion representation [Murray et al., 2000, Stallo, 2000], others expose that VoQ can improve emotion recognition in the user [Heuft et al., 1996, Monzo et al., 2010]. These contradictions have been partially explained by Schöder [Schröder, 1999] who explains that people have their own strategies to express emotions; thus someone may focus on speech rate while another speaker may exploit VoQ. This keeps the VoQ topic opened which is one of the explored topics by the author with his collaboration with other members of the GTM in [Monzo et al., 2010] explained in more detail in chapter 3. Other research work focusing on VoQ includes [Drioli et al., 2003, Türk et al., 2005, Türk and Schröder, 2008].

Expressive speech synthesis research can also be classified depending on the used synthesis technique (see section 2.1). For instance, corpus-based techniques mainly rely on the corpus expressiveness for conveying the desired emotion in the final synthesized speech; on the other side, we can find conversion techniques which rely on a powerful signal parametrisation and accurate conversion functions. Next, a brief listing of various approaches for expressive speech synthesis found in the literature will be described, focusing on the capabilities and constraints for conveying emotions.

2.2.1 Corpus-based approaches

Corpus based approaches use unit selection synthesis systems. As explained in section 2.1.3, these systems mainly rely on the corpus to get high quality synthesis, trying to minimise the signal modifications required, which could degrade the overall quality. This principle is also used when trying to apply speech transformation or voice conversion. Different corpora for each emotion are usually recorded following an acted speech strategy, thus the main problems to solve are the validation of the conducted recordings (usually based on subjective recognition rates) and searching for the most suitable unit when synthesising a specific emotion [Eide et al., 2004, Campbell, 2004, Erro et al., 2009]. An example of the relevance of the corpus for unit selection based techniques can be found in [Bulut et al., 2002] where different prosody strategies and types of units used for synthesis are explored. In this work, Bulut et al. explore the multiple combinations of prosody and units selected from neutral and three emotional corpora. In table 2.1 the recognition rate table in his study corroborates that there is high influence of the corpus characteristics in the final synthesis, leaving prosody in a second plane.



Pros. - Inv. Combination	Recognition Rate – Average Success			
	Angry-L	*Sad-L*	*Happy-L*	*Neutral-L*
ApAi	**86.1 - 4.1**	1.2 - 3.0	6.1 - 3.1	6.7 - 2.7
NpAi	**63.0 - 3.7**	3.6 - 3.2	1.2 - 3.0	32.1 - 3.2
HpAi	**59.4 - 3.4**	15.8 - 2.7	11.5 - 2.7	13.3 - 2.7
SpSi	2.4 - 3.3	**89.1 - 3.7**	4.8 - 2.6	3.6 - 2.8
SpNi	0.0 - 0.0	**89.1 - 3.6**	6.7 - 2.7	4.2 - 2.3
SpHi	1.8 - 3.0	**82.4 - 3.2**	11.5 - 3.3	4.2 - 2.1
SpAi	28.5 - 3.3	**61.8 - 3.2**	3.0 - 2.4	6.7 - 2.3
HpSi	15.2 - 3.3	**46.7 - 3.3**	23.6 - 3.2	14.5 - 3.1
ApSi	32.1 - 3.2	**37.6 - 3.0**	7.9 - 2.8	22.4 - 2.8
ApNi	15.2 - 2.9	**35.8 - 2.7**	17.0 - 2.8	32.1 - 3.0
HpNi	7.3 - 3.5	**35.2 - 3.2**	34.6 - 3.2	24.2 - 3.2
HpHi	10.3 - 2.9	27.3 - 3.0	**44.2 - 3.0**	18.2 - 3.1
ApHi	20.6 - 3.0	25.5 - 3.1	**29.7* - 3.2**	24.2 - 3.0
NpNi	3.0 - 3.2	10.9 - 3.3	4.2 - 3.0	**81.8 - 3.5**
NpHi	10.3 - 2.8	9.7 - 2.7	8.5 - 2.9	**71.5 - 3.3**
NpSi	13.3 - 3.5	17.6 - 3.1	2.4 - 3.7	**63.7 - 3.2**

Table 2.1 — Recognition rates in percentage and Average Success ratings (5=excellent and 1=bad) for the 16 possible prosody and inventory combinations from [Bulut et al., 2002]. The nomenclature used is $XpYi$ where p stands for prosody, X is the corpus which prosody model was used, i stands for the input instance and Y is the corpus from where the units were selected during synthesis. For instance, paying attention to angry emotion, selecting prosody from angry corpus ($ApYi$) the emotion conveyed in the synthesis was that of the corpus where the units were selected.

2.2.2 Mapping codebooks

This techniques was first used for voice conversion in [Abe et al., 1988]. The proposed conversion system is based on generating codebooks for spectrum parameters, pitch and energy which will then be used for converting the source speaker's characteristics into the target speaker's ones. The algorithm for generating the codebooks is as follow:

1. Two speakers, source and target, pronounce a learning word set. Then all words are vector-quantized frame-by-frame.

2. Correspondence between vectors of the same words from the two speakers is determined using Dynamic Time Warping (DTW) [Sakoe and Chiba, 1978].

3. The vector correspondences are accumulated as histograms.

4. Using each histogram as a weighting function, the mapping codebook is defined as a linear combination of the target's vectors.

5. Steps 2,3, and 4 are repeated to refine the mapping codebook.

In the conversion step, the source speech is analysed by linear prediction methods, the spectrum parameters are vectors-quantized using its own codebook, and also the pitch and energy are analysed. Synthesis is next done by decoding mapping the source's coded parameters with codebooks between source and target, and finally the waveform is generated.

In 1999, Arslan proposed a new codebook-based conversion method called *Speaker Transformation Algorithm using Segmental Codebooks (STASC)* [Arslan, 1999]. The speech frames are classified by phones or by HMM states. The alignment is carried out by *"sentence HMM"* method which requires phonetically balanced utterances, a more detailed explanation of the method can be found in his study. Linear Spectral Frequencies (LSF) vectors are calculated for both source and target speakers. These vectors are used as the entry for a one-to-one mapping codebook.

An improvement of the method was proposed by Turk and Arslan in [Türk and Arslan, 2006], where some refinements were done in the codebooks to eliminate the outliers. In 2008 another improvement was presented by Turk in [Türk and Schröder, 2008] where a comparison of the previous codebook mapping with a new version based on frame codebook mapping which resulted in quality improvements avoiding averaging of LSF vectors was reported. The vectors used were also LSF directly extracted from the original signals for each phonemes, which results in a increase of the codebook size.

2.2.3 Frequency warpping functions

The dynamic frequency warping (DFW) technique was first introduced in [Valbret et al., 1992]. Given a pair of spectra of a speech frame from source $X(\omega)$ and target $Y(\omega)$ speakers, modelled by their log spectral samples, a function $\omega'(\omega)$, called frequency warping function, is to be found following the criteria of minimizing the spectral distance between $X(\omega'(\omega))$ and $Y(\omega)$. Since the spectrum varies for each phone, a vector-quantization is carried out first and then independent warping functions are defined for each acoustic class. In synthesis the most suitable warping function is selected and applied to the source's log-spectrum.

A combination of DFW and HMM was proposed by [Rentzos et al., 2004]. Before training the system, the phoneme boundaries are extracted by HMM forced-alignment segmentation. Then the formant candidates of each HMM are modelled by other HMM which associated feature vectors contain the frequency, bandwidth, and intensity of each formant. The equation for a converted frame t is

$$Y(\omega,t) = \gamma(\omega,t)\ X[\alpha(\omega,t) * \beta(\omega,t) * \omega, t] \tag{2.7}$$

, where the warping function includes mapping for both formant frequency $\alpha(\omega, t)$ and bandwidth $\beta(\omega, t)$ defined by subbands, and γ maps spectral magnitude.

Although the concept of frequency warping functions is quite simple the resulting quality is very high though conversion properties still have to be improved.

2.2.4 Artificial Neural Networks

Artificial neural networks (ANN) were also used for voice conversion purposes. Their use did not expanded due to the appearance of Gaussian Mixture Models (GMM) techniques which really became a revolution in the voice conversion field [Erro, 2008]. Also a comparative study carried out by [Bandoin and Stylianou, 1996] reported that neural networks performed worse than GMM. This made ANN almost disappear from the speech synthesis field for a long period. However, their use has increased thanks to the recent improvements in the training stage and fast-forward algorithm. With these improvements, the Artificial Neural Networks (ANN)-based voice conversion methods have been evaluated again against Gaussian Mixture Models (GMM) showing a better performance [Laskar et al., 2012, Mohammadi and Kain, 2014]. The use of ANN in speech synthesis has growth, specially in the statistical speech synthesis systems. However, this is out of the scope of this thesis which is focused on detecting the attributes or parameters susceptible to being modified in speech with the HNMs) in order to improve the speech expressiveness beyond voice conversion.

2.2.5 Probabilistic Linear transformations

As already mentioned in the previous section, Gaussian Mixture Models (GMM) became a revolution in the speech conversion field. A first approach was carried out by [Stylianou and Moulines, 1998]. The conversion function is defined as a linear combination of linear mappings between each mixture target component and the input source vector “x”, where the main goal is to establish an efficient relation between feature vectors concerning the vocal tract filter in the excitation filter approach. Mathematically it is defined as

$$F(\bar{x}) = \sum_i p_i(\bar{x}) \left[\bar{v}_i + \bar{\bar{\Gamma}}_i \bar{\bar{\Sigma}}_i^{-1} \left(\bar{x} - \bar{\mu}_i \right) \right] \tag{2.8}$$

, where $p_i(\bar{x})$ is the probability that $\bar{x}$ belongs to the i^{th} Gaussian component. Vectors $\bar{v}_i$ and matrices $\bar{\bar{\Gamma}}_i$ are calculated during the training phase by minimizing the least squares error given by the distance between transformed vectors ($F(\bar{x}_k)$) and the corresponding target vector ($\bar{y}_k$).

An improvement in GMM was proposed by [Kain, 2001] which concatenates the LSF parameters of source and target speakers and then feeds the GMM training system with them. This is commonly known as joint GMM. In this case the conversion function is similar to the one proposed by Stylianou but information of both source and target is gathered in the same model variables obtaining

$$F(\bar{x}) = \sum_i p_i^x(\bar{x}) \lfloor \bar{\mu}_i^y + \bar{\bar{\Sigma}}_i^{yx} \bar{\bar{\Sigma}}_i^{xx^{-1}} (\bar{x} - \bar{\mu}_i^x) \rfloor, \qquad \bar{\mu}_i = \begin{bmatrix} \bar{\mu}_i^x \\ \bar{\mu}_i^y \end{bmatrix} \; \bar{\bar{\Sigma}}_i = \begin{bmatrix} \bar{\bar{\Sigma}}_i^{xx} & \bar{\bar{\Sigma}}_i^{xy} \\ \bar{\bar{\Sigma}}_i^{yx} & \bar{\bar{\Sigma}}_i^{yy} \end{bmatrix} \tag{2.9}$$

GMM outperform all other methods in the balance between quality and conversion degree, however over-smoothing effects are one of their main drawbacks. Many studies have been done in order to solve this problem [Toda et al., 2001]. Also, systems which try to model the temporal and spectral evolution with Gaussian Mixture Bigram Model (GMBM) [Hsia et al., 2007] and a combination of GMBM with k-means and prosody conversion resulted in very good quality and conversion. Other approaches try to combine GMM with HNM, such as [Drioli et al., 2003] which also improved their quality but they did not model the stochastic part which degraded the overall quality. They agree that modelling the stochastic part could bring more quality.

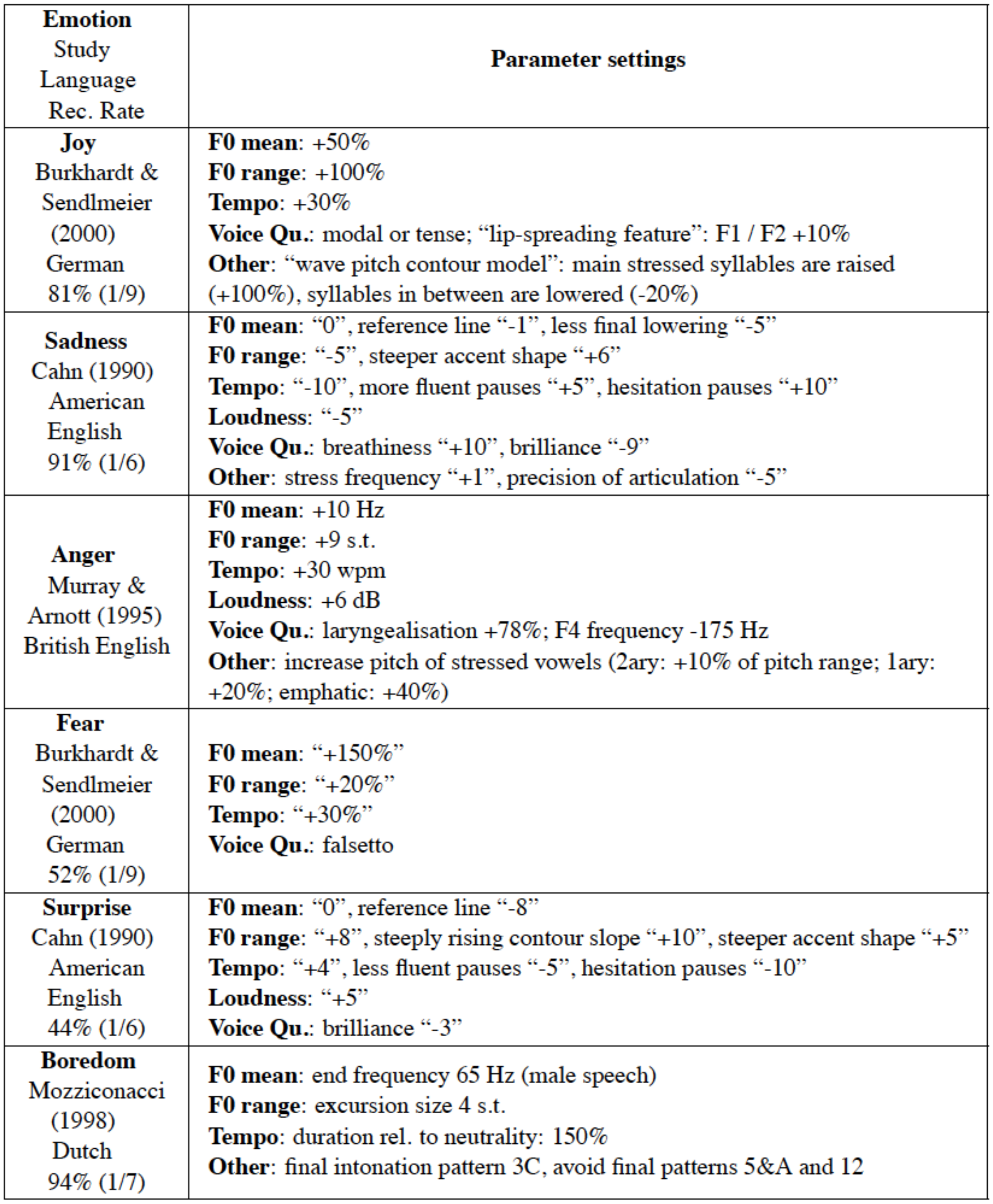
2.2.6 Rule based conversion

These methods try to extract general rules from analysing large amounts of data. Some examples can be [Stallo, 2000, Murray et al., 2000, Iriondo et al., 2000]. Once rules have been extracted, signal modifications are carried out in order to get the expressive speech synthesis. These methods try to concentrate on specific parameters, which should be able to correctly define the given emotion. In table 2.2 a list of different studies and some of the rules they have are shown.

[Monzo et al., 2010] used prosody rules as well from a Case Based Reasoning (CBR) system, and also rules for defining VoQ parameters which best conveyed a specific emotion. VoQ rules were extracted from a previous study carried out about VoQ parameters in the discrimination of expressive speech styles, and heuristic tests performed during the design process to adjust the final parameters involved in the different emotions. Further explanation of the parameters and their use is given in section 3.2.1.

2.2.7 Speaker interpolation

The idea of speaker interpolation was originally proposed by [Iwahashi and Sagisaka, 1994]. Iwahashi and Sagisaka state that it is possible to extract a general purpose speaker model from the combination



Emotion Study Language Rec. Rate	**Parameter settings**
Joy Burkhardt & Sendlmeier (2000) German 81% (1/9)	**F0 mean**: +50% **F0 range**: +100% **Tempo**: +30% **Voice Qu.**: modal or tense; “lip-spreading feature”: F1 / F2 +10% **Other**: “wave pitch contour model”: main stressed syllables are raised (+100%), syllables in between are lowered (-20%)
Sadness Cahn (1990) American English 91% (1/6)	**F0 mean**: “0”, reference line “-1”, less final lowering “-5” **F0 range**: “-5”, steeper accent shape “+6” **Tempo**: “-10”, more fluent pauses “+5”, hesitation pauses “+10” **Loudness**: “-5” **Voice Qu.**: breathiness “+10”, brilliance “-9” **Other**: stress frequency “+1”, precision of articulation “-5”
Anger Murray & Arnott (1995) British English	**F0 mean**: +10 Hz **F0 range**: +9 s.t. **Tempo**: +30 wpm **Loudness**: +6 dB **Voice Qu.**: laryngealisation +78%; F4 frequency -175 Hz **Other**: increase pitch of stressed vowels (2ary: +10% of pitch range; 1ary: +20%; emphatic: +40%)
Fear Burkhardt & Sendlmeier (2000) German 52% (1/9)	**F0 mean**: “+150%” **F0 range**: “+20%” **Tempo**: “+30%” **Voice Qu.**: falsetto
Surprise Cahn (1990) American English 44% (1/6)	**F0 mean**: “0”, reference line “-8” **F0 range**: “+8”, steeply rising contour slope “+10”, steeper accent shape “+5” **Tempo**: “+4”, less fluent pauses “-5”, hesitation pauses “-10” **Loudness**: “+5” **Voice Qu.**: brilliance “-3”
Boredom Mozziconacci (1998) Dutch 94% (1/7)	**F0 mean**: end frequency 65 Hz (male speech) **F0 range**: excursion size 4 s.t. **Tempo**: duration rel. to neutrality: 150% **Other**: final intonation pattern 3C, avoid final patterns 5&A and 12

Table 2.2 — Examples of successful prosody rules for emotion expression in synthetic speech. From [Schröder, 2004]

of previously recorded speakers as

$$\bar{Y}_{ij} = \sum_{k=1}^{N} \omega_k \bar{x}_{kij} \tag{2.10}$$

with the constrain

$$\sum_{k=1}^{M} \omega_1 = 1, \tag{2.11}$$

where $\bar{x}_{ikj}$ represents the j^{th} spectral parameter of the i^{th} frame in the utterance of the k^{th} speaker. Y_{ij} is the j^{th} spectral parameter of the i^{th} frame of the interpolated spectrum.

This technique has also been used for interpolating VoQ parameters in [Türk et al., 2005]. But they only use a linear interpolation of LSF vectors of two corpora. One of the conclusions is that interpolation of residuals is also necessary but it requires good modelling to avoid introducing distortions in the synthesized signal.

2.2.8 Hidden Markov Models

The main disadvantage of HMM speech synthesis is that the synthetic speech signal naturalness is not as high as that achieved by unit selection techniques. But its main advantage is that modifications in the synthesis can be carried out easily, modifying the model parameters without severely affecting the quality [Erro, 2008].

Other proposals in stochastic voice conversion are recently appeared, modelling a generic voice from different voices in the corpus and then tuning model parameters to a personalized target voice that can be synthesized [Toda et al., 2007]. A similar approach is done by extracting eigenvectors to model prosody in [Mori et al., 2006].

2.3 Vocal effort

Vocal effort is a hot research topic in voice disorders and usually used for dysfunction identification and treatment design. In the voice disorders field, vocal effort is often understood as a “strained” voice quality [Kempster et al., 2009], and is also considered a component of vocal hyperfunction, which entails increased muscle stiffness and activity in the vocal tract system [Boone et al., 2013]. Regarding physiology affection, vocal hyperfunction entails an increased adduction of vocal folds [Askenfelt and Hammarberg, 1986], increased constriction of structures above the vocal folds [Mayerhoff et al., 2014], and other muscle activations. The muscle tension associated with high vocal effort can be a result of trying to compensate the dysfunction of some parts of the vocal tract system [Rosenthal et al., 2014].

Despite the important role of vocal effort in voice disorders, this falls outside the scope of this thesis, which applies the concept of vocal effort to expressive speech synthesis. Although the physiological cues of vocal effort are out of the scope of this thesis, it is interesting to point out that the association of vocal effort with muscle tension can also be related with the arousal degree of the emotion to synthesize in the ESS. For instance, anger and happiness entail high activation [Cowie, 2000], which could be related with increased vocal effort level, so henceforth we will refer to it as high vocal effort. On the other hand, sadness or relaxed speaking registers have low activation [Cowie, 2000], which could be related with more relaxed musculature and thus, will be referred as low vocal effort levels. This relation between vocal effort and emotions led setting hypothesis *H2*. It is important to remark that in this dissertation vocal effort will be used for speaking styles entailing tense (**high** VE) and relaxed (**low** VE) vocal tract musculature.

2.3.1 Definition

Vocal effort (VE) is a subjective physiological quantity [Traunmüller and Eriksson, 2000] trying to quantify the amount of effort involved with speech production. Thus, a common VE definition is the effort when producing speech in order to increase vocal intensity aiming to compensate a large distance from the receiver [Traunmüller, 1997,Traunmüller and Eriksson, 2000,Cushing, 2010]. The experiment setting used in initial studies analysing VE levels in anechoic conditions record the speaker's voice with a microphone placed at different distances. It is important to note that the effects of VE for compensating distances between speaker and interlocutor are not only volume or sound pressure level [Traunmüller, 1997,Traunmüller and Eriksson, 2000]. Indeed, Traunmüller demonstrates that listeners are able to identify different VE levels independently from the sound pressure level (SPL) [Traunmüller, 1997,Traunmüller and Eriksson, 2000] which indicates that VE also affects other aspects than SPL.

Vocal effort modifications can also be applied to speech systems in which environmental noise is a problem. Some research is currently focused on analysing the well known Lombard effect, that is, how speakers modify their speaking in noisy environments [Lau, 2008,Jokinen et al., 2015] (e.g., the LISTA project [LISTA, 2012]). Although vocal effort and the Lombard effect share some similarities because both entail variation in speech production to increase intelligibility, VE aims to cope with distance whereas the Lombard effect is focused on compensating environmental noise conditions [Traunmüller and Eriksson, 2000]. This implies that for the latter case, the speaker analyses the noise and then adapts his/her speaking style accordingly with the characteristics of environmental noise [Traunmüller and Eriksson, 2000,Garnier and Henrich, 2014]. An example of this behaviour can be seen in [Garnier et al., 2006, Garnier and Henrich, 2014] where the type of noise had significant effect on the speech acoustic parameters; the speaker used higher effort for cocktail party noise than white noise. [Jokinen et al., 2015] try to apply this strategy for enhancing intelligibility in telephone speech, designing new post-processing techniques based on Gaussian process regression and GMMs.

2.3.2 Spectral behavior and Voice Quality (VoQ)

Due to the variation of vocal tract muscle tension, air pressure, and vocal fold vibration at VE levels, the spectral characteristics of the produced speech signal vary. The main variations among different VE levels are: pitch, segment duration, formant frequencies, formant amplitudes, and energy distribution in the spectrum [Beller et al., 2008, Garnier et al., 2008, Tasko and Greilick, 2010, Cushing, 2010].

In [Cushing, 2010], a description of the setup for VE recording in anechoic conditions is presented. Cushing used a five microphone arrangement around the speaker: two microphones in front of the speaker at 0.5 and 1 meters distance, one on each side at 1 meter and one behind the speaker, also at 1 meter. In his experiment, Cushing obtained interesting results confirming an increase of energy in certain bands of the spectrum for higher VE levels (see figures 2.5 and 2.6 for male and female speakers, respectively). In both cases an increase of VE entails an increase of the energy in the central band of the spectrum around 600-800 Hz for males and 800-1000 Hz for female speakers. Another interesting finding in [Cushing, 2010] is the VE variation between subjects. For the male speakers, this variation increases for higher VE levels, whereas for female gender the variation is almost maintained equal along the different VE levels evaluated.

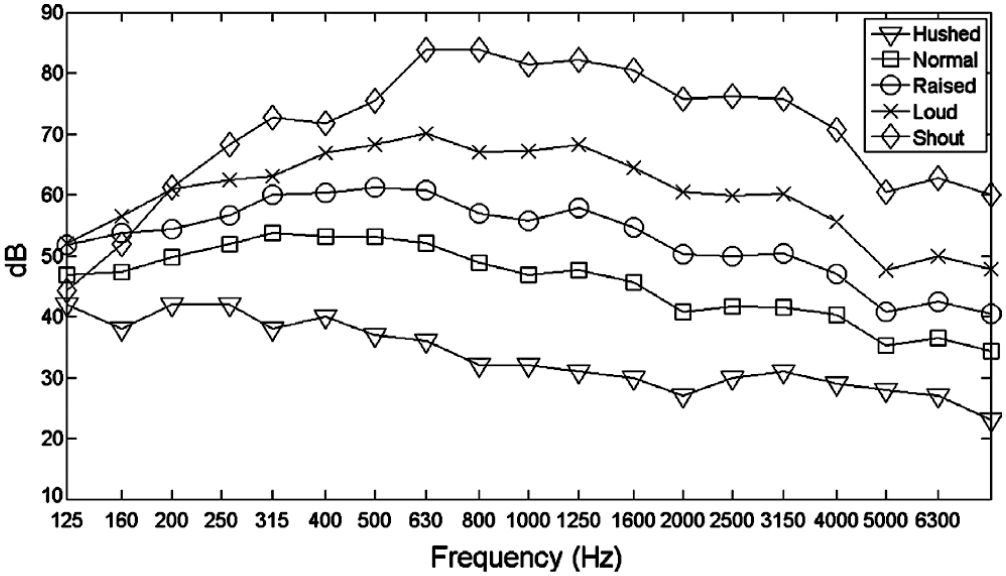


Figure 2.5 — Speech spectra for male speakers at different VE levels. From [Cushing et al., 2011]

Regarding formant variations, the major variations take place in the two first formants. Figure

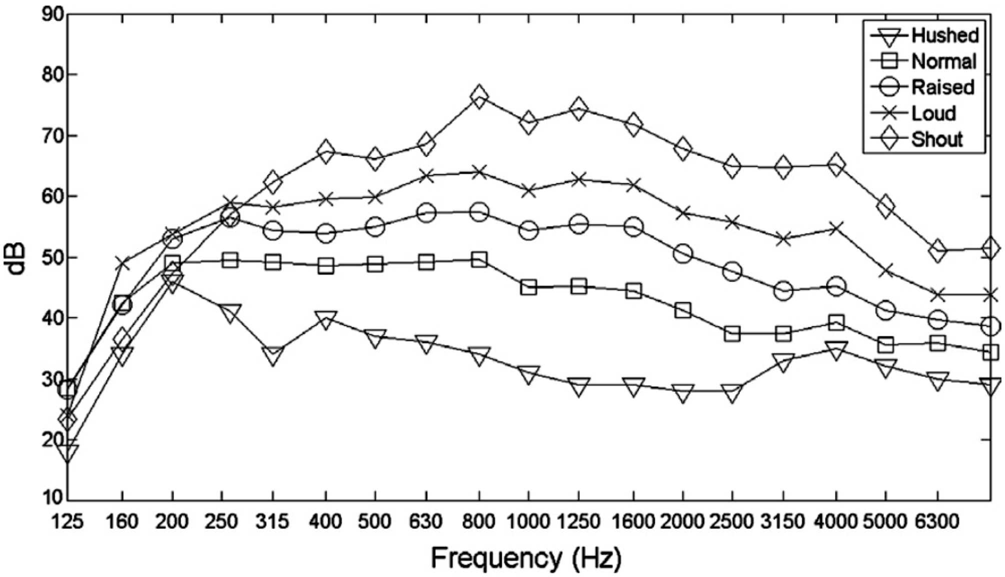


Figure 2.6 — Speech spectra for female speakers at different VE levels. From [Cushing et al., 2011]

2.7 shows the variation in formant frequencies for higher VE levels [Garnier et al., 2008, Tasko and Greilick, 2010] comparing the vowel space in regular speech register with shouted speech.

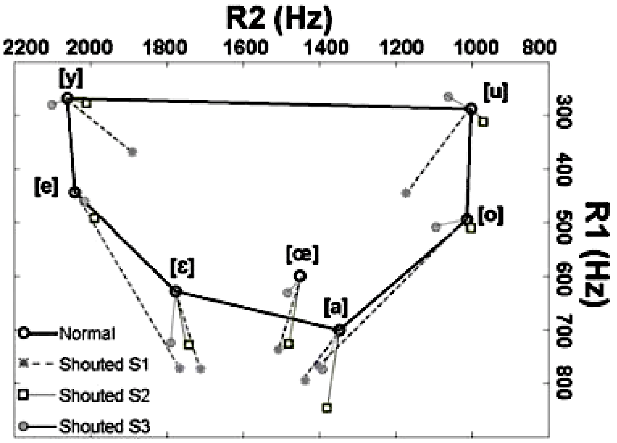


Figure 2.7 — Formant variation in shouted speech in three different conditions: (*S1*) no instructions given to speaker, (*S2*) maintaining constant pitch, (*S3*) maintaining constant pitch and articulatory position. From [Garnier et al., 2008]

Finally, [Cushing, 2010] also points out the impact of VE on duration. Cushing states that as VE increases, the speaker relies more on voiced segments increasing their duration related to unvoiced segments. This means that VE not only influences the spectral envelope, but also the prosody [Garnier et al., 2006, Cushing, 2010]. In fact, considering VE as articulation effort, some studies consider it to be an additional prosodic dimension [d'Alessandro, 2006, Beller et al., 2008].

For the purpose of this thesis, VE will be analysed from the spectral domain focusing on the spectrum envelope of the resulting speech conveying certain levels of VE. Thus, VE modifications will be related with VoQ modifications.

In order to express different VE levels in the synthesized speech while maintaining naturalness, the spectral envelope has to be modified considering all spectral characteristics (formants, energy distribution) and prosody (pitch, duration, signal energy). Modifying some prosodic features might affect spectral characteristics. An example is the articulation rate, which is tightly related with VoQ [Grichkovtsova et al., 2012]. It is important to maintain coherence between prosody and VoQ parameters. Indeed, some studies claim VoQ to be another dimension of prosody [Campbell and Mokhtari, 2003, Garnier et al., 2008]. The term VoQ was defined by [Laver, 1980] as the total vocal image of the speaker, including the phonatory part, the articulatory part and overall muscular tensions [Grichkovtsova et al., 2012].

The next section explains a methodology for modifying VE presented by [Nordstrom et al., 2008] which was the base for the experiments conducted in chapter 4 and 5.

2.4 Adaptive Pre-Emphasis Linear Prediction for VE modification

The reference technique that has been used as a baseline in the experiments reported in chapters 4 and 5 is the Adaptive Pre-emphasis Linear Prediction (APLP). This method was developed by [Nordstrom et al., 2008] and was designed to modify the vocal effort to transform high-effort voices into breathy voices.

APLP is based on the excitation-filter approach, and as shown in figure 2.8(a), the speech signal $S(z)$ is separated into a spectrally flattened excitation $E(z)$, a spectral emphasis filter $H_E(z)$ and a formant filter $V_F(z)$.

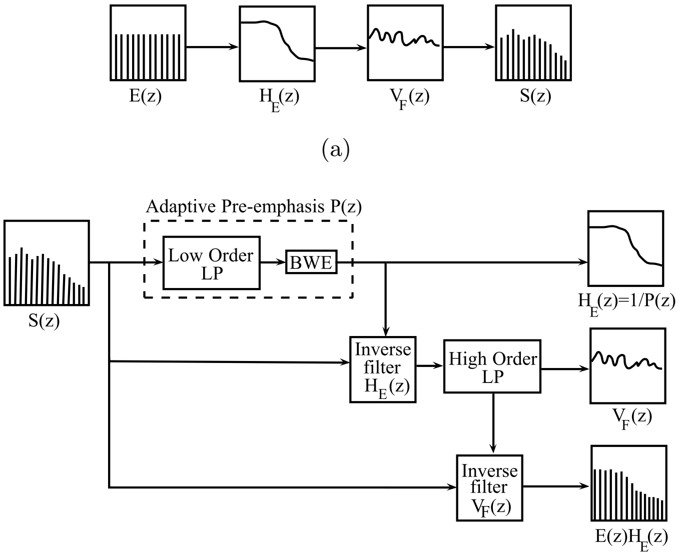
Figure 2.8(b) presents the analysis procedure used by APLP. To obtain the spectral emphasis filter $H_E(z)$, a low-order linear prediction (LP) filter is used. When performing the time-varying frame-based filtering process, for each signal frame the final state of the filter is used to initialise the filter's buffer for the next frame in order to avoid filter discontinuities. This initialisation is carried out for both the analysis and synthesis filters. This LP filter attempts to capture the slope of the signal spectra and the lowest frequency component of the spectral envelope; thus, low orders must be chosen to prevent $H_E(z)$ from capturing any formant information. Here, $H_E(z)$ is of order 3, following the recommendations and experiments carried out in [Nordstrom et al., 2008]. If $H_E(z)$ contains too many sharp peaks, it might contain formant information, rather than leaving it to be represented by the formant filter $V_F(z)$. Therefore, in addition to using low-order LP, bandwidth expansion (BWE) is applied to the obtained LPC. This BWE procedure consists of applying radial scaling to the filter coefficients with $\alpha = 0.9$ taken from [Nordstrom et al., 2008].

Once the $H_E(z)$ has been computed, the spectral emphasis is subtracted from the speech signal $S(z)$ by inverse filtering with the spectral emphasis filter $H_E(z)$, thus obtaining a more flattened spectrum than the original, that theoretically contains only formants from the vocal tract filter but not the spectral emphasis that is produced by vocal effort variations. In this sense, we will use the term of "flattened spectrum" although the result still contains spectral details regarding the vocal tract formants. Once the spectral emphasis has been removed from the original signal, the formant filter $V_F(z)$ is computed with a higher-order LP filter. Removing the spectral emphasis before computing the second filter ensures that $V_F(z)$ is more spectrally flat than if it had been extracted directly from the original signal $S(z)$. Here, the order used for the second LP analysis was 30.

Finally, once the spectral emphasis and formant filters ($H_E(z)$ and $V_F(z)$, respectively) have been computed, the signal excitation can be obtained by inverse filtering $S(z)$ with $H_E(z)$ and $V_F(z)$. The spectral emphasis $H_E(z)$ corresponds to the inverse of the pre-emphasis filter $P(z) = \frac{1}{H_E(z)}$ (see figure 2.8(b)). The bandwidth expansion technique [Kabal, 2003] was used for computing the Linear Prediction Coefficient (LPC) of $H_E(z)$ in order to prevent it from capturing the peaks corresponding to the formants, which should be captured by the formant filter ($V_F(z)$). This technique consists of moving the poles of the filter inside the unit circle to prevent the all-pole filter from becoming too peaky.

Figure 2.9 illustrates the procedure for transferring a target signal's vocal effort to a source signal using the APLP system. First, the spectral emphasis of the source $S(z)$ and target $T(z)$ signals is computed ($H_E^s(z)$ and $H_E^t(z)$), respectively. The spectral emphasis of the source signal $H_E^s(z)$ is subsequently removed from this source signal, thus obtaining a spectrally flattened source signal $S_f(z)$. Next, the desired spectral emphasis $H_E^t(z)$ is applied to $S_f(z)$, thus obtaining the converted version, $Sc(z)$.

In [Nordstrom et al., 2008] the main goal was to obtain voices with lower vocal effort, and so introducing aspiration noise can help to obtain that result. However, in the work presented in this dissertation, the goal was to also explore the conversion for voices with increased vocal effort levels. Thus, in the presented study no aspiration noise was added to the converted signals, contrary to what was used in [Nordstrom et al., 2008]. This aspiration signal is only relevant for breathy voices and could prevent APLP from achieving good results when increasing the vocal effort in the transformation procedure.



(a)

(b)

Figure 2.8 — (a) Linear model of the voice resulting from APLP. (b) APLP analysis diagram. (BWE stands for Bandwidth expansion). Adapted from [Nordstrom et al., 2008].

The entire procedure is based on applying the filters in the temporal domain, using the output signal of the filters as input for the next filter, without any speech model.

$$Sc(z) = \overbrace{S(z)\frac{1}{H_E^s(z)}}^{S_f(z)} \frac{1}{V_F^s(z)} H_E^t(z) V_F^s(z) = \overbrace{S(z)\frac{1}{H_E^s(z)}}^{S_f(z)} H_E^t(z) \tag{2.12}$$

2.5 Sinusoidal and harmonics models overview

Although trying to represent speech signal with sinusoidal functions has been studied for a long time [Almeida and Tribolet, 1982,McAulay and Quatieri, 1986,Marques et al., 1990,Depalle and Hélie, 1997,George and Smith, 1992], this dissertation begins with the contribution made by Stylianou in his PhD dissertation [Stylianou, 1996]. In [Stylianou, 1996], Stylianou presented a new model for analysis, synthesis, and speech signal modification based on a decomposition of the speech signal spectrum in two bands: the lower spectrum is used to model deterministic events and uses a sum of harmonics for generating this part of the signal. Conversely the upper band of the spectrum is modelled as an autoregressive process (a non-deterministic part formed by white Gaussian noise filtered with an all-pole filter). This representation of the speech signal helped to improve the overall synthetic speech quality when performing speech signal modifications, compared with Time Domain Pitch Synchronous

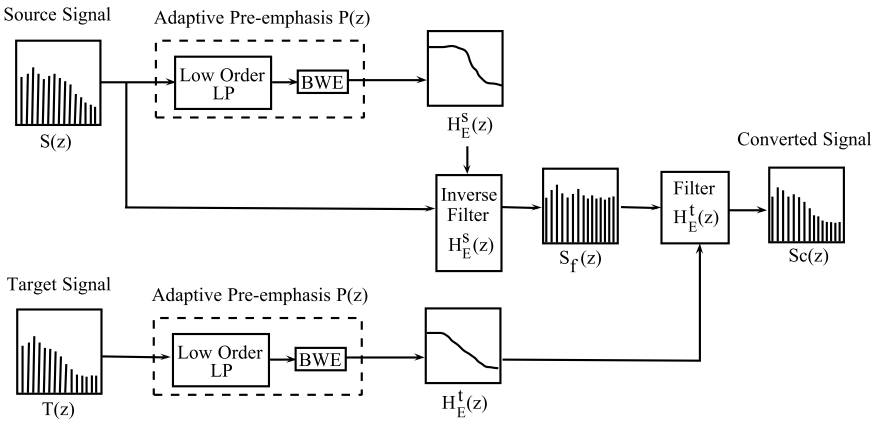


Figure 2.9 — Block diagram for transferring vocal effort with APLP. From [Nordstrom et al., 2008]

Overlap-and-Add (TD-PSOLA).

Since then, sinusoidal and harmonic models have been widely used for several applications: speech coding and synthesis [Stylianou, 2001], speech enhancement [Jensen and Hansen, 2001], voice transformation [Erro, 2008, Degottex et al., 2013], hearing aids [Hu and Loizou, 2010] or for glottal source estimation [Degottex et al., 2011].

In most of the works referenced previously, the model parameters, sinusoid amplitudes and frequencies, were supposed to be constant along the analysis window. However, [Pantazis et al., 2008], considered the model parameters, amplitudes and frequencies, to vary within the analysis window period, and thus proposed a time-varying quasi-harmonic model. One of the advantages of this model over other alternatives using quasi-harmonic approaches [Godsill and Davy, 2002] is the reduced number of model parameters to estimate. The quasi-harmonic term used in [Pantazis et al., 2008] is due to using sinusoidal modelling only for the lower band of the speech spectrum, which was considered related to deterministic events in speech signal. The models aimed to improve the estimation of harmonic parameters, complex amplitudes and frequencies, compared with typical harmonic models which considered the signal stationary, by letting the parameters vary inside the analysis window, usually 2 or 3 pitch periods.

In order to allow complex amplitudes and frequencies to vary inside the analysis window, the proposed model represents the speech signal as

$$s(t) = \left(\sum_{k=-L}^{L} (a_k + t\, b_k)\, e^{j\, 2\pi\, k f_0\, t} \right) w(t), \tag{2.13}$$

where a_k are the complex amplitudes, b_k the complex slope for the complex amplitudes and $w(t)$

is the analysis windows. Typical harmonic plus noise models set the term $b_k = 0$, thus keeping the complex amplitudes (a_k) constant during the whole analysis window. However, in the quasi-harmonic models a_k, and b_k are complex variables, which makes the instantaneous phases and instantaneous frequency variable functions over time. Moreover, the $t\ b_k$ term helps correct any deviations in phase and frequency [Pantazis et al., 2008]. The study also demonstrates the dependency between frequency estimation and amplitude information. Thus, taking advantage of the properties of the $t\ b_k$ term for correcting phase and frequency estimations, [Pantazis et al., 2010a] propose an iterative algorithm for updating and improving the estimation of frequencies. The iterative procedures is explained in detail in [Pantazis et al., 2010a]. There the Quasi-Harmonic Model (QHM) is defined and an iterative method is mathematically derived. The study also demonstrates the validity of HNMs, concluding that it provides better parameter estimations than the Fast Fourier Transform (FFT)-based frequency approaches. It also highlights QHMs restrictions regarding initialization of f_0 and the impact the analysis window on frequency estimation errors. Although the $t\ b_k$ term helps to correct the frequency estimation errors, the method requires the f_0 to be relatively close to the real value, meaning that the error should be less than one third of the bandwidth of the squared analysis window. Thus, the larger the main lobe of the window spectrum, the larger the allowed frequency mismatch.

As mentioned earlier, the term quasi-harmonic is used because the harmonics represent the lower band of the speech spectrum, leaving the upper band still to be modelled. Based on the QHMs [Pantazis et al., 2008] and the iterative method for readjusting the estimated frequencies [Pantazis et al., 2010a], in [Pantazis et al., 2010b] a model for representing the full band speech spectrum was presented. The lower spectrum is modelled by an improved version of the QHM and the upper band is modelled by filtering Gaussian noise by a time-varying 18 order AR filter. In the proposed methodology, the signal was analysed synchronously with pitch, and the analysis window covered three pitch periods. Although the QHM presented in [Pantazis et al., 2008] allowed complex amplitudes to vary inside the analysis window, frequencies were kept constant, making the $t\ b_k$ term capture variations of amplitudes, phases and frequencies. In the proposed model in [Pantazis et al., 2010b] the frequencies are also allowed to vary inside the analysis window. In order to permit frequencies to vary in the analysis window, the linear phase corresponding to the relation between time evolution and the frequencies is estimated as the integral of the frequency variation along the analysis window. The computation of linear phase (ϕ_l^k) for a given harmonic (l) in the frame k is performed as

$$\tilde{\phi}_l^k(t) = 2\,\pi \int_{t_k}^{t_k+t} f_l(u)du\ ,\ \text{with}\ \ t \in [-T, T], \tag{2.14}$$

where t_k is the time of the centre of the analysis window with length $2T$ and f_l is the instantaneous frequency for the lth harmonic.

Applying equation (2.14) into equation (2.13), to permit frequency variations, leads to the adaptive quasi-harmonic model aQHM, which expresses a single frame k as

$$s(t+t_k) = \left(\sum_{l=-L}^{L} \left(a_l^k + t\, b_l^k \right) e^{j \tilde{\phi}_l^k(t)} \right) \omega(t). \tag{2.15}$$

An extension of the Adaptive Quasi-Harmonic Model (aQHM) which includes not only frequency corrections, as in the original QHM, but also amplitude corrections, was proposed in [Kafentzis et al., 2012] adding a new term to the model for correcting amplitude mismatches. The new method is denoted as Extended Adaptive Quasi-Harmonic Model (eaQHM). The speech signal is represented by

$$s(t+t_k) = \left(\sum_{l=-L}^{L} \left(a_l^k + t\, b_l^k \right) \frac{A_l(t+t_k)}{A_l(t_k)}\, e^{j \tilde{\phi}_l^k(t)} \right) \omega(t)\ ,\ \text{with } t \in [-T, T], \tag{2.16}$$

where $A_l(t)$ is the instantaneous amplitude of the lth harmonic component. Note that the instantaneous amplitude $A_l(t)$ is divided by $A_l(t_k)$, which makes the correction factor equal to 1 at the centre of the analysis window. This extra correction on the amplitude estimation makes eaQHM outperform aQHM in terms of convergence speed, and also reduces synthesis errors.

Up to now it was a common approach to represent the lower band of the speech spectrum with harmonic models, and model the upper band with stochastic processes due to the lack of clear harmonic trajectories in the upper band of the Discrete Fourier Transform (DFT) representation. However, [Dunn and Quatieri, 2007] showed that the DFT representation of the speech signal masks the harmonic structure of the signal for higher frequencies, the problem being the frequency stationariness of the DFT's basis functions. Due to stationary frequencies, the basis functions cannot resolve the speech harmonics which have rapid frequency modulation and are closely spaced in frequency. They also state that this effect is persistent regardless of the analysis window length. Instead of using the DFT, they proposed using the Fan-Chirp Transform (FChT) where the basis sinusoid functions of the model vary their frequency over time, thus being able to capture the frequency modulation that happens in the speech signal harmonics.

Figure 2.10 shows a graphical representation of the evolution of the basis functions used in the Fourier, Chirplet, fractional Fourier, and fan-Chirp transforms. As can be seen, the basis functions of the Fourier Transform (FT), the sinusoids, are constant along the time axis, whereas the FChT sets a focus point from which the basis sinusoids' frequencies diverge. [Dunn and Quatieri, 2007] and [Weruaga and Képesi, 2007] demonstrated the better representation of the speech signal with FChT than regular DFT.

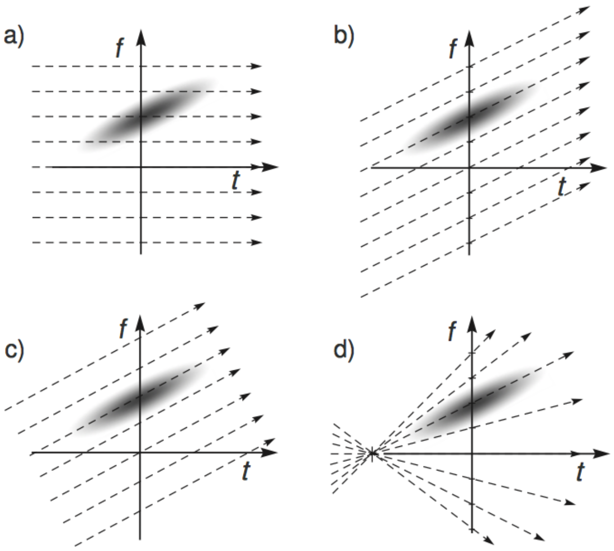
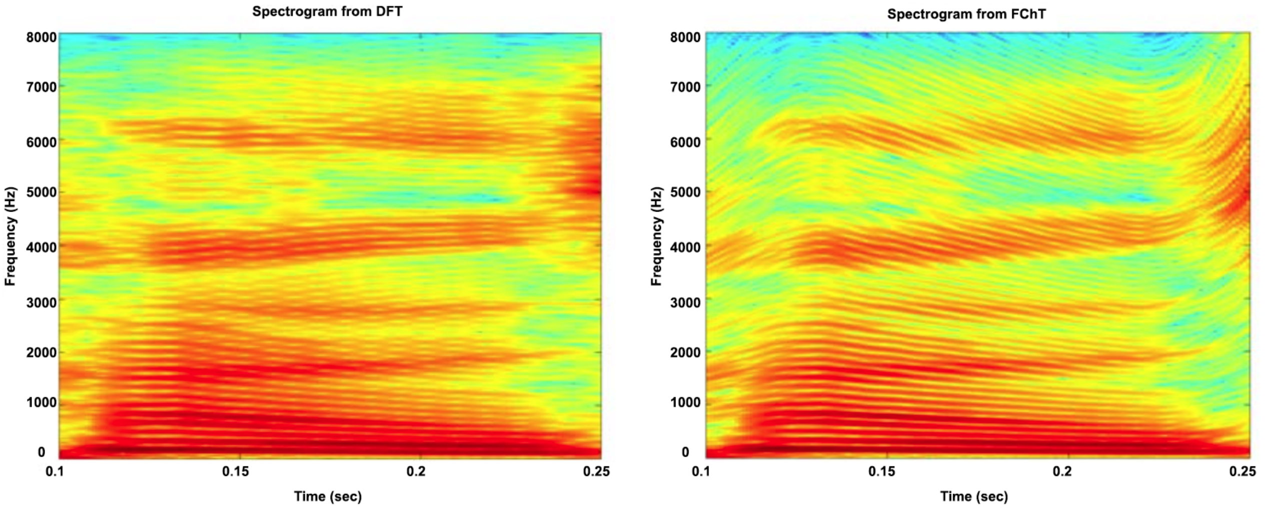


Figure 2.10 — Graphical representation of the time evolution for basis function in the: (a) Fourier, (b) Chirplet, (c) fractional Fourier and (d) fan-Chirp transforms. From [Weruaga and Képesi, 2007]

Figure 2.11 compares two spectrograms corresponding to the same speech signal segment using the DFT and the FChT. As can be seen, in the latter case, the evolution of the harmonics is more clear than in the DFT representation. It can also be seen the erroneous behaviour of the harmonics around 3000 Hz going up and down the spectrum in the DFT representation. Due to this lack of frequency modulation tracking, the DFT tends to present blurry spectrograms, specially in the upper band of the spectrum. This fact led to represent this band with stochastic processes leaving harmonic representations only for the lower frequencies. However, in the FChT representation it is easily seen that the speech signal also has strong harmonics in the upper band of the spectrum. Having harmonic structure in the upper band of the spectrum is coherent with the fact that the glottal excitation signal spectrum presents a slow decay along the overall spectrum [Cabral et al., 2014, Degottex et al., 2011, d'Alessandro and Sturmel, 2011].

This appearance of harmonic structure in the upper band of the spectrum, led Degottex et al. to use the previously presented Adaptive Quasi-Harmonic Model (aQHM) for modelling the full band instead of restricting it to the lower frequencies, naming this approach Adaptive Harmonic Model (aHM) [Degottex and Stylianou, 2012]. The article also presents a new iterative method for computing harmonic parameters for the full spectral band. This was required in order to ensure correct parameter estimation. In all quasi-harmonic models (QHM [Pantazis et al., 2008], aQHM [Pantazis et al., 2010a], aQHNM [Pantazis et al., 2010b]) the initial values of the frequencies should be close to the original frequency values. However, due to the harmonic structure of the model, any potential error of f_0 is multiplied by the harmonic number, thus obtaining a significant initialization error for the higher



(a) Spectrogram using DFT

(b) Spectrogram using FChT

Figure 2.11 — Narrowband spectrogram of the same large-bandwidth speech segment, with substantial frequency modulation, using the discrete Fourier and fan-chirp transforms. From [Dunn and Quatieri, 2007]

frequencies. For instance, an error of 2 Hz at the harmonic of 100 Hz results in an error of 100 Hz for the 50th harmonic (5KHz), which means an error equal to f_0. Thus, in order to control the error, an iterative method for computing the model parameter values was applied starting the parameter estimation at the lower band of the spectrum, where the error is assumed to be reasonably small, and iteratively increasing the number of harmonics to compute. The new method is referred to as Adaptive Iterative Refinement (AIR), thus naming the overall method Adaptive Harmonic Model with Adaptive Iterative Refinement (aHM-AIR). In [Degottex and Stylianou, 2013] the aHM-AIR is compared against the sinusoidal and harmonic model in noisy conditions; the proposed methodology presents a slight tendency for better signal quality results over its predecessors. Nevertheless, the study demonstrates that the AIR algorithm produces a fine estimation of the model parameters and that the full speech spectrum can be accurately modelled with aHM.

As stated in [Morfi et al., 2015], although aHM-AIR is precise, it lacks the computational efficiency that would make it feasible for large databases. The least squares solution used in the original aHM-AIR entails high computational cost. In [Morfi et al., 2014], a Peak Picking (PP) approach is presented as a substitution for the LS solution for computing f_0 in the AIR algorithm. Also, an Adaptive Discrete Fourier Transform (aDFT), whose frequency basis can fully follow the variations of the f_0 curve was presented for integrating the adaptability scheme of aHM in the PP approach. [Morfi et al., 2015], evaluated the above methods for the whole analysis process of a speech signal, showing an average four fold time reduction using PP and aDFT compared to the least squares solution.

Additionally, based on listening evaluations, when using PP and aDFT, the quality of the re-synthesis is preserved compared to the original least squares based approach.

Finally [Kafentzis et al., 2013, Kafentzis et al., 2014a] explain time and pitch modifications using the full-band aHM. The procedure is quite the same as used for the deterministic part in HNMs, which is explained in the appendix sections A.3, for time modifications, and A.4 for pitch modifications.

2.6 GTM's current US-TTS implementation and scope of this dissertation

This section presents the specific implementation of the US-TTS system for GTM group at la Salle based on the work conducted in [Iriondo, 2008] and highlights the components that require modifications in order to include the capabilities for modifying vocal effort in the synthesised speech using the techniques presented in the following chapters. Section 6.2 proposes modifications to be made to the proposed methodology in chapter 5 for including vocal effort control capabilities to the US-TTS system presented in this section.

Figure 2.12 shows the general pipeline of a US-TTS system. The current system is drawn in grey, and the modifications are in blue. The input data is a text, which is analysed with a Natural Language Processing (NLP) component that predicts phonetic labels and prosodic information. All the information extracted by the NLP component will be used for the selection of the units to be used for synthesis. For instance, the unit selection component will use phonetic information for finding the best candidate units, which are those units available in the speech corpus that permit the generation of the speech signal from the input text. When there are multiple candidate units available, the unit selection block has to choose the best sequence of units based on the cost metrics defined in the system, such as the ones explained in section 2.1.3. To evaluate the suitability of each candidate unit using the cost functions, acoustic information such as prosody is used.

In order to be able to incorporate VE modifications, the NLP component also should generate VE information that will help the digital signal processing (DSP) component apply the proper signal modifications.

Figure 2.13 shows a schematic diagram of the NLP component, in which the existing components are in black, while the modifications are highlighted in blue. The latter correspond to the modifications considered necessary to make in order to obtain a US-TTS able to control the degree of VE in the output speech.

One of the main functions of the NLP component in the context of ESS is to deliver a suitable sequence of acoustic parameters to the *unit selection* component in order to synthesize natural sounding

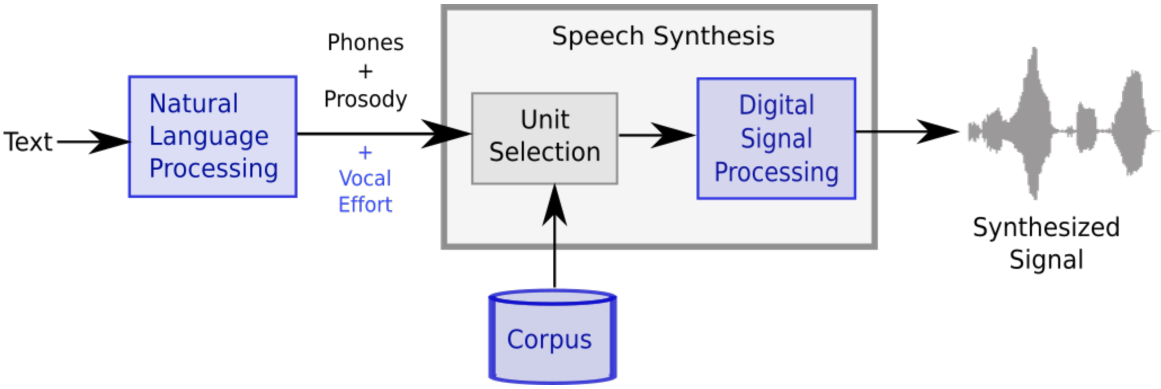


Figure 2.12 — General schematic of a US-TTS system. In blue, the newly added parameter regarding VE and the components that need to be modified according to the proposal.

and expressive speech. The acoustic parameters are obtained by prediction performed by a machine learning (ML) technique. Thus the prosody prediction process is composed of two stages. First a *training stage* where the ML is presented with a set of speech signal attribute and value pairs that represent multiple speech realizations in the contexts present in the speech corpus, with the aim of learning the relation between attributes and their values for any possible input. The context can be described by any combination of information related with the specific sample cases to be presented to the Machine Learning (ML) that help to discriminate those sample cases in some way. For instance a context could be the surrounding (preceding and following) phonemes, the part of speech of the word the phoneme forms part, whether the phoneme is stressed or not and so on. Secondly, the *prediction stage* where a set of attributes of an usually non previously seen context are feed to the ML to produce a prediction of the prosody speech values based on the learned patterns.

Figure 2.13 shows the blocks used on each stage of the prosodic plus VE learning and prediction steps of the modified NLP process. The first step in the training stage is the extraction of the attribute and value pairs. For this the context information is firstly extracted from the original text. *Text Processing* obtains the part of speech information. Next, *Phonetic Transcription* block obtains the proper phonemes sequence given the language phonetic rules and context. Finally the *Accent Group Labelling* block obtains the stressed parts of sentence and words. All this information describe the context for a given unit, for instance a phoneme or a diphoneme. Then the speech signal corresponding to the phoneme analysed in the previous blocks is passed together with its context information to the *Sample Generation* block which extracts the acoustic information (pitch , energy and duration) from the speech signal and associates it with the contextual information. The last step of the training stage is to introduce different attribute-value pairs of each prosodic parameter (energy, duration and pitch) to a ML so it can learn the proper relation between contextual attributes and acoustic information

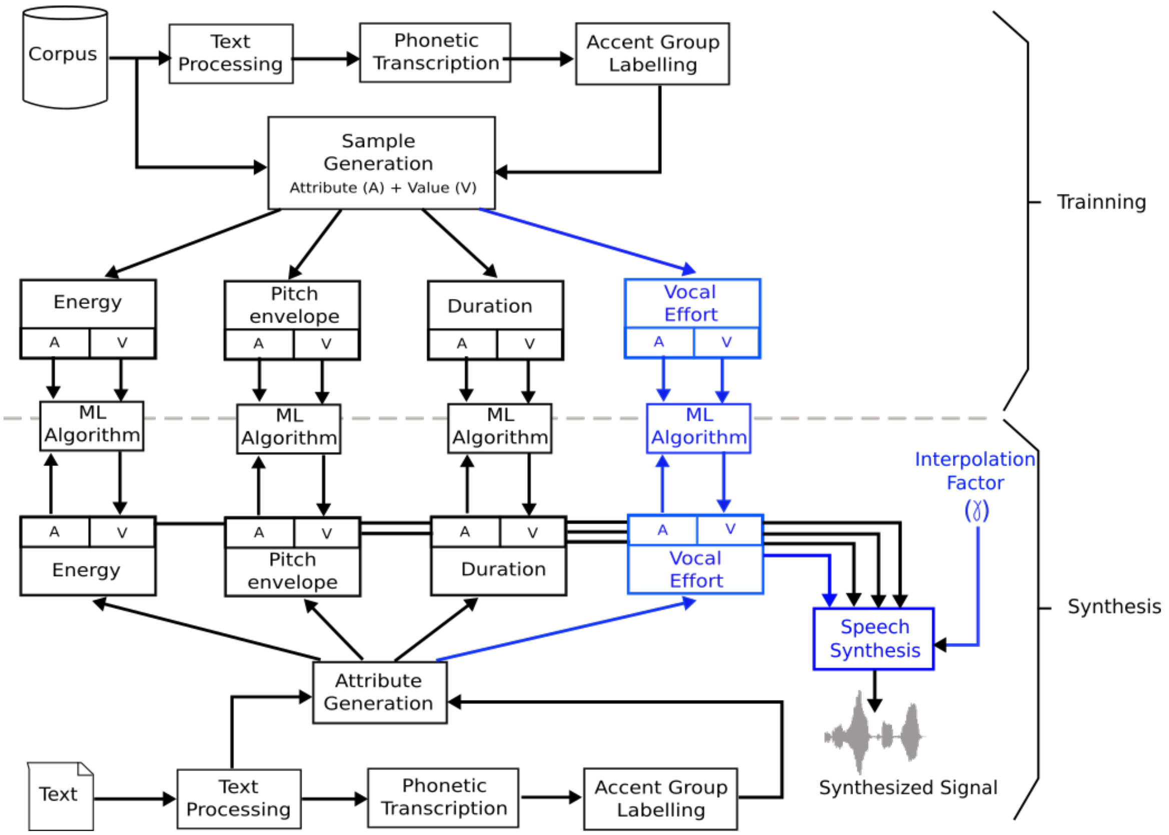


Figure 2.13 — Proposed general scheme modifications of the current US-TTS system [Iriondo, 2008] to include VE modification capabilities. Colour blue indicates additional data flows and components, such as VE predictors via a ML algorithm, and modified components, such as the speech synthesis component. Color black depicts the components already present in the current system (black). In the ML component, "A" stands for attribute and "V" for value for the given attribute. Adaptation from [Iriondo, 2008]

values.

The current US-TTS of GTM uses Case Based Reasoning (CBR) as the ML algorithm of the NLP module. The CBR consists of solving new problems (such as finding out the proper acoustic parameters -pitch, duration and energy- for the current unit to synthesize a natural sounding speech) by recovering previous similar situations (taking into account contextual information that answers the following questions: which are the surrounding phonemes? current phoneme is stressed? which part of speech belongs the current phoneme?) and by reusing information and knowledge of that situation [Aamodt and Plaza, 1994]. For example, using the type of sentence (declarative, interrogative, imperative and exclamatory) as a simplified context information, most of the cases interrogative sentences will present an increment of pitch value towards the end of the sentence. The CBR can gather those examples where this phenomenon takes place and the next time it is presented an interrogative sentence to synthesize, based on previous cases the system will respond that the sentence to synthesize must present a rising pitch envelope towards the end of the sentence.

At synthesis time, the same sort of contextual information is extracted from the text and this is presented to the ML block for each prosodic attribute (energy, pitch and duration). For the CBR ML technique a matching search of the context to synthesize with all the examples stored in its memory of cases is performed, each context the algorithm has stored, and selects the closest one. Then, for the selected case in the case database, the corresponding prosody attribute values are taken and used as predicted values to be associated with the unit to be synthesized.

In the current US-TTS schema the prediction is performed only for prosodic parameters (pitch, duration and energy). In order to be able to perform VE modifications in the synthesized speech, VE prediction should also be carried out. In figure 2.13 the blocks and data flows to be added in order to extend the flexibility of the system for conducting VE modifications in the speech synthesis are highlighted in blue. As can be seen, an additional ML block should be introduced which requires the *Sample Generation* block to generate extra data (new attribute-value pairs) regarding VE.

The *Speech Synthesis* is the component responsible of generating the synthetic speech signal based on the information given by the NLP component. The modifications to perform into this component in order to let the system control de VE conveyed in the synthesised speech are explained in section 6.2 based on the methodology presented in chapter 5. Based on research question *Q3*: *Is it possible to synthesize VE levels different from the ones available in the speech recordings used for modelling purposes and controlling the VE level using a simple parameter?*, an *Interpolation Factor* could be used to easily control the VE level in the synthesised speech.

This dissertation focuses on the speech signal modification techniques required to be implemented in the *Speech Synthesis* component, which are presented in the following chapters.

CHAPTER 3

Expressivity modification based on prosody and VoQ low-level parameters using HNM

Contents

3.1 VoQ measurements from HNM parameters 49

3.2 Expressive transformation based on VoQ modifications 52

3.2.1 Expressive speech corpus 52

3.2.2 HNM parameter modifications 55

3.3 Evaluation and results 56

3.4 Conclusions 58

This chapter aims to answer the first research question *Q1* focusing on the speech modifications applied to the speech signal in order to convey multiple expressions in the synthesized speech using only a single neutral corpus to produce the synthetic speech. Moreover, combining VoQ with prosody modifications could improve the expressiveness in the final synthetic speech signal compared to applying only prosody modifications. Thus, the following hypothesis (*H1*) was formulated: HNM is a suitable speech representation for conducting VoQ modifications for expressive speech synthesis, and VoQ can be added as extra acoustic features to improve the expressiveness in synthetic speech.

In order to validate this hypothesis this chapter pursues the following objectives (*O1*):

- ▷ Conduct experiments for transforming neutral speech into different expressive speech styles using the HNMs.

- ▷ In these experiments, compare transformations produced modifying only prosodic attributes against transformations where a preselected set of low-level VoQ parameters are added to the transformation pipeline (in addition to prosodic parameters).

As explained in section 2.5, the Harmonics plus Noise Model (HNM) allows to easily perform prosody modifications on speech signals, thereby maintaining high level sound quality in the synthesised signal [Gu and Liau, 2008, Syrdal et al., 1998]. However, there was no evidence that the same system could be used for modifying VoQ parameters. This chapter explains the work conducted in [Monzo et al., 2010] where a first attempt to modify low-level VoQ parameters using the HNM representation is presented, comparing two speech style transformation methods, using prosody only and prosody combined with VoQ, in order to assess the contribution of VoQ parameters in defining specific emotions, and the feasibility of using HNM for conducting the expressive modifications. The obtained results validate the hypothesis *H1*.

In [Monzo et al., 2010] low-level VoQ parameters are computed frame by frame and then averaged for the whole speech utterance. The conducted signal modifications aim to transform the neutral speech (without emotion) of the original utterance to a set of predefined speech styles: sad, happy, sensual, or aggressive. The modifications are performed by applying a constant multiplicative factor to a selected set of VoQ parameters. The constant multiplicative values were chosen after analysing multiple expressive speech corpora obtaining a value for each expressive style. Thus, the proposed methodology did not consider the VoQ dynamics along the speech signal, but applied a constant factor for every frame in the synthesized utterance. On the other hand, proposed methodologies in chapters 4 and 5 do consider these VoQ dynamics.

This chapter is structured as follow: Section 3.1 defines the VoQ parameters used in the study and how they were measured from the HNM parameters. Next, section 3.2 explains the experiments conducted for validating the proposed procedures and the obtained results. Only VoQ modifications based on HNM are explained. For prosody modifications and VoQ parameter modifications not performed via the HNMs refer to [Monzo et al., 2010].

3.1 VoQ measurements from HNM parameters

This section presents and defines the following low-level VoQ parameters: *Jitter*, *Shimmer*, *Harmonic-to-noise ratio*, *Hammarberg Index*, and *Relative amount of energy above 1000 Hz.* Since their definition is based on pitch periods and spectral band energy ratios, their computation from HNM parameters is quite straightforward. For more detail on their computation refer to [Monzo, 2010].

This work was done with an early version of the HNM system where the harmonic part was pitch synchronously analysed. This means that the analysis time instants are set at every pitch period. On the other hand, the stochastic part was analysed at a constant frame rate. This asynchrony between harmonic and stochastic parts conditioned the study.

Jitter

Jitter measures the short time variations of the fundamental period T_0. It describes the cycle-to-cycle variation of the fundamental period, which appears as a frequency modulated noise. The expression used for computing the jitter from the original signal is

$$jitter(absolute) = \frac{1}{N-1}\sum_{i=2}^{N}|T_0(i) - T_0(i-1)|, \tag{3.1}$$

where N is the number of signal periods, i is the period index and $T_0(i)$ is the fundamental period value for period i. However in the conducted work a relative measure of *jitter* expressed as a percentage (%) was used instead of the absolute measure. It was calculated dividing the period difference, or absolute *jitter*, (3.1) by the average period as

$$jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|T_0(i) - T_0(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}T_0(i)} \times 100 \qquad [\%]; \tag{3.2}$$

Shimmer

This parameter measures the short time amplitude variations. It describes the cycle-to-cycle variations of the waveform amplitude. This appears as an amplitude modulated noise on the speech waveform defined

$$shimmer(absolute) = \frac{1}{N-1}\sum_{i=2}^{N}|U_0(i) - U_0(i-1)|, \tag{3.3}$$

where N is the number of signal periods, i is the period index and $U_0(i)$ is the peak-to-peak amplitude in period i. In the presented study *shimmer* was also used as a relative measure as a percentage (%). The computation is based on dividing the difference between peak-to-peak amplitudes between periods (3.3) by the average peak-to-peak amplitude as

$$shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=2}^{N}|U_0(i) - U_0(i-1)|}{\frac{1}{N}\sum_{i=1}^{N}U_0(i)} \times 100 \qquad [\%]; \tag{3.4}$$

Harmonic-to-noise ratio - (HNR)

HNR describes the ratio between the energies of the harmonic (E_h) and stochastic (E_s) parts of a speech segment as

$$HNR = 10\,log\left(\frac{E_h}{E_s}\right). \tag{3.5}$$

The use of a segment as the reference unit is due to restrictions from the asynchrony between analysis time instants of the harmonic and stochastic parts of the HNM implementation.

Hammarberg index - ($HammI$)

This parameter expresses the ratio between maximum energies of two frequency bands which in turn are computed from the harmonic part. The lower band frequency range is [0,2000] Hz and the second band range is (2000,5000] Hz. Due to working with HNM parameters, the maximum energy has been computed considering variations between energies of each harmonic component, which would be the same as having used a smoothing filter with bandwidth equal to the minimum pitch. Thus, the hammI parameter is expressed as

$$\begin{aligned} HammI^{(k)} &= \frac{\max\left(E_l^{(k)} \mid_{l\in[0,l_{2000}^{(k)}-1]}\right)}{\max\left(E_l^{(k)} \mid_{l\in[l_{2000}^{(k)},L^k]}\right)} \\ HammI &= 10\,log\left(\frac{1}{K}\sum_{k=1}^{K} HammI^{(k)}\right), \end{aligned} \tag{3.6}$$

where $E_l^{(k)}$ is the energy of the harmonic l at frame k, $l_f^{(k)}$ is the harmonic integer index with the minimum frequency f_l, from frame k, but f_l being greater or equal to the frequency corresponding to index $l_f^{(k)}$. For example, numerator and denominator in setting equation (3.6) to the specific case

$$HammI^{(k)} = \frac{\max\limits_{l\in\omega[0,2000)}\left(E_l^{(k)}\right)}{\max\limits_{l\in\omega[2000,5000]}\left(E_l^{(k)}\right)}, \tag{3.7}$$

where the numerator only considers harmonic indices corresponding to frequencies in the range [0,2000)Hz, and the denominator harmonic indices corresponding to frequencies in the range (2000,5000] Hz.

Thus, $l_{2000}^{(k)}$ stands for the index of the harmonic in frame k with a frequency of 2000 Hz, or the next nearest one in case of no exact match.

Relative amount of energy above 1000 Hz - ($pe1000$)

This parameter is the relative amount of energy in the high (above 1000 Hz) versus the low frequency range (below 1000 Hz) of the voice spectrum. It is computed by

$$\begin{aligned} pe1000^{(k)} &= \frac{\sum\limits_{l=1}^{l_{1000}^{(k)}-1} E_l^{(k)}}{\sum\limits_{l=l_{1000}^{(k)}}^{L^{(k)}} E_l^{(k)}} \\ pe1000 &= 10\,log\left(\frac{1}{K}\sum_{k=1}^{K} pe1000^{(k)}\right). \end{aligned} \tag{3.8}$$

As for *HammI* index (3.6), $l_{1000}^{(k)}$ stands for the index of the harmonic at 1000 Hz, or the next nearest one in case of no exact match.

Equation (3.8) can be rewritten as

$$pe1000^{(k)} = \frac{\sum\limits_{l \in \omega[0,1000)} E_l^{(k)}}{\sum\limits_{l \in \omega(1000,5000]} E_l^{(k)}}. \tag{3.9}$$

With the VoQ parameters and their computation equations from HNM parameters defined, we proceed to define the modification procedure and describe the obtained results.

3.2 Expressive transformation based on VoQ modifications

The main goals of the conducted work are the following: **i)** Verify that including VoQ allows to improve the perception of the conveyed emotion compared to using only prosodic modifications in the transformation from neutral to expressive speech styles. **ii)** the second objective aims to evaluate to which extent the proposed low-level VoQ parameters within the HNM framework are acceptable to enrich the transformation between different expressive speech styles with the required quality. **iii)** Verify that HNM are a good speech signal representation not only for prosody modifications, but for VoQ conversion as well.

3.2.1 Expressive speech corpus

The expressive corpus used in this experiment was created by other GTM members, produced using a mix of acted speech and induced emotion approaches. It was recorded in Spanish and divided in 5 discrete emotion categories: neutral, happy, sensual, aggressive, and sad. The corpus contains 4638 phrases, (total corpus duration is 5 h and 12 minutes), performed by a professional female speaker. The texts were extracted from an advertisement database. Due to the relation between semantics and expression, the texts for the utterances were chosen to be semantically related with the desired expression in order to get utterances with a better defined expression. Some studies extend this influence by building up a scenario, generating a semantic environment around the utterance to be recorded in the corpus, in order to minimize the interpretation variations that may result from speaker to speaker [Bulut et al., 2002]. Moreover it has been demonstrated that in concatenation synthesis, the corpus emotion is present in the synthesized signal [Schröder, 2001, Eide et al., 2004, Eide, 2002], so for these reasons, specific emotion corpus were built. Also a relation between topics from the advertisement database and specific expressiveness was established leading to the following [topic→emotion] assignments: [New technology→neutral], [Education→Happy], [Cosmetics→Sensual], [Automobiles→Aggressive] and [Trips→Sad].

Firstly, as the corpus was produced following an acted speech approach (not using real emotions but acted speech by a professional speaker) a subjective test was done to validate the database emotion labelling. In this test the subjects had to decide which emotion was conveyed in the synthesized speech. Possible options were the five emotions plus two more options: *Do not know* and *Another*. The resulting confusion matrix (see [Monzo et al., 2010]) showed that the intended emotion recognition was: 86,4% for neutral, 81% for happy, 86,8% for sensual, 82,7% for aggressive, and 98,8% for sad. The major confusion was between happy and aggressive emotions.

VoQ vs. HNM parameter modification

Figure 3.1 depicts the block diagram of the proposed VoQ and prosody modification system. As can be seen, the system is divided into three stages. First, HNM analysis and re-synthesis blocks perform the signal parametrisation for later manipulation and the reconstruction of the modified signal's waveform, respectively. The HNM representation was chosen for its flexibility which allowed us to modify more VoQ parameters than PSOLA systems (PSOLA does not make a frequency-based analysis nor separate the aperiodic and periodic components of the speech as HNM does).

Next, prosody and VoQ are modelled. Prosody is predicted by a Case Based Reasoning prediction system (CBR) obtaining the target information for each phoneme: fundamental frequency f_0 and energy contours and segmental durations [Iriondo, 2008]. VoQ parameters are computed as described in section 3.1. It is important to mention that while prosody parameters were used in all expressive style conversions, specific VoQ parameters were chosen for a given target style.

Finally, the expressive modifications were carried out by applying the estimated prosody and modifying the measured VoQ parameters to the original neutral style speech signal, and using the HNM signal representation. VoQ modifications were based on a multiplicative factor (with a constant value for each of the four expressive target styles and VoQ parameters) that was previously computed as the ratio between mean VoQ parameter values of the target and the original neutral corpora.

Table 3.1 lists the different VoQ parameters used for the transformation and their percentage variation from the original neutral parameter value. Note that the "–" symbol designates parameters not used in the transformation.

Some VoQ parameters (*hammI, pe1000* and *do1000*) modify the energy of specific frequency bands, thus altering the overall energy in the signal. A restriction for maintaining the overall signal's energy

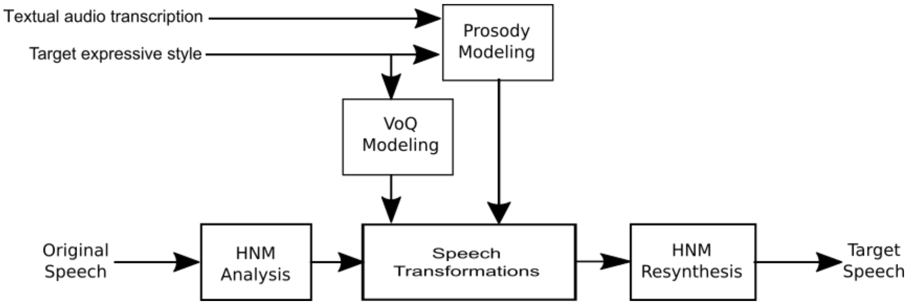


Figure 3.1 — Block diagram of the proposed VoQ transformation system. From [Monzo et al., 2010]

%factor	jitter	shimmer	HNR	hammI	pe1000	do1000
Happy	–	–	–	-60	110	–
Sensual	–	85	-50	155	-50	–
Aggressive	-20	-45	–	-70	220	–
Sad	-45	90	–	655	-75	–

Table 3.1 — Set of selected VoQ parameters for the natural-to-target expressive speech style transformation and the % factor to be applied for every involved parameter. ('–' parameters not involved in the transformation) From [Monzo et al., 2010]

is applied. Consider a VoQ parameter (P_{VoQ}) as a relation between two spectral bands (E_1) and (E_2), thus we have

$$P_{VoQ} = \frac{E_1}{E_2}. \tag{3.10}$$

VoQ modifications will be done by a given multiplicative factor β, resulting a modified parameter P^m_{VoQ}, which can be expressed as

$$P^m_{VoQ} = \frac{E^m_1}{E^m_2} = \beta \frac{E_1}{E_2}. \tag{3.11}$$

As can be seen in equation (3.11) the modified signal has the spectral sub-band's energies modified (E^m_1) and (E^m_2). These energies are modified by a multiplicative factor associated for each subband,

$$\begin{aligned} E^m_1 &= \beta_1 E_1 \\ E^m_2 &= \beta_2 E_2. \end{aligned} \tag{3.12}$$

Combining the processes we have

$$P^m_{VoQ} = \frac{E^m_1}{E^m_2} = \frac{\beta_1 E_1}{\beta_2 E_2} = \underbrace{\left(\frac{\beta_1}{\beta_2}\right)}_{\beta} \frac{E_1}{E_2}. \tag{3.13}$$

So considering (3.13) in order to apply the equal energy constraint we establish the following relation

$$E_T = E^m_1 + E^m_2 = E_1 + E_2, \tag{3.14}$$

forcing the energy of the original and modified signals to be the same. Combining the previous equations, the required multiplicative factor for each sub-band in order to modify the VoQ parameter while presenting the original signal energy can be found isolating β_1 and β_2 from

$$E^m_1 + E^m_2 = \beta_1 E_1 + \beta_2 E_2 = \beta\beta_2 E_1 + \beta_2 E_2 = \beta_2(\beta E_1 + E_2) = E_1 + E_2$$

$$\begin{aligned} \beta_2 &= \frac{E_1 + E_2}{\beta E_1 + E_2} \\ \beta_1 &= \beta\beta_2 = \frac{E_1 + E_2}{E_1 + \left(\frac{1}{\beta}\right) E_2}. \end{aligned} \tag{3.15}$$

Then, since the energy of either band, E_1 and E_2, can be computed with the HNM parameters, and β is the multiplicative factor for modifying the VoQ parameter and is already known, the multiplication factors for each band (β_1 and β_2) can be computed with equation (3.15).

3.2.2 HNM parameter modifications

Due to the definition of each parameter, it is important to note that modifying one parameter affects others which model shared characteristics, especially those that apply modifications directly to spectral bands. For instance, modifying the hammI parameter, changing bands below and above 1000 Hz may have effects on the $pe1000$ parameter, which works at frequencies around 2000 Hz. In order to minimize the impact among them, and considering the modifications that each one requires, the following modification sequence was proposed:

(1) Jitter, (2) HNR, (3) $pe1000$, (4) hammI, (5) Shimmer

For further details of the VoQ parameter modification expressions, see [Monzo, 2010].

Jitter and Shimmer parameters are both modified based on analysis and modification of short time pitch and signal amplitude variations, respectively. This procedure is completely decoupled from the HNM parametrization as jitter is related by fundamental frequency, then the modification procedure is that of a simple pitch modification (the pitch modification procedure is explained in section A.4).

Harmonic-to-Noise Ratio (HNR) Taking the amplitudes and frequencies of the harmonic part and the energies from the stochastic part, the mean energy for each part is computed. Next the modifying factor (β) is computed from the relation between each part's energy in the original signal and the desired target energy balance between harmonic and stochastic part.

Hammarberg Index ($HammI$) The modifications are carried out for each frame k, computing the energies of the bands below and above 2000 Hz, ensuring with an α factor that the total speech signal energy is unaltered after the modification. Finally the value modified is that frequency which has the maximum amplitude in each sub-band for each frame k. If the whole band were modified, that would entail a bigger side effect in the $pe1000$ parameter.

Relative amount of energy above 1000 Hz ($pe1000$) The energy for each frame is computed for the bands below and above 1000 Hz and adjusted according to the target energy distribution. In this case the overall signal's energy is also maintained using a regularization factor $\alpha^{(k)}$.

3.3 Evaluation and results

In order to evaluate whether the proposed methodology for expressive speech transformation based on prosodic plus VoQ modifications improves the perceptual quality compared with an only prosody-based approach, subjective tests were carried out using a web platform designed for this type of experiment [Planet et al., 2008]. In this evaluation, eight neutral utterances were collected from the corpus and converted to each of the four target expressive styles (happy, sensual, aggressive, and sad) using both methods, prosody only, and prosody combined with VoQ modifications, resulting in a total of 32 new utterances for each method. For each pair of utterances the question *"From which utterance* *is the target emotion best perceived?"* was asked. Listeners were allowed to choose one of the following seven possible answers; based on the Comparative Mean Opinion Scale (CMOS) test: *"much more"*, *"more"* or *"slightly more than the other"*, or *"the same"*, with scores of 3, 2, 1, 0, -1, -2 and -3. Positive values were assigned to cases in which the combination of VoQ and prosody improved the ability to perceive the expressive style. A total of 15 people completed the evaluation, of which 40% were experts in speech technology and 60% were no experts.

The obtained results, depicted in figure 3.2, demonstrate the preference of applying VoQ with prosody over only prosody modifications. This can be seen in the positive polarisation of the CMOS values; medians are situated around 1 (above 0), with 50% of the central values inside the range [0-2].

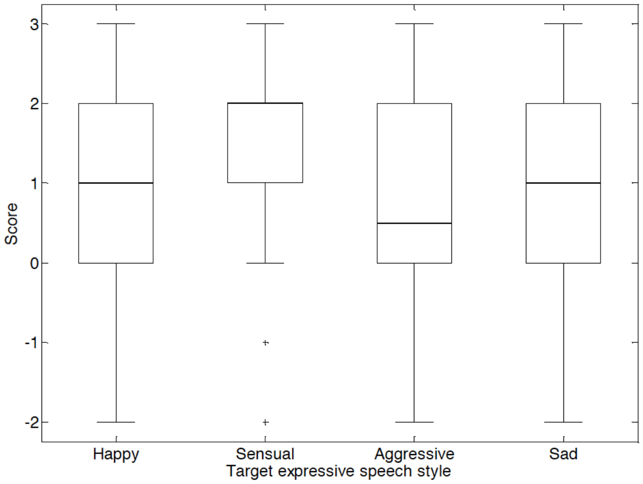


Figure 3.2 — CMOS evaluation for using VoQ with prosody or only prosody. Positive tendency indicate a preference of the synthesis using VoQ and prosody together, whereas negative trend correspond to a preference for using prosody only. From [Monzo et al., 2010]

Moreover, in the way of having a statistical support of the previous reported conclusions, an analysis of the median and confidence intervals was performed using the *Wilcoxon test*. The test corroborated the significance of the obtained results, confirming that using VoQ along with prosody is preferred over using prosody alone.

Then, the modelling and modification of VoQ together with prosody for conveying expressiveness in text-to-speech synthesis can avoid the need of using multiple emotion corpora. This expressiveness can be conveyed through the use of only one neutral speech corpus and an HNM-based technique that modifies the generated signal (VoQ and prosody attributes). In this chapter VoQ modifications using multiplicative factors have been proposed based on ratios of mean values for the whole expressive style. In order to be able to convey VoQ variations inside a single synthesized utterance, the multiplicative factor should be able to capture the VoQ parameter time dynamics.

On the other hand, the synthesized speech signal quality was significantly degraded when introducing VoQ modifications. Figure 3.3 presents the signal quality evaluation for emotion conversion using prosody only and when using prosody and two VoQ parameters (*Jitter* and *Shimmer*). As can be seen, when adding *Jitter* and *Shimmer*, the MOS rating for each expressive style decreases with compared to the case of using only prosody.

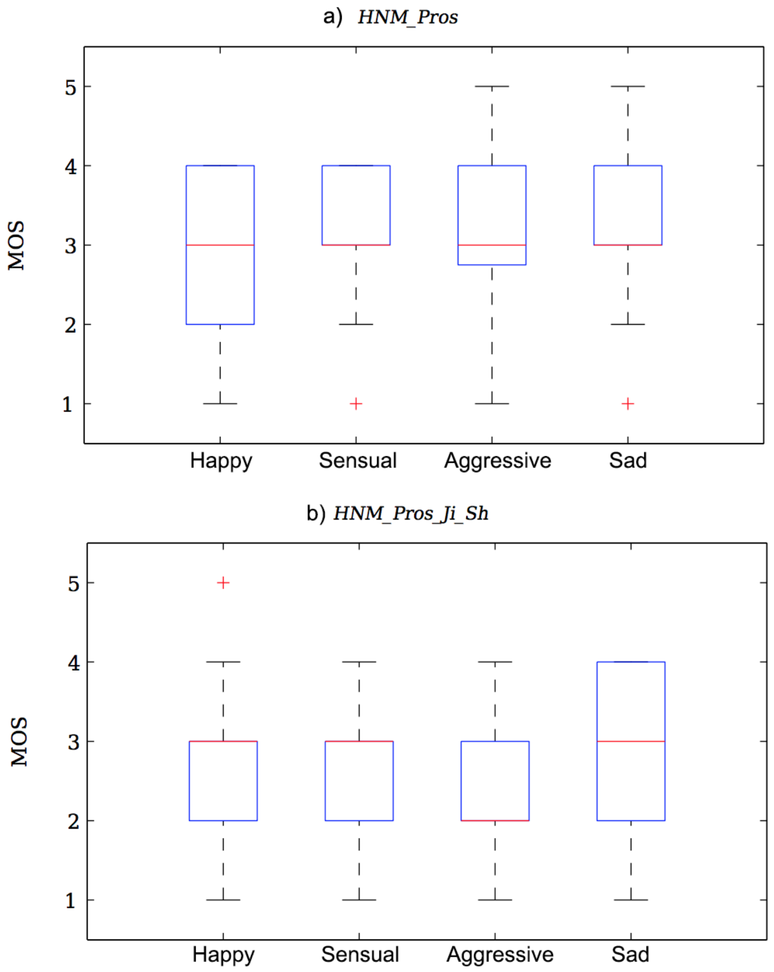


Figure 3.3 — MOS signal quality evaluation for emotion conversion with HNM using: a) Prosody only, b) prosody, Jitter and Shimmer. From [Monzo, 2010]

3.4 Conclusions

This chapter presented the work conducted in [Monzo et al., 2010] where following the objective *O1*, experiments for transforming neutral speech into different expressive styles using HNMs were conducted. These experiments also compared two speech modification approaches, one modifying only prosody, and the other combining prosody with low-level VoQ parameter modifications. The conducted experiment results confirmed the relevance of VoQ in the expressive style perceived by the listeners; figure 3.2 shows the preference for introducing VoQ modifications for conveying expressiveness in synthetic speech, and also confirmed the feasibility of using the HNMs for modifying low-level VoQ parameters, thus validating the hypothesis (*H1*): (HNM is a suitable representation for conducting VoQ modifications for expressive speech synthesis and VoQ can be added as an extra acoustic feature to improve the expressiveness in synthetic speech). Because the presented work does not compare HNM against other speech signal representations, it cannot be concluded that HNM is the best model, However, harmonic models are among the best speech signal representations regarding synthesized speech signal quality [King and Karaiskos, 2013]. Thus, HNM can be considered a suitable speech signal representation for conducting VoQ modifications, answering the research question (*Q1*).

Despite achieving good results in terms of emotion recognition, the quality of the synthesised signal was seriously degraded compared with applying only prosody modifications (see figure 3.3). The causes for this signal quality degradation are: **i)** The number of signal manipulations performed; up to five parameters (jitter, shimmer, HNR, Hammarberg Index (hammI) and the relative amount of energy above 1000 Hz (pe1000)) were modified separately, based on a five-stage procedure in which each stage was specifically designed to modify a unique VoQ parameter; **ii)** the existing interdependence of some spectral parameters (i.e., the Hammarberg Index (hammI) and the relative amount of energy above 1000 Hz (pe1000)), where several parameters represent overlapping bands of the speech spectrum; **iii)** use of a constant modification factor β for the whole utterance.

The modification of VoQ parameters was done with a multiplicative factor β obtained from analysing different expressive style speech corpora and obtaining a multiplicative factor transforming from neutral to each of the available expressive styles. For a given target expressive style, the proper multiplicative factor β was selected, and this was constant for the whole synthesized utterance. Thus, the converted voice quality was obtained by multiplying the original VoQ parameters by a constant factor β corresponding to the desired target emotion. Having the multiplicative factor β constant along sentences precludes the possibility os varying the expressive style inside each utterance, reducing the system's flexibility.

As was explained in section 2.3, a tight relation exists between prosody and spectral envelope, or in this case, VoQ [Campbell and Mokhtari, 2003], so when performing prosody modifications it might be necessary to adjust the spectral envelope accordingly [Grichkovtsova et al., 2012]. This supports the conclusion that if prosody is allowed to vary over time, the VoQ parameter modifications should also introduce time dynamic variations. Thus, VoQ and prosody should be modified accordingly when applying speech signal modifications. Due to applying a constant VoQ along the sentence, the resulting synthesised speech quality was degraded. In order to vary the VoQ parameters along an utterance, a more flexible representation capable of capturing these VoQ time dynamics is necessary, leading to the work presented in the following chapters 4 and 5.

To conclude, the experiments presented in this chapter verified the feasibility of using HNMs for conducting VoQ modifications and apply it for converting a neutral style speech signal more expressive styles. The conducted work demonstrates the feasibility of HNM for conducting expressive speech conversions modifying prosody plus VoQ, which could be used for extending TTS system flexibility and relaxing the need for a larger speech corpus for conveying multiple expressive speaking styles.

However, in order to avoid the possible degradation of the output signal due to transforming multiple and interdependent VoQ parameters, in chapter 4 a new approach is proposed; instead of using multiple low-level VoQ parameters, a single parameter is used for modifying the overall spectral envelope at once. And next, building on the concepts in chapter 4, chapter 5 presents a new methodology for conducting VoQ modifications, which not only allows to convey expressiveness observed in the corpus, but generate new ones interpolating model parameters that represent the smooth low frequency component of the spectral envelope.

CHAPTER 4

Transferring vocal effort with LPC

Contents

4.1 NECA database and HNM parameterization 62

4.2 Spectral emphasis modification with HNM 65

4.3 Evaluation and results 67

4.4 Discussion 73

4.5 Conclusions 76

Section 3.3 presented the results of [Monzo et al., 2010]; demonstrating the impact of combining VoQ modifications together with prosodic modifications, and the feasibility of doing it with HNM parameters. However, in section 3.4 some problems causing signal degradation were highlighted, mainly the amount of signal modification performed without considering the relations between the parameter. To simplify the procedure and focus on high-quality modifications, the number of parameters to be modified were reduced to just one, vocal effort (explained in section 2.3); this was chosen for its salient role in expressive speech characterisation as reported in [Schröder and Grice, 2003]. This choice led to the hypothesis (*H2*) -*HNM can be used for transferring VE between two original recordings with different VE levels.*-.

To validate this hypothesis the following specific objectives (*O2*) were set:

- ▷ Represent VE with a low-order AR filter applied to the HNM's harmonic component parameters.
- ▷ Use the low-order AR filter representation for exchanging VE between two parallel utterances using HNMs.
- ▷ Verify that the VE transfer can be conducted for all VE levels available in the corpus (low, neutral, and high).

These objectives were reached in the work presented in [Calzada and Socoró, 2011] and [Calzada and Socoró, 2013] which proposed an algorithm for vocal effort (VE) modification using HNM parametrisation, based on an adaptation of the Adaptive Pre-emphasis Linear Prediction (APLP) [Nordstrom and Driessen, 2006] technique (briefly introduced in section 2.4).

The proposed methodology in this chapter was validated through perceptual tests that demonstrated the ability of the procedure to modify the VE level perceived in the synthesized speech signal, to not only decrease the vocal effort level, as proposed in [Nordstrom and Driessen, 2006], but also to increase the perceived level (e.g., from low to high vocal effort level).

This chapter is organised as follows. Section 4.1 briefly describes the speech database used in the experiments. Section 4.2 presents the proposed HNM-based vocal effort modification system based on APLP. Finally, sections 4.3 and 4.5 present the results and conclusions, respectively.

4.1 NECA database and HNM parameterization

The speech material used to validate the proposed vocal effort transformation methodology was a German diphone set recorded with three degrees of vocal effort (vocal effort levels were labelled as *high*, *neutral*, and *low*), [Schröder and Grice, 2003]. The corpus was divided into six datasets containing nonsense words three syllables in length, with voiced recorded with a constant pitch. Recordings of the three vocal effort levels from one male and one female speaker were available. As explained in [Schröder and Grice, 2003], the data was automatically labelled and subsequently hand-corrected.

The proposed method is based on mapping smooth frequency variations along the spectral envelope of a reference signal, in the conducted experiment *neutral* vocal effort level, to the target signals, *low* and *high* vocal effort levels. *Neutral* VE level is between *low* and *high* VE levels; thus, *neutral* VE was chosen as the reference point in order to minimise the degree of modification required to cover all three VE levels. Thus, from the synthesis point of view, and considering the general goal proposed

(*O4*), it makes sense to select the *neutral* VE level as a reference for modifying the neutral speech (close to *neutral* VE level) to more expressive speech styles (*high* and *low* VE levels).

In [Nordstrom et al., 2008] the voice samples used were isolated vowels, so the spectral characteristics corresponding to the phone were very similar between utterances. However, in the experiment presented in this chapter, the utterances contained multiple phones (consonants and vowels) which have very different spectra. So in order to map spectral characteristics corresponding to the same phone, time mapping was an essential step. If no time mapping was conducted, the spectral envelope of two different phones could be erroneously mapped together and mistake the spectral differences to be different vocal effort levels. Thus with proper time mapping, VE contribution to the spectrum could be isolated from the phone spectral characteristics. In order to take into consideration the effects of co-articulation, spectral variation of a phone due to its proximity to the following phone, the time alignment was applied to all analysis time instants. This means that for every set of HNM parameters from the reference signal (neutral VE), the corresponding set of HNM parameters for the same points were extracted from the low and high VE signals.

Pitch marks from the reference signal (neutral VE) were used for initialising the analysis time instants. Then, given a reference word from the neutral vocal effort version, the corresponding time alignment functions for the high and low VE versions were computed using hand-corrected phonetic boundary marks for each utterance using linear interpolation in between (see figure 4.1).

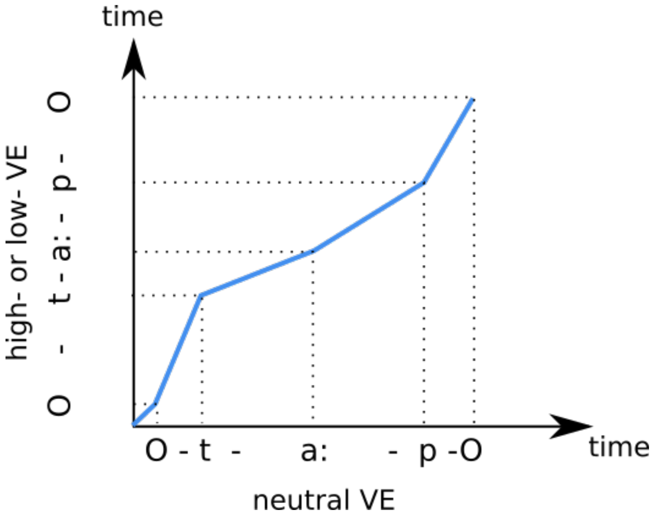


Figure 4.1 — Time mapping function construction from phoneme boundary marks for the logatom */O/-/t/-/a:/-/p/-/o/*.

In order to perform a more straightforward vocal effort transfer (using a copy synthesis strategy), all parallel utterances should contain the same number of frames i.e., analysis time instants. To compute the time instants for high and low VE versions, the reference time instants were mapped according to the time mapping functions. Due to differences in speech rates, this mapping can lead to consecutive analysis time instants with lapses greater than the period in which the signal characteristics are usually supposed to remain stationary [Depalle and Hélie, 1997,Quatieri and McAulay, 1998,Stylianou, 1996, Degottex and Stylianou, 2013], approximately a couple of pitch periods (see figure 4.2). These lapses correspond to an undersampled signal, which might cause quality degradation when applying signal modifications. To prevent leaving any pitch period unanalysed, the mapped analysis time instants were post-processed adding new analysis time instants, such that the distance between each pair of consecutive analysis time instants became smaller than two pitch periods. Referring to the case depicted in figure 4.2, new analysis time instants were inserted between the mapped times marked with an exclamation mark for the high or low vocal effort time instants. This process was performed for both the high and low VE versions, thus obtaining a new array of analysis time instants usable with prior mapping for all parallel utterances. This process ensured that all parameterised versions, each version corresponding to a different vocal effort level, for a given word contain the same number of aligned frames and are properly sampled.

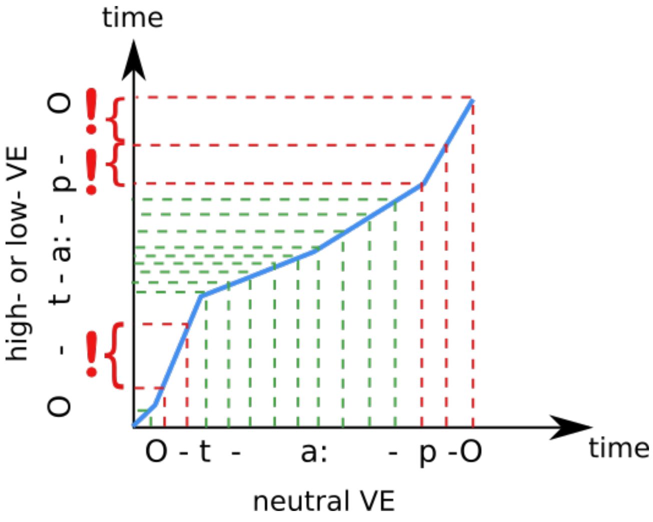


Figure 4.2 — Reference analysis time instants mapped to high or low VE version. The red exclamation marks indicate the analysis time instants with time lapses larger than the period where the signal parameters are stationary, causing undersampling of the signal.

Once the array of reference analysis time instants was computed for each utterance, the corpus was analysed and proper HNM parameters were extracted and saved. Thus, the analysis of the signals was done at a variable frame rate, but the length of the analysis windows was always two pitch periods. Amplitudes, phases, and angular frequencies for each frame could vary according to the relation between instantaneous pitch and maximum voiced frequency, which was set to $5KHz$. For the stochastic part, the analysis time instants were also computed following the same process, with the difference that the initial time instants were equidistant with a $10ms$ period. And finally, the 30 order LPC filter was computed.

This computation and mapping of the reference analysis time instants was chosen instead of using a constant frame rate analysis on time aligned and length normalized utterance (applying time modifications to the original signals) to avoid having to perform resynthesis twice, which would introduce more errors into the signal model and require temporal modification of the signal prior to the analysis for extracting the vocal effort.

4.2 Spectral emphasis modification with HNM

The transformation methodology proposed here was inspired by [Nordstrom et al., 2008] (explained in section 2.4), where the APLP algorithm was presented and validated for transforming high-effort voices into breathy (low-effort) voices. In our proposal, this methodology was adapted to work within the HNM parameterisation framework by modifying the HNM parameters from the harmonic part during the conversion stage. Moreover, the methodology was validated not only for reducing vocal effort, as in [Nordstrom et al., 2008], but also for conversions that entail an increase of the vocal effort level.

In [Nordstrom et al., 2008], a spectral emphasis filter was used as a tool to produce both glottal source and vocal tract changes due to vocal effort variations. This idea was reused in this work but employing HNM parameters rather than LPC residuals. In this technique, after the HNM parameters of the source and target signals are computed, the vocal effort conversion procedure is fully carried out in the HNM parameter space. The complete transformation methodology is depicted in figure 4.3, in which both spectral emphasis and prosody are modified through incorporating time, pitch and energy modification stages. Source and target HNM parameters must be time aligned in order to guarantee a perfect time mapping between the two signals, as the objective is to transfer the low frequency, or smooth component variations, of the spectral envelope from one signal to the other. Thus, for a given pair of signals, the harmonic parts have the same number of frames, as do the stochastic components (see section 4.1 where this synchronisation is explained). Unlike the work presented in [Nordstrom et al., 2008], the methodology proposed in this dissertation does not require the estimation of the formant filter $V_F^s(z)$ of the source signal $S(z)$. The formant filter $V_F(z)$ would have been removed before applying the desired vocal effort with $H_E^t(z)$ and later reintroduced in the final step (see equation (2.12)).

The proposed methodology depicted in figure 4.3 is based on transforming the source signal in order to transfer from the target signal: its time marks (time modification), its pitch curve (pitch modification), its spectral emphasis evolution (with the aforementioned APLP-based approach), and its energy time evolution (energy transformation).

- ▷ The **time modification step** computes time modification factors (ρ^k) using the time-aligned source and target analysis instants (t_s^k and t_t^k, respectively). The time modification factors are necessary for adjusting the linear-in-frequency phase term of the HNM signal phases during re-synthesis (see section A.3).

- ▷ The **pitch modification step** replaces the source pitch curve by the target pitch curve and interpolates source amplitudes and phases at the given new harmonic frequencies (see section A.4). The corresponding pitch modification factors (λ^k), which also affect the linear-in-frequency phase term (see section A.4), are then computed.

- ▷ The **spectral emphasis modification step** (SE Modification in figure 4.3) produces the main transformation effects regarding the signals' vocal effort variations. As shown in figure 4.4, the frame-based source and target spectral emphasis functions (denoted in the diagram as $H_s^k(f)$ and $H_t^k(f)$, respectively) are computed, for each frame k, using pitch frequencies (${f_0}_s^k$ and ${f_0}_t^k$) and the amplitude vectors of the harmonics ($\boldsymbol{A}_s^k$ and $\boldsymbol{A}_t^k$).

  As in [Nordstrom et al., 2008], in this study, the spectral emphasis functions are calculated using a low-order LPC model [Nordstrom et al., 2008] (order 3 in the experiments) using the simple and efficient procedure described in section A.6. Due to working with speech signals containing CV syllables instead of sustained vowels, the coarticulations of vowels and consonants might affect the stability of the SE model estimation. Thus, in this model, a smoothing step is applied to the spectral emphasis to reduce audible artifacts in the final modified signal. Linear Spectral Frequency (LSF) coefficients are used instead of LPC for the smoothing process to prevent all-pole filter instabilities. The smoothing is applied to each LSF coefficient time trajectory, and a local robust linear regression is used. The smoothing is performed along each coefficient time trajectory with a span of 60 frames for computing the smoothed value. After the smoothing step is conducted, the LSF coefficients are converted back to LPC. Then, the spectral emphasis of the time and pitch-aligned source signal is subtracted from its complex spectrum, dividing its complex harmonic amplitudes[1] ($\boldsymbol{C}_s^k$) by a sampled version using the source harmonic frequencies of its spectral emphasis function ($H_s^k(f)$), thus producing $\tilde{\boldsymbol{A}}_s^k$ and $\tilde{\boldsymbol{\Phi}}_s^k$. The target spectral emphasis function is then applied to the time and pitch-aligned signal by sampling $H_t^k(f)$ at the source harmonic frequencies (${f_0}_s^k$) and subjected to posterior multiplication with the time and

[1] $\boldsymbol{C}_s^k = \{a_1^k e^{j\varphi_l^k}, a_2^k e^{j\varphi_2^k}, \cdots, a_l^k e^{j\varphi_l^k}\}$, where $l \in [1, .., L_k]$

pitch-aligned signal's complex harmonic amplitudes[2] ($\tilde{\boldsymbol{C}}_s^k$) as

$$\begin{aligned} a_l^{k\prime} &= a_l^k \, \frac{|H_t^k(f_l^k)|}{|H_s^k(f_l^k)|} \\ \varphi_l^{k\prime} &= \varphi_l^k + \angle H_t^k(f_l^k) - \angle H_s^k(f_l^k) \end{aligned} \tag{4.1}$$

where $a_l^{k\prime}$ and $\varphi_l^{k\prime}$ are the modified amplitudes and phases for the harmonic l in frame k. Thus, in this work, vocal effort is transformed in the harmonic part of the speech signal.

▷ Finally, the **energy modification step** (see figure 4.3) adjusts the source-modified signal amplitudes ($\hat{\boldsymbol{A}}_s^k$) and the original source noise variances (p_s^k) to match the total target frame energies computed with them own harmonic amplitudes ($\boldsymbol{A}_t^k$) and noise variances (p_t^k). In this case, the same smoothing procedure as previously described for spectral emphasis is applied to the multiplicative frame-based energy conversion factor obtained as the quotient of the target and source frame energies.

As shown in figure 4.3, the vocal effort modification system considers both prosodic (time, pitch, and energy) and voice quality (spectral emphasis or vocal effort) modifications. The presented block order (time → pitch → SE → energy) was chosen based on the alignment restrictions of the spectral emphasis step. Energy modification was decided to be at the end of the process to ensure that the energy of the modified signal matches that of the target because some energy fluctuations may occur when pitch and spectral emphasis modifications are carried out.

4.3 Evaluation and results

Two perceptual evaluations were designed with the online testing platform for multimedia evaluation TRUE [Planet et al., 2008]. The first evaluation with 41 users (27 male and 14 female) focused on comparing the overall quality of the proposed method based on the HNM, against the original APLP. The second evaluation with 46 users (31 male and 15 female) evaluated the proposed methodology against the original recordings. Among the participants, 26.83% had expertise in speech synthesis technology or special musical and singing training. This subgroup is later referred to as the experts.

Each evaluation consisted of six nonsense words (see table 4.1) with six vocal effort conversions covering all possible vocal effort transitions (from to target signal) for the given corpus (neutral to low (*N2L*), high to low (*H2L*), low to neutral (*L2N*), high to neutral (*L2M*), low to high (*L2H*), and neutral to high (*N2H*)).

[2] $\tilde{\boldsymbol{C}}_s^k = \{\tilde{a}_1^k e^{j\tilde{\varphi}_1^k}, \tilde{a}_2^k e^{j\tilde{\varphi}_2^k}, \cdots, \tilde{a}_l^k e^{j\tilde{\varphi}_l^k}\}$, where $l \in [1, .., L_k]$

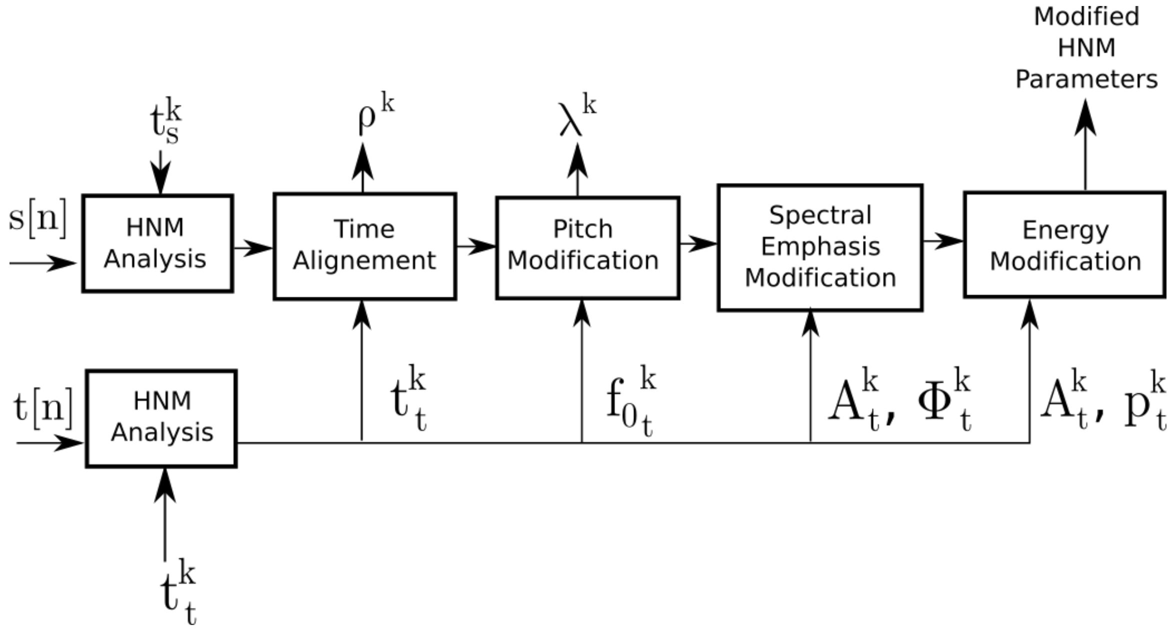


Figure 4.3 — Block diagram of the proposed vocal effort transformation methodology. SE stands for spectral emphasis. From [Calzada and Socoró, 2011]

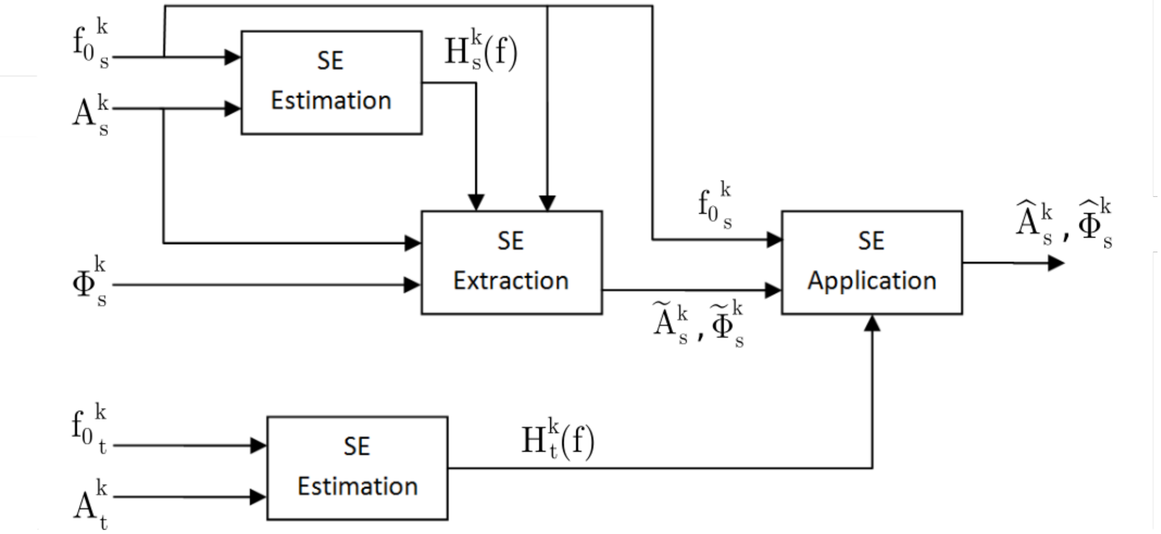


Figure 4.4 — Detail of the Spectral Emphasis (SE) transformation block depicted in figure 4.3. From [Calzada and Socoró, 2011]

Table 4.1 — SAMPA [Wells, 1997] transcription of the nonsense words used in the evaluation.

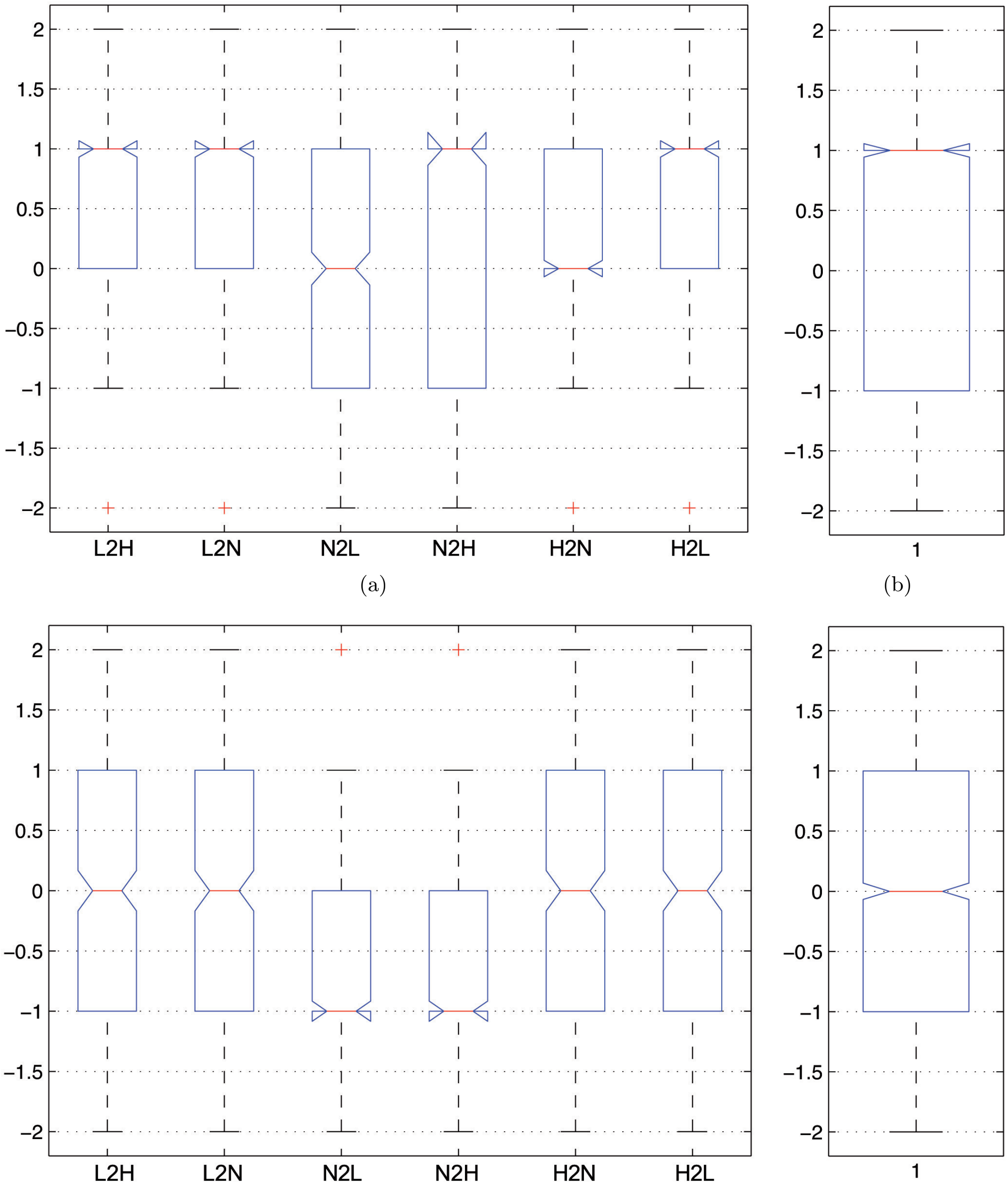
[aI - t - a: - p - aI]	[t - a: - f - u: - f - a:]	[b - U - t - ae - t - a]
[t - a: - tS - aI - tS - a:]	[t - a: - m - U - m - a:]	[O - t - a: - p - O]

When evaluating the overall quality of the proposed method against APLP [Nordstrom et al., 2008], subjects were presented with a reference audio, an original recording expressing a desired (target) vocal effort level, and two synthetic stimuli labelled as A and B corresponding to either of the methods HNM and APLP respectively (where the source signal was transformed following the corresponding methodology to imitate the target signal). At each step, the audio samples being evaluated were randomly switched to prevent a pattern from arising in the subjects responses. Participants were requested to evaluate two aspects (1) the vocal effort distance to the reference and (2) the quality of the signal from the audio stimuli by answering the following questions: *Obviating the signal's quality and focusing only on vocal effort, evaluate which audio (A or B) expresses a closer vocal effort to the reference.* and *Obviating the vocal effort, which audio has better signal quality?*.

Both questions were answered based on a five-point Likert scale. Regarding vocal effort distance to the reference, the options were: *A is much closer*, *A is slightly closer*, *Both are more or less equally close*, *B is slightly closer*, *B is much closer*. For rating the signal's quality, the options were: *A is much better than B*, *A is slightly better than B*, *A and B are more or less the same*, *B is slightly better than A*, and *B is much better than A*.

As shown in figure 4.5, there was a tendency toward the proposed methodology based on HNM when evaluating the distance to the vocal effort of the original target reference (figure 4.5(b)) without suffering serious signal degradation compared to its counterpart APLP (figure 4.5(d)). Apart from conversions from the neutral vocal effort, which appeared to present slightly greater quality with APLP, there was no strong evidence of serious signal degradation for any of the conversions, even for those requiring a higher degree of modification, such as *L2H* or *H2L* (figure 4.5(c)). In addition, there was a clear tendency towards the proposed methodology regarding vocal effort proximity to the reference utterance for nearly all of the conversions except *N2L* and *L2N* (figure 4.5(a)), where both systems were perceived as performing similarly. Figure 4.6 presents the statistical results considering only data from the experts subgroup, where the preference for the presented methodology over APLP was more noticeable.

To evaluate the performance of the proposed methodology based on HNM against the original recordings, only utterances from the male gender were used. In this test, users were presented with two audio stimuli, one of which was used as a reference and the other, as the test audio. The reference was always an original recording expressing neutral vocal effort. The audio used for evaluation included a mixture of original recordings of the three types of vocal effort and synthesised audio, which were generated with the proposed HNM-based schema and covered all possible vocal effort conversion combinations. The subjects were requested to evaluate two properties of the audio under evaluation the expressed vocal effort and the signal quality compared against the reference by answering the



(a)

(b)

(c)

(d)

Figure 4.5 — Boxplots for (a) vocal effort approximation to the reference audio by conversion, (b) averaged value of the proximity to the reference, (c) the signal's quality preference by conversion and (d) average of the signal's quality preference. Positive values correspond to the proposed HNM methodology, and negative values correspond to APLP. Conversion type legend: neutral to low (*N2L*), high to low (*H2L*), low to neutral (*L2N*), high to neutral (*H2N*), low to high (*L2H*), and neutral to high (*N2H*). From [Calzada and Socoró, 2013]

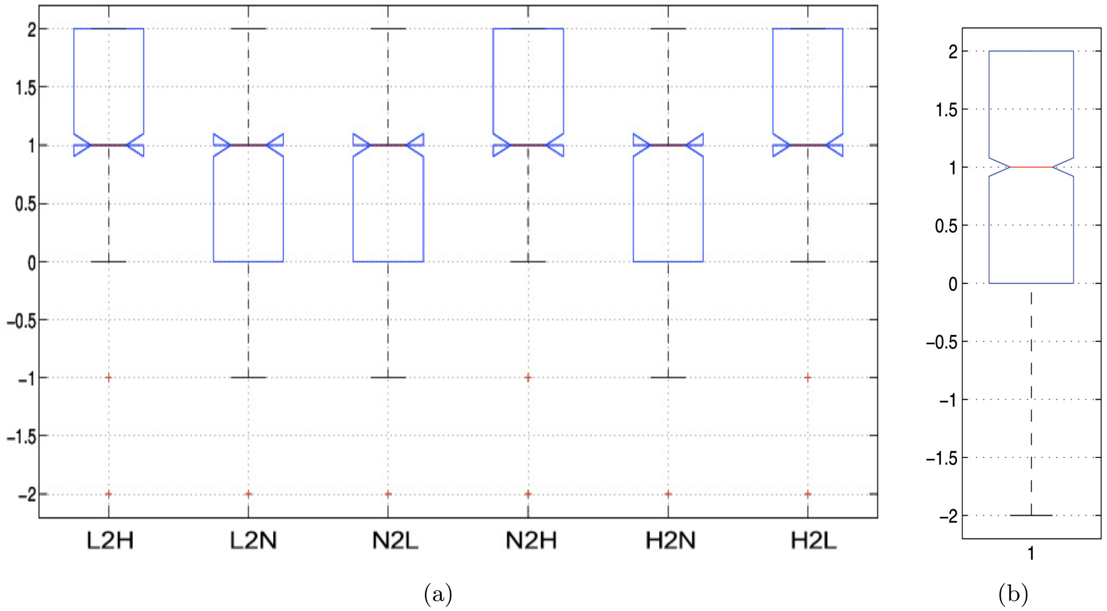


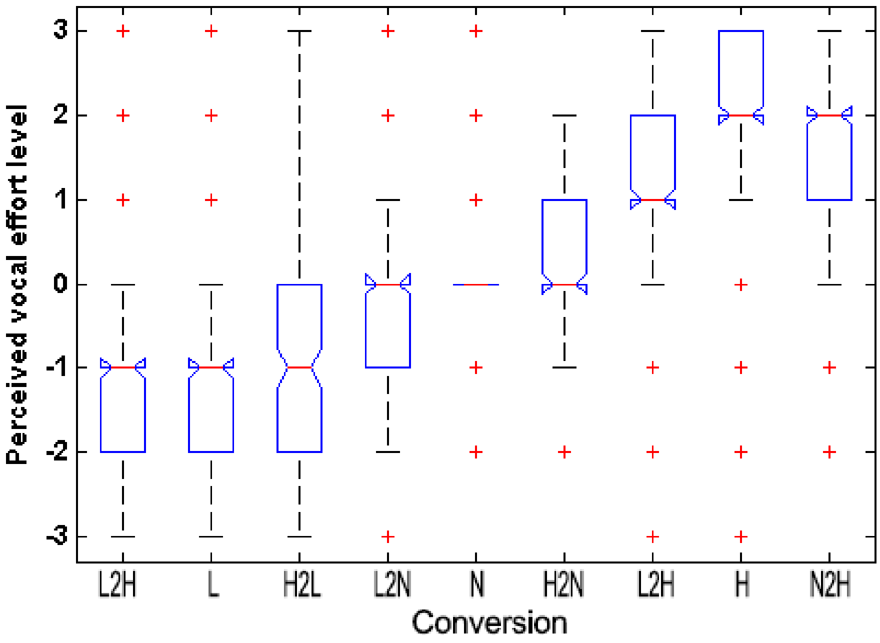
Figure 4.6 — Boxplots from the expert subgroup data. (a) Vocal effort approximation to the target (positive values correspond to HNM, and negative values correspond to APLP). (b) Averaged system preference for the experts. Conversions legend: neutral to low (*N2L*), high to low (*H2L*), low to neutral (*L2N*), high to neutral (*H2N*), low to high (*L2H*), and neutral to high (*N2H*). From [Calzada and Socoró, 2013]

following questions: *Considering the Reference audio as a Modal vocal effort, grade the vocal effort of the audio to evaluate.* and *Considering the Reference audio as Excellent quality, what is the quality of the evaluation audio?.*

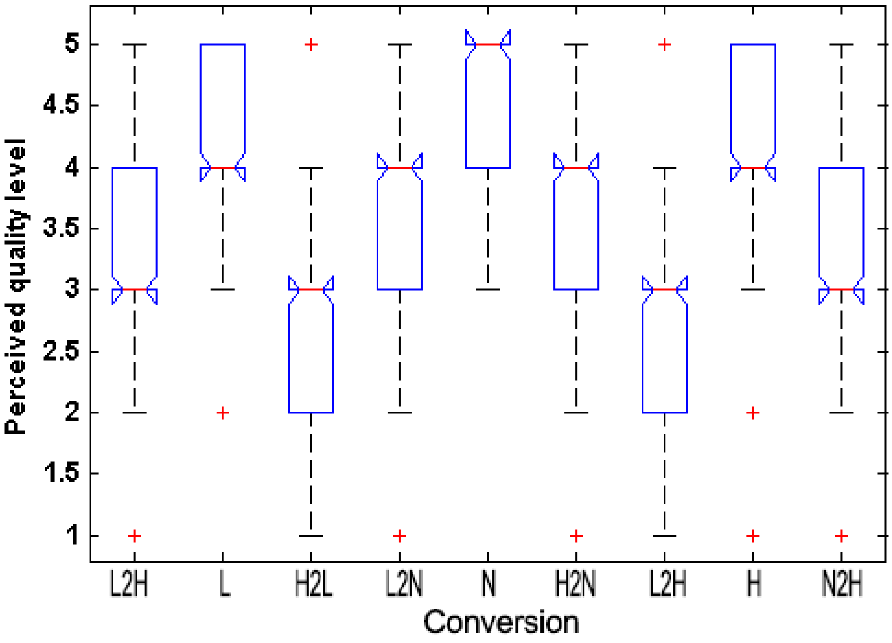
For each of the six words, subjects evaluated six vocal effort conversions and the corresponding three original recordings, one for each vocal effort. These 36 audio stimuli entailed an average test length of 13 minutes. The expressed vocal effort was rated on a seven-point scale labelled as follows: *Completely Loud, Very Loud, Slightly Loud, Modal, Slightly Soft, Very Soft, and Completely Soft.* These labels were assigned the following values, respectively: 3, 2, 1, 0, -1, -2, and -3. In terms of signal quality evaluation, the mean opinion score (MOS) was used. Rating labels were *Excellent, Good, Fair, Poor* and *Bad* and assigned the values 5, 4, 3, 2, and 1, respectively.

The results of the perceptual experiment were statistically analysed with the Kruskal-Wallis test and the Bonferroni correction for multiple pairwise comparison purposes with $\alpha = 0.05$ (figure 4.7).

As shown in figure 4.7(a), the stimuli with converted vocal effort tended to be recognised as the expected target vocal effort. For instance, the medians of the *N2L* and *H2L* perceived vocal effort punctuations had the same value as the original low stimuli (L) evaluations. No significant differences were found when comparing vocal effort conversions sharing the same target. However,



(a)



(b)

Figure 4.7 — Boxplots for (a) perceived vocal effort and (b) quality MOS for each conversion (neutral to low (*N2L*), high to low (*H2L*), low to neutral (*L2N*), high to neutral (*H2N*), low to high (*L2H*), neutral to high (*N2H*)) and also for original unmodified stimuli (low (*L*), neutral (*N*) and high (*H*)). From [Calzada and Socoró, 2013]

significant differences were observed when comparing conversions toward different vocal effort levels[3]. For example, no significative differences of the median values of the distributions between *N2L, S*, and *H2L* were found, and these median values were significantly different from the other distributions obtained (*L2N, M, L2M, L2H, L, N2H*). In figure 4.7(b), the MOS results reveal that the most important signal degradation appears when large conversions are applied, such as converting from low to high levels of vocal effort (*L2H*) and vice versa (*H2L*). Conversely, better signal quality was perceived when applying conversions towards adjacent vocal effort levels, such as between low and neutral (*L2N*) or neutral and high (*N2H*) vocal effort levels.

4.4 Discussion

Current TTS systems present high intelligibility, but improving their naturalness and expressiveness remains an open research topic. Many US-TTS systems attempt to overcome this problem with extensive corpora to include sufficient diversity to meet the demands of the synthesis process. For instance, synthesising such expressive styles as happiness or sadness would entail replicating the corpus for each of these expressive styles, thus requiring large amounts of data and big corpus sizes. However, although the proposed system for transferring VE improves the flexibility of the system and can help to reduce the corpus size replacing speech signal utterances expressing different VE levels by their corresponding SE models, this procedure continues to require having an example signal for each expressive style to synthesize, in order to be able to analyse the signal and extract the proper SE models to use for the final synthesis. In order to use the proposed system, parallel corpora are required. However, the utterances do not need to be aligned or have the same length, since the first step is computing the reference time instants for multiple VE versions. On the other hand, it is required to have well-delimited phone boundaries for the utterances.

The direct applicability of the proposed method in the context of a US-TTS system might produce unnatural VE dynamics. In the presented experiment, the VE remained almost constant over the entire words. Within the speech corpus used for synthesis in US-TTS, VE may vary along the recorded utterances, each unit from the same utterance with a different VE level. This could cause unnatural VE evolution in the synthesized signal.

The proposed procedure is based on the results of previous work [Nordstrom et al., 2008] in which vocal effort was modified by varying the spectral slope of the speech signal using the APLP procedure. In the presented work, the idea of modifying vocal effort from the spectral slope has been reused and applied within the HNM model. Despite the fact that HNM has proven to offer high quality

[3] A threshold of 0.05 for the obtained p-value was chosen in the comparisons.

and flexibility for speech modification, to the knowledge of the author, no reports have obtained such promising results in modifying vocal effort by applying the HNM.

The study carried out on APLP by [Nordstrom et al., 2008] only applied this technique to isolated vowels and for conversions from high levels of vocal effort to breathy ones; in contrast, the experiments carried out in this work extend the method not only for isolated vowels but also for nonsense words containing several phonemes (vowels and consonants). In addition, the conducted experiments have demonstrated that APLP can be applied not only for converting high vocal effort registers to breathy voices, as in [Nordstrom et al., 2008], but also for increasing the vocal effort level, and this capability has been confirmed for both APLP and HNM-based systems.

The results presented in this chapter confirm that the proposed procedure for vocal effort modification with HNM is able to modify the vocal effort level of speech signals to any desired target without severe degradation of the signal quality (see figure 4.7). When comparing the proposed method using the HNM against APLP, modified utterances with the proposed method have been considered to approximate the desired target vocal effort levels slightly better than APLP (see figures 4.5(a) and 4.5(b)). This preference for the proposed procedure over APLP might be due to the periodic and impulsive permanent noise that appears in all signals generated with APLP. Regarding signal quality, there are no observed advantages between the compared methods (figure 4.5(d)), although some conversions present slightly better results with APLP. This result might be caused by the presence of some audible artifacts that were present in three of the audio stimuli that were generated with the proposed procedure. Despite its good signal quality, subjects strongly penalised the proposed method when unexpected artifacts appeared, even when evaluating the proximity of the converted voice quality level to the target. This finding explains the difference between the general results (figure 4.5) and those in which only speech technology experts and participants with special music and singing training were considered (see figure 4.6). The artifacts that appeared in utterances generated with the proposed method were typically located at coarticulation boundaries with affricate consonants, such as */tS/* in [t - a: - tS - aI - tS - a:]. These artifacts might be avoided by applying spectral emphasis modifications only for stable regions of vowels for voiced sounds, and disabling modifications for unvoiced sounds. Further work is needed to improve the estimation of spectral emphasis at coarticulation boundaries because this feature has been found to be the main cause of audible synthesised signal artifacts.

The proposed procedure exhibits special potential for TTS systems where it could be applied to allow the system to express several degrees of vocal effort in the synthesised speech without requiring the production of widely diverse expressive speech corpora. Moreover, the proposed method is fully scalable because it is compatible with other VoQ and prosody modification techniques; this compatibility arises from the fact that all speech modifications are directly applied to the HNM parameters.

In the present work, the spectral slope was modelled with a low-order LPC estimation function to emulate the reference system. In contrast to the reference method (APLP), the proposed procedure is not constrained to use LPC to carry out vocal effort modifications. All spectral emphasis modifications have been directly used to modulate the amplitude and phase parameters of the HNM model. Any function can be inserted into the proposed procedure in place of LPC to modulate the complex amplitudes, thereby achieving the desired vocal effort modifications [Fraiha Machado et al., 2013, Jokinen et al., 2014, Jokinen et al., 2015]. It would be good to study the use of simpler functions to model vocal effort, such as low-order polynomials, because these functions could be easily interpolated without suffering from instability problems that as might occur with LPC. These simpler functions would make vocal effort modification more straightforward. The next chapter studies the possibility of using polynomial models for modelling the spectral emphasis of speech signals, and allowing fine vocal effort modifications. It also relaxes the restriction of needing sample utterances for every VE level to synthesize, allowing interpolation of the VE models.

In the conducted experiments, vocal effort was transferred from target to source utterances using parallel corpora. However, to design flexible and expressive TTS systems, it would be desirable to have a vocal effort level prediction system capable of driving the vocal effort modification procedure without requiring parallel corpora. This capability would entail the analysis of vocal effort dynamics in expressive speech and the training of machine-learning algorithms to produce the optimal prediction for a given input text.

The results obtained here are encouraging considering that APLP applies the spectral slope modification to the entire spectrum, whereas the proposed method only applies modifications to the harmonic components that cover the range from 0 to 5 kHz, leaving the upper band of the spectrum unmodified. Applying vocal effort modifications to the upper band of the spectrum, as well, and smoothing the transition between the harmonic and stochastic component spectra, might improve the system modification range and signal quality.

To focus only on vocal effort and isolate it from other external factors, such as prosody or lexical meaning, the conducted experiments used nonsense words only, obtained from a specially designed corpus. However, further work should also consider using full sentences while evaluating the presented procedure combined, for instance, with prosody modifications to achieve more advanced speech modifications. Such combined speech signal modifications could be applied in expressive style conversion environments.

4.5 Conclusions

This chapter has introduced an adaptation of the spectral emphasis conversion method proposed in [Nordstrom et al., 2008] to modify vocal effort directly within an HNM parametrisation framework. According to objective *O2*, the proposed methodology used a low-order AR filter, in the form of an order 3 all-pole filter, for modelling the smooth or low frequency component of the spectrum envelope. This representation allows shifting the overall energy distribution along the harmonic part of the speech signal, preserving the spectral envelope details related to prosody, such as formant information.

In order to isolate VE from other external factors, such as prosody or lexical meaning, a specially designed speech database with three levels of vocal effort was used to validate the proposed methodology. This speech corpus contains nonsense words in three different VE levels (low, neutral, and high). Given an utterance expressed in the three VE levels, each version of the utterance is analysed at multiple time instants, obtaining the proposed VE model parameters along the whole utterance. Thus, the dynamics of the VoQ along time are captured, instead of having a single constant value for the whole sentence to synthesize, as proposed in [Monzo et al., 2010]. This brings more flexibility, allowing variations of VoQ inside a single utterance.

The conducted experiment compared the proposed methodology with APLP. The results reveal that the proposed HNM-based speech modification procedure can transform the perceived vocal effort while maintaining sufficient quality in a copy synthesis experiment, validating hypothesis (*H2*) (HNM can be used for transferring Vocal Effort (VE) between two original recordings with different VE levels).

In accordance with the second part of objective *O2*, conversions between all VE levels present in the NECA database for each utterance were performed. However, the methodology introduced signal degradations correlated with the degree of modification. Figure 4.7(b) shows this correlation between the distance from source to target VE levels, or the degree of modification to perform, and the perceived signal quality in the synthesized utterances. For larger conversions, such as *low-to-high* and *high-to-low*, subjects rated the quality of the converted signal with lower ratings than conversions between adjacent VE levels, such as *low-to-neutral* or *high-to-neutral.* This finding makes it more desirable to work with neutral vocal effort as a reference (which allow for conversion towards other vocal effort levels with higher quality) than using low or high vocal-effort signals. In a TTS system this would allow synthesising multiple VE levels, thus conveying different expressions in the synthesized speech, using only a single neutral corpus (without expressiveness) for the synthesis and applying the proper VE models, which comes up with the general goal of this dissertation.

Comparing MOS ratings for the signal quality of the first approach for modifying VoQ with low-level parameters presented in [Monzo, 2010] (see figure 3.3) and the signal quality ratings for the proposed method in this chapter (see figure 4.7(b)), it can be concluded that for conversions performed based on the intermediate vocal effort level (starting from neutral VE), the ratings tend to be higher. However, in order to properly compare the quality of both systems, a subjective test comparing both systems should be performed.

Unlike to the original APLP method [Nordstrom et al., 2008], which work directly with residuals and LPC filter for analysis and synthesis, HNM adds more flexibility by allowing for time, pitch, and energy transformations. However, although the HNM does not suffer from stability problems, special care must be taken to prevent the appearance of signal discontinuities that may introduce artifacts into the modified signal. To prevent these effects and ensure gradual transitions in both vocal effort and signal modifications, the computed dynamics were smoothed in the time domain.

The proposed method was compared with the original proposal [Nordstrom et al., 2008], demonstrating that HNM is a feasible model for vocal effort modification. Moreover, in this work, conversions from lower to higher vocal efforts (which were not analysed in [Nordstrom et al., 2008]) have been studied, demonstrating that the HNM and APLP can be used for this type of modification.

This chapter has answered the research question (*Q2*) (Is it possible to transfer Vocal Effort (VE) levels between two parallel signals using a spectral envelope representation based on HNM?) by presenting a methodology for transferring VE between two parallel utterances with different VE levels. However, the proposed methodology in this chapter presented several limitations. The system was based on LPC, which could become unstable when performing model interpolations. With the aim of having more flexible signal modification systems, a method that permits generating new VE levels is required. Thus, it is desirable to have a parametric representation which permits easily interpolating models to obtain new VE levels (see objective *O3*). To address this objective, chapter 5 presents a parametric VE model that allows controlling the synthesized vocal effort levels without requiring parallel corpora at the synthesis stage. Parametric vocal effort models, such as those presented in the next chapter, can be combined with artificial intelligence algorithms to predict vocal effort dynamics, and thereby, improve the expressiveness and naturalness of TTS systems. In this regard, section 2.6 proposes a new US-TTS schema, incorporating the methodologies proposed in chapter 5, aiming to improve the US-TTS flexibility by incorporating VE modification. The proposed methodology can help US-TTS lead the synthesis process with new expressive models that would be capable of predicting vocal effort information, thereby allowing the synthesised signal to express other vocal effort levels than those present in the original speech database.

CHAPTER 5

Vocal effort interpolation

Contents

5.1 Parametric spectral model and codebooks 81

5.2 Vocal effort modification with the polynomial model 83

5.3 Evaluation and results 89

5.4 Discussion 92

5.5 Conclusions 95

This chapter extends the work presented in chapter 4, where a method based on low-order LPC for transferring VE using a copy-synthesis approach was presented. However, the proposed methodology in chapter 4 based on LPCs is very sensitive to interpolation of filter coefficients, which can lead to filter instabilities and thus, cause acoustic artifacts in the generated signal [Nakagawa et al., 1995, Paliwal, 1995, Islam, 2000, Peinado and Segura, 2006, Inžinerija et al., 2007]. So it does not permit generating new vocal effort levels other that the ones present in the parallel corpora, which limits the flexibility of the system for conveying VE levels. Thus, the system flexibility still depends on the corpus size, although in order to extend the corpus expressive capabilities, VE models are stored in the corpus instead of speech signals with different VE levels. The procedure presented in this chapter is based on generating a set of VE model databases for low, neutral, and high VE levels obtained from analysing the original nonsense words in the speech corpus [Schröder and Grice, 2003]. The models stored in the database are referred to as codebooks. In line with the main goal of this dissertation, synthesizing expressive speech with a single neutral corpus, the conducted experiments aim to use modal VE level utterances, which corresponds to a neutral speech style, as the source signal and transform them towards the desired target VE level. The signal modifications are performed applying frame-based multiplicative factors to the HNM amplitudes. The multiplicative factors are computed based on the models stored in the database, f.e., the codebooks. For instance, in order to increase the VE level, the neutral and high VE codebooks are used for computing the corresponding multiplication factors. The computation process is performed by interpolating the codebook models according to the degree of conversion desired, taking the selected codebooks as the extreme points of the conversion (neutral codebook meaning no modification, and high VE codebook meaning maximum transformation to be performed). Thus, the system is capable of producing any intermediate VE level between the *low* and *neutral*, and *neutral* and *high* VE levels.

In order to validate the hypothesis proposed in this chapter (*H3*) (A parametric VE model based on the HNM could be used for synthesizing other VE levels than the ones present in the utterances available in the speech corpus.) the following specific objectives are set:

- ▷ Propose a parametric model with easy and stable interpolation properties for modelling and modifying the spectral envelope.
- ▷ Use the parametric model for modifying VE, transferring VE levels between different speech utterances, and evaluate the obtained results.
- ▷ Finally, generate intermediate VE models by means of using the utterances with the available VE levels in the speech database, interpolating the model's parameters, and use them for synthesis, and evaluation.

This chapter presents the work conducted in [Calzada et al., 2013] where these objectives were addressed, presenting a new model based on ninth order polynomials for representing the full band of the harmonic spectral envelope; which not only allows transferring vocal effort among template signals available in the corpus, but also allows generating intermediate vocal effort levels not present in the speech corpus. This would also increase the flexibility of the TTS system by permitting easily adjusting the synthesized expressiveness level by using only a single neutral corpus at the synthesis stage. The desired flexibility for the signal modification procedure was pursued with the combination of HNMs and using a parametric model to represent the VE.

The conducted perceptual evaluation demonstrates the effectiveness of the proposed technique for performing vocal effort interpolations while preserving the signal quality in the final synthesis. The speech database used for the experiments was the same as in experiments conducted in chapter 4.

This chapter is organized as follows. Section 5.1 presents the polynomial model and details how the model codebooks were built from the original corpus. Next, the proposed vocal effort modification procedure is detailed in section 5.2. In section 5.3, the conducted perceptual experiments are outlined. Section 5.3 discusses the proposed procedure, the obtained results, and future work. Finally, section 5.5 provides conclusions obtained from the experiments carried out in this chapter.

5.1 Parametric spectral model and codebooks

The speech database used in the conducted experiments is the NECA corpus (see section 4.1), which is the same one used in the experiments conducted in chapter 4. It was specifically designed for conducting expressive speech synthesis conveying multiple levels of vocal effort. For more information regarding the corpus see section 4.1.

The aim is to generate a set of parametric VE models, named codebooks, which will be used for computing multiplicative factors to be applied to the HNM amplitudes in order to conduct the VE modification. A codebook contains a set of VE model parameters corresponding to a single phone in a certain position inside each word expressed in a certain VE level for a given gender. Figure 5.1 shows a visual representation of a codebook used in this dissertation.

The entire corpus was represented using HNM parameters. Informal tests conducted prior to the proper evaluation presented in this work, highlighted some speech signal variations in the syllables due to syllable position inside the utterance. Thus, the acoustic characteristics for a given phone depend on its position inside the word. For the sake of obtaining more accurate models without losing too much generalization, the syllable position inside the word were considered together with phone labels for indexing the models in the codebooks. Therefore, each dataset (corresponding to a specific vocal effort level, syllabe position and gender) was divided into three subsets, one per syllable position in the word. Words are formed by three syllables and the position in the word refers to the position of the syllable containing the sound; thus, *init* refers to the first syllable, *middle* to the second syllable, and *final* to the last syllable. For each subset, all HNM parameters from multiple realizations of a common phone were combined. Only parameters from the stable part of the voiced phones were used to prevent coarticulation effects in the final computed models. The stable part was considered to be the second and third quartiles of the full phone duration (figure 5.3). At this point the HNM parameters of all realizations of the stable part of same phone for a given vocal effort level, gender, and syllable position in the word are combined. Next, the coefficients of the parametric model were computed to fit all data points formed by the harmonic amplitudes and frequencies for a given phone, VE level, and gender using a least squares approximation. Finally, the model parameters for all subsets corresponding to the same gender and vocal effort level were combined to form the codebook for that vocal effort level and gender. Thus, each codebook contains as many VE polynomial models as phones present in the

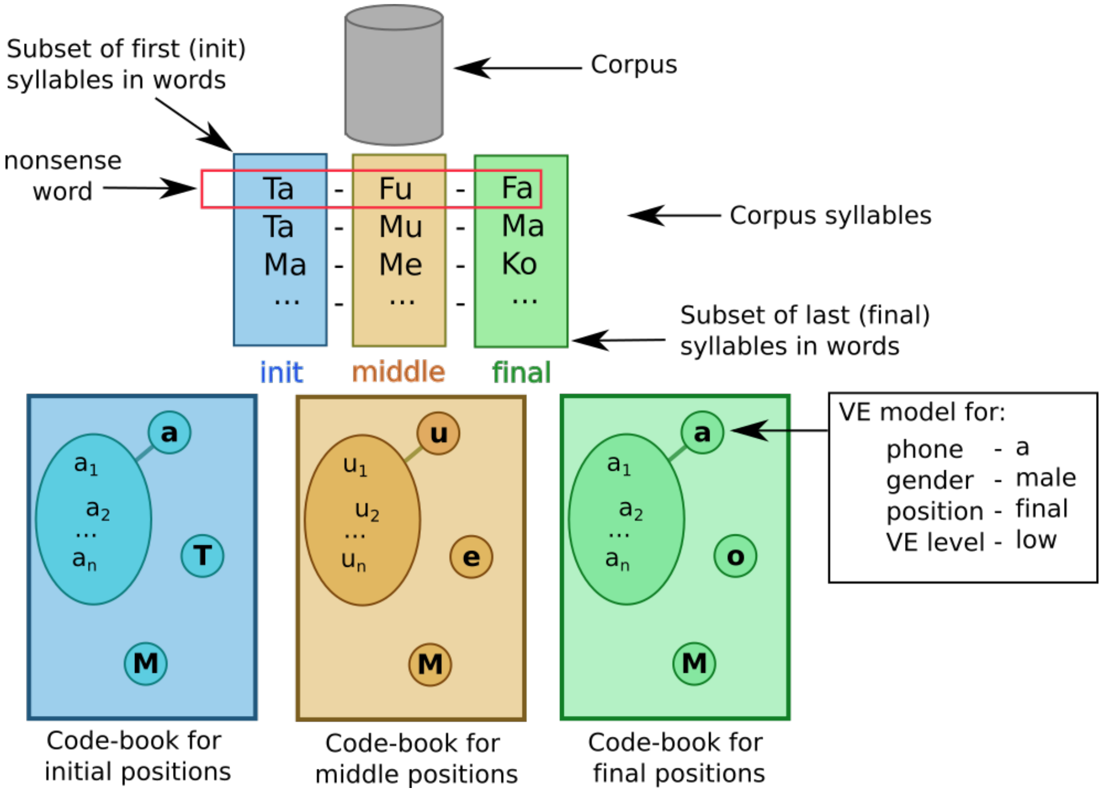


Figure 5.1 — Codebooks generation from the NECA coprus. This composition is replicated for each VE level and gender.

corpus for that VE level and syllable position within the word. Six codebooks were generated covering all vocal effort levels and gender combinations. For a given synthesis, only the three codebooks from the gender to synthesize are used to carry out the vocal effort modification and synthesis procedure.

Codebooks are used to retrieve extreme vocal effort models, which in our case are labelled *High* and *Low*. The third codebook, labelled *Neutral*, is used as the reference level. Apart from the extreme *Low* and *High* vocal effort levels, the proposed methodology aims at allowing the TTS system to also synthesize intermediate vocal effort levels. A polynomial model was chosen because it is a linear model, and it is possible to obtain intermediate functions simply by using interpolation techniques over the model coefficients of two extreme functions (e.g., given two extreme polynomial models $g_1(f) = a_1 + a_2 f + \cdots + a_N\, f^N$ and $g_2(f) = b_1 + b_2 f + \cdots + b_N\, f^N$ the interpolated function $g_i(x) = c_1 + c_2 f + \cdots + c_N\, f^N = 0.5(f_1(f) + f_2(f))$ can be obtained by computing each coefficient of the final model as $c_k = (a_k + b_k)/2)$. In order to be able to capture the fourth formant peaks and valleys, the proposed methodology uses ninth order polynomials as

$$\widehat{ampl}(f) = a_0 + a_1 f + a_2 f^2 + a_3 f^3 + \cdots + a_9 f^9, \tag{5.1}$$

where $\widehat{ampl}(f)$ is the harmonic's amplitude envelope which is a function of the harmonic's frequencies f, and a_i for $i \in [0, 9]$ are the model coefficients. Table 5.1 presents an excerpt from the codebook corresponding to the *High* vocal effort level for the female speaker.

As explained in section 2.3.2, VoQ modifications should be performed simultaneously with prosody modifications [Grichkovtsova et al., 2012]. Although the speech corpus was created trying not to introduce prosody variations, some acoustic characteristics were modified according to the position of the syllable inside the logatom. Moreover, the proposed methodology was aimed to be used in US-TTS systems where prosody modifications will be performed requiring this coherence between VoQ and prosody. Section 2.6 presents the necessary schema modifications of the GTM's US-TTS system for integrating the procedure introduced in this chapter.

Table 5.1 — Excerpt of the codebook corresponding to *High* vocal effort level for the female speaker. Phone labels follow the SAMPA notation [Wells, 1997]. Part of the coefficients have been removed. The semicolon ";" is the delimiter character used to separate fields when reading data from the codebook, which is a plain text file.

Phone	Syllable position	Model coefficients
⋮	⋮	⋮
U;	init;	1.941e-31;-4.810e-27;$\cdots$;-0.067;
U;	middle;	1.501e-31;-3.761e-27;$\cdots$;-0.039;
U;	final;	1.071e-31;-2.664e-27;$\cdots$;-0.030;
o;	init;	2.046e-31;-5.298e-27;$\cdots$;-0.118;
o;	middle;	3.044e-32;-9.366e-28;$\cdots$;-0.040;
o;	final;	-1.502e-31;3.287e-27;$\cdots$;0.045;
⋮	⋮	⋮

5.2 Vocal effort modification with the polynomial model

The proposed methodology uses the neutral vocal effort level data as the starting point for the modifications. This decision was based on the results obtained in previous work [Calzada and Socoró, 2011], where it was found that the signal quality degradation was directly related with the amount of signal modification. Thus, in order to minimize the amount of signal modification for all cases, raising and lowering the vocal effort level, we decided to use the neutral vocal effort level as the source for all signal modifications. For this reason only the HNM parameters from the neutral corpus are used for synthesis. HNM parameters from the high and low datasets are used only for building the respective (*High* and *Low*) codebooks. Figure 5.2 depicts the general schema for the proposed methodology.

The vocal effort synthesis procedure conducted in this work begins with a given phonetic tran-

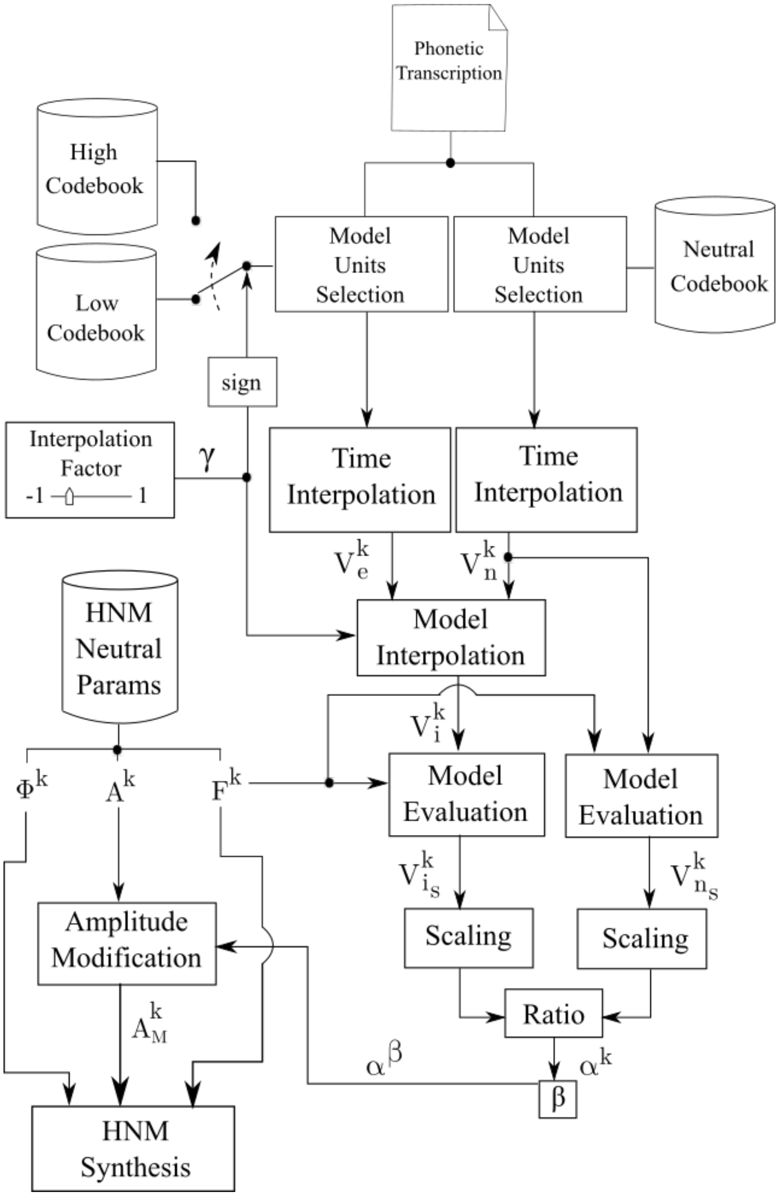


Figure 5.2 — Schematic diagram of the proposed vocal effort interpolation method. From [Calzada et al., 2013]

scription of the text to be synthesized. The transcription is used to retrieve the corresponding model parameters from the codebooks. The neutral codebook is always used because the spectral envelope from this vocal effort level serves as the baseline for the subsequent modifications. However, the *High* and *Low* codebooks are used only when necessary. The decision is taken based on the target vocal effort to be synthesized. An *interpolation factor* (γ) is defined in the range $[-1\ ,\ +1]$ to control the degree of VE modification, its sign being the direction in which the vocal effort modification is performed. Negative values correspond to lowering the vocal effort, whereas positive values are used for increasing it. Thus, the extreme values (-1 and 1) indicate using the *Low* and *High* vocal effort parameter models as retrieved from the corresponding codebook.

Once the two codebooks to use are identified (the neutral and the low or high VE, depending on the sign of γ), the transcription is divided into three regions, where each region corresponds to a syllable from the word (regions were labelled as *init, middle*, and *final*). This information is used in combination with the phone label for finding the model units in the codebooks. For instance, given the transcription $/t - \{ - m - u - t - \{/$ the first unit to find is the phone $/\{/$ with an indicator of initial (*init*) position. However, the model parameters for the second $/a/$ will be different due to its position in the word (*final*). The anchor phones for which the proper models will be retrieved from the codebook are $/a/, /m/, /u/$ and the second $/a/$. Phone $/t/$ is discarded because only voiced phones are used for generating the VE model. Once the proper units for the whole sentence are selected from the corresponding codebooks, the model coefficients are linearly interpolated in order to have model parameters for each frame to synthesize. However, the linear interpolation is carried out only in the unstable parts of the phones, where coarticulation effects are present. For the central regions of the phone the original model parameters obtained from the codebook are preserved (figure 5.3).

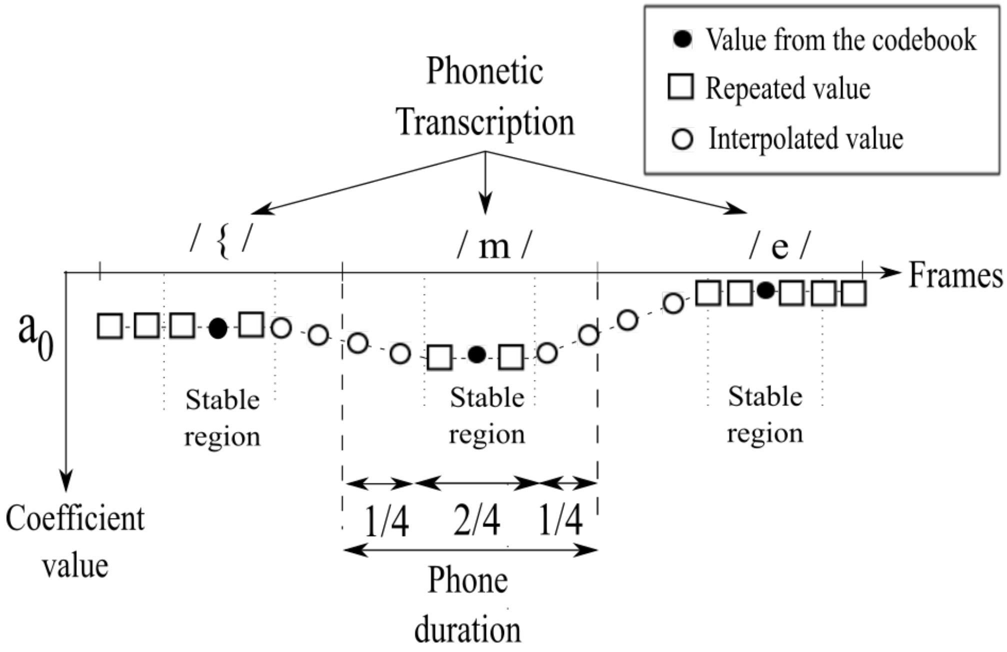


Figure 5.3 — Temporal linear interpolation of model coefficients. From [Calzada et al., 2013]

Taking into consideration that the corpus used is formed by isolated words, so there is no coarticulation effect between words; then, this preservations was also performed at the beginning and end of the utterance to synthesize. This process results in two vectors for each frame, $\boldsymbol{V}_e^k$ and $\boldsymbol{V}_n^k$ of length

R being the number of coefficients of the polynomial model. Thus, the vectors contain the model parameters for each frame. $\boldsymbol{V}_e^k$ contains the extreme vocal effort parameters obtained from the *Low* or *High* vocal effort codebooks, depending on γ sign, and $\boldsymbol{V}_n^k$ contains the neutral vocal effort model parameter values for each frame.

The next step is to obtain the vector corresponding to the interpolated vocal effort level ($\boldsymbol{V}_i^k$) from $\boldsymbol{V}_e^k$ and $\boldsymbol{V}_n^k$. The interpolated model parameters for a frame k are computed by

$$\boldsymbol{V}_i^k = (1 - |\gamma|)\ \boldsymbol{V}_n^k + |\gamma|\ \ \boldsymbol{V}_e^k \quad \text{with} \ \ \gamma \in [-1, 1], \tag{5.2}$$

where $\boldsymbol{V}_i^k$ is the interpolated model coefficients corresponding to the final desired vocal effort level for the k^{th} frame, γ is the interpolation factor, and $\boldsymbol{V}_n^k$ and $\boldsymbol{V}_e^k$ are the coefficients from the k^{th} frame for the neutral and extreme (*Low* or *High*, depending on $sign(\gamma)$) vocal effort levels, respectively. Figure 5.4 depicts the process of interpolating VE models according to different values of γ.

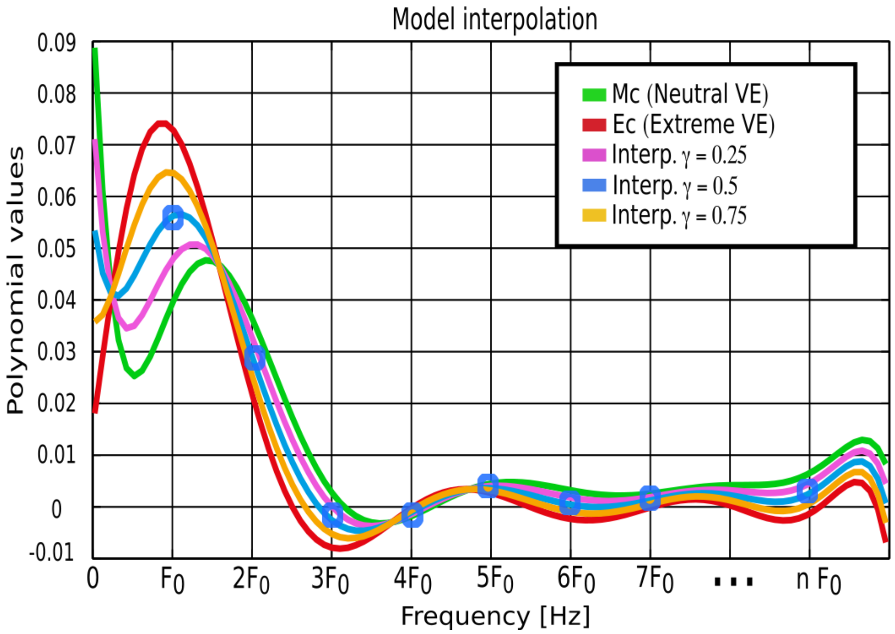


Figure 5.4 — VE model interpolation for different interpolation factor (γ) values at multiples of f_0. The graphic shows how easy is to intuit an interpolated model given corresponding neutral and extreme VE models. The graphic shows multiple interpolation models starting from the neutral VE model (green), which corresponds to $\gamma = 0$, towards the extreme VE model (red), which corresponds to $\gamma = 1$. The maximum frequency is $nF_0 < 5000Hz$ because it only modifies the harmonic part of the signal. From [Calzada et al., 2013]

Once the final desired model coefficients are computed, the models ($\boldsymbol{V}_i^k$ and $\boldsymbol{V}_n^k$) are evaluated at the original signal's harmonic frequencies ($\boldsymbol{F}^k = f_0^k \cdot [1, 2, \cdots, L^k]$) using expression (5.1) to obtain

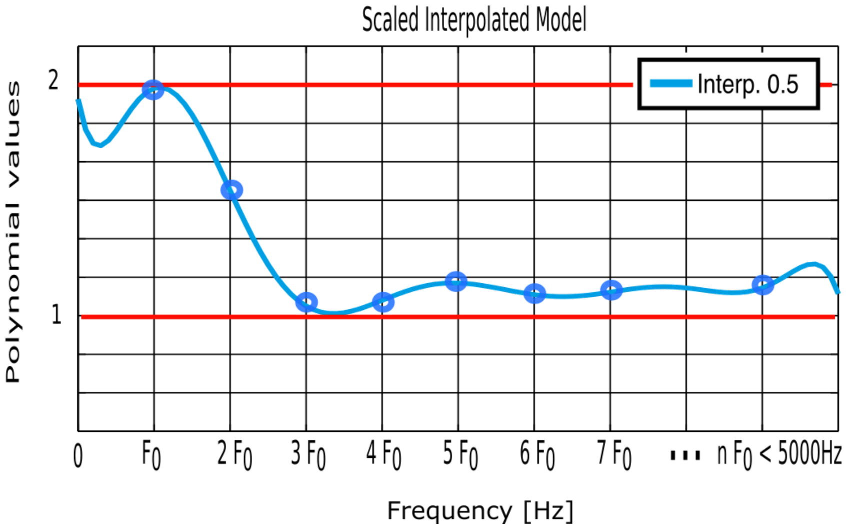


Figure 5.5 — The model, in this case an interpolated model with $\gamma = 0.5$, is scaled to the range [1, 2] in order to prevent numerical instabilities when performing the division operation with the model coefficients. The maximum frequency is $nF_0 < 5000Hz$ because it only modifies the harmonic part of the signal. Blue round markers indicate the points where the model will be evaluated in order perform the comparison of the models to obtain the final multiplicative factors (α). From [Calzada et al., 2013]

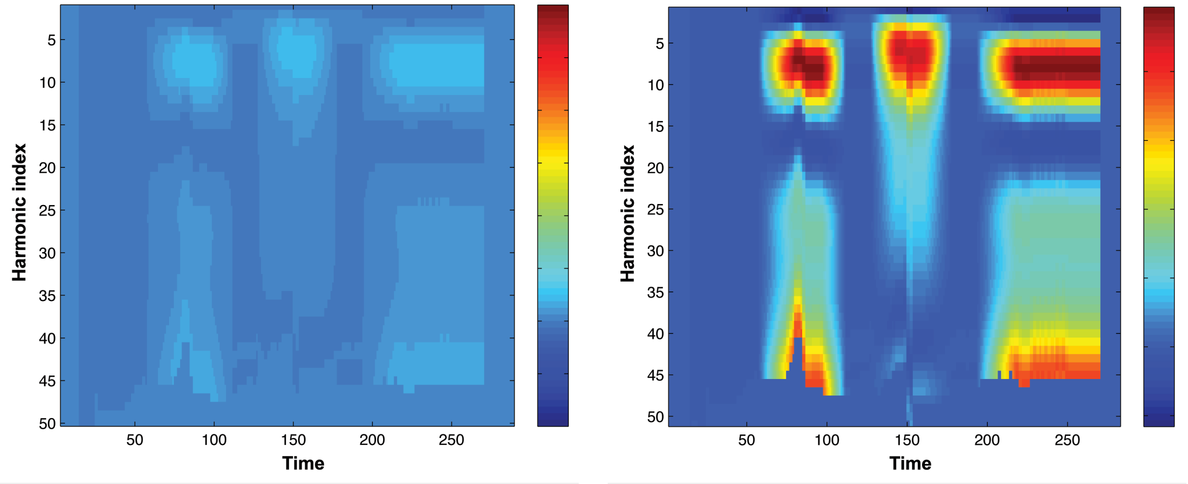
the harmonic spectral envelopes. The harmonic spectral envelope is computed for both the neutral ($\boldsymbol{V}_n^k$) and the desired vocal effort level ($\boldsymbol{V}_i^k$) coefficient vectors, obtaining $\boldsymbol{V}_{n_s}^k$ and $\boldsymbol{V}_{i_s}^k$, respectively. These two vectors contain the samples of the polynomial models at the harmonic frequencies of the speech signal. The multiplication factors ($\boldsymbol{\alpha}^k$) that are applied to the HNM amplitudes ($\boldsymbol{A}^k$) from the original signal will be obtained from the harmonic spectral envelopes quotient for each frame by

$$\boldsymbol{\alpha}^k = \left[\frac{\boldsymbol{V}_{i_s}^k(f_0^k)}{\boldsymbol{V}_{n_s}^k(f_0^k)}, \frac{\boldsymbol{V}_{i_s}^k(2f_0^k)}{\boldsymbol{V}_{n_s}^k(2f_0^k)}, \cdots, \frac{\boldsymbol{V}_{i_s}^k(L^k f_0^k)}{\boldsymbol{V}_{n_s}^k(L^k f_0^k)} \right] = [\alpha_1^k, \alpha_2^k, \cdots, \alpha_{L^k}^k], \tag{5.3}$$

where $\boldsymbol{V}_{i_s}^k(l\ f_0^k)$ and $\boldsymbol{V}_{n_s}^k(l\ f_0^k)$ correspond to the interpolated and neutral VE models evaluated at the frequency of the harmonic l, which has frequency $f_l^k = l\ f_0^k$.

However, the magnitude of the harmonic spectral envelopes might contain values close to zero which could introduce numerical instabilities. In order to prevent this effect and focus on the energy distribution over the spectrum, the envelopes are scaled to fit into the range [1, 2] prior to carrying out the ratio for computing the multiplicative factors ($\boldsymbol{\alpha}^k$). Figure 5.5 depicts an example of a scaled interpolated model inside the range [1, 2].

Scaling the VE model envelopes to the range [1, 2] forces the final multiplicative factors ($\boldsymbol{\alpha}$) to take values in the range [0.5, 2], where values < 1 correspond to attenuations, and values > 1



(a)

(b)

Figure 5.6 — Multiplicative factors ($\boldsymbol{\alpha}$) used for VE modification for each frame (time) and frequency (harmonic index) analysed. 5.6(a) original $\boldsymbol{\alpha}$ values. 5.6(b) $\boldsymbol{\alpha}$ values raised to the power of β. The legend correspond to the scalar values of the parameter. The scale is the same in both images, which allows to see the increased difference between valleys and peaks. From [Calzada et al., 2013]

to amplifications, of the amplitudes. For $\alpha = 1$, no modification is applied to the corresponding amplitude.

In order to emphasize the effect of the energy distribution, the $\boldsymbol{\alpha}^k$ factors are raised to the power of β, obtaining $\boldsymbol{\alpha}^\beta$ (in this case the superindex k indicating the frame is omitted for clearer notation). This factor depends on the desired magnitude of the signal modification. Raising the factors to the power of β amplifies the difference between peaks and valleys in the harmonic spectral envelope after the modification is applied. Figure 5.6 shows an example illustrating the effects of raising the modification factors $\boldsymbol{\alpha}$ to the power of β. As can be seen, after applying the exponentiation the differences between peaks and valleys are emphasized.

Next, the HNM amplitudes ($\boldsymbol{A}^k = [a_1^k, a_2^k, \cdots, a_{L^k}^k]$) are modified applying the exponentiated factors ($\boldsymbol{\alpha}^\beta = [\alpha_1^{k\beta}, \alpha_2^{k\beta}, \cdots, \alpha_{L^k}^{k\;\beta}]$); obtaining the modified amplitudes ($\boldsymbol{A}_{\boldsymbol{M}}^k = [a_1^{k\prime}, a_2^{k\prime}, \cdots, a_{L^k}^{k\;\prime}]$) with the expression

$$a_l^{k\prime} = a_l^k\ (\alpha_l^k)^\beta. \tag{5.4}$$

Finally, the energy of each frame is adjusted in order to preserve the original energy magnitude of the frame before applying any signal modification.

The last step consists of synthesizing the signal with the regular HNM resynthesis procedure using the original frequencies ($\boldsymbol{F}^k$) and phases ($\boldsymbol{\Phi}^k$), and the modified amplitudes ($\boldsymbol{A}^k_{\boldsymbol{M}}$).

5.3 Evaluation and results

Two perceptual evaluations were designed with the online testing platform for multimedia evaluation (TRUE) [Planet et al., 2008]. The first test (with 22 subjects) focused on comparing the overall quality of the proposed method using ninth order polynomials against the previous proposal using low order LPC (see chapter 4). The second test (with 21 subjects) evaluated the performance of the proposed method to interpolate vocal effort levels between the extreme levels available in the corpora (*Low* and *High*). At the beginning of each evaluation, the participants were presented with a set of example audio stimuli expressing several different vocal effort levels. In both evaluations the participants were forced to decide between two responses (A or B). In order to prevent introducing any bias in the participant's responses, for each pair of audio stimuli evaluated, the labels (A or B) were randomly assigned. All statistical significance (p-values) were computed using a one-tailed test. The values used for β parameter, used for exponentiating the multiplicative factors ($\boldsymbol{\alpha}$), in the evaluations were set according to informal evaluations carried out prior to creating the audio samples used. For the male gender, β was set to 10, whereas for the female speaker, it was set to 7.

The first evaluation consisted of 3 words (see table 5.2) uttered by both speakers (male and female). For each word, two vocal effort conversions were applied, from neutral to low (*N2L*) and from neutral to high (*N2H*). The conversions were carried out using either one of the two methods under evaluation. For the proposed method in this chapter, the interpolation factor was given values in the range $[-1,\ 1]$. Thus, the evaluation composed a total of 12 audio stimuli presented to the subject in pairs. For a given pair of stimuli, both stimuli corresponded to the same vocal effort conversion performed by either one of the two methods, the one presented in chapter 4 and the one proposed in this chapter. The subject was asked to answer the following two questions for each pair:

1. Omitting the signal quality, which of the following files (A or B) transmits a **Higher Vocal Effort?**

2. Which of the following samples (A or B) has better signal quality?

The first question evaluated the performance of both methods for modifying VE. In order to minimise the errors due to confusion by asking for a different task in each pair under evaluation, the subject was always requested to perform the same task, select the stimuli conveying **higher** vocal effort no matter the intention of the conversion under evaluation. When evaluating conversions intending

Table 5.2 — Transcription of the stimuli used in the first evaluation using SAMPA [Wells, 1997] notation.

[	aI	-	t - a:	-	p - aI	]
[	t - a:	-	f - u:	-	f - a:	]
[	t - a:	-	m - U	-	m - a:	]

to increase the VE level, the method generating the stimuli chosen by the subject is the one scoring for that evaluated pair. However, the scoring criteria is reversed when evaluating conversions aiming to lower the VE level, for instance from high to neutral, because the subject was always requested to identify the stimuli with higher VE; thus, in those cases the method generating the stimuli that was not chosen by the subject is the one scoring for the stimuli pair under evaluation.

Table 5.3 — Preference of the proposed method with the reference (APLP with HNM, presented in Chapter 4) according to vocal effort performance in first evaluation. p-value was computed indicating no preference between the methods as the null hypothesis (H_0)

Parameter evaluated	Preference for proposed method [%]	p-value
Vocal effort conversion	53.7879	0.1093
Converted signal quality	82.9545	< 0.0001

Table 5.3 presents the results of the first test where the performance of the proposed method was compared with the previous proposal (see chapter 4). The results show a slight global preference for the new proposed method based on polynomial models interpolation. Regarding vocal effort modification there is a $53,79\%$ preference, whereas in terms of signal quality, this preference is more pronounced reaching 82.95%. For obtaining the p-values of the results, the null hypothesis (H_0) was formulated as: *There is no preference between the proposed method and the reference method* [Nordstrom et al., 2008]. The obtained p-values (0.1093) state that in regards to vocal effort modification, there is no strong preference for the proposed methodology. On the other hand, regarding the signal quality, the preference for the proposed methodology is statistically significant ($p - value < 0.0001$). Statistics were computed using one-tailed significance tests on the sampling distribution.

With the results obtained from the first evaluation, we conclude that despite not presenting relevant improvements for extreme vocal effort modifications when compared with the previous approach combining APLP with HNM (see chapter 4), the method proposed in this chapter performed better in terms of signal quality. This result supports the suitability of the proposed method for transferring vocal effort.

The purpose of the second evaluation was to verify the feasibility of using polynomial models

[	t - a:	-	s - i:	-	s - a:	]
[	t - a:	-	j - a	-	j - a:	]
[	t - a:	-	l - i:	-	l - a:	]
[	t - a:	-	t - o:	-	t - a:	]
[	t - a:	-	r - @:	-	r - a:	]
[	t - a:	-	p - Y	-	p - a:	]

Table 5.4 — Transcriptions of the stimuli used for the second evaluation using the SAMPA [Wells, 1997] notation.

to interpolate vocal effort levels. In order to investigate its flexibility, several vocal effort levels were generated from the same neutral vocal effort level source utterance. Thus, using the neutral (N) vocal effort level as a reference, the following four vocal effort levels were synthesized: low (L), intermediate low (IL), intermediate high (IH), and high (H). Samples labelled as IH correspond to a linear interpolation between vocal efforts levels with interpolation factor $\gamma = 0.5$, using the expression (5.2). Thus, samples labelled as IH were expected to be perceived between high (H) and neutral (N) vocal efforts. On the other hand, samples labelled as IL correspond to a linear interpolation between low (L) and neutral (N) vocal effort levels with an interpolation factor of $\gamma = -0.5$, applying equation (5.2). Likewise, samples labelled as IL were expected to be perceived between neutral (N) and low (L) vocal effort levels.

The perceived vocal effort level for each synthesized utterance was compared with the samples corresponding to the surrounding vocal effort levels. Extreme vocal effort levels were also compared with the neutral reference. Thus, the participants evaluated the following vocal effort level pairs: *L-IL*, *IL-N*, *N-IH*, *IH-H*, *L-N*, and *N-H*. Extreme vocal effort levels L and H were synthesized using the models from their respective codebooks using $\gamma = -1$ for L, and $\gamma = 1$ for H. The question for each pair of samples was: *Which of the following stimuli (A or B) conveys a* ***Higher Vocal Effort?***. There was no option for equal, subjects were forced to choose one of the two stimuli. Each pair presented to the subject corresponded to two vocal effort levels synthesized from the same neutral reference utterance for the same gender. The stimuli presented to the subjects were randomized to prevent biases in the answers.

Three words were taken for each gender, yielding six different utterances used for the second evaluation (see table 5.4). Each subject evaluated each conversion once per utterance, producing a total of 132 responses for each vocal effort level comparison.

Tables 5.5 and 5.6 present the results from the second evaluation which evaluated the interpolation of vocal effort levels of the proposed methodology. Table 5.5 presents the results from the comparison of the synthesized versions for high (H) and low (L) vocal effort levels with the neutral (N) version. Results indicate the general ranking for low (L), neutral (N), and high (H) vocal effort levels.

VE level pair	[%]	p-value
$H > N$	84.0909	< 0.01
$N > L$	90.9091	< 0.01

Table 5.5 — Perception of the extreme vocal effort levels synthesized against neutral level. The null hypothesis (H_0) stated that subjects can not perceive any vocal effort level difference in each pair of stimuli.

VE level pair	[%]	P-value
$H > IH$	81.8182	< 0.01
$IH > N$	56.8182	0.0594
$N > IL$	81.0606	< 0.01
$IL > L$	76.5152	< 0.01

Table 5.6 — Perception of the interpolated vocal effort levels synthesized against high, neutral and low VE levels. The null hypothesis (H_0) stated that subjects can not perceive any vocal effort level difference in each pair of samples.

These results indicate that utterances synthesized with low vocal effort (L) are perceived as expected compared to neutral (N), and those utterances synthesized with high vocal effort (H) level are also perceived as expected when compared against the neutral (N) reference.

Thus, results from table 5.5, indicate that subjects perceived the synthesized extreme vocal effort levels according to the following ranking: low (L) < neutral (N) < high (H). The analysis of interpolated vocal effort levels (IL and IH) can be found in table 5.6. As can be seen, IL synthesized utterances, which are supposed to represent vocal effort levels between neutral and low, were perceived as expected. When comparing IL with N, the success rate was 81.06%, while comparisons between IL and L presented a success rate of 76.51%. For both cases $p < 0.01$. These results support the capability for interpolating vocal effort levels.

On the other hand, IH samples comprise those samples generated from interpolating vocal effort levels between high H and neutral N. Comparisons between IH and H were successfully recognized 81.82% of the time with $p < 0.01$. However, when comparing IH with N, the success rate was slightly less favourable at 56.81% with $p = 0.0594$.

Results from table 5.5 and 5.6 demonstrate the capability of the proposed methodology to generate interpolated vocal effort levels with the following ranking: $L < IL < N \leq IH < H$.

5.4 Discussion

Previous work presented a parametric model based on low order LPC (see chapter 4); however the model itself is sensitive to interpolation artefacts, which can lead to filter instabilities [Nakagawa et al., 1995, Paliwal, 1995, Islam, 2000, Peinado and Segura, 2006, Inžinerija et al., 2007], thereby reducing its suitability to generate intermediate vocal effort levels. Other approaches could be based on adding extra speech data to the corpus to cover the desired vocal effort levels to synthesize, but this creates a dependency between the model's flexibility and the corpus size [Schröder and Grice, 2003]. This study, presented a methodology using parametric models based on ninth order polynomials, instead of the low order LPC model, not only for transferring vocal effort, but also for generating new interpolated vocal effort levels not present in the speech recordings. The proposed methodology was compared to previous work, presented in chapter 4, in terms of vocal effort modification and synthesized signal quality. The results obtained from this comparison show that the presented methodology can reach the same degree of vocal effort modification as previous work while resulting in an improved signal quality in the final synthesis. The second evaluation demonstrated that the presented method can be used for interpolating vocal effort levels. This was possible due to linearity properties of the polynomial expressions used to represent spectral characteristics related to VE. Despite presenting clear performance differences for most conversions, it is necessary to note the case of comparing *IH* against *N*, where the effect is less clear. This could be a consequence of associating a wider vocal effort range to neutral speech. The fact that statistical confidence increased back to $p < 0.01$ when comparing *IH* against *H* makes us discard the possibility of the system not being able to represent high vocal efforts. Thus, this uncertainty in intermediate high (IH) vocal effort with neutral (N) levels could also be caused by the nonlinear nature of vocal effort perception or production.

These findings extend the previous conducted work presented in chapter 4, not only in overcoming the problem for generating interpolated vocal effort levels, but also achieving better performance in terms of signal quality.

In the methodology proposed in this chapter, vocal effort models were adapted not only for each voiced phone identity but also for each phone position in the recorded word. This decision was made based on informal inspection of the corpus, which led us to realise that the speakers of the corpus performed different speaking patterns based on the syllable position within the utterance. Thus, this distinction was used to prevent effects of the position of the syllable in the extraction of the harmonic spectral envelope models. In some words containing the same phone in several positions, the achieved vocal effort modification varied from one position to the other. The fact of obtaining different harmonic spectral envelope models which produced different vocal effort degrees depending on the syllable position could be related with attack, decay, sustain, and release. Whether this position distinction enhances the procedure, or degrades its performance, has not been evaluated in these experiments. However, when applying the model to sentences with semantic meaning, the position of the syllables in the whole phrase should be considered. Moreover, when applying the model to expressive corpora with multiple expressive styles, the vocal effort modifications to be carried out could depend on contextual information (grammatical, etc.) conditions, such as whether the phone is stressed or not, position inside a stressed word or using accent group information [Erro et al., 2010]. The vocal effort model could be improved by adding these additional features into the codebooks. This could also be used to perform prediction of VE model parameters by Maximum Likelihood (ML) techniques, as GTM's TTS system does with the CBR machine learning technique for predicting prosody parameters (duration, energy, and pitch).

The proposed method could also be combined with prosodic modifications such as pitch, energy, or speech rate. The combination of these signal modifications could be used synthesise expressive speech conveying different expressive styles.

The parameter β was introduced into the system as a result of noticing that the multiplicative factors ($\boldsymbol{\alpha}$), despite achieving vocal effort modifications towards the expected target, did not produce sufficient gain. This can be the consequence of scaling the harmonic spectral envelopes to fit into the range $[1, 2]$ before applying the quotient to obtain $\boldsymbol{\alpha}$, which in turn was needed to prevent possible numerical instabilities in the estimation of the frequency envelope transformation function. The use of the parameter β allows increasing the intensity of the vocal effort transformation and compensates for its decrease due to the scaling procedure. Multiplication factor ($\boldsymbol{\alpha}$) values are constrained to the range $[0.5, 2]$. $\boldsymbol{\alpha}$ values in the range $(1, 2]$ increase the harmonic energy, whereas values in the interval $[0.5, 1)$ decrease the harmonic energy. To increase the modification magnitude $\boldsymbol{\alpha}$, values were exponentiated, increasing the magnitude of the difference between amplification ($\boldsymbol{\alpha} \in (1, 2]$) and attenuation ($\boldsymbol{\alpha} \in [0.5, 1)$) values. The β values used in the experiments were heuristically chosen in order to produce noticeable signal modifications. Two values were chosen, one for each gender, and they were held constant for all synthesised utterances. Some improvement could be made to have better control of the magnitude of the modifications applied by the multiplicative factors matrix ($\boldsymbol{\alpha}$).

The conducted experiments applied the computed VE models to the same speaker analysed when computing them. It would be desirable to try to transfer VE between different speakers. Future work should focus on computing the VE models from data of several speakers and attempt to learn the variation patterns that the model experiences when the speaker's vocal effort varies around different vocal effort levels. This would lead to generalized versions of the VE models which could be applied to multiple speakers.

5.5 Conclusions

This chapter has addressed the research question *Q3* regarding the capability for synthesizing VE levels different from the ones present in the corpus used for synthesis.

According to the objective *O3*, this chapter has used a ninth order polynomial as the parametric model for spectral emphasis in order to capture the conveyed VE and transfer it to another speech signal. The advantage of the proposed model in this chapter over the one used in chapter 4 is the capability of generating additional VE levels to those present in the corpus by interpolating the model parameters. The conducted experiments have evaluated the capabilities for generating intermediate VE models, but the possibility for extrapolating new unseen VE levels beyond the extreme VE level has not been tested.

The conducted experiments used the NECA corpus from DFKI-LT, which is a corpus especially designed for vocal effort research. In chapter 4, a 3-order all-pole polynomial was used, which allowed capturing general spectral energy distribution; however, as specified in section 2.3.2, VoQ is tightly related to prosody [Grichkovtsova et al., 2012], thus using the ninth-order polynomials helps capture part of the prosodic information represented in the spectrum. The results present compelling evidence of the proposed system performing better than the one based on HNM-based APLP presented in chapter 4. Moreover, the results of a second evaluation statistically support the proposed system's capability for generating interpolated vocal effort levels.

The methodology proposed in chapter 4 was based on transferring VE from specific source and target signals, a one-to-one mapping, where a copy-synthesis approach was very suitable. However, the presented methodology used the concept of codebooks, a context dependent generalized VE model, which permits applying the desired VE models to any speech signal without the one-to-one match restriction in the corpus representations.

According to the main goal of this thesis for improving the flexibility of TTSs, section 2.6 in the next chapter proposes several modifications of the US-TTS scheme from GTM for improving its flexibility based on the methodology proposed in this chapter.

CHAPTER 6

Conclusions

Contents

6.1 Thesis review 97
6.2 Future work 104
6.3 Research contributions 109

This chapter presents the conclusions derived from the work conducted in this thesis; and, following the general objective of this dissertation, a scheme for improving the flexibility of a current US-TTS system by including the modification of VE is presented. Section 6.1 reviews the goals of this thesis, compares them with the main findings presented in each chapter; and finally presents the conclusions for the conducted work. Next, in section 2.6, the impact of the presented findings on the design of Unit-Selection Text-To-Speech (US-TTS) systems are discussed, concluding the section with the proposal of a scheme that incorporates Vocal Effort (VE) modification within the pipeline of a typical US-TTS system, in order to synthesize expressive speech when a neutral speech database is used. Finally, research contributions to the scientific community are outlined and participation in research events during the process of the thesis are listed.

6.1 Thesis review

In section 2.1, multiple speech synthesis techniques were listed, but this thesis has focused on the signal modifications to be performed by the final synthesis block that can be used by most of the main Text-To-Speech (TTS) systems. As explained in section 2.1.3, the size of the speech corpus has a strong impact on the systems performance regarding naturalness, signal quality, and flexibility [Kumar and Kishore, 2004, Black et al., 2007]:

- ▷ Naturalness: The units available in the speech corpus should cover multiple acoustic feature combinations in order to be able to safeguard a natural evolution of the acoustic parameters in the synthesized speech. Sudden changes of acoustic parameters due to the lack of proper templates in the corpus can lead to the production of unnatural synthetic speech. So in this respect, increasing the number and diversity of speech units can improve the synthetic speech naturalness [Kumar and Kishore, 2004].
- ▷ Signal Quality: US-TTS generates the synthesized speech signal by concatenating short speech segments (also called units). Each unit used for synthesizing a sentence presents certain acoustic characteristics that can be significantly different from the surrounding units. This mismatch of acoustic parameters of two consecutive units can introduce acoustic artifacts in the synthesised speech. Thus, the selection of the proper units that minimizes concatenation costs (see section 2.1.3 for explanation of costs in US-TTS) is a crucial step [Black et al., 2007]. Moreover, having several realisations of the same unit produced with different acoustic characteristics increases the chances of using units with acoustic parameters closer to the target, reducing the target costs (also explained in 2.1.3), which can enhance the overall signal quality as well. Furthermore, the more units available in the speech corpus, the easier it is to find consecutive units with similar acoustic characteristics.
- ▷ Flexibility: The flexibility of US-TTS systems for synthesising multiple speech styles is closely related with the presence of samples for those speech styles available in the speech corpus to be used for the synthesis process. Thus, the flexibility of these systems can be improved by increasing the amount of speech styles in the corpus.

The underlying motivation for conducting this thesis was to improve the quality of human-computer interfaces using speech. According to this motivation, the main goal of this thesis was to contribute to the process of improving the flexibility and naturalness of Expressive Speech Synthesis (ESS) systems. With this global objective, a set of research questions were proposed in 1.3.

Regarding question *Q1*, HNM has been proven to be a good model for conducting VoQ modifications. It has also been demonstrated that VE can be transferred using HNMs, answering the research question *Q2*, and moreover, a methodology has been presented for interpolating VE models tackling question *Q3*. Despite having accomplished to some degree all the objectives for this thesis, the global objective of improving flexibility and naturalness of US-TTS systems still requires further development.

Expressivity modification based on prosody and low-level VoQ parameters using HNM

In order to address the first research question (*Q1*), this thesis focused on using HNMs because of its good performance [Banos et al., 2008]. Several studies on expressive speech synthesis used VoQ for conducting expressive speech synthesis, which led us to consider verifying the impact of VoQ combined with prosody versus using only prosody modifications for conveying expressiveness in synthetic speech. Thus the following hypothesis was stated: *HNM is a suitable speech representation for conducting VoQ modifications for expressive speech synthesis, and VoQ can be added as extra acoustic features to improve the expressiveness in synthetic speech.* Based on this hypothesis, chapter 3 presented a first attempt for synthesizing expressive speech combining prosodic modifications with low-level Voice Quality (VoQ) parameters modifications using the Harmonics plus Noise Model (HNM). The low-level VoQ parameters presented in this work (*Harmonic-to-noise ratio, Hammarberg Index, Relative amount of energy above 1000 Hz energy*) were selected because of their ability to discriminate expressive speech styles [Iriondo et al., 2009].

The VoQ parameters were obtained by analysing an expressive speech corpus produced using a mixture of acted speech with an induced expressive styles approach. The text scripts used for the corpus utterances were obtained from advertisements in different domains, and then a relation between topic and a set of prototypical speech styles (happy, aggressive, sad, and sensual) was set. The VoQ modification was performed using multiplicative factors computed as the ratio between the averaged low-level VoQ parameter values for the corpus of the source and target speech styles. The multiplicative factor used for VoQ modification was kept constant during all synthesised utterances. Thus, this system was not able to produce changes in the VoQ dynamics of the modified signal compared to the source signal.

The conducted experiment compared two methodologies for expressiveness conversion: one based only on prosody modifications, and another combining prosody and VoQ modifications. The conducted experiments validated that using VoQ combined with prosody modifications improves the expressiveness in the synthetic speech. However, the proposed procedure suffered from significant signal quality degradation. The causes for this signal quality loss were considered to be: **i)** the number of signal modifications performed; up to five parameters (jitter, shimmer, HNR, hammI and pe1000) were modified independently one from each other; **ii)** the relations between the low-level VoQ parameters used (for instance, Hammarberg Index (hammI) and the relative amount of energy over 1000 Hz (pe1000) represent a common range of the speech spectrum); **iii)** using a constant multiplicative factor for the VoQ parameters modifications over the whole sentence. As explained in section 2.3.2, VoQ has a close relationship with prosody [Grichkovtsova et al., 2012]. Thus, prosody and VoQ must vary coherently. In the presented experiment, prosody was allowed to vary frame by frame, whereas VoQ modification was held constant for the whole signal.

These experiments demonstrated the capability of HNM for carrying out VoQ modifications, it was also verified that the expressiveness level conveyed in the synthesized signal is improved when combining VoQ with prosody modifications, compared with modifying only prosody. However, the cause for the signal quality degradation was not clearly identified. After the conducted experiments and the obtained results, it was found that modifying only a single speech signal attribute closely related to VoQ could led to better results.

Also, the feasibility of HNM for conducting VoQ modifications does not mean that it is the best speech signal representation for this purpose. Proper comparison with other methods such as full-band Adaptive Harmonic Model (aHM) should be performed.

It is important to recall that, as was explained in section 2.5, current state-of-the-art techniques, and also the HNM implementation modification proposed in section A.1.1, force the estimated frequencies to be harmonic.

In order to avoid the low-level VoQ parameter interdependency problem, the idea of conducting a full band spectral envelope modification at once was proposed in chapter 4.

Transferring vocal effort with LPC

Once the ability of HNM for conducting VoQ modifications was verified, and that modifying VoQ combined with prosody improve the expressiveness in the synthesized speech, the research question *Q2* (*Is it possible to transfer Vocal Effort (VE) levels between two parallel signals using a spectral envelope representation?*) was raised. This approach to transferring acoustic parameter values (e.g., in a prosodic perspective [Iriondo, 2008]) from one utterance to another in order to transform the expressiveness of the final synthesis is a common practice in US-TTS for reducing corpus size requirements while maintaining system flexibility.

In chapter 4, the main goal was to transfer VoQ features of speech between two parallel utterances. However, in contrast to the approach presented in chapter 3, where a subset of low-level VoQ parameters (usually used for speech analysis purposes, rather than for transformation, in the literature) were first measured and subsequently modified with the support of HNM of speech independently, the research conducted in this chapter aimed to validate a different approach where only one (highly relevant) VoQ-based feature was modified dynamically along a speech sentence. This proposal was conceived in order improve the synthesized signal quality, while maintaining good transformation results. To that end, Vocal Effort (VE) was chosen (closely related with VoQ), due to its salient role in expressive speech [Schröder and Grice, 2003], leading to the following hypothesis: *HNM can be used* *for transferring VE between two parallel original recordings.*

This chapter proposed a scheme for transferring VE between two parallel utterances using HNMs, based on an adaptation of the Adaptive Pre-emphasis Linear Prediction (APLP) technique proposed by [Nordstrom and Driessen, 2006]. As explained in section A.1.1, the harmonic structure restriction was introduced into the model frequencies estimation step in the HNM parameter estimation algorithm proposed in [Depalle and Hélie, 1997]. This restriction allowed reducing the amount of model parameters, discarding all frequencies but the first and the maximum harmonic frequency, which was always fixed at 5000 Hz. Thus, because all frequencies are harmonically related, the number of harmonics was computed by dividing the maximum harmonic frequency by the model's estimated fundamental frequency or pitch (the first harmonic). Then, each frequency is computed by multiplying the harmonic index by the value of the first harmonic. As explained in section 2.5, the current state-of-the-art in sinusoidal modelling of speech, such as the Adaptive Harmonic Model (aHM), has also adopted this approach of enforcing a harmonic relation among the model frequencies. The implementation used in chapter 4 divides the speech spectrum into two bands, following the convention of HNMs. However, as explained in section 2.5, the Fan-Chirp Transform (FChT) for the speech spectrum reveals harmonic patterns also for the upper band of the spectrum. This has led new approaches to model the full speech spectrum with harmonics, such as full-band aHM [Degottex and Stylianou, 2013]. These new models, which represent the full spectrum with harmonics have the advantage of not having to set a division between harmonic and stochastic bands, which is one critical point. Moreover, considering the fact that the voice source consists of glottal pulses, which are wideband signals whose amplitude spectrum decays smoothly [Cabral et al., 2014, d'Alessandro and Sturmel, 2011], it suggests applying harmonics to the full band might be a better spectrum representation.

For the conducted experiments, a specially designed corpus for vocal effort synthesis in diphone synthesis system was used [Schröder and Grice, 2003]. A brief description of the corpus was given in section 4.1. The corpus utterances were 3-syllable nonsense words at three different levels of VE (low, neutral, and high). The proposed methodology exploited having represented each utterance using the HNM in the three VE levels for conducting a copy-synthesis approach. However, in order to perform a more straightforward procedure, the number of analysis time instants used to extract VE model parameters were forced to be the same in all realisations of a given utterance. Moreover, time alignment of the analysis time instants was performed in order to have the corresponding VE models according to contextual phonetic information for each of the VE level realisations of the utterance. Thus, given a certain utterance produced at the three VE levels, the number of VE models extracted for each realisation (low, neutral, and high VE) was the same, and their VE were time aligned, in order to match the same acoustic context (for instance, the middle of a phone or in coarticulation between two phones).

Vocal effort was modelled using a low-order Linear Prediction Coefficient (LPC) filter, and then, the low frequency variation of the spectral envelope was mainly represented. To that end, the order of the LPC filter modelling VE was set to 3 as in [Nordstrom et al., 2008]. The VE transplantation procedure was defined as follows: **i)**performing time and pitch modification in order to match the same prosody in both utterances, **ii)**next computing the aligned VE model for source and target signals from the HNM parameters (as explained in section 4.2), **iii)** then the VE from the source signal was removed, **iv)** the target signal VE was applied. The final step consists in correcting the energy of the signal to better match the target signal.

The conducted experiment compared the APLP method proposed in [Nordstrom et al., 2008] with the proposed methodology in chapter 4, in terms of perceived VE compared with the original signals, and the quality of the converted signal. The method proposed in this dissertation obtained better results than the one presented in [Nordstrom et al., 2008] in terms of VE conversion. However, in terms of signal quality, both systems were similar, except for conversions from *neutral-to-low* and *neutral-to-high*, where original APLP was preferred. In the results, presented in figure 4.7(b), it can be seen that the greater the distance between source and target VE levels, the more signal degradation occurs. In further studies presented in chapter 5 this was addressed.

Although the quality of the presented HNM-based methodology in chapter 4 cannot be strictly compared with the results obtained in chapter 3 (because the former system presented experiments on nonsense words, whereas the latter was applied to meaningful sentences) the improved performance over the original APLP technique, and the signal quality (Mean Opinion Score (MOS)) indicate that the proposed methodology in chapter 4 leads to better signal quality than the one presented in chapter 3. However, in order to be able to validate this, both systems should be evaluated under the same conditions. Moreover, it would be desirable to verify the proposed system based on VE for converting different speaking styles (i.e., from neutral to sad or sensual speech styles) as was done in chapter 3.

The results presented in chapter 4 confirmed the possibility to impose a desired VE level onto a synthesized speech signal answering the research question *Q2* and achieving objective *O2*. Despite establishing that VE can be transferred between parallel utterances using HNM models, the use of LPC for modelling VE makes it difficult to generate intermediate VE levels, thus having some limitations regarding the system ability to convey VE levels not present in the corpus (objective *O3*).

Vocal effort interpolation

In order to improve the flexibility for conveying VE levels other than those present in the speech database, and also to have a more stable VE modification procedure, chapter 5 proposed a new VE model based on polynomials.

In this chapter, ninth order polynomials were used for modelling VE. This allowed us to better model the overall spectral envelope. Moreover, linear polynomials have the interesting property that they do not present instabilities when interpolating their coefficients. Unlike to the copy-synthesis approach taken in chapter 4, the new methodology proposed here used a codebook of VE models dependent on gender, phone, vocal effort level, and position in the word. This new approach was closer to the speech synthesis framework, and represented a step towards the main objective of this work. Thus, the database was analysed, obtaining one model for each phone in each position (initial, middle, or final) inside the three-syllable word produced at a VE level (low, neutral, or high). Each model parameter was computed based on all realisations of the same phone in the same position for a given VE, thus the restriction of having parallel utterances in order to produce the VE transformation was no longer required here.

The insights taken from the experiments conducted in chapter 4, where it was observed that the distance between source and target VE was proportional to the signal quality degradation introduced when applying signal modifications, were used for the design of the methodology presented in chapter 5. Thus, in the new proposed system, all VE conversions used neutral VE as the source VE level. The proposed methodology also introduced a new degree of flexibility allowing other VE levels than the ones available in the speech corpus, by means of the interpolation of VE model coefficients.

The conducted experiments compared the HNM-based APLP method presented in chapter 4, based on an all-poles filter, to the method based on ninth order polynomials and VE model codebooks. The obtained results show that while there is a slightly preference for the new approach in terms of VE conversion degree, in terms of signal quality, the new proposed method clearly outperformes the one based on APLP.

However, the conducted experiments did not assess the impact of the criteria used for the clustering applied in the definition of the VE model codebooks. For instance, the impact of the phone position inside the word in the VE model estimation was not evaluated. Also, the experiments were designed using a VE corpus where the VE level was kept constant within each sentence [Schröder and Grice, 2003], as mentioned in Chapter 4. It would be desirable to see how the generation of the VE codebooks would perform in an expressive corpus where utterances are formed by sentences where the VE might have higher variation along the utterance. Also, in this situation, the context cases are much more diverse and require more sophisticated clustering schemes.

The proposed technique could be used in Unit-Selection Text-To-Speech (US-TTS) systems to increase the flexibility in terms of expressiveness without the need to increase the speech corpus size, especially for synthesizing speech signals expressing a range of vocal effort levels different from those in the original recordings.

6.2 Future work

Based on the current GTM's US-TTS system presented in section 2.6 and the VE modification and interpolation technique presented in chapter 5, the following modifications of the US-TTS should be implemented and evaluated.

For the Natural Language Processing (NLP) component based on Case Based Reasoning (CBR), the attributes to work could be the coefficients of the VE polynomial models presented in chapter 5. Thus, when parametrizing the speech signal, it is also analysed to obtain the VE model parameters following the analysis procedure explained in chapter 5. As it is done in that chapter, for a given set of examples that correspond to the same context (a context is represented here as a case), the VE models are gathered and the best fit (using also the least squares criterion) for all the realizations would be selected as the model for that context.

At the synthesis stage, the same procedure would be performed: the context information would be extracted and the *Attribute Generation* component will enter it to each ML block to retrieve from its cases database the VE model coefficients corresponding to the case that best match the context to synthesize. Thus, additional to the prosodic information (pitch, energy and duration), the VE model parameters could also be passed to the *Digital Signal Processing* block (see figure 2.12) together with the desired *Interpolation Factor* (γ). The interpolation factor would be used to be able to convey VE levels different from the ones available in the memory of cases of the CBR by interpolating the predicted models as was presented in section 5.2 between neutral and extreme VE models. In this case the interpolation would be done with the original polynomial model corresponding to the unit to use for the synthesis and the predicted polynomial model by the CBR in the NLP block. Figure 6.1 shows an example of applying an interpolation factor $\gamma = 0.5$ for obtaining an intermediate model between the original VE level of the unit to use and the proposed VE model predicted by the CBR from the NLP block. However, the range values for the interpolation factor (γ) to be used would differ from the ones presented in chapter 5. In the previous chapter the interpolation factor range was $[-1\ ,\ 1]$ and the sign helped to distinguish which VE level (high or Low) to select. For the schema proposed in figure 2.13 the interpolation factor range is $[0\ ,\ 1]$ where 0 means maintaining the original unit's

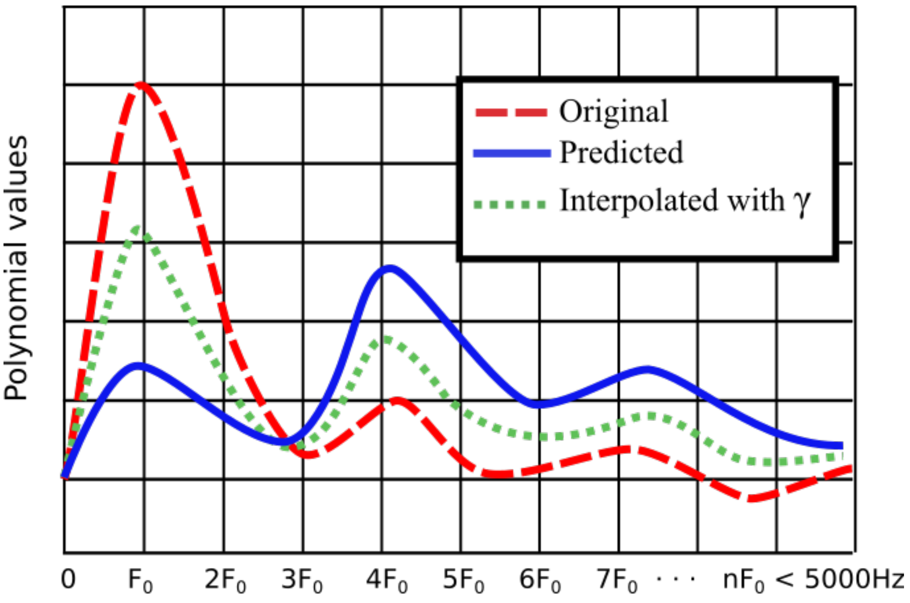


Figure 6.1 — Example of a possible VE model interpolation (green dotted line) of the original VE (red dashed line) and the predicted VE model (blue continuous line) using an interpolation factor ($\gamma = 0.5$).

VE and 1 completely apply the VE model predicted by the CBR.

Figure 6.2 presents the block diagram of the *Speech Synthesis* module (see figure 2.13 for the US-TTS general pipeline) based on the proposed methodology in chapter 5. The input data for this component are the predicted prosody (duration, pitch and energy) and VE (vocal effort model coefficients) parameters predicted by the CBR and the interpolation factor (γ). The interpolation factor could be generated automatically based on any desired criteria, for instance based on the expressive style to convey it could be more desirable to apply the CBR predicted $\boldsymbol{V}_{\boldsymbol{i_s}}^{k}$ VE model ($\gamma = 1$) or an intermediate version ($\gamma = 0.5$). On the other hand, the interpolation factor could be set by the user in order to let to manually control the amount of VE to introduce in the synthesized speech signal.

As can be seen in figure 6.2 the speech corpus would be parametrized, thus working with HNM parameters (amplitudes, frequencies, phases, LPC, noise variance power and analysis time instants) instead of directly with the speech signal. Moreover to storing the HNM parameters for each unit, their corresponding VE polynomial model coefficients would also be stored. Thus, during the corpus parametrization, the VE models for each unit would also be computed and saved. The original VE model for the unit to be used for the synthesis ($\boldsymbol{V_n}$) would be used to compute the VE modification factors ($\boldsymbol{\alpha}$) performing the VE models ratio (dividing the evaluation at the harmonic frequencies of the time interpolated models of the target and the selected vocal effort polynomials frame by frame, $\boldsymbol{V_e^k}$ and $\boldsymbol{V_n^k}$, respectively).

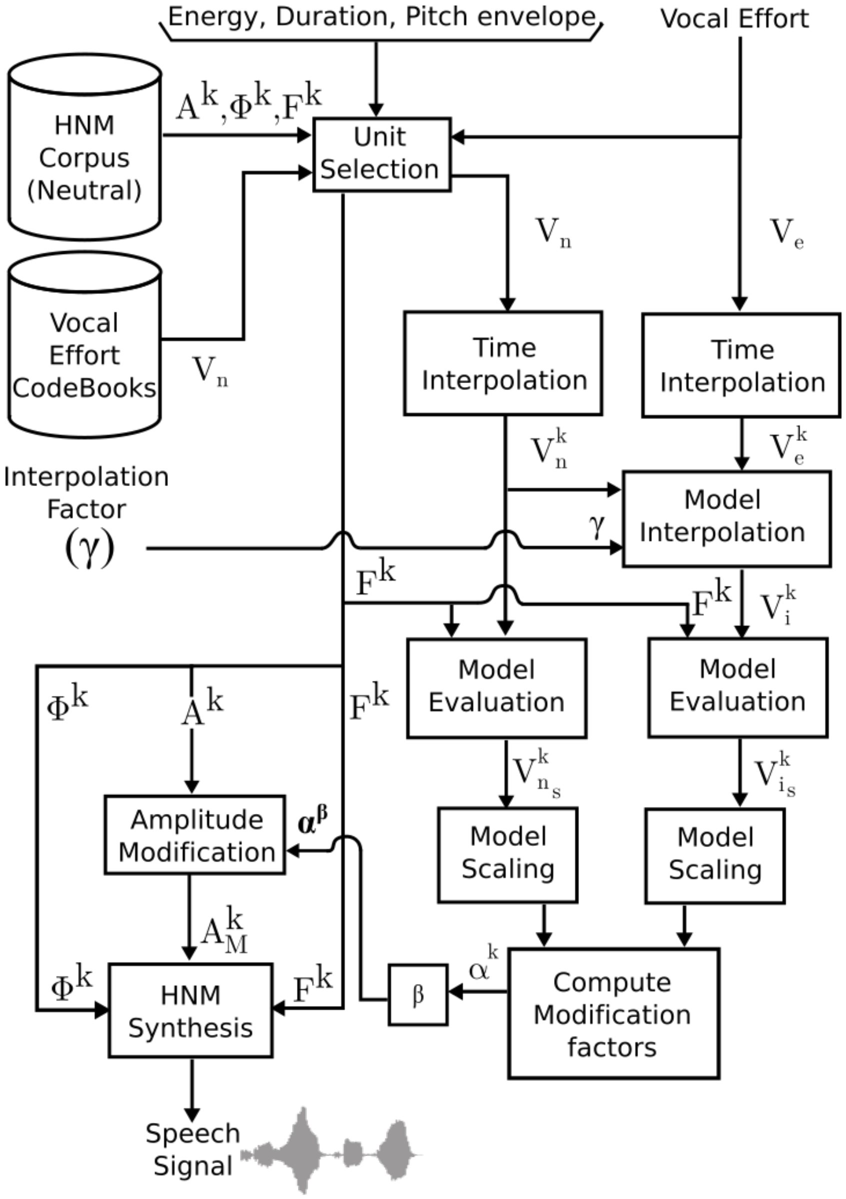


Figure 6.2 — Proposed US-TTS speech synthesis block adaptation to include the VE modification methodology proposed in chapter 5. Adaptation from [Calzada et al., 2013]

With the input information (prosodic and VE parameters), the unit selection block searches in the speech corpus the best units for synthesizing the speech signal. The criteria could be the same as currently used, minimizing concatenation and target cost functions based on acoustic parameters, as explained in section 2.1.3. However, it might be desirable to introduce VE as another parameter to be taken into consideration in the selection criteria. The euclidean distance of the VE model polynomials could be used as a measure between original and target VE models to be incorporated into the target costs functions. In the same way, the euclidean distance between the VE models of the units to be concatenated could be incorporated as an extra term in the concatenation cost function. However, obtaining to what extent the VE must influence the *unit selection* process in the context of a more flexible and expressive US-TTS represents another non trivial challenge.

Once a sequence of units has been selected by the *Unit Selection* block, the original ($\boldsymbol{V_n}$) and original target ($\boldsymbol{V_e^k}$) VE models would be time interpolated in order to compute the evolution frame-by-frame of the VE model along the unit as explained in section 5.2 and depicted in figure 5.3. Next, using the interpolation factor (γ), the time-interpolated VE models for the original unit ($\boldsymbol{V_n^k}$) and the prediction from the CBR ($\boldsymbol{V_e^k}$) are used to perform the linear interpolation of its model's coefficients obtaining the interpolated model ($\boldsymbol{V_i^k}$), as described in section 5.2 and shown in figure 6.1.

Once we have the original unit's VE model ($\boldsymbol{V_n^k}$) and the desired target VE model ($\boldsymbol{V_i^k}$) for every frame to synthesize, the signals are scaled and then the multiplication factors ($\boldsymbol{\alpha}^k$), for VE modification, are computed as the ratio of the two models frame by frame. Then, in order to boost the modification effect, the multiplication factors are powered, thus increasing the difference between peaks and valleys.

Finally, the VE modification is performed applying the multiplication factors to the amplitudes of the HNM model ($\boldsymbol{A}^k$) obtaining the new set of amplitudes ($\boldsymbol{A_M^k}$). The speech signal is generated with the regular process for HNM synthesis as described in section A using the unit's original analysis time instants, phases ($\boldsymbol{\Phi}^k$) and frequencies ($\boldsymbol{F}^k$) and the modified amplitudes ($\boldsymbol{A_M^k}$). In case that prosodic modifications are also desired, additionally to the proposed schema modifications regarding vocal effort, there should also be added the corresponding time (modifying the synthesis time instants -$\boldsymbol{t}_s$- using the phoneme duration predictions and the unit's duration to synthesize) and the pitch modification blocks (modification of $\boldsymbol{F}^k$ and also amplitudes and phases of the vocal tract). This would bring a complete Digital Signal Processing (DSP) schema for bringing major flexibility into the synthesis obtaining multiple expressive speech styles from the prosodic and VoQ point of view.

Although, the proposed schema has not been fully implemented, the different parts of the *Speech Synthesis* block have been implemented and verified separately in chapter 5. However, it would be desirable to evaluate the benefits of the proposed system against the current implementation in terms of performance, naturalness, flexibility and speech signal quality.

Based on the this proposed schema some future tasks arise and new research topics are opened. First, the proposed methodology in figure 6.2 should be implemented and compared with the current US-TTS at GTM in terms of performance, flexibility, naturalness and signal quality of the synthesized speech.

Moreover, the proposed methodology for VE interpolation is based on simple polynomial models, which have been chosen mainly because they offer suitable properties to produce interpolation of vocal effort levels, but other representations might bring better VE representations such as sums of general radial functions [Fraiha Machado et al., 2013], 1/3-octave band energy fit [Jokinen et al., 2014], or stabilised weighted linear prediction coefficients [Jokinen et al., 2015]. Further research on VE models should be conducted in order to find the best representation while maintaining flexibility for performing VE modifications.

Also, it would be desirable to combine the proposed vocal effort modifications in this dissertation based on the proposed methodology in chapter 5 and the scheme presented in section 2.6 with prosody modification for conveying expressive styles in synthetic speech and evaluate the contribution of VE against performing only prosodic modifications similar to the studies presented in chapter 3.

The proposed methodology presented in chapter 5 and the scheme proposed in chapter 6 introduce a single parameter that permit controlling the level of vocal effort to convey in the final synthetic speech signal. This opens the possibility for exploring reactive speech synthesis such as in the *HandSketch* musical instrument [d'Alessandro et al., 2014] where, with a simple user interface, a user can control the speech synthesis in real time. Proper interfaces allowing us to easily and conveniently control certain parameters of speech, such as VE, could be integrated in for portable speech synthesis systems used in medical applications for speech impaired people [Matsui et al., 2013].

Also, chapter 5 used the codebook approach, storing VE models dependent of the context: gender, VE level, phone, and phone position inside the tri-syllabic word. Further research should be conducted in order to determine which contextual parameters are the best for VE predictions used in the natural language processing (NLP) component of the proposed scheme in chapter 6. Regarding the NLP component, the implementation suggested is based on the CBR algorithm, but other ML components could be considered and a proper comparison could highlight benefits of the different approaches. In this regard, if the VE approach presented in this dissertation is to be applied in natural speech, it is important to remark that the corpus used (see section 4.1) in the proposed methodology is a specially designed corpus for VE research, where the utterances contained nearly constant VE levels along the utterance. In natural speech, a single utterance may contain VE alterations, which may produce an average VE model when performing the VE analysis. Thus, it is important to identify which speech segments inside the utterance present a representative VE model. In order to perform this discrimination, it is required to define a VE scale that permits quantifying the degree of VE conveyed in a speech segment.

Also, the proposed methodology used linear interpolation to obtain the interpolated models corresponding to conversions to intermediate VE levels. Although it makes the interpolation very straightforward and it is very intuitive to imagine what will be the resulting VE model to be applied, other interpolations, such as quadratic or cubic, should be compared with the presented approach in this dissertation and evaluate their benefits over linear interpolation.

6.3 Research contributions

International Conferences

1. Monzo, C., Calzada, A., Iriondo, I., and Socor, J. C. (2010). Expressive speech style transformation: voice quality and prosody modification using a harmonic plus noise model. In *Speech Prosody*, (Chicago)

2. Calzada, A., Socoró, J. C., and Clark, R. (2013). Parametric model for vocal effort interpolation with Harmonics Plus Noise Models. In *8th ISCA Workshop on Speech Synthesis*, pages 45–50, (Barcelona, Spain)

International journals

1. Calzada, A. and Socor, J. C. (2011). Vocal effort modification through harmonics plus noise model representation. In Travieso-Gonzalez, C. M. and Alonso-Hernandez, J., editors, *Advances in Nonlinear Speech Processing*, volume 7015 of *Lecture Notes in Computer Science*, pages 96–103. Springer, (Berlin Heidelberg)

2. Calzada, A. and Socor, J. C. (2013). Voice quality modification using a harmonics plus noise model. *Cognitive Computation*, 5(4):473–482

Participation in research projects

CreaVeu
From any text to any voice
(`http://lasallerd.salleurl.edu/CreaVeu`)

The project was conducted in the Human Computer Interaction (HCI) area of La Salle R&D and consisted in the development of the first software to generate customisable synthetic voice messages through the combination of Text-To-Speech (TTS) synthesis and Voice Transformation (VT). This technology produces synthetic speech from any input text using a simple user interface designed to intuitively modify voice personalization parameters. Different user-defined voice presets can be generated and saved for future use. The technology is a multi-platform software implemented using a client-server architecture. Speech synthesis is done using a TTS system installed on a server that receives requests from the client application, while voice customization is performed by the voice transformation module running in the client application. A non-expert user can easily define, modify and save, the specific settings of each voice preset using a Graphical User Interface (GUI). All synthesis and voice conversion was implemented by the author of this dissertation using the presented HNM implementation presented in chapter 2, section A.

Participation in events

- ▷ eNTERFACE'11. The 7th International Summer Workshop on Multimodal Interfaces. Hosted by the University of West Bohemia, Department of Cybernetics, Faculty of Applied Sciences. Pilsen, Czech Republic. 2011.
- ▷ Applications of Speech Technologies. Hosted by Centro mediterráneo, Universidad de Granada. Granada, Spain. 2011.

International stays at research centers via public fundings

- ▷ German Research Center for Artificial Intelligence - Language Technology Lab (Deutsches Forschungszentrum für Künstliche Intelligenz GmbH - DFKI-LT). Saarbrücken, Germany. From September 2011 to March 2012.
- ▷ (The) Centre for Speech Technology Research. University of Edinburgh. Edinburgh, United Kingdom. From November 2012 to May 2013.

Obtained grants

1. 2008-2009: La Salle University grant for research fellowship and project development granted by Funitec.
2. 2010-2013: FI grant, programa de ayudas para la contratación de personal investigador novel from Agencia de Gestió dAjuts Universitaris i de Recerca (AGAUR). (2010 FI B01083)
3. 2010-2011: BE-DGR grant for research stays abroad Spain from AGAUR. (2010 BE 100503)
4. 2011-2012: BE-DGR grant for research stays abroad Spain from AGAUR. (2011 BE 101084)

Appendices

APPENDIX A

Harmonics plus Noise Model (HNM) implementation

Harmonic-plus-Noise Model (henceforth HNM) is a method based on decomposition of the speech spectrum into two subbands (see eq. (A.1)), and the frequency delimiting the subbands boundary is called **maximum voicing frequency** (ω_m). The part of the spectrum below ω_m is known as the **harmonic part** while the upper band is usually referred to as the **stochastic part**. The harmonic part is modelled by a sum of harmonically related sinusoids while the stochastic part is modelled with an AutoRegressive (AR) model.

$$s[n] = s_h[n] + s_n[n], \tag{A.1}$$

where $s_h[n]$ and $s_n[n]$ correspond to the harmonic and stochastic parts, respectively. The model considers the harmonic part stationary in time windows of two to four pitch periods, and the stochastic parts, in windows of 5ms. Thus, the harmonic and stochastic parts can be analysed at a different frame rate. A single harmonic frame ($\hat{s}_h^k[n]$), k being the frame index, is modelled as

$$\hat{s}_h^k[n] = \sum_{l=1}^{L^k} a_l^k \ \cos(\omega_l^k \ n + \varphi_l^k), \tag{A.2}$$

where L^k the number of harmonics (that depends on the relation between ω_m and the fundamental harmonic angular frequency (ω_0^k), which is the inverse of the pitch period in the frame k); a_l^k, ω_l^k, and φ_l^k correspond to the amplitude, angular frequency, and phase estimated for the l^{th} harmonic respectively, and usually the following condition is applied $\omega_l^k = l \cdot \omega_0^k$.

And a stochastic frame ($\hat{s}_n^j[n]$), being j the frame index, is modelled as

$$\hat{s}_n^j[n] = \sqrt{p^j}\ \sigma^j[n] * \boldsymbol{H}_{LPC}^j, \tag{A.3}$$

where p^j is the power to be applied to the noise contribution, $\sigma^j[n]$ is White Gaussian Noise (WGN) and $\boldsymbol{H}_{LPC}^j$ are the AR filter coefficients for the frame j.

The complete signal is obtained synthesising each part ($s_h[n]$ and $s_n[n]$) separately and then adding them together. The synthesis of each part of the signal is done frame by frame using the equations (A.2) and (A.3) to generate each resynthesized frame for the harmonic ($\hat{s}_h^k[n]$) and the stochastic ($\hat{s}_n^j[n]$) parts, respectively. Then, the frames for each part are delayed and summed with the Overlap-and-Add (OLA) technique to obtain the whole harmonic and stochastic parts separately. Finally, the complete signal is obtained adding the harmonic and stochastic parts together following equation (A.1). Thus, for a given speech signal, the harmonic part is composed of K frames in which every frame k is described by the following set of parameters:

- ▷ $\boldsymbol{A}^k$: a column vector of L^k elements with the amplitudes a_l^k of the L_k harmonics of the frame k.
- ▷ $\boldsymbol{F}^k$: a column vector of L_k elements with the angular frequencies ω_k^l divided by 2π of the L_k harmonics of the frame k (they contain the normalized frequencies).
- ▷ $\boldsymbol{\Phi}^k$: a column vector of L_k elements with the phases φ_l^k of the L_k harmonics of the frame k.
- ▷ $\boldsymbol{t}_{a|s}^k$: the analysis ($\boldsymbol{t_a^k}$) or synthesis ($\boldsymbol{t_s^k}$) time instants, expressed in samples, defining the beginning and end sample for each harmonic frame k.

The stochastic part is composed of J frames in which every frame j is described by the following set of parameters:

- ▷ p^j: the power of the input white Gaussian noise at frame k, $\sqrt{p^j}$, that is fed to the LPC AR synthesis filter.
- ▷ $\boldsymbol{H}_{LPC}^j$: the LPC filter coefficients column vector for the frame j.
- ▷ $\boldsymbol{t}_{a|s}^j$: the analysis ($\boldsymbol{t_a^j}$) or synthesis ($\boldsymbol{t_s^j}$) time instants array, expressed in samples, defining the beginning and end sample for each stochastic frame j.

It is important to mention that each part, the harmonic ($s_h[n]$) and stochastic ($s_n[n]$), is analysed and synthesized independently, which means that each part may have its own analysis and synthesis time instants, and thus, different amount of analysis windows. In that case, the signal would have two arrays of time instants, one for each part of the signal.

The harmonic part models the periodic events present in the speech waveform which are mainly produced by the vibration of the vocal folds. On the other hand, the stochastic part models the non-periodic events such as breath, aspiration, and noise turbulences (e.g., due to lips constriction, etc.) usually present in obstruent sounds. The value of maximum voiced frequency (ω_m) is usually fixed to a constant value, 5 Khz being the most widely used [Erro, 2008, McAulay and Quatieri, 1986, Quatieri and McAulay, 1986], but other implementations compute ω_m for each frame instead of fixing it before computing other model parameters [Stylianou et al., 1997, Stylianou, 2001, Kim et al., 2006]. For the experiments presented in this dissertation ω_m was fixed to 5 Khz.

Figure A.1 shows the main blocks of the HNM implementation. The process begins by analysing the signal for extracting the harmonic part ($s_h[n]$) parameters (amplitudes ($\boldsymbol{A}^k$), frequencies ($\boldsymbol{F}^k$), and phases ($\boldsymbol{\Phi}^k$)) for $1 \leq k \leq K$. With these the harmonic part signal is synthesized, producing $\hat{s}_h[n]$. Next, in order to obtain the noise part ($s_n[n]$) parameters, the harmonic part synthesized ($\hat{s}_h[n]$) is subtracted from the original signal $s[n]$, yielding the original noise part ($s_n[n]$). The original noise part is analysed in order to obtain the corresponding parameters ($\boldsymbol{H}^j_{LPC}$ and p^j) for $1 \leq j \leq J$. Finally, in order to reconstruct the original signal, a noise generator block generates white Gaussian noise with variance $\sqrt{p^j}$, which is introduced into the LPC filter to reconstruct the noise part of the signal ($\hat{s}_n[n]$) which is finally added to the synthesized harmonic part ($\hat{s}_h[h]$), thus obtaining the complete synthesized signal ($\hat{s}[n]$).

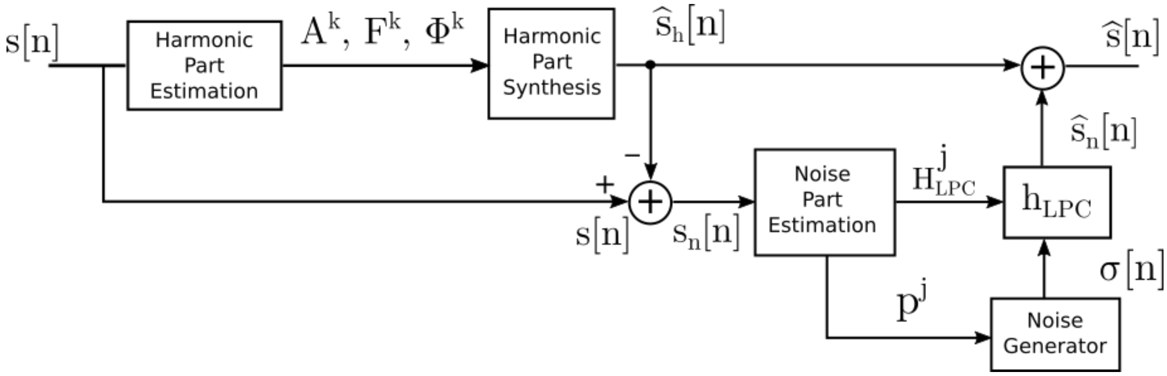


Figure A.1 — Overview of the HNM schema implemented.

As seen in equation (A.2), the harmonic part models the signal as a sum of harmonics. The estimation of the harmonic parameters can be done in the frequency domain with the iterative algorithm proposed by [Depalle and Hélie, 1997]. There are also alternatives for parameter extraction in time domain, which require less computational effort [Erro, 2008]; however this approach does not perform correction of the estimated frequencies in case there are errors in the pitch marking used as a reference. The method used allows reducing the analysis window size required to two periods of the fundamental frequency of the signal frame to be analysed [Erro, 2008, Stylianou et al., 1997]; thus reducing the window size and overcoming the problem of smoothing rapid variations in the signal [Depalle and Hélie, 1997]. For the harmonic part analysis, the signal is analysed pitch synchronously, which means that the analysis time instants ($\boldsymbol{t}_a$) match the pitch marks. The length of the analysis window was set to $2T_0$, where T_0 is the distance between the pitch mark coinciding with the analysis time instant (t_a^k) and the following pitch mark. Figure A.2 shows a graphic representation of the pitch synchronous analysis carried out for the harmonic part. As can be seen, there is an overlap between the analysis time windows.

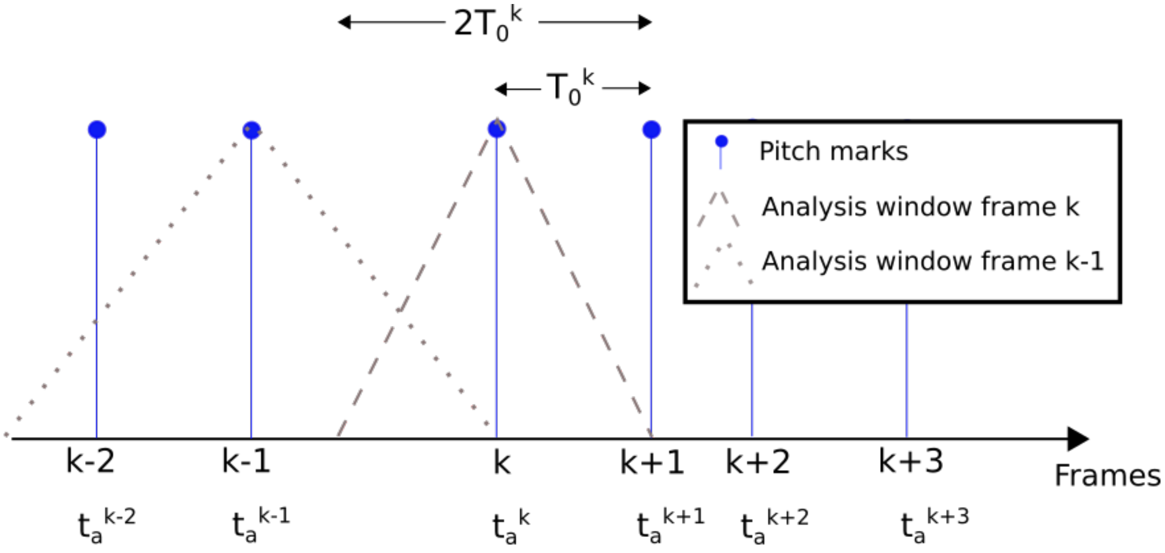


Figure A.2 — Graphical representation of the analysis windows used for frame k at analysis time t_a^k and frame $k-1$ centred at t_a^{k-1}.

A.1 Harmonic part modelling

The spectral method used for extracting the spectral parameters is based on the Short Time Fourier Transform (henceforth STFT), and usually the stochastic component is omitted (or relegated to a second place). Using the STFT to extract the spectral parameters, we assume that amplitudes and frequencies remain constant along the analysed frame. Then we can express the Fourier Transform (FT) $\hat{S}(f)$ of a windowed frame $s[n]$ as

$$\hat{S}(f) = \frac{1}{2}\sum_{l=1}^{L} a_l \left(e^{j\varphi_l} W(f - f_l) + e^{-j\varphi_l} W(f + f_l)\right), \tag{A.4}$$

where $W(f)$ is the Fourier Transform of the analysis window, and a_l, φ_l, f_l are the amplitudes, phases, and frequencies of the harmonics in the speech signal to be analysed. Here the dependence on the frame number has been omitted for the sake of clarity.

We could consider $\hat{S}(f)$ a model of the Fourier Transform of a windowed frame of the observed signal $s[n]$. Our goal is to find the parameters for which the model $\hat{S}(f)$ best fits the observation $S(f)$ according to the least squares criterion. Considering that $\hat{S}(f)$ is sampled at N equidistant frequencies $F_j = (j-1)\frac{1}{N}$ for $j = 1 \cdots N$, and defining $\boldsymbol{S}$ as an N-size column vector that contains the DFT of the windowed frame, the best estimate $\hat{\boldsymbol{S}}$ of $\boldsymbol{S}$ is obtained by minimizing the euclidean distance

$$\hat{\boldsymbol{S}} = \underset{a_l, f_l, \varphi_l}{\operatorname{argmin}}(\|\boldsymbol{S} - \hat{\boldsymbol{S}}\|_2). \tag{A.5}$$

As we can see if we substitute equation (A.4) in (A.5), it is not a linear problem, due to the analysis window spectrum dependency on the frequencies (see the terms $W(f \mp f_l)$ in equation (A.4)). In order to overcome this non-linearity, the analysis window spectrum is approximated by its Taylor series and truncated at the first element. The parameter estimation is performed in two steps. At the first step, amplitudes (a_l) and phases (φ_l) are considered unknown and frequencies (ω_l) a known parameter. In the next step the opposite is done, amplitudes and phases are considered as given and frequencies as variables. The estimation process is performed iteratively, repeating the two steps, thus obtaining an adjusted estimation of the parameters (a_l, φ_l and f_l).

Amplitude and phase estimation

Developing equation (A.4) we get

$$\begin{aligned}\hat{S}(f) &= \frac{1}{2}\sum_{l=1}^{L} a_l \left(e^{j\varphi_l} W(f-f_l) + e^{-j\varphi_l} W(f+f_l)\right) \\ &= \frac{1}{2}\sum_{l=1}^{L} a_l \left\{\left[\left(\cos\varphi_l + j\sin\varphi_l\right) W(f-f_l)\right] + \left[\left(\cos\varphi_l - j\sin\varphi_l\right) W(f+f_l)\right]\right\} \\ &= \frac{1}{2}\sum_{l=1}^{L} a_l \left\{\cos\varphi_l \left[W(f-f_l) + W(f+f_l)\right] + j\sin\varphi_l \left[W(f-f_l) - W(f+f_l)\right]\right\}.\end{aligned} \tag{A.6}$$

From (A.6) we define the following terms

$$\begin{cases} \wp_l = a_l\ \cos\varphi_l & l \in [1..L] \\ \wp_{l+L} = a_l\ \sin\varphi_l & l \in [1..L] \end{cases} \tag{A.7a}$$

$$\begin{cases} \mathcal{H}_l(f) = W(f-f_l) + W(f+f_l) & l \in [1..L] \\ \mathcal{H}_{L+l}(f) = j\left[W(f-f_l) - W(f+f_l)\right] & l \in [1..L] \end{cases} \tag{A.7b}$$

We can use (A.7b) to define the matrix $\boldsymbol{\mathcal{H}}$ of size Nx$2L$, being the element $\mathcal{H}_{j,l} = \mathcal{H}_l(F_j)$, where F_j corresponds to one of the sampled frequencies by the DFT. Also a vector $\boldsymbol{\wp}$ with $2L$ elements can be defined using equation (A.7a). We can now rearrange a discretisation of equation (A.6) in matrix form with $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\wp}$ as

$$\hat{\boldsymbol{S}} = \boldsymbol{\mathcal{H}}\boldsymbol{\wp}, \tag{A.8}$$

where

$$\boldsymbol{\mathcal{H}} = (\mathcal{H})_{j,l} = \mathcal{H}_l(F_j) = \begin{bmatrix} \mathcal{H}_1(0) & \mathcal{H}_2(0) & \cdots & \mathcal{H}_{2L}(0) \\ \mathcal{H}_1(\frac{1}{N}) & \mathcal{H}_2(\frac{1}{N}) & \cdots & \mathcal{H}_{2L}(\frac{1}{N}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{H}_1(\frac{N-1}{N}) & \mathcal{H}_2(\frac{N-1}{N}) & \cdots & \mathcal{H}_{2L}(\frac{N-1}{N}) \end{bmatrix} \tag{A.9}$$

and

$$\boldsymbol{\wp} = \begin{bmatrix} \wp_1 \\ \wp_2 \\ \vdots \\ \wp_L \\ \wp_{L+1} \\ \vdots \\ \wp_{2L} \end{bmatrix} \qquad \hat{\boldsymbol{S}} = \begin{bmatrix} S[F_1] \\ S[F_2] \\ \vdots \\ S[F_N] \end{bmatrix}. \tag{A.10}$$

Paying attention to the dimensions of $\boldsymbol{\mathcal{H}}$ and $\boldsymbol{\wp}$, we can see that it is an overdetermined system, as usually $N > 2L$. The least squares solution is the proposed approach to be used.

Equation (A.11) resumes the least square solution derivation based on gradients of the cost function. The orthogonality principle can be alternatively used for deriving the same solution [Moon, 1999]. The least squares solution

$$\begin{aligned} \varepsilon = \|\boldsymbol{S} - \hat{\boldsymbol{S}}\|_2^2 = (\boldsymbol{S} - \hat{\boldsymbol{S}})^H(\boldsymbol{S} - \hat{\boldsymbol{S}}) &= \boldsymbol{S}^H\boldsymbol{S} - \boldsymbol{S}\hat{\boldsymbol{S}} - \hat{\boldsymbol{S}}^H\boldsymbol{S} + \hat{\boldsymbol{S}}^H\hat{\boldsymbol{S}} \\ &= \boldsymbol{S}^H\boldsymbol{S} - \boldsymbol{S}\boldsymbol{\mathcal{H}}\boldsymbol{\wp} - \boldsymbol{\wp}^H\boldsymbol{\mathcal{H}}\boldsymbol{S} + \boldsymbol{\wp}^H\boldsymbol{\mathcal{H}}^H\boldsymbol{\mathcal{H}}\boldsymbol{\wp} \\ \frac{\delta\varepsilon}{\delta\boldsymbol{\wp}^H} = 0 = -\boldsymbol{\mathcal{H}}^H\boldsymbol{S} + \boldsymbol{\mathcal{H}}^H\boldsymbol{\mathcal{H}}\boldsymbol{\wp} &\rightarrow \boldsymbol{\mathcal{H}}^H\boldsymbol{\mathcal{H}}\boldsymbol{\wp} = \boldsymbol{\mathcal{H}}^H\boldsymbol{S} \end{aligned}, \tag{A.11}$$

$$\boldsymbol{\wp} = (\boldsymbol{\mathcal{H}}^H\boldsymbol{\mathcal{H}})^{-1}\boldsymbol{\mathcal{H}}^H\boldsymbol{S}, \tag{A.12}$$

coincides with the Moore-Penrose pseudo-inverse of a non-squared matrix, where A^H is the Hermitian (self-adjoin or conjugate transpose) matrix of A, where the element $A^H_{i,j} = A^*_{j,i}$.

Inversion of $\boldsymbol{\mathcal{H}}^H\boldsymbol{\mathcal{H}}$ matrix can be computed using the robust pseudo-inverse method [Moon, 1999], thus avoiding some possible practical issues.

As we can see in equation (A.7a), $\boldsymbol{\wp}$ contains the amplitudes and phases, so working with the expression we can get their values as

$$\begin{cases} \wp_l = a_l \ \cos\varphi_l \\ \wp_{l+L} = a_l \ \sin\varphi_l \end{cases},$$

$$\wp_l^2 + \wp_{l+L}^2 = a_l^2 \overbrace{[\cos^2(\varphi_l) + \sin^2(\varphi_l)]}^{=1} \rightarrow \wp_l^2 + \wp_{l+L}^2 = a_l^2, \tag{A.13}$$

$$a_l = \sqrt{\wp_l^2 + \wp_{l+L}^2}.$$

Phase can be estimated from the following trigonometric expression

$$\arctan\left(\frac{\wp_{l+L}}{\wp_l}\right) = \arctan\left(\frac{a_l \sin(\varphi_l)}{a_l \cos(\varphi_l)}\right) = \arctan\left(\frac{\sin(\varphi_l)}{\cos(\varphi_l)}\right) = \varphi_l \tag{A.14}$$

Frequency estimation

To extract the frequency values, amplitudes (a_l) and phases (φ_l) are considered to be known. As we pointed out in page 117, the minimization process shown in equation (A.5) is non-linear with respect to frequency. In order to be able to solve the system easily, a linear dependency is enforced. This is achieved by representing the spectrum of the analysis window centred at each harmonic by its Taylor series truncated at the first term as

$$W(f \mp f_l) = W(f \mp \mathcal{F}_l) \mp W'(f \mp \mathcal{F}_l)\Delta_l, \tag{A.15}$$

where $\Delta_l = f_l - \mathcal{F}_l$ is the distance between the estimated frequencies in the last iteration ($\mathcal{F}_l$) and the real frequencies in the signal (f_l).

As can be seen when substituting equation (A.15) into (A.4), the obtained expression becomes linearly dependent on the frequencies (f_l). Now the goal is to find the values of f_l and thus, we need to know the distance between the previous rough approximation of the frequencies ($\mathcal{F}_l$) and the actual harmonic's frequency value (f_l) with Δ_l. The $\mathcal{F}_l$ values were previously assigned with a preliminary rough initialisation.

In order to obtain a linear expression of $\hat{\boldsymbol{S}}$ with $\boldsymbol{\Delta}$, the frequency dependence is linearised by limiting the expansion of the FT of the analysis window around $f - \mathcal{F}_l$ at its first-order term as expressed in (A.15).

Substituting (A.15) into (A.4) and discretising the variable frequency on the DFT values $f = F_j$ for $j = 1..N$ we obtain

$$\hat{\boldsymbol{S}} = \tilde{\boldsymbol{S}} + \boldsymbol{\Omega\Delta}, \tag{A.16}$$

where $\tilde{\boldsymbol{S}} = \boldsymbol{\mathcal{H}}\boldsymbol{\wp}$ is the estimation of the DFT of the speech frame in first step of the iterative algorithm obtained from the last iteration, $\hat{\boldsymbol{S}}$ represents the same estimation after taking into account a correction term due to estimation of amplitudes and phases in the first step of the algorithm, and $\boldsymbol{\Omega}$ is a matrix with dimensions NxL defined as

$$\Omega_{j,l} = a_l \left[-e^{j\varphi_l} W'(F_j - \mathcal{F}_l) + e^{-j\varphi_l} W'(F_j + \mathcal{F}_l) \right] \qquad 1 \leq j \leq N, 1 \leq l \leq L, \tag{A.17}$$

where F_j are the frequencies evaluated by the DFT, $W'(f)$ is the derivative of the window in the frequency domain, and $\mathcal{F}_l$ the estimated harmonic frequencies in the last iteration.

To compute the updating correction factor for frequencies Δ_l which allow getting a more accurate estimation of $S(f)$, the least squares error criterion is used. We begin expressing the least squares error using equation (A.16) as

$$\begin{aligned} \varepsilon = \|\boldsymbol{S} - \hat{\boldsymbol{S}}\|_2^2 &= \left[\boldsymbol{S} - \left(\tilde{\boldsymbol{S}} + \boldsymbol{\Omega\Delta} \right) \right]^H \left[\boldsymbol{S} - \left(\tilde{\boldsymbol{S}} + \boldsymbol{\Omega\Delta} \right) \right] \\ &= \boldsymbol{S}^H\boldsymbol{S} - \boldsymbol{S}^H\tilde{\boldsymbol{S}} - \boldsymbol{S}^H\boldsymbol{\Omega\Delta} - \boldsymbol{S}^H\boldsymbol{S} + \boldsymbol{S}^H\tilde{\boldsymbol{S}} + \boldsymbol{S}^H\boldsymbol{\Omega\Delta} - \boldsymbol{\Delta}^H\boldsymbol{\Omega}^H\boldsymbol{S} + \boldsymbol{\Delta}^H\boldsymbol{\Omega}^H\tilde{\boldsymbol{S}} + \boldsymbol{\Delta}^H\boldsymbol{\Omega}^H\boldsymbol{\Omega\Delta}. \end{aligned} \tag{A.18}$$

Setting the gradient of the previous expression error with respect to $\boldsymbol{\Delta}$ to zero and substituting $\Delta_l = f_l - \mathcal{F}_j$ the least squares solution is found,

$$f_l = \mathcal{F}_j + \left(\boldsymbol{\Omega}^H \boldsymbol{\Omega} \right)^{-1} \boldsymbol{\Omega}^H \left(\boldsymbol{S} - \tilde{\boldsymbol{S}} \right). \tag{A.19}$$

The estimation of the frequencies f_l in this second step is based on the following assumptions:

- ▷ The harmonic's amplitudes and phases are assumed to be known (from equations (A.13) and (A.14)).

- ▷ An initialisation of the frequencies is required at the beginning of the iterative procedure, and it could be based on pitch mark analysis of the speech signal.

- ▷ The analysis window spectrum is approximated by the Taylor series truncated at the first coefficient, which implies linearising its variations. This fact leads to some considerations that must be taken into account in the selection of the type of the analysis window [Depalle and Hélie, 1997].

A.1.1 Forcing harmonic frequencies constraint with Lagrange multiplyiers

The algorithm used for computing the HNM harmonic parameters (amplitudes, frequencies and phases) is based on [Depalle and Hélie, 1997], where the frequencies obtained where not forced to have a harmonic structure. In order to enforce the harmonic structure of the estimated frequencies, a constrained optimization based on Lagrange multipliers was introduced in the harmonic frequencies estimation step.

The harmonic structure in the frequencies means that the frequencies array for a given frame has the following structure

$$\boldsymbol{f}_k = f_0 \times [1\ 2\ 3\ \cdots L_k] = f_0\ \boldsymbol{f}_h, \tag{A.20}$$

, where f_0 is the fundamental frequency, or pitch, in the frame, which in expression (A.20) is considered a scalar and can be obviated in the following, and $\boldsymbol{f}_h$ is a vector that enforces the harmonic dependency between the frequencies, and thus is the constraint to impose.

In order to apply the constrained optimization procedure, first the restrictions ($\boldsymbol{h}(x)$) have to be defined as equalities. Thus, to enforce the harmonic structure, an equations system can be used to compare the element pairs of $\boldsymbol{f}_h$ ensuring that the second element doubles the first, three times the second element to be double of the third, four times the third element triples the third, and so on. From a geometric point of view it is an orthogonal vector space to $\boldsymbol{f}_h$, so the next system of $K-1$ linear equations and K variables is obtained

$$\begin{aligned}
&[\begin{matrix} -2 & 1 & 0 & \cdots & 0 & 0 \end{matrix}] \cdot \boldsymbol{f}_h = 0\\
&[\begin{matrix} 0 & -3 & 2 & \cdots & 0 & 0 \end{matrix}] \cdot \boldsymbol{f}_h = 0\\
&\qquad\vdots\\
&[\begin{matrix} 0 & 0 & 0 & \cdots & -K & K-1 \end{matrix}] \cdot \boldsymbol{f}_h = 0
\end{aligned} \tag{A.21}$$

Putting (A.21) into a matrix form, the system becomes

$$\begin{bmatrix} -2 & 1 & 0 & \cdots & 0 & 0 \\ 0 & -3 & 2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -K & K-1 \end{bmatrix} \boldsymbol{f}_h = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} = \boldsymbol{R}_h \boldsymbol{f}_h = \boldsymbol{h}(x) \tag{A.22}$$

At this point, we can define the error function that incorporates the constraints through the corresponding Lagrange multiplier, as

$$\varepsilon = (\boldsymbol{S} - \tilde{\boldsymbol{S}} - \boldsymbol{\Omega f}_h)^H (y - \boldsymbol{\Omega f}_h) + \lambda(\overbrace{\boldsymbol{R}_h \boldsymbol{f}_h}^{\boldsymbol{h}(x)}) \tag{A.23}$$

and deriving equation equation (A.23) with respect to $\boldsymbol{f}_k$ and λ, and making it equal to zero we obtain

$$\begin{aligned}
\frac{\delta\varepsilon}{\delta f_k} &= -\boldsymbol{\Omega}^H(\boldsymbol{S} - \boldsymbol{\Omega} f_h) + \boldsymbol{R}_h^H \lambda = 0\\
\boldsymbol{\Omega}^H \boldsymbol{\Omega} f_h &= \boldsymbol{\Omega}^H \boldsymbol{S} - \boldsymbol{R}_h^H \lambda\\
f_h &= (\boldsymbol{\Omega}^H \boldsymbol{\Omega})^{-1}(\boldsymbol{\Omega}^H \boldsymbol{S} - \boldsymbol{R}_h^H \lambda),
\end{aligned} \tag{A.24}$$

and

$$\begin{aligned}
\frac{\delta\varepsilon}{\delta\lambda} &= \boldsymbol{R}_h f_h = 0\\
\boldsymbol{R}_h(\boldsymbol{\Omega}^H \boldsymbol{\Omega})^{-1}(\boldsymbol{\Omega}^H \boldsymbol{S} - \boldsymbol{R}_h^H \lambda) &= 0\\
\boldsymbol{R}_h(\boldsymbol{\Omega}^H \boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}^H \boldsymbol{S} - \boldsymbol{R}_h(\boldsymbol{\Omega}^H \boldsymbol{\Omega})^{-1}\boldsymbol{R}_h^H \lambda &= 0\\
\lambda = [-\boldsymbol{R}_h(\boldsymbol{\Omega}^H \boldsymbol{\Omega})^{-1}\boldsymbol{R}_h^H]^{-1}\ \boldsymbol{R}_h(\boldsymbol{\Omega}^H \boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}^H \boldsymbol{S}&.
\end{aligned} \tag{A.25}$$

Then, substituting equation (A.25) into (A.24), the expression

$$\boldsymbol{f}_h = (\boldsymbol{\Omega}^H\boldsymbol{\Omega})^{-1}\left\{\boldsymbol{\Omega}^H\boldsymbol{S} - \boldsymbol{R}_h^H\overbrace{[-\boldsymbol{R}_h(\boldsymbol{\Omega}^H\boldsymbol{\Omega})^{-1}\boldsymbol{R}_h^H]^{-1}\ \boldsymbol{R}_h(\boldsymbol{\Omega}^H\boldsymbol{\Omega})^{-1}\boldsymbol{\Omega}^H\boldsymbol{S}}^{\lambda}\right\} =$$

$$(\boldsymbol{\Omega}^H\boldsymbol{\Omega})^{-1}[I - \boldsymbol{R}_h^H(\boldsymbol{R}_h(\boldsymbol{\Omega}^H\boldsymbol{\Omega})^{-1}\boldsymbol{R}_h^H)^{-1}\boldsymbol{R}_h(\boldsymbol{\Omega}^H\boldsymbol{\Omega})^{-1}]\boldsymbol{\Omega}^H\boldsymbol{S} = \boldsymbol{f}_h \tag{A.26}$$

is obtained. Equation (A.26) allows to find the frequencies, forcing them to be harmonically related, and it is the equation used for frequency estimation, instead of equation (A.19) from [Depalle and Hélie, 1997], in the speech parametrisation carried out along this dissertation (see chapters 4. 5).

A.1.2 Summary of the iterative algorithm for harmonic part computation

```
$n = 0$
Initial frequencies are $\boldsymbol{f}^k = f_0^k\ [1, 2, \cdots, L_k]$ where $f_0^k = \frac{1}{t_a^k - t_a^{k-1}}$
while $(n < n_{max})$ or (Relative Error $<$ threshold) do
    Step 1: Computation of amplitudes and phases using eq. (A.12)-(A.14).
    Step 2: Refinement of the harmonic frequencies using equation (A.26).
end
```

The relative error is computed in the frequency domain as the norm of the difference between the DFT of the frame to model and the sampling of the theoretical model using equation (A.6) and the DFT frequencies, normalized with the norm of the DFT for the frame.

For the speech parametrization throughout this work using HNM, the maximum number of iteration was set to 40, the error threshold to -40 dB's and the DFT length to 1024 points. In Step 2, the options implemented using the harmonic frequencies constraint for the model frequencies using eq. (A.26) was used in chapters 3, 4 and 5.

A.1.3 Harmonic part signal reconstruction

Based on equation (A.2), a single frame k is synthesised as

$$s_h^k[n] = \sum_{l=1}^{L^k} a_l^k\ \cos(\omega_l^k\ n + \varphi_l^k) \quad , \quad t_s^{k-1}\text{-}t_s^k \le n \le t_s^{k+1}.\text{-}t_s^k \tag{A.27}$$

It is important to emphasize that the synthesis window is not symmetrical to the centre of the frame (see figure A.3). For a given frame k, the synthesised signal corresponding to the harmonic component ($s_h^k[n]$) has two parts: the left, part which corresponds to the signal compressed between the synthesis time instants t_s^{k-1} and t_s^k, and the right part, which is the signal between the synthesis instants t_s^k and t_s^{k+1}.

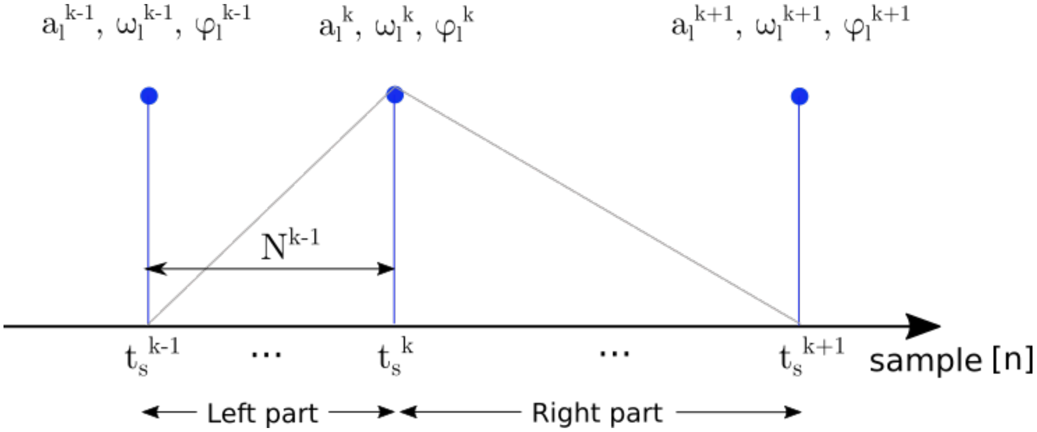

$a_l^{k-1}, \omega_l^{k-1}, \varphi_l^{k-1}$
$a_l^{k}, \omega_l^{k}, \varphi_l^{k}$
$a_l^{k+1}, \omega_l^{k+1}, \varphi_l^{k+1}$
N^{k-1}
t_s^{k-1}
...
t_s^{k}
...
t_s^{k+1}
sample [n]
Left part
Right part

Figure A.3 — Frame synthesis Overlap-and-Add for the harmonic part reconstruction (A.27).

However, the synthesis for the whole harmonic part of the signal is done using the Overlap-and-Add (OLA) technique, which means overlapping the current frame (k) with the previous ($k-1$) and the next ($k+1$) frame. In this work, a linear interpolation is performed during each time period between two synthesis marks, considering the frames that use the parameters of the HNM that correspond to these time marks. Considering the plot depicted in figure A.3, two steps are envisaged. First, the signal between t_s^{k-1} and t_s^k is generated with the right part of $s_h^{k-1}[n]$ and the left part of $s_h^k[n]$. Next, the signal between t_s^k and t_s^{k+1} is generated with the right part of $s_h^k[n]$ and the left part of $s_h^{k+1}[n]$. The synthesis for the whole harmonic part signal is performed with

$$\hat{s}_h[t_s^{k-1}+m] = \left(\frac{N^{k-1}-m}{N^{k-1}}\right)\ s_h^{k-1}[m] + \left(\frac{m}{N^{k-1}}\right)\ s_h^k[m-N^{k-1}] \qquad 0 \leq m \leq N^{k-1}, \tag{A.28}$$

where

$$N^{k-1} = t_s^k - t_s^{k-1} \tag{A.29}$$

corresponds to the number of samples between the two consecutive synthesis times. If no time modifications are performed to the signal, the analysis time instants (t_a^k) are the same as the synthesis time instants (t_s^k) and no corrections to the model parameters are required.

A.2 Stochastic part modelling

The stochastic part is computed once the harmonic part has been subtracted from the original signal. Figure A.4 shows examples of signals used in the process. With the harmonic part parameters, the harmonic part signal is resynthesized with equations (A.27) and (A.28) (see figure A.4(b)) and subtracted from the original signal (figure A.4(a)). The resulting signal (see figure A.4(c)) can be used as an estimation of the noisy part of the signal $s_n[n]$, that is

$$\tilde{s}_n[n] = s[n] - \hat{s}_h[n], \tag{A.30}$$

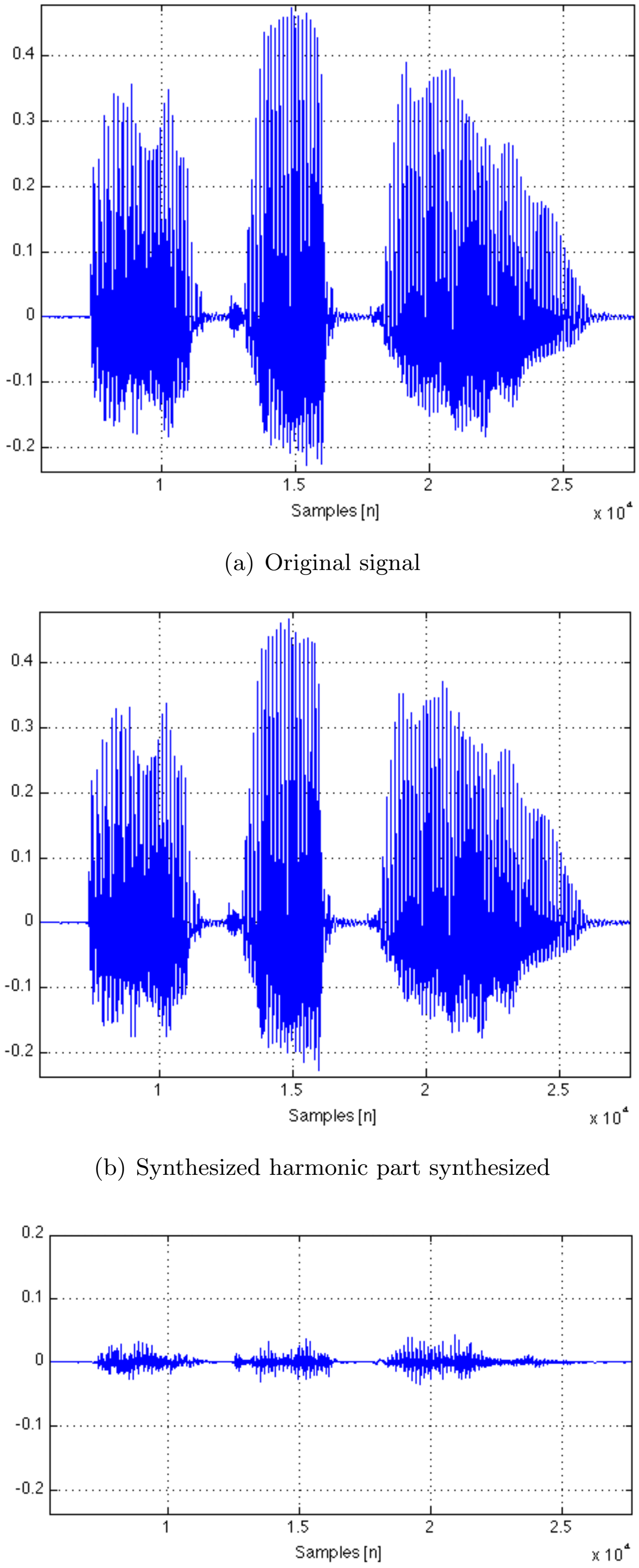
where $s[n]$ is the original signal, $\hat{s}_h[n]$ is the resynthesized harmonic part signal, using only the harmonic part parameters (amplitudes ($\boldsymbol{A}^k$),frequencies ($\boldsymbol{F}^k$) and phases ($\boldsymbol{\Phi}^k$)), and $\tilde{s}_n[n]$ is an estimation of the original signal's noise component.

In figure A.6, the spectrogram of the original signal, the spectrogram of the synthesized harmonic part signal ($\hat{s}_h[n]$), and the spectrogram of the resulting noise signal after subtracting the synthesized harmonic part from the original signal are shown. As can be seen, after the subtraction, there is still some information in the lower band of the spectrum. This difference can be attributed to different aspects, among which are the difficulty in obtaining a harmonic signal which represents all the quasi-periodic components of the speech signal (in [Erro, 2008], the error in the harmonic component estimation is highlighted, especially in the time region between two analysis marks). In this work, this phenomenon has not been tackled and all the experiments have been performed without trying to reduce the presence of errors due to modelling inaccuracies.

The stochastic part of an input signal frame is modelled by an Auto-Regressive (AR) model which is represented by an amplitude modulated white Gaussian noise source fed to an all-pole filter of order 15 [Stylianou, 1996]. The filter coefficients in the analysis stage are computed as the Linear Prediction Coefficients (LPCs), and the source noise variances ($\sqrt{p^j}$) are derived from the residual energy.

Linear prediction systems try to predict the current sample from a linear combination of previous observations. Given the signal $\tilde{s}_n[n]$, the AR model of $\hat{s}_n[n]$ may be represented as:

$$\hat{s}_n[n] = \sum_{k=1}^{P} \alpha_k \tilde{s}_n[n-k] \tag{A.31}$$

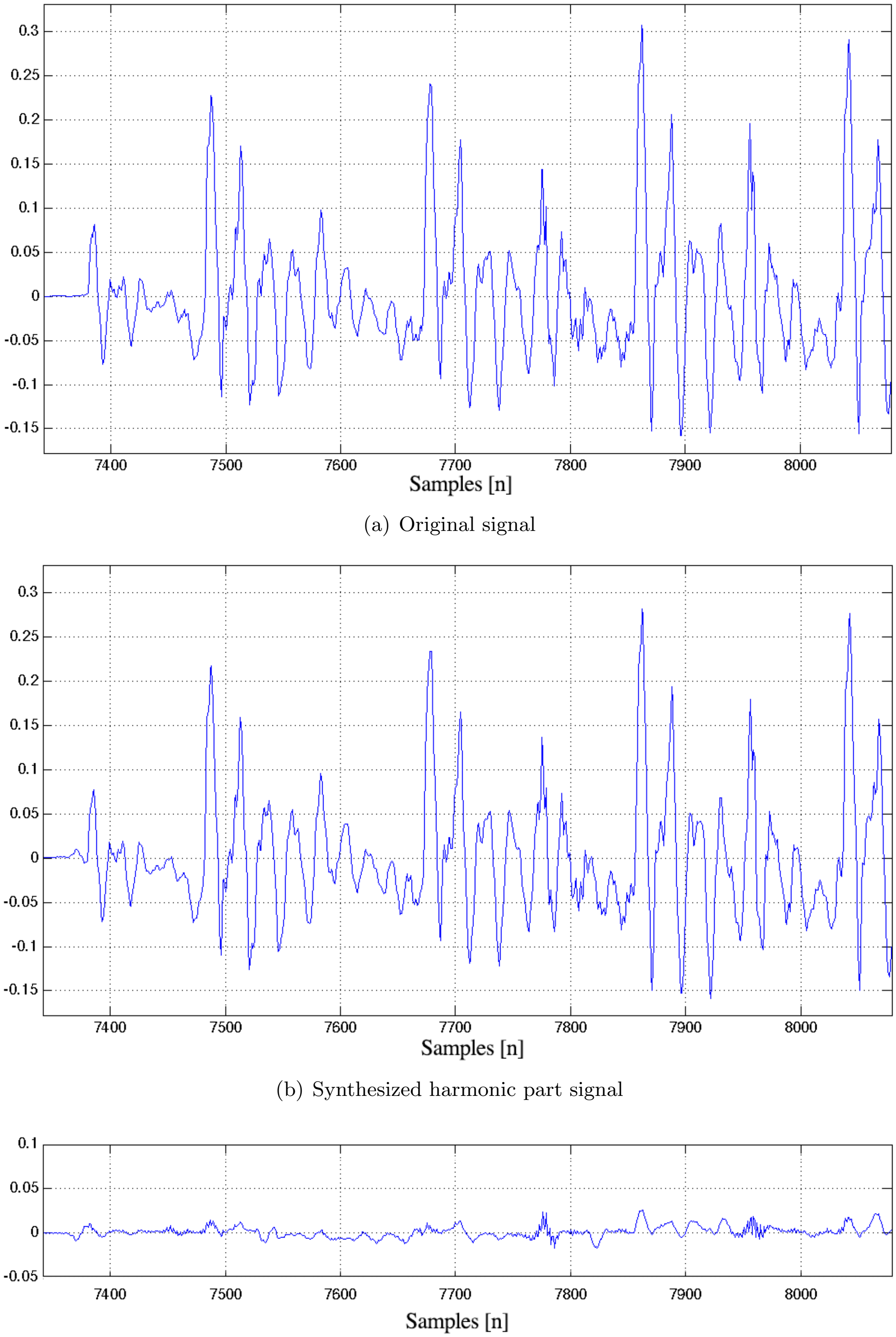


(a) Original signal

(b) Synthesized harmonic part synthesized

(c) Original noise part

Figure A.4 — Example of subtracting the synthesized harmonic part ($\hat{s}_h[n]$) from the original signal ($s[n]$). (a) Original signal, (b) signal synthesized only with harmonic parameters, and (c) result of applying equation (A.30)

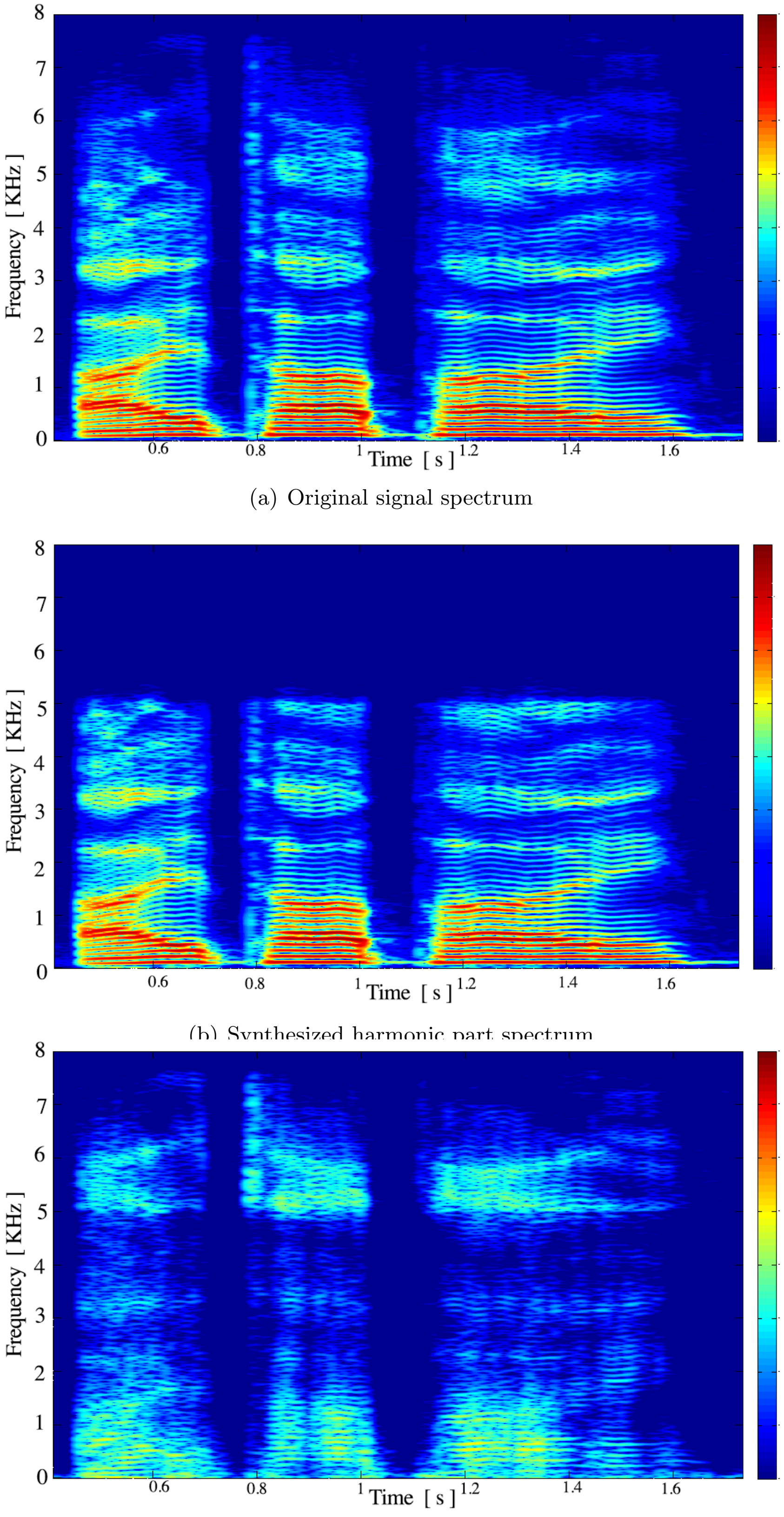


(a) Original signal

(b) Synthesized harmonic part signal

(c) Original noise part

Figure A.5 — Amplified time scale of the example shown in figure A.4. (a) Original signal, (b) signal synthesized only with harmonic parameters and (c) result of applying equation (A.30)



(a) Original signal spectrum

(b) Synthesized harmonic part spectrum

(c) Original noise part spectrum

Figure A.6 — Spectrum using 4096-point DFT of the signal presented in figure A.4. (a) Original signal spectrum, (b) spectrum of signal synthesized only with harmonic parameters, and (c) spectrum of the residual signal after subtracting the harmonic part from the original signal using equation (A.30)

Signal reconstruction

With the parameters obtained from the HNM analysis, the jth frame of the stochastic signal is reconstructed by

$$s_n^j[n] = \sqrt{p}_j \sigma^j[n] * \boldsymbol{H}_{LPC}^j \qquad 0 \leq n \leq t_s^j - t_s^{j-1}, \tag{A.32}$$

where t_s^j is the synthesis time stamps of the stochastic component.

The complete stochastic part signal is reconstructed as

$$\hat{s}_n[n] = \sum_{j=1}^{J} s_n^j[n - jL_n], \tag{A.33}$$

where L_n is the number of samples in a stochastic frame and J is the total number of stochastic frames in the speech signal.

In the filtering process, the memory of the IIR filter is passed from one frame to the next in order to maintain output signal continuity, as the only things that change from frame to frame are input noise variances and the filter coefficients.

As will be seen, the stochastic component parameters of HNM related to power and spectrum (LPC) are left unchanged in time and frequency modifications, and only its time marks are modified in the time modification.

A.3 Time modifications

The parameters obtained from the signal analysis are linked to specific times, the analysis instants t_a^k. Time modifications can be performed by modifying these time references. ρ^k defines the signal's time modification factor for a given frame k which can be different from frame to frame. As can be seen in figure A.7, this time modification factor (ρ^k) is defined as a multiplication factor that modifies the duration of each signal frame (N^k) by

$$N^{k\prime} = N^k \rho^k. \tag{A.34}$$

Thus, the synthesis time instants (t_s^k) can be directly computed using equation (A.34) recursively as

$$t_s^k = t_s^{k-1} + N^k \rho^k \tag{A.35}$$

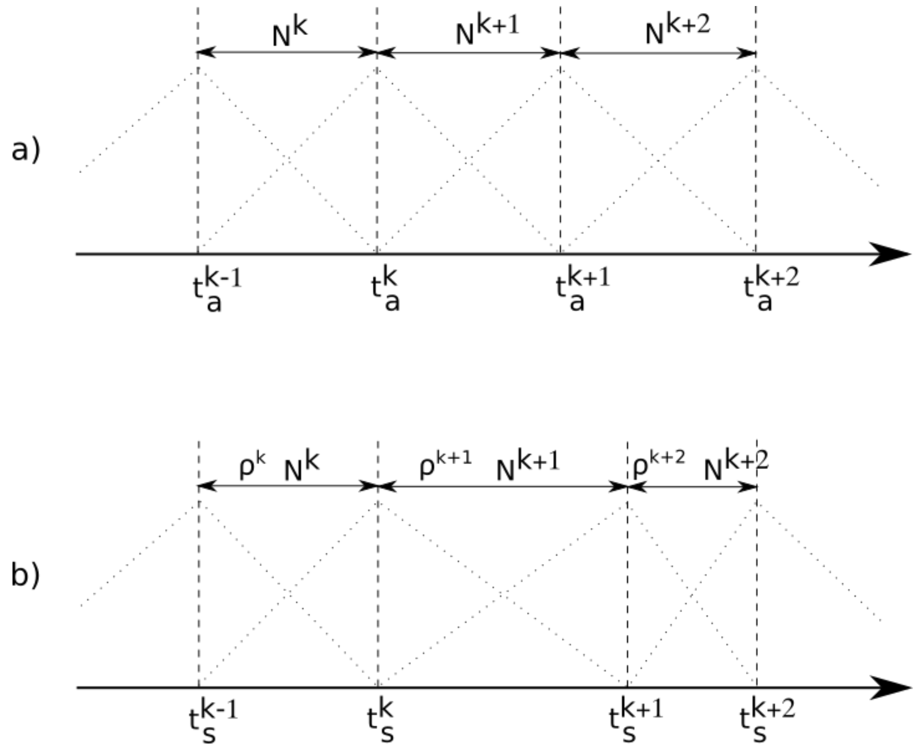


Figure A.7 — Time modifications. (a) original and (b) modified time instants and signal OLA. From [Erro, 2008]

Harmonic frequencies ($\boldsymbol{F}^k$) and amplitudes ($\boldsymbol{A}^k$) are not modified during the time modification procedure, so the original trajectory is preserved and mapped to the new time line drawn by new time marks ($\boldsymbol{t_s}$). Synthesis time marks $\boldsymbol{t}_s$ (defined as the array of the time marks of the modified frames for both harmonic and stochastic signal parts, t_s^k and t_s^j) are derived from the analysis time marks $\boldsymbol{t}_a$ and the applied time modification factors ρ^k with the equation (A.35). On the other hand, if the harmonic phases are kept unchanged for the new time instants, phase mismatches may arise at the frame boundaries. These mismatches can produce variations in the fundamental frequency due to the strong relation between phase and frequency. In figure A.8 we can see how the frequency may vary as a consequence of phase mismatch when a single stationary sinusoid is processed by a change of tempo. For more detailed explanation, refer to [Calzada, 2008].

So phase correction must be done. [Erro, 2008] proposes a simple method for correcting the phase mismatches. He assumes that the estimated phases φ_l^k can be decomposed in two terms: a **linear-in-frequency phase** term $l\ \alpha^k$, which varies from one sample to the next according to the fundamental

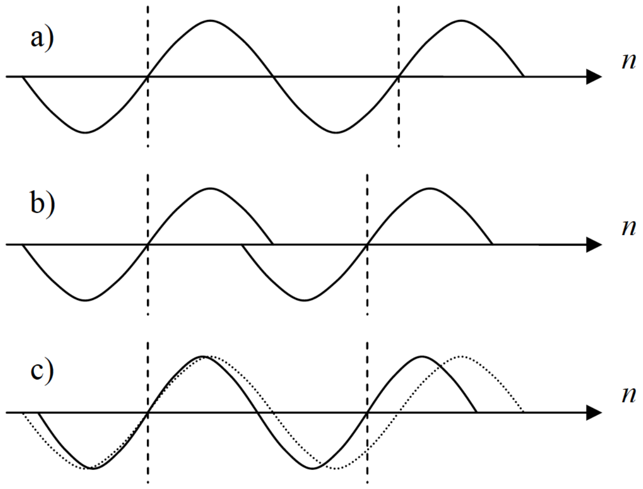


Figure A.8 — (a) Single tone analysed at two pitch instants (b) Frame centres are moved without correcting the phases. (c) After OLA the frequency of the signal has changed with respect to the original signal (dotted line). From [Erro, 2008]

frequency, and the **vocal tract phase** contribution θ_l^k at the frequency of the harmonic, as

$$\varphi_l^k = l\ \alpha^k + \theta_l^k. \tag{A.36}$$

As can be seen, the linear phase term is linearly dependent on the harmonic index l. This is because this term is produced by the measurement of instantaneous phases at the analysis instants (t_a^k), which can be measured as the integral of instantaneous frequencies when performing time modifications. The linear-in-frequency phase term is closely related to the phase due to the voiced excitation signal (glottal pulse) whereas the vocal tract phase represents the phase variations mainly produced by articulatory movement in the oral cavity.

For a given harmonic frame k, the vocal tract is defined by the array of amplitudes of the harmonics ($\boldsymbol{A}^k$) and the array of vocal tract phases ($\boldsymbol{\Theta}^k$). Because it is not desirable to change the speaker identity, only the speech rate, the vocal tract must remain unchanged. For this reason, a linear phase correction is applied to the original estimated phases in order to maintain a global coherence between the original estimated parameters relative to the new time locations provided by the time modification. The linear phase increase from frame $k-1$ to frame k can be estimated as

$$\begin{aligned} \alpha^k - \alpha^{k-1} &\cong \int_0^{N^k} \left[\omega_0^{k-1} + \frac{n}{N^k}\left(\omega_0^k - \omega_0^{k-1}\right)\right] dn = \\ &= \frac{1}{2}\left(\omega_0^{(0)} + \omega_0^{k-1}\right) N^k = \Psi\left(\omega_0^k, \omega_0^{k-1}, N^k\right), \end{aligned} \tag{A.37}$$

where Ψ is the name to designate the function that will give the α increase between two adjacent frames. As can be seen in (A.37), the fundamental frequency is supposed to have linear variations between consecutive frames. When applying a time modification, the analysis time instant differences ($N^k = t_a^k - t_a^{k-1}$) are multiplied by a certain factor ρ^k. The α increase in this case would be

$$\alpha'^k - \alpha'^{k-1} \cong \Psi\left(\omega_0^k, \omega_0^{k-1}, \rho^k N^k\right) \tag{A.38}$$

Thus the phase correction that will maintain this global coherence in the synthesized signal can be expressed as:

$$\begin{aligned} &\Delta\alpha^k = \Psi\left(\omega_0^k, \omega_0^{k-1}, \rho^k N^k\right) - \Psi\left(\omega_0^k, \omega_0^{k-1}, N^k\right) = \frac{1}{2}\left(\omega_0^k + \omega_0^{k-1}\right) N^k(\rho^k - 1) \\ &\varphi_l'^k = \varphi_l^k + l \sum_{q=2}^{k} \Delta\alpha^{(q)} \qquad\qquad \forall k > 1, l = 1 \cdots, L^k. \end{aligned} \tag{A.39}$$

So the final frame reconstruction of both harmonic and noise parts after the time modification would be

$$\hat{s}_h^k[n] = \sum_{l=1}^{L^k} a_l^k \cos\left(l\, \omega_0^k n + \varphi_l'^k\right) \qquad t_s^{k-1} - t_s^k \le n \le t_s^{k+1} - t_s^k \tag{A.40}$$

$$\hat{s}_n^j[n] = p^j \sigma^j[n] * \sqrt{\boldsymbol{H_{LPC}^j}} \qquad 0 \le n \le t_s^j - t_s^{j-1}, \tag{A.41}$$

where each component is generated using its own synthesis time marks, obtained from the transformation time map.

A.4 Pitch modifications

Pitch modification aims to modify the pitch of the speaker without modifying the vocal tract envelope or the time scale. Omitting the influence of the glottal source, the estimated HNM parameters within a signal frame k can be interpreted as a sampled version of the vocal tract at the harmonic frequencies ($\boldsymbol{F}^k$), being the amplitudes ($\boldsymbol{A}^k$) of the samples filter's magnitude envelope and the phase envelope of their vocal tract phases ($\boldsymbol{\Theta}^k$). Modifying the pitch means changing the harmonic frequencies and thus the samples positions of this vocal tract. If the influence of the source on the measured spectrum is also considered, in this case both source and vocal tract spectral variations are supposed to be maintained during pitch modifications. However, in this work we will talk only in terms of vocal tract to simplify the discourse. In order to keep the original speaker's vocal tract, amplitudes and phases must be recalculated by resampling the vocal tract at the new sample points. As mentioned in section A.3, the vocal tract can be represented by complex amplitudes (B_l^k), where $B_l^k = a_l^k e^{j\theta_l^k}$. So the new amplitude and phase values are found by interpolating the vocal tract from the complex amplitude trajectories as depicted in figure A.9 along the frequency axis.

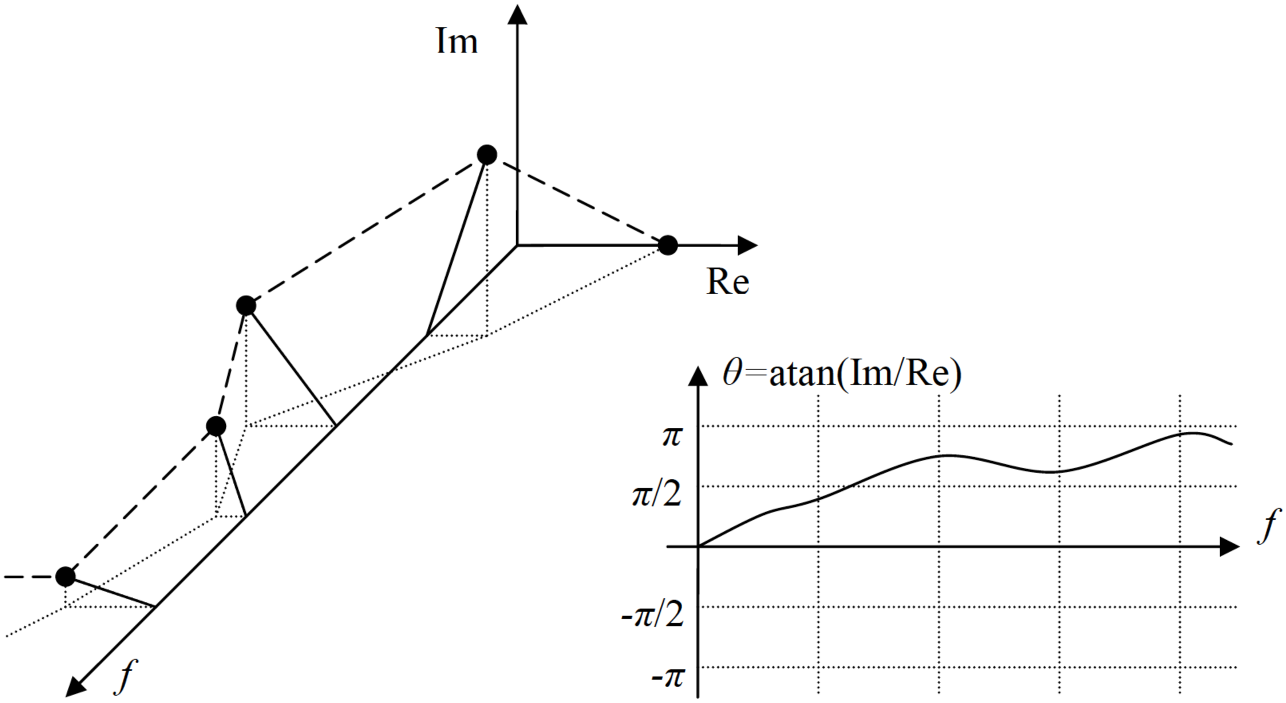


Figure A.9 — Vocal tract phase envelope obtained by linear interpolation of the complex amplitudes. From [Erro, 2008]

But to be able to interpolate the complex amplitudes we first have to find the vocal tract's phase from the HNM parameters.

Vocal tract phase extraction

In the previous section, we found that the phase φ estimated by HNM analysis could be decomposed into two parts, described in equation (A.36). The same procedure used in time scale modifications for linear-in-frequency correction (A.39) can be applied here in order to obtain the linear phase evolution when signal modifications are performed. Thus, extracting this linear phase term from the original estimated phases obtained from HNM analysis,

$$\begin{aligned} \varphi_l^k &= l\alpha^k + \theta_l^k \\ \tilde{\alpha}^k &= \alpha^{k-1} + \Psi\left(\omega_0^j, \omega_0^{j-1}, N_j\right) = \sum_{j=1}^{k} \Psi\left(\omega_0^j, \omega_0^{j-1}, N_j\right) \\ \tilde{\theta}_l^k &= \varphi_l^k - l\tilde{\alpha}^k = l\alpha^k + \theta_l^k - l\tilde{\alpha}^k \end{aligned} \tag{A.42}$$

we can obtain the vocal tract phase, as can be seen in (A.42), where $\tilde{\alpha}^k$ is the estimation of the linear phase in frame k and N_j is the frame length.

Pitch modification

Once we have the vocal tract phases θ_l^k we can linearly interpolate the vocal tract at the new frequency points to obtain the new amplitudes (a_l^k) and phases (θ_l^k) (see the example in figure A.9) by using the complex vocal tract amplitudes $B_l^k = a_l^k \; e^{j\,\theta_l^k}$. The interpolation process can be applied by interpolating the real and imaginary parts of the complex amplitudes B^k, and the obtained complex amplitudes are expressed in polar notation obtaining the new amplitudes a_k' and the new vocal tract phases θ_k'. But changing the pitch, and thus the fundamental period, for a given frame can lead to phase mismatches in the frame boundaries causing undesirable pitch modification as depicted in figure A.10, where a single stationary sinusoid is decreased in pitch by a factor of $\lambda < 1$.

Then, considering that pitch frequency for a given frame k is modified by a multiplication factor λ^k, the linear phase correction is

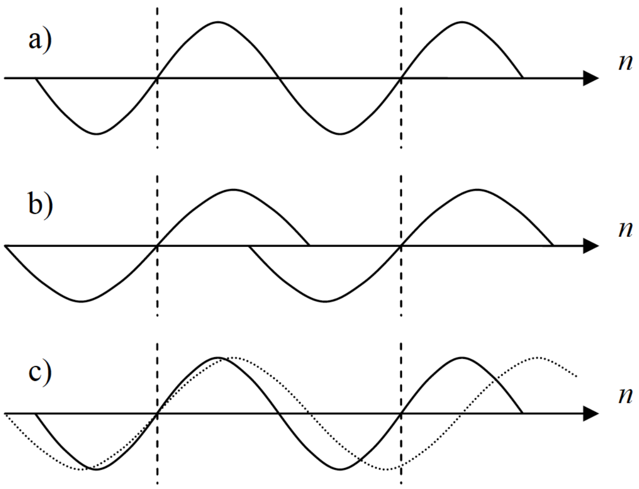


Figure A.10 — (a) Single tone analysed at two instants (b) Pitch is decreased at each frame without correcting the phases. (c) After OLA the pitch is different from the desired one(dotted line). From [Erro, 2008]

$$\begin{aligned}\Delta\alpha^k &= \Psi\left(\lambda^k\omega_0^k, \lambda^{k-1}\omega_0^{k-1}, N^k\right) - \Psi\left(\omega_0^k, \omega_0^{k-1}, N^k\right) = \\ &= \frac{1}{2}\left[\omega_0^k\left(\lambda^k - 1\right) + \omega_0^{k-1}\left(\lambda^{k-1} - 1\right)\right] N^k \\ \alpha'^k &= \alpha^k + \sum_{q=2}^{k} \Delta\alpha^{(q)} \qquad \forall k > 1, l = 1, \cdots, L^k.\end{aligned} \tag{A.43}$$

Thus the total new phases are

$$\varphi_l'^k = l\alpha'^k + \theta_l'^k. \tag{A.44}$$

The modified harmonic component of the resynthesized signal can be reconstructed using equation (A.27), where the frames to be overlapped are

$$\hat{s}_h^k[n] = \sum_{l=1}^{L^k} a_l'^k \cos\left(l\lambda^k\omega_0^k n + \varphi_l'^k\right) \qquad t_s^{k-1} - t_s^k \leq n \leq t_s^{k+1} - t_s^k. \tag{A.45}$$

A.5 Window characteristics

[Depalle and Hélie, 1997] designed a new window with no sidelobes in order to improve the precision of the algorithm for tracking the frequencies in the speech signal. The proposed window consists of a Gaussian function, which presents no sidelobes but is not time limited, multiplied by the power of a triangular window, whose Fourier Transform (FT) is always real and positive. When the variance of the Gaussian window is large enough, the smoothing in the spectral domain is capable of removing the sidelobes of the triangular window [Depalle and Hélie, 1997].

Thus the proposed window is

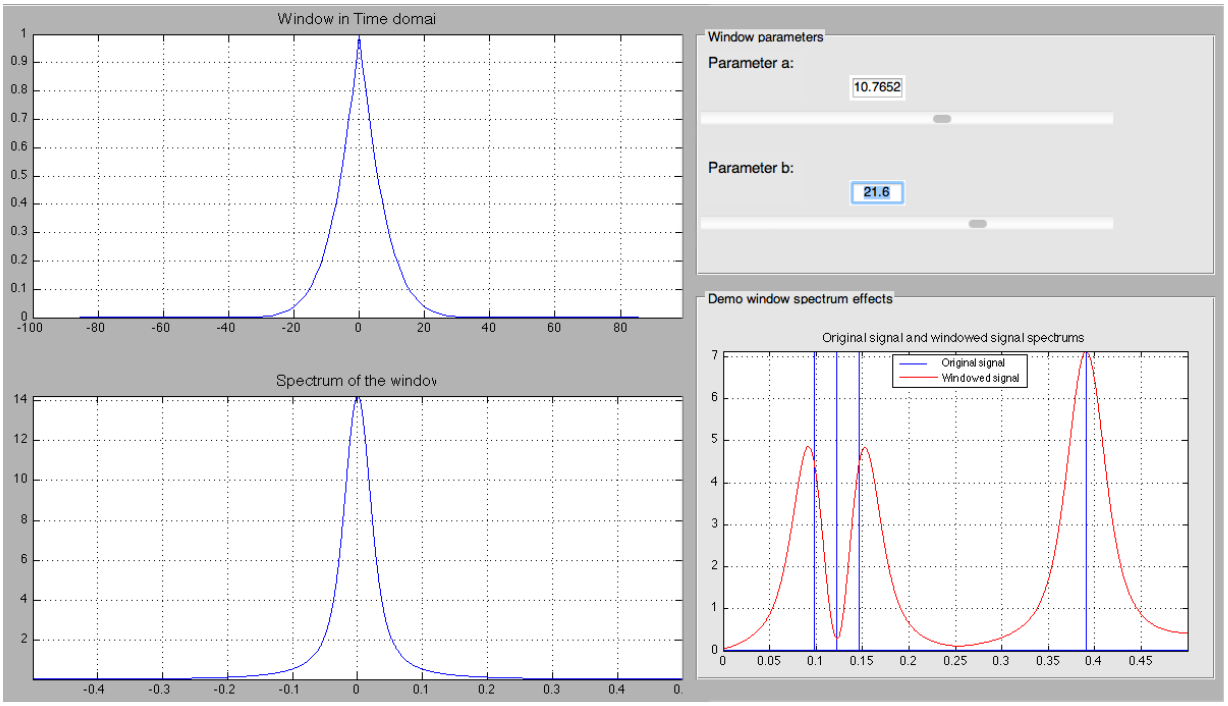
$$w[n] = \left(1 - \frac{|n|}{\frac{N}{2}}\right)^a \times\ e^{-\frac{1}{2}\left[b\left(\frac{n}{\frac{N}{2}}\right)\right]^2} \qquad -\frac{N}{2} \leq n \leq \frac{N}{2} - 1, \tag{A.46}$$

where N is the length of the window in samples. The analysis window length was set to $2T_0$, T_0 being the absolute difference between the pitch mark where the analysis window is centred and the following one (see figure A.2).

According to [Harris, 1978], two factors must be considered regarding the window function when windowing harmonic signals: the main lobe bandwidth and the side lobe amplitudes in the window's spectrum. The main lobe of the window might mask harmonics for low pitch frequencies, or when the harmonics are close. Figure A.11 shows an interactive GUI developed in Matlab that permits control of the window parameters (a and b) and observing the generated window in the time and frequency domains. The bottom right plot shows in blue the spectrum of a synthetic signal formed by 4 harmonics and in red, the spectrum of the signal windowed with the analysis window with parameters $a = 10.7652$ and $b = 21.6$. As can be seen, the second harmonic (at the normalized frequency 0.125) is almost lost due to the spectral energy contribution of the spectrum of the analysis window centred at the surrounding harmonics (0.1 and 0.15). In order to avoid this effect it is important to select an analysis window with a narrow-bandwidth main lobe.

In order to obtain a window with no sidelobes, multiple combinations of values for the window parameters were tested. Figure A.12 shows the result for multiple combinations, where blue indicates a,b value pairs with a window spectrum with no sidelobes, and red, are the value pair combinations producing sidelobes.

The following criteria for selecting the a,b values is the bandwidth of the main lobe. Figure A.13 shows the bandwidth computed at the middle of the main lobe (-6dB with respect to its maximum) of the analysis windows for multiple combinations of a,b values. As can be seen, the best combinations



Window parameters

Parameter a: 10.7652

Parameter b: 21.6

Figure A.11 — Effects of a wide main lobe in the window spectrum. Window parameters ($a = 10.7652$,$b = 21.6$). (Top left) window in time domain,(Bottom left) window spectrum, (Top right) parameter value selection and (Bottom right) ideal signal spectrum and windowed signal spectrum superimposed. From [Calzada, 2008]

(without sidelobes and with lower main lobe bandwidth) of parameters are those having $b \leq 0$. Figure (b) shows more detail of the area corresponding to values $a \in [2, 2.04]$ and $b \in [-0.9, 0]$. As can be seen, there appears to be range of values where the main lobe bandwidth is constant (eg., for $b \in [-0.5, -0.1]$).

Based on the information shown in figure A.12, showing the space of (a, b) pair values generating windows with no sidelobes, and figure A.13, showing the main lobe bandwidth for different (a, b) pair values, the values ($a = 2.02$, $b = -0.6$) were selected for generating the window with equation (A.46). Figure A.14 shows the window and its spectrum for a few combinations of parameter values. From top to bottom, the first is the window with the parameter values used for the experiments conducted in this dissertation ($a = 2.02$, $b = -0.6$), next the triangular window ($a = 0, b = 1$), next the Gaussian window ($a = 1, b = 0$), and the last two correspond to two combinations presented in [Depalle and Hélie, 1997].

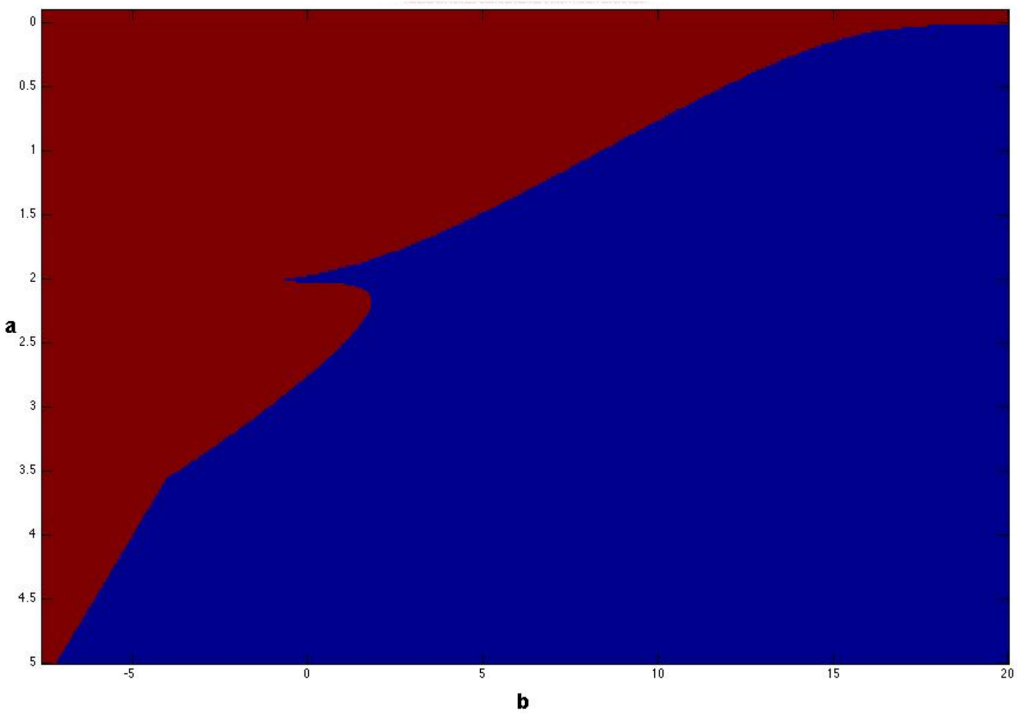


Figure A.12 — Verification of sidelobe presence in the analysis window for multiple combinations of a,b pairs. In red, the parameter space where the window has sidelobes, while the no sidelobe space is shown in blue. $a \in [0\ ,\ 5]$ and $b \in [-6\ ,\ 20]$. From [Calzada, 2008]

A.6 Linear prediction analysis with HNM

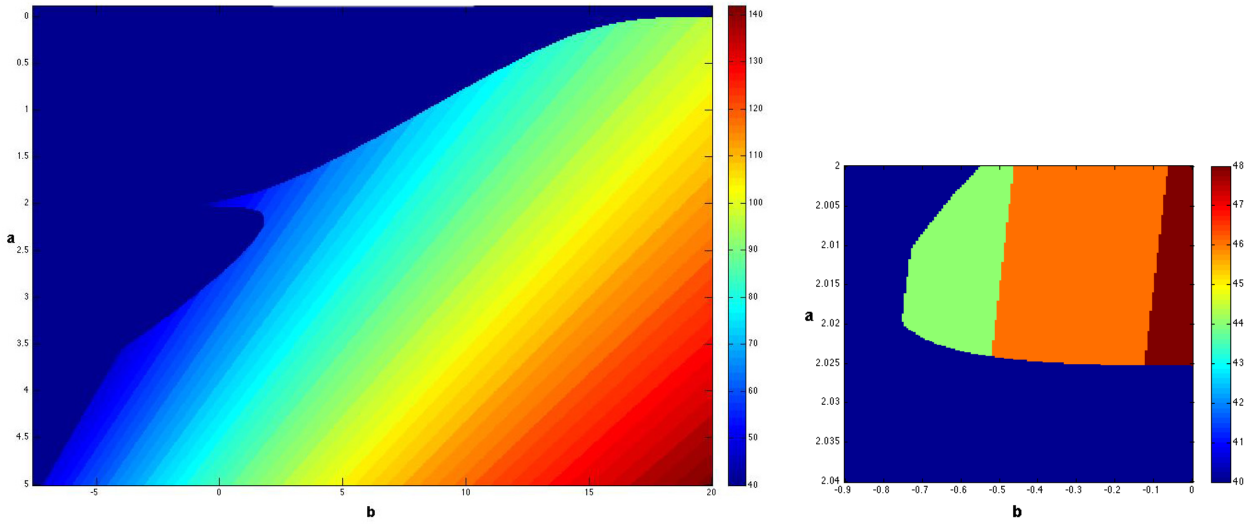
In chapter 4, a methodology for transferring VE based on the APLP (see section 2.4) technique is presented. The proposed methodology requires the computation of low order LPC filter of the signal spectrum based on the HNM parameters. This section explains the procedure for conducting this computation, presented in [Erro, 2008].

A particularised frequency domain implementation of the LPC technique presented in [Makhoul, 1975] can be applied to obtain the all-pole representation of a set of harmonics from the HNM parameters.

Considering the all-pole filter as

$$\frac{1}{A(z)} = \frac{1}{1 + a_1 z^{-1} + \cdots + a_p z^{-p}}, \tag{A.47}$$

the LPC coefficients can be computed by solving the equation system



(a) General tested space

(b) Zoom

Figure A.13 — Verification of main lobe bandwidth for multiple combinations of a,b pairs. $a \in [0,\ 5]$ and $b \in [-6,\ 20]$. (a) Shows the overall space for a,b combinations. (b) Shows a zoom performed at the area surrounding $a = 2$,$b = 0$. From [Calzada, 2008]

$$\begin{bmatrix} R_0 & R_1 & \cdots & R_{p-1} \\ R_1 & R_0 & \cdots & R_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ R_{p-1} & R_{p-2} & \cdots & R_0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} -R_1 \\ -R_2 \\ \vdots \\ -R_p \end{bmatrix}, \tag{A.48}$$

where R_n, for $n = 1, \cdots, p$, corresponds to the first $p + 1$ values of the autocorrelation sequence of the speech signal. The system matrix is a Toeplitz matrix, also known as the Yule-Walker AR equations, and it can be be efficiently inverted using the Levinson-Durbin recursion [Levinson, 1949, Durbin, 1960]. The values of R_n can be computed using the inverse Fourier Transform of the power spectrum as

$$R_n^k \propto \int_{-\pi}^{\pi} |S^k(\omega)|^2 \, e^{j\,\omega\,n} d\omega = \sum_l \frac{1}{4} {A_l^k}^2 \left(e^{j\,\omega_l^k\,n} + e^{-j\,\omega_l^k\,n} \right) = \frac{1}{2} \sum_l {A_l^k}^2 \cos(\omega_l^k\,n), \tag{A.49}$$

where R_n^k is the autocorrelation values from the equation system (A.48) for the frame k, A_l^k and ω_l^k are the HNM amplitudes and frequencies, respectively for frame k.

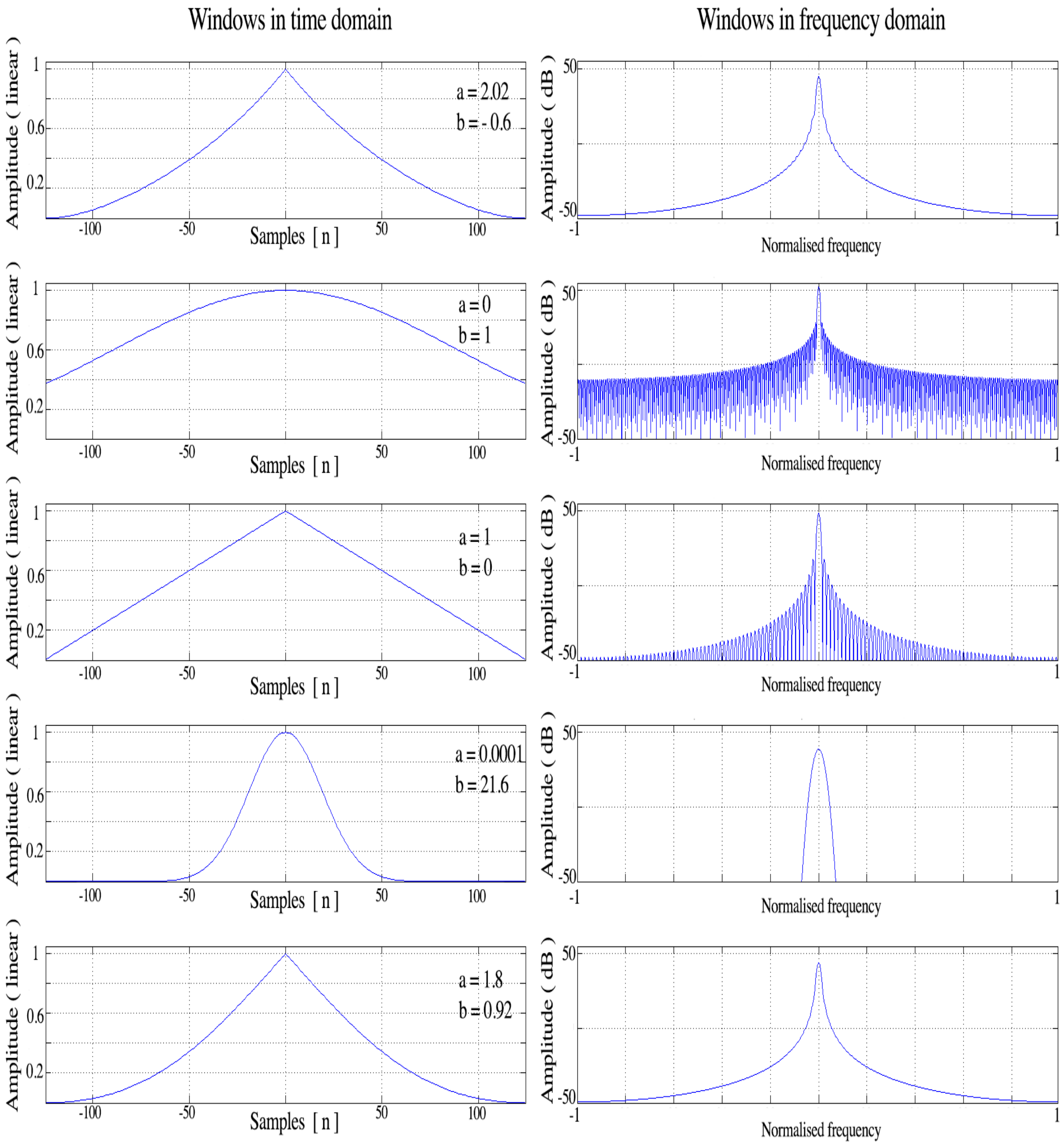


Figure A.14 — Window used for signal analysis. (Left) Windows in time domain. (Right) Window spectrum. Parameters from top to bottom: ($a = 2.02$,$b = -0.6$), ($a = 0$,$b = 1$), ($a = 1$,$b = 0$), ($a = 10^{-4}$,$b = 21.6$), ($a = 1.8$,$b = 0.92$)

Bibliography

[Aamodt and Plaza, 1994] Aamodt, A. and Plaza, E. (1994). Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Commun.*, 7(1):39–59.

[Abdel-Hamid et al., 2006] Abdel-Hamid, O., Abdou, S., and Rashwan, M. (2006). Improving Arabic HMM based speech synthesis quality. In *Proc. Interspeech*, pages 1332–1335.

[Abe et al., 1988] Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1988). Voice conversion through vector quantization. In *Int. Conf. on Acous., Speech, and Signal Proc. ICASSP88*, volume 1, pages 655–658.

[Adell, 2009] Adell, J. (2009). *Prosodic analysis and modelling of conversational elements for speech synthesis*. PhD thesis, Universitat Politecnica de Catalunya (UPC).

[Agüero et al., 2006] Agüero, P. D., Adell, J., and Bonafonte, A. (2006). Prosody generation for speech-to-speech translation. In *Proc. of ICASSP*, pages 557–560.

[Alías, 2008] Alías, F. (2008). Storyteller 2.0. Available at: `http://lasallerd.salleurl.edu/CreaVeu` Last accessed: December 2015.

[Alías, 2011] Alías, F. (2011). Reuniones virtuales de nueva generación con telepresencia. Available at: `http://lasallerd.salleurl.edu/CreaVeu` Last accessed: December 2015.

[Alías and Iriondo, 2002] Alías, F. and Iriondo, I. (2002). La evolución de la síntesis del habla en ingenería la salle. In *II Jornadas en Tecnología del Habla*, Granada, Spain.

[Almeida and Tribolet, 1982] Almeida, L. and Tribolet, J. (1982). Harmonic coding: A low bit-rate, good-quality speech coding technique. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP*, volume 7, pages 1664–1667.

[Alías et al., 2005] Alías, F., Iriondo, I., Formiga, L., Gonzalvo, X., Monzo, C., and Sevillano, X. (2005). High quality Spanish restricted-domain TTS oriented to a weather forecast application. In

9th European Conference on Speech Communication and Technology (Interspeech), pages 2573–2576, (Lisbon, Portugal).

[Alías et al., 2006] Alías, F., Llora, X., Formiga, L., Sastry, K., and Goldberg, D. (2006). Efficient interactive weight tuning for tts synthesis: Reducing user fatigue by improving user consistency. In *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*, volume 1, pages I –I.

[Alías et al., 2008] Alías, F., Sevillano, X., Socoro, J., and Gonzalvo, X. (2008). Towards high-quality next-generation text-to-speech synthesis: A multidomain approach by automatic domain classification. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(7):1340–1354.

[Anberbir and Takara, 2009] Anberbir, T. and Takara, T. (2009). Development of an Amharic text-to-speech system using cepstral method. In *Proc. of the First Workshop on Language Technologies for African Languages*, AfLaT, pages 46–52, (Stroudsburg, PA, USA). Association for Computational Linguistics.

[Arslan, 1999] Arslan, L. M. (1999). Speaker transformation algorithm using segmental codebooks (STASC). *Speech Communication*, 28(3):211 – 226.

[Askenfelt and Hammarberg, 1986] Askenfelt, A. G. and Hammarberg, B. (1986). Speech waveform perturbation analysis a perceptual-acoustical comparison of seven measures. *Journal of Speech, Language, and Hearing Research*, 29(1):50–64.

[Aylett and Pidcock, 2007] Aylett, M. P. and Pidcock, C. J. (2007). The CereVoice characterful speech synthesiser SDK. In *Artificial and ambient intelligence*, pages 174–178.

[Aylett and Yamagishi, 2008] Aylett, M. P. and Yamagishi, J. (2008). Combining statistical parameteric speech synthesis and unit-selection for automatic voice cloning. In *Proc. LangTech.*, Brisbane, Australia.

[Bailly et al., 2003] Bailly, G., Campbell, N., and Mobius, B. (2003). ISCA special session: Hot topics in speech synthesis. In *In EUROSPEECH*, pages 37–40, Geneva, Switzerland.

[Bandoin and Stylianou, 1996] Bandoin, G. and Stylianou, Y. (1996). On the transformation of the speech spectrum for voice conversion. In *Proc. of 4th Int. Conf. on Spoken Language (ICSLP)*, volume 3, pages 1405–1408.

[Banos et al., 2008] Banos, E., Erro, D., Bonafonte, A., and Moreno, A. (2008). Flexible harmonic/stochastic modelling for HMM-based speech synthesis. In *Actas de las V Jornadas en Tecnologías del Habla, VJTH.*

[Barra-Chicote et al., 2010] Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., and Macias-Guarasa, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication*, 52(5):394–404.

[Beller et al., 2008] Beller, G., Obin, N., and Rodet, X. (2008). Articulation degree as a prosodic dimension of expressive speech. In *4th Int. Conf. on Speech Prosody.*

[Beskow, 2015] Beskow, J. (accessed on December 2015). KTH vowel formant synthesizer. Accessible at: `http://www.speech.kth.se/wavesurfer/formant/` Last accessed: January 2016.

[Black and Lenzo, 2000] Black, A. W. and Lenzo, K. A. (2000). Limited domain synthesis. In *Proc. ICSLP*, pages 411–414.

[Black et al., 2007] Black, A. W., Zen, H., and Tokuda, K. (2007). Statistical parametric speech synthesis. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, volume 4, pages IV–1229. IEEE.

[Bonafonte et al., 2008] Bonafonte, A., Adell, J., Esquerra, I., Gallego, S., Moreno, A., and Pérez, J. (2008). Corpus and voices for Catalan speech synthesis. In *The 6th Int. Language Resources and Evaluation Conf. (LREC2008).*

[Boone et al., 2013] Boone, D. R., McFarlane, S. C., Von Berg, S. L., and Zraick, R. I. (2013). *The Voice and Voice Therapy.* Pearson, 9th edition.

[Bulut et al., 2002] Bulut, M., Narayanan, S. S., and Syrdal, A. K. (2002). Expressive speech synthesis using a concatenative synthesizer. In *InterSpeech*, pages 1265–1268, (Vencer, CO.).

[Bulyko et al., 2002] Bulyko, I., Ostendorf, M., and Bilmes, J. (2002). Robust splicing costs and efficient search with BMM models for concatenative speech synthesis. In *ICASSP*, volume 1, pages 461–464. IEEE.

[Burkhardt, 2005] Burkhardt, F. (2005). Emofilt: the simulation of emotional speech by prosody-transformation. In *Proc. of Interspeech*, pages 509–512, (Lisbon).

[Burkhardt, 2009] Burkhardt, F. (2009). Rule-based voice quality variation with formant synthesis. In *Proc. Interspeech.*, pages 2659–2662, (Bristol).

[Burkhardt, 2015b] Burkhardt, F. (last web access December 2015b). EmoSyn (emotion-synthesizer). Available at: `http://emosyn.syntheticspeech.de/` Last accessed: December 2015.

[Burkhardt, 2015a] Burkhardt, F. (last web access on December 2015a). Emofilt synthesizer. Available at: `http://emofilt.syntheticspeech.de/`; Last accessed: December 2015.

[Burkhardt and Sendlmeier, 2000] Burkhardt, F. and Sendlmeier, W. F. (2000). Verification of acoustical correlates of emotional speech using formant synthesis. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 151–156, (Northern Ireland).

[Cabral et al., 2014] Cabral, J., Richmond, K., Yamagishi, J., and Renals, S. (2014). Glottal spectral separation for speech synthesis. *Selected Topics in Signal Processing, IEEE Journal of*, 8(2):195–208.

[Cahn, 1990] Cahn, J. (1990). The generation of affect in synthesized speech. *Journal of the American Voice I/O Society*, page 1–19.

[Calzada, 2008] Calzada, A. (2008). *Estudi d'esquemes de modificació de les característiques vocals*. PhD thesis, Universitat Ramon Llull.

[Calzada et al., 2013] Calzada, A., Socoró, J. C., and Clark, R. (2013). Parametric model for vocal effort interpolation with Harmonics Plus Noise Models. In *8th ISCA Workshop on Speech Synthesis*, pages 45–50, (Barcelona, Spain).

[Calzada and Socoró, 2011] Calzada, A. and Socoró, J. C. (2011). Vocal effort modification through harmonics plus noise model representation. In Travieso-Gonzalez, C. M. and Alonso-Hernandez, J., editors, *Advances in Nonlinear Speech Processing*, volume 7015 of *Lecture Notes in Computer Science*, pages 96–103. Springer, (Berlin Heidelberg).

[Calzada and Socoró, 2013] Calzada, A. and Socoró, J. C. (2013). Voice quality modification using a harmonics plus noise model. *Cognitive Computation*, 5(4):473–482.

[Campbell, 2004] Campbell, N. (2004). Specifying affect and emotion for expressive speech synthesis. In *Computational Linguistics and Intelligent Text Processing*, volume 2945 of *Lecture Notes in Computer Science*, pages 395–406. Springer Berlin Heidelberg.

[Campbell, 2005] Campbell, N. (2005). Expressive speech synthesis: What is the goal?

[Campbell, 2007] Campbell, N. (2007). Approaches to conversational speech rhythm: Speech activity in two-person telephone dialoges. In *16th Int. Congress of Phonetic Sciences*, pages 343–348.

[Campbell and Mokhtari, 2003] Campbell, N. and Mokhtari, P. (2003). Voice quality: the 4th prosodic dimension. *15th ICPhS*, pages 2417–2420.

[Camps et al., 1992] Camps, J., Bailly, G., and Martí, J. (1992). Synthèse à partir du texte pour le Catalan. In *19èmes Journées d'Études sur la Parole.*

[Carlson, 2002] Carlson, R. (2002). Data-driven formant synthesis. *Proceedings of Fonetik, TMH-QPSR*, 44(1):121–124.

[Chen et al., 2014] Chen, L.-H., Ling, Z.-H., Zu, Y.-Q., Yan, R.-Q., Jiang, Y., Xia, X.-J., and Wang, Y. (2014). The USTC System for Blizzard Challenge 2014. In *Blizzard Challenge 2014*.

[Cornelius, 2000] Cornelius, R. R. (2000). Theoretical approaches to emotion. In *Proceedings of the ISCA Workshop on Speech and Emotion*, pages 3–10, (Northern Ireland).

[Cowie, 2000] Cowie, R. (2000). Describing the emotional states expressed in speech. In *SpeechEmotion-2000*, pages 11–18, (Northern Ireland, UK).

[Creaveu, 2013] Creaveu (2013). CreaVeu: From any text to any voice. Available at: `http://lasallerd.salleurl.edu/CreaVeu` Last accessed: December 2015.

[Cushing, 2010] Cushing, I. R. (2010). *Vocal effort levels and underlying acoustic phonetic characteristics*. PhD thesis, Acoustics Research Centre, School of Computing, Science and Engineering, University of Salford, UK.

[Cushing et al., 2011] Cushing, I. R., Li, F. F., Cox, T. J., Worrall, K., and Jackson, T. (2011). Vocal effort levels in anechoic conditions. *Applied Acoustics*, 72(9):695 – 701.

[d'Alessandro, 2006] d'Alessandro, C. (2006). Voice source parameters and prosodic analysis. In Sudhoff, S., Lenertová, D., Meyer, R., Pappert, S., Augursky, P., Mleinek, I., Richter, N., and Schliesser, J., editors, *Methods in Empirical Prosody Research*, volume 3 of *Language, Context, and Cognition*. Walter de Gruyter.

[d'Alessandro and Sturmel, 2011] d'Alessandro, C. and Sturmel, N. (2011). Glottal closure instant and voice source analysis using time-scale lines of maximum amplitude. *Sadhana*, 36(5):601–622.

[Danieli, 2006] Danieli, M. (2006). Current topics in speech synthesis.

[Degottex et al., 2013] Degottex, G., Lanchantin, P., Roebel, A., and Rodet, X. (2013). Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Commun.*, 55(2):278–294.

[Degottex et al., 2011] Degottex, G., Roebel, A., and Rodet, X. (2011). Phase minimization for glottal model estimation. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1080–1090.

[Degottex and Stylianou, 2012] Degottex, G. and Stylianou, Y. (2012). A Full-Band Adaptive Harmonic Representation of Speech. In *Proc. Interspeech*, (Portland, USA). International Speech Communication Association (ISCA).

[Degottex and Stylianou, 2013] Degottex, G. and Stylianou, Y. (2013). Analysis and synthesis of speech using an adaptive full-band harmonic model. *Audio, Speech, and Language Processing, IEEE Transactions on*, 21(10):2085–2095.

[Depalle and Hélie, 1997] Depalle, P. and Hélie, T. (1997). Extraction of spectral peak parameters using s STFT modeling and no-sidelobe windows. In *Proc. of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*

[DFKI, 2015] DFKI, u. (2015). MARY TTS. Available at: `http://mary.dfki.de/download/index.html` Last accessed: December 2015.

[Drioli et al., 2003] Drioli, C., Tisato, G., Cosi, P., and Tesser, F. (2003). Emotions and voice quality: Experiments with sinusoidal modeling. In *Proc. of VOQUAL*, pages 127–132, (Geneva, Switzerland).

[Dunn and Quatieri, 2007] Dunn, R. and Quatieri, T. (2007). Sinewave analysis/synthesis based on the fan-chirp tranform. In *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.*, pages 247–250.

[Durbin, 1960] Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 3(28):233–244.

[d'Alessandro et al., 2014] d'Alessandro, N., Tilmanne, J., Astrinaki, M., Hueber, T., Dall, R., Ravet, T., Moinet, A., Cakmak, H., Babacan, O., and Barbulescu, A. (2014). Reactive statistical mapping: Towards the sketching of performative control with data. In *Innovative and Creative Developments in Multimodal Interaction Systems*, volume 425 of *IFIP Advances in Information and Communication Technology*, pages 20–49. Springer Berlin Heidelberg.

[Ebert, 2011] Ebert, R. (2011). TED talk: Remaking my voice. Available at: `https://www.ted.com/talks/roger_ebert_remaking_my_voice` Last accessed: January 2016.

[Eide, 2002] Eide, E. (2002). Preservation, identification, and use of emotion in a text-to-speech system. In *Proc. of IEEE Workshop on Speech Synthesis*, pages 127–130.

[Eide et al., 2004] Eide, E., Aaron, A., Bakis, R., Hamza, W., Picheny, M., and Pitrelli, J. (2004). A corpus-based approach to ¡ahem/¿ expressive speech synthesis. In *5th ISCA Workshop on Speech Synthesis*, pages 79–84.

[Erro, 2008] Erro, D. (2008). *Intra-lingual and cross-lingual voice conversion using Harmonic plus Stochastic Models.* PhD thesis, Universitat Politècnica de Catalunya.

[Erro et al., 2009] Erro, D., Navas, E., Hernáez, I., and Saratxaga, I. (2009). Emotion conversion based on prosodic unit selection. *Audio, Speech, and Language Processing, IEEE Transactions on*, 18(5):974–983.

[Erro et al., 2010] Erro, D., Sainz, I., Luengo, I., Odriozola, I., Sánchez, J., Saratxaga, I., Navas, E., and Hernáez, I. (2010). HMM-based speech synthesis in Basque language using HTS from AhoTTS to Aho-HTS. In *FALA 2010*, pages 67–70, (Vigo, Spain).

[Formiga and Alías, 2007] Formiga, L. and Alías, F. (2007). Extracting user preferences by GTM for aiGA weight tuning in unit selection text-to-speech synthesis. *Computational and Ambient Intelligence*, 4507/2007:654–661.

[Formiga et al., 2010] Formiga, L., Trilla, A., Alías, F., Iriondo, I., and Socoró, J. (2010). Adaptation of the URL-TTS system to the 2010 Albayzin evaluation campaign. In *Jornadas en Tecnología del Habla and Iberian SLTech Workshop*, pages 363–370.

[Fraiha Machado et al., 2013] Fraiha Machado, A., Bonafonte, A., and Queiroz, M. (2013). Parametric decomposition of the spectral envelope. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 571–574.

[Garnier et al., 2006] Garnier, M., Bailly, L., Dohen, M., Welby, P., and Loevenbruck, H. (2006). An acoustic and articulatory study of Lombard speech: global effects on the utterance. In *9th Int. Conf. on Spoken Language Processing (ICSLP)*, pages 2246–2249, (Pittsburgh, PA, USA).

[Garnier and Henrich, 2014] Garnier, M. and Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise? *Computer Speech and Language*, 28(2):580–597.

[Garnier et al., 2008] Garnier, M., Wolfe, J., Henrich, N., and Smith, J. (2008). Interrelationship between vocal effort and vocal tract acoustics: a pilot study. In *Proceedings of the International Conference on Speech and Language Processing*, (Brisbane, Australia).

[Gauvain and Lee, 1994] Gauvain, J. and Lee, C. (1994). Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. In *IEEE Transactions on Speech and Audio Processing*, volume 2, pages 291–298.

[George and Smith, 1992] George, E. B. and Smith, M. J. T. (1992). Analysis-by-synthesis/overlap-add sinusoidal modeling applied to the analysis and synthesis of musical tones. *J. Audio Eng. Soc.*, 40(6):497–516.

[Godsill and Davy, 2002] Godsill, S. and Davy, M. (2002). Bayesian harmonic models for musical pitch estimation and analysis. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 2, pages 1769–1772.

[Golipour et al., 2013] Golipour, L., Conkie, A., and Syrdal, A. (2013). Prosodically modifying speech for unit selection speech synthesis databases. In *8th ISCA Workshop on Speech Synthesis*, pages 275–279, (Barcelona, Spain).

[Gonzalvo et al., 2006] Gonzalvo, X., Socoró, J. C., Iriondo, I., Monzo, C., and Martínez, E. (2006). Linguistic and mixed excitation improvements on a HMM-based speech synthesis for Castilian Spanish. In *Proc. ISCA Workshop on Speech Synthesis (SSW-6)*, page 362–367.

[Gonzalvo Fructuoso, 2010] Gonzalvo Fructuoso, J. (2010). *HMM-based speech synthesis applied to Spanish and English, its applications and a hybrid approach.* PhD thesis, Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle - Ramon Llull.

[Grichkovtsova et al., 2012] Grichkovtsova, I., Morel, M., and Lacheret, A. (2012). The role of voice quality and prosodic contour in affective speech perception. *Speech Communication*, 54(3):414–429.

[Gu and Liau, 2008] Gu, H.-Y. and Liau, H.-L. (2008). Mandarin singing voice synthesis using an hnm based scheme. In *Congr. on image and signal process.*, volume 5, pages 347–351.

[Guaus et al., 1996] Guaus, R., Gudayol, F., and Martí, J. (1996). Conversión texto-voz mediante síntesis psola. In *Jornadas Nacionales de Acústica*, pages 355–358, Barcelona.

[Guner and Demiroglu, 2012] Guner, E. and Demiroglu, C. (2012). A small footprint hybrid statistical/unit selection text-to-speech synthesis system for agglutinative languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4537–4540.

[Hamon et al., 1989] Hamon, C., Mouline, E., and Charpentier, F. (1989). A diphone synthesis system based on time-domain prosodic modifications of speech. *International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, 1:238 –241.

[Harris, 1978] Harris, F. (1978). On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83.

[Heuft et al., 1996] Heuft, B., Portele, T., and Rauth, M. (1996). Emotions in time domain synthesis. In *Proc. of the 4th Int. Conf. on Spoken Language ICSLP*, volume 3, pages 1974–1977.

[Hsia et al., 2007] Hsia, C., Wu, C., and Jian, Q. (2007). Conversion function clustering and selection using linguistic and spectral information for emotional voice conversion. *IEEE Trans. Comput.*, 56(9):1225–1254.

[Hu and Loizou, 2010] Hu, Y. and Loizou, P. C. (2010). On the importance of preserving the harmonics and neighboring partials prior to vocoder processing: Implications for cochlear implants. *Journal of the Acoustical Society of America*, 127(1):427–434.

[Huang et al., 2001] Huang, X., Acero, A., and Hon, H. (2001). *Spoken Language Processing*. Prentice Hall.

[Hunt and Black, 1996] Hunt, A. and Black, A. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP.*, volume 1, pages 373–376. IEEE.

[Inžinerija et al., 2007] Inžinerija, T., Paulikas, Š., and Karpavičius, R. (2007). Application of linear prediction coefficients interpolation in speech signal coding. *Elektronika ir Elektrotechnika*, 80(8):39–42.

[Iriondo, 2008] Iriondo, I. (2008). *Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva*. PhD thesis, Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle - Ramon Llull.

[Iriondo et al., 2002] Iriondo, I., Alías, F., and Melenchón, J. (2002). Un modelo híbrido orientado a la síntesis multimodal del habla. In *Procesamiento del Lenguaje Natural*, volume 29, pages 159–163, Valladolid.

[Iriondo et al., 2000] Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., and Longhi, L. (2000). Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 161–166, Northern Ireland.

[Iriondo et al., 2009] Iriondo, I., Planet, S., Socoró, J.-C., Martínez, E., Alías, F., and Monzo, C. (2009). Automatic refinement of an expressive speech corpus assembling subjective perception and automatic classification. *Speech Communication*, 51(9):744 – 758.

[Islam, 2000] Islam, T. (2000). *Interpolation of linear prediction coefficients for speech coding*. PhD thesis, McGill University, (Montreal, Canada).

[Iwahashi and Sagisaka, 1994] Iwahashi, N. and Sagisaka, Y. (1994). Speech spectrum transformation by speaker interpolation. In *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing ICASSP*, volume 1, pages 461–464.

[Jensen and Hansen, 2001] Jensen, J. and Hansen, J. (2001). Speech enhancement using a constrained iterative sinusoidal model. *Speech and Audio Processing, IEEE Transactions on*, 9(7):731–740.

[Jokinen et al., 2015] Jokinen, E., Remes, U., and Alku, P. (2015). Comparison of {G}aussian process regression and {G}aussian mixture models in spectral tilt modelling for intelligibility enhancement of telephone speech. (Dresden, Germany).

[Jokinen et al., 2014] Jokinen, E., Remes, U., Takanen, M., Palomaki, K., Kurimo, M., and Alku, P. (2014). Spectral tilt modelling with extrapolated GMMs for intelligibility enhancement of narrowband telephone speech. In *14th Int. Workshop on Acoustic Signal Enhancement (IWAENC)*, pages 164–168.

[Jreige et al., 2009] Jreige, C., Patel, R., and Bunnell, H. T. (2009). Vocalid: Personalizing text-to-speech synthesis for individuals with severe speech impairment. In *Proc. of the 11th International ACM SIGACCESS Conf. on Computers and Accessibility*, pages 259–260, (New York, NY, USA). ACM.

[Kabal, 2003] Kabal, P. (2003). Ill-conditioning and bandwidth expansion in linear prediction of speech. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing ICASSP*, volume 1, pages 824–827.

[Kafentzis et al., 2013] Kafentzis, G., Degottex, G., Rosec, O., and Stylianou, Y. (2013). Time-scale modifications based on a full-band adaptive harmonic model. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8193–8197.

[Kafentzis et al., 2014a] Kafentzis, G., Degottex, G., Rosec, O., and Stylianou, Y. (2014a). Pitch modifications of speech based on an adaptive harmonic model. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7924–7928.

[Kafentzis et al., 2012] Kafentzis, G., Pantazis, Y., Rosec, O., and Stylianou, Y. (2012). An extension of the adaptive quasi-harmonic model. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4605–4608.

[Kafentzis et al., 2014b] Kafentzis, G., Rosec, O., and Stylianou, Y. (2014b). Robust full-band adaptive sinusoidal analysis and synthesis of speech. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6260–6264.

[Kain, 2001] Kain, A. B. (2001). *High resolution voice transformation.* PhD thesis, B.A. Computer science and mathematics, Rockfort College.

[Kempster et al., 2009] Kempster, G. B., Gerratt, B. R., Verdolini Abbott, K., Barkmeier-Kraemer, J., and Hillman, R. E. (2009). Consensus auditory-perceptual evaluation of voice: development of a standardized clinical protocol. *American Journal of Speech-Language Pathology*, 18(2):124–132.

[Kim et al., 2006] Kim, S.-J., Kim, J.-J., and Hahn, M. (2006). HMM-based Korean speech synthesis system for hand-held devices. *IEEE Trans. on Consumer Electronics*, 52(4):1384–1390.

[King and Karaiskos, 2013] King, S. and Karaiskos, V. (2013). The Blizzard Challenge 2013. In *8th Speech Synthesis Workshop*, (Barcelona).

[Klatt, 1980] Klatt, D. H. (1980). Software for a cascade/parallel formant synthesizer. *Journal Acoustical Society of America*, 67(3):971–955.

[Klatt, 1987] Klatt, D. H. (1987). Review of text-to-speech conversion for English. *Journal of the acoustical Society of America*, 82(3):737–793.

[Kuhn et al., 2000] Kuhn, R., Junqua, J.-C., Nguyen, P., and Niedzielski, N. (2000). Rapid speaker adaptation in eigenvoice space. *IEEE Transactions on Speech and Audio Processing*, 8:695–707.

[Kumar and Kishore, 2004] Kumar, R. and Kishore, S. P. (2004). Automatic pruning of unit selection speech databases for synthesis without loss of naturalness. In *International Conference on Spoken Language Processing*, (Korea).

[Laskar et al., 2012] Laskar, R. H., Chakrabarty, D., Talukdar, F. A., Rao, K. S., and Banerjee, K. (2012). Comparing ANN and GMM in a voice conversion framework. *Applied Soft Computing*, 12(11):3332–3342.

[Lau, 2008] Lau, P. (2008). The Lombard Effect as a Communicative Phenomenon. *UC Berkley Phonology Lab Annual Report*, pages 1–9.

[Laver, 1980] Laver, J. (1980). *The phonetic description of voice quality.* Cambridge University Press.

[Law and Lee, 2000] Law, K. M. and Lee, T. (2000). Using cross-syllable units for Cantonese speech synthesis. In *Proc. of the Int. Conf. on Spoken Language Processing.*

[Leggetter and Woodland, 1995] Leggetter, C. J. and Woodland, P. C. (1995). Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. In *Computer Speech & Language*, volume 9, pages 171–185.

[Levinson, 1949] Levinson, N. (1949). The Wiener RMS (root mean square) error criterion in filter design and prediction. In *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: with Engineering Applications*, pages 129–148. MIT Press.

[LISTA, 2012] LISTA (2012). The Listening Talker Project. Accessible at: `http://listening-talker.org/`.

[Loquendo, 2015] Loquendo, T. (last web access December 2015). Loquendo tts. Accessible at: `http://www.nuance.com/for-business/by-solution/customer-service-solutions/solutions-services/inbound-solutions/loquendo-small-business-bundle/tts-demo/index.htm` Last accessed: January 2016.

[Macchi, 1998] Macchi, M. (1998). Issues in text-to-speech synthesis. In *Proc. of IEEE Int. Joint Symposia on Intelligence and Systems*, pages 318–325.

[Makhoul, 1975] Makhoul, J. (1975). Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580.

[Marques et al., 1990] Marques, J., Almeida, L., and Tribolet, J. (1990). Harmonic coding at 4.8 kb/s. In *Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 17–20.

[Martí, 1985] Martí, J. (1985). *Estudi acústic del català i síntesi automàtica per ordinador.* PhD thesis, Universitat de València.

[Masuko et al., 1997] Masuko, T., Tokuda, K., Kobayashi, T., and Imai, S. (1997). Voice characteristics conversion for hmm-based speech synthesis system. In *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 3, pages 1611–1614.

[Matoušek et al., 2005] Matoušek, J., Hanzlíček, Z., and Tihelka, D. (2005). Hybrid syllable/triphone speech synthesis. In *Proc. of 9th European Conference on Speech Communication and Technology*, pages 2529–2532, (Lisbon, Portugal).

[Matsui et al., 2013] Matsui, K., Kimura, K., Nakatoh, Y., and Kato, Y. O. (2013). Development of electrolarynx with hands-free prosody control. In *8th ISCA Workshop on Speech Synthesis*, pages 293–297, (Barcelona, Spain).

[Mayerhoff et al., 2014] Mayerhoff, R. M., Guzman, M., Jackson-Menaldi, C., Munoz, D., Dowdall, J., Maki, A., Johns, M. M., Smith, L. J., and Rubin, A. D. (2014). Analysis of supraglottic activity during vocalization in healthy singers. *The Laryngoscope*, 124(2):504–509.

[McAulay and Quatieri, 1986] McAulay, R. and Quatieri, T. (1986). Speech analysis/synthesis based on a sinusoidal representation. *IEEE Trans. on Acoust., Speech and Signal process*, 34(4):744–754.

[Mohammadi and Kain, 2014] Mohammadi, S. and Kain, A. (2014). Voice conversion using deep neural networks with speaker-independent pre-training. In *IEEE Int. Spoken Language Technology Workshop (SLT)*, pages 19–23.

[Montero et al., 1999] Montero, J. M., Gutiérrez-Arriola, J., Colás, J., Enríquez, E., and Pardo, J. M. (1999). Analysis and modelling of emotional speech in Spanish. In *14th International Congress of Phonetic Sciences (ICPhS)*, page 957–960, (San Francisco, USA).

[Monzo, 2010] Monzo, C. (2010). *Modelado de la cualidad de la voz para la síntesis del habla expresiva.* PhD thesis, Escola Tècnica Superior d'Enginyeria Electrònica i Informàtica La Salle - Ramon Llull.

[Monzo et al., 2010] Monzo, C., Calzada, A., Iriondo, I., and Socoró, J. C. (2010). Expressive speech style transformation: voice quality and prosody modification using a harmonic plus noise model. In *Speech Prosody*, (Chicago).

[Monzo et al., 2008] Monzo, C., Formiga, L., Adell, J., Iriondo, I., Alías, F., and Socoró, J. C. (2008). Adaptación del CTH-URL para la competición Albayzin 2008. In *V Jornadas en Tecnología del Habla*, pages 87–89, (Bilbao, Spain).

[Moon, 1999] Moon, T. K. (1999). *Mathematical Methods and Algorithms for Signal Processing.* Addison-Wesley Educational Publishers Inc.

[Morfi et al., 2014] Morfi, V., Degottex, G., and Mouchtaris, A. (2014). A computationally efficient refinement of the fundamental frequency estimate for the adaptive harmonic model. In *IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1478–1482.

[Morfi et al., 2015] Morfi, V., Degottex, G., and Mouchtaris, A. (2015). Speech analysis and synthesis with a computationally efficient adaptive harmonic model. *IEEE/ACM Trans. Audio, Speech and Lang. Processing*, 23(11):1950–1962.

[Mori et al., 2006] Mori, S., Moriyama, T., and Ozawa, S. (2006). Emotional speech synthesis using subspace constraints in prosody. *Multimedia and Expo, IEEE International Conference on*, 0:1093–1096.

[Moulines and Charpentier, 1990] Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453 – 467.

[Murray, 2008] Murray, I. (2008). Hamlet (helpful automatic machine for language and emotional talk) synthesizer. Available at: `http://staff.computing.dundee.ac.uk/irmurray/hamlet.asp` Last accessed: January 2016.

[Murray et al., 2000] Murray, I. R., Edgington, M. D., Campion, D., and Lynn, J. (2000). Rule-based emotion synthesis using concatenated speech. In *ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion*, pages 173–177, (Northern Ireland).

[Nakagawa et al., 1995] Nakagawa, S., Shikano, K., and Tohkura, Y. (1995). *Speech, Hearing and Neural Network Models*. Ohmsha, Tokyo, Japan, Japan.

[Nordstrom et al., 2008] Nordstrom, K., Tzanetakis, G., and Driessen, P. (2008). Transforming perceived vocal effort and breathiness using adaptive pre-emphasis linear prediction. *IEEE Trans. on Audio, Speech, and Lang. Process.*, 16(6):1087–1096.

[Nordstrom and Driessen, 2006] Nordstrom, K. I. and Driessen, P. F. (2006). Variable pre-emphasis LPC for modeling vocal effort in the singing voice. *Proc. of the 9th Int. Conf. on Digit. Audio Eff. (DAFx)*, pages 18–20.

[Paliwal, 1995] Paliwal, K. K. (1995). Interpolation properties of linear prediction parametric representations. In *European Conference on Speech Communication and Technology.*

[Pantazis et al., 2008] Pantazis, Y., Rosec, O., and Stylianou, Y. (2008). On the properties of a time-varying quasi-harmonic model of speech. In *Annual Conference of the International Speech Communication Association*, pages 1044–1047.

[Pantazis et al., 2010a] Pantazis, Y., Rosec, O., and Stylianou, Y. (2010a). Iterative estimation of sinusoidal signal parameters. *Signal Processing Letters, IEEE*, 17(5):461–464.

[Pantazis et al., 2010b] Pantazis, Y., Tzedakis, G., Rosec, O., and Stylianou, Y. (2010b). Analysis/synthesis of speech based on an adaptive quasi-harmonic plus noise model. In *IEEE Int. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, pages 4246–4249.

[Pearson et al., 1998] Pearson, S., Kibre, N., and Niedzielski, N. (1998). A synthesis method based on concatenation of demisyllables and a residual excited vocal tract model. In *Proc. of the Int. Conf. on Speech and Language Processing.*

[Peinado and Segura, 2006] Peinado, A. M. and Segura, J. C. (2006). Alternative representations of the LPC coefficients. In *Speech Recognition Over Digital Channels*, pages 225–226. {John Wiley \& Sons}.

[Phung et al., 2013] Phung, T.-N., Luong, C. M., and Akagi, M. (2013). A hybrid {TTS} between unit selection and {HMM}-based {TTS} under limited data conditions. In *8th ISCA Workshop on Speech Synthesis*, pages 299–304, (Barcelona, Spain).

[Planet et al., 2008] Planet, S., Iriondo, I., Martínez, E., and Montero, J. A. (2008). TRUE: an online testing platform for multimedia evaluation. In *Proc. 2nd Int. Workshop on Emot.: Corpora for Res. on Emot. and Affect. at the 6th conf. on lang. resour. & eval. (LREC)*, pages 61–65, (Marrakech, Morocco).

[Quatieri and McAulay, 1998] Quatieri, T. and McAulay, R. (1998). Audio signal processing based on sinusoidal analysis/synthesis. In Kahrs, M. and Brandenburg, K., editors, *Applications of Digital Signal Processing to Audio and Acoustics*, volume 437 of *International Series in Engineering and Computer Science*, pages 343–416. Springer.

[Quatieri and McAulay, 1986] Quatieri, T. F. and McAulay, R. J. (1986). Speech transformations based on a sinusoidal representation. In *IEEE Trans. on Accous., Speech, Signal Process.*, volume ASSP-34.

[Rank and Pirker, 1998] Rank, E. and Pirker, H. (1998). Generating emotional speech with a concatenative synthesizer. In *Proc. of Int. Conf. on Spoken Language Processing (ICSLP)*, pages 671–674, (Sydney, Australia).

[Rentzos et al., 2004] Rentzos, D., Vaseghi, S., Yan, Q., and Ho, C.-H. (2004). Voice conversion through transformation of spectral and intonation features. In *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 21–24.

[Rodríguez et al., 1999] Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernadas, D., Oliver, J., and Longhi, L. (1999). Modelización acústica de la expresión emocional en el español. In *Procesamiento del Lenguaje Natural*, volume 25, pages 159–166, Lleida, Spain.

[Rosenthal et al., 2014] Rosenthal, A. L., Lowell, S. Y., and Colton, R. H. (2014). Aerodynamic and acoustic features of vocal effort. *Journal of Voice*, 28(2):144–153.

[Rupal, 2013] Rupal, P. (2013). TED talk: Synthetic voices, as unique as fingerprints. Available at: `https://www.ted.com/talks/rupal_patel_synthetic_voices_as_unique_as_fingerprints` Last accessed: January 2016.

[Saito et al., 1996] Saito, T., Hashimoto, Y., and Sakamoto, M. (1996). High-quality speech synthesis using context-dependent syllabic units. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 1, pages 381–384.

[Sakoe and Chiba, 1978] Sakoe, H. and Chiba, S. (1978). Dynamic programming algorithm optimization for spoken word recognition. In *IEEE Transactions on Acoustics, Speech and Signal Processing*, pages 43–49.

[Schröder, 1999] Schröder, M. (1999). Can emotions be synthesized without controlling voice quality? In *In: Phonus 4, Research Report of the Institute of Phonetics, University of the Saarland*, pages 37–55.

[Schröder, 2001] Schröder, M. (2001). Emotional speech synthesis: A review.

[Schröder, 2004] Schröder, M. (2004). *Speech and Emotion Research.* PhD thesis, Saarland University.

[Schröder and Grice, 2003] Schröder, M. and Grice, M. (2003). Expressing vocal effort in concatenative synthesis. In *Proc. 15th International Conference of Phonetic Sciences*, pages 2589–2592.

[Schröder and Trouvain, 2003] Schröder, M. and Trouvain, J. (2003). The German text-to-speech synthesis system MARY: A tool for research, development and teaching. *International Journal of Speech Technology*, 6:365–377.

[Schwartz et al., 1984] Schwartz, R., Chow, Y., Roucos, S., Krasner, M., and Makhoul, J. (1984). Improved hidden Markov modeling of phonemes for continuous speech recognition. In *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, volume 9, pages 21–24.

[Sharma et al., 2013] Sharma, K., Suryakanthi, T., and Prasad, T. V. (2013). Exploration of speech enabled system for English. *CoRR*, abs/1304.8013.

[Shichiri et al., 2002] Shichiri, K., Sawabe, A., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2002). Eigenvoices for HMM-based speech synthesis. In *Proc. of Int. Conf. on Spoken Language Processing (ICSLP)*, pages 1269–1272.

[Sproat, 1998] Sproat, R. (1998). *Multilingual Text-To-Speech Synthesis: The Bell Labs Approach.* Kluwer Academic Publishers.

[Stallo, 2000] Stallo, J. (2000). *Simulating emotional speech for a talking head.* PhD thesis, School of Computing, Curtin University of Technology, Australia.

[Stylianou, 1996] Stylianou, I. (1996). *Harmonic plus noise models for speech combined with statistical methods for speech and speaker modification.* PhD thesis, Ecole Nationale Supérieure des Télécomunications.

[Stylianou, 2001] Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. In *IEEE Trans. on Speech and Audio Process.*, volume 9, pages 21–29.

[Stylianou et al., 1997] Stylianou, Y., Dutoit, T., and Schroeter, J. (1997). Diphone concatenation using a harmonic plus noise model of speech. In *Proc. Eurospeech*, pages 613–616.

[Stylianou and Moulines, 1998] Stylianou, Y. and Moulines, E. (1998). Continuous probabilistic transform for voice conversion. *IEEE Transactions on Speech and Audio Processing*, 6:131–142.

[Syrdal et al., 1998] Syrdal, A., Stylianou, Y., Garrison, L., Conkie, A., and Schroeter, J. (1998). TD-PSOLA versus harmonic plus noise model in diphone based speech synthesis. In *Proc. of IEEE Int. Conf. on Acoust., Speech and Signal Process.*, volume 1, pages 273–276.

[Tachibana et al., 2008] Tachibana, M., Izawa, S., Nose, T., and Kobayashi, T. (2008). Speaker and style adaptation using average voice model for style control in HMM-based speech synthesis. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4633–4636.

[Tamura et al., 2001] Tamura, M., Masuko, T., Tokuda, K., and Kobayashi, T. (2001). Adaptation of pitch and spectrum for hmm-based speech synthesis using mllr. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 805–808.

[Tasko and Greilick, 2010] Tasko, S. M. and Greilick, K. (2010). Acoustic and articulatory features of diphthong production: A speech clarity study. *Journal of Speech, Language, and Hearing Research*, 53(1):84–99.

[Taylor, 2007] Taylor, P. (2007). *Text-to-speech Synthesis*. Cambridge university press.

[Tiomkin et al., 2011] Tiomkin, S., Malah, D., Shechtman, S., and Kons, Z. (2011). A hybrid text-to-speech system that combines concatenative and statistical synthesis units. *Audio, Speech, and Language Processing, IEEE Transactions on*, 19(5):1278–1288.

[Toda, 2003] Toda, T. (2003). *High-Quality and Flexible Speech Synthesis with Segment Selection and Voice Conversion*. PhD thesis, Graduate School of Information Science, Nara Institute of Science and Technology.

[Toda et al., 2007] Toda, T., Ohtani, Y., and Shikano, K. (2007). One-to-many and many-to-one voice conversion based on eigenvoices. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 4, pages 1249–1252.

[Toda et al., 2001] Toda, T., Saruwatari, H., and Shikano, K. (2001). Voice conversion algorithm based on Gaussian mixture model with dynamic frequency warping of STRAIGHT spectrum. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 2, pages 841–844.

[Tokuda et al., 1995] Tokuda, K., Kobayashi, T., and Imai, S. (1995). Speech parameter generation from HMM using dynamic features. In *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, volume 1, pages 660–663.

[Tokuda and Zen, 2015] Tokuda, K. and Zen, H. (2015). Directly modeling speech waveforms by neural networks for statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4215–4219.

[Traunmüller, 1997] Traunmüller, H. (1997). Perception of speaker sex, age, and vocal effort. *Umeå University, Department of Phonetics*, 4:183–186.

[Traunmüller and Eriksson, 2000] Traunmüller, H. and Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *Journal of the Acoustical Society of America*, 107(6):3438–3451.

[Türk and Arslan, 2006] Türk, O. and Arslan, L. M. (2006). Robust processing techniques for voice conversion. *Computer Speech & Language*, 20(4):441–467.

[Türk and Schröder, 2008] Türk, O. and Schröder, M. (2008). A comparison of voice conversion methods for transforming voice quality in emotional speech synthesis. In *Interspeech*, pages 2282–2285. Int. Speech Communication Association (ISCA).

[Türk et al., 2005] Türk, O., Schröder, M., Bozkurt, B., and Arslan, L. M. (2005). Voice quality interpolation for emotional text-to-speech synthesis. In *Proceedings Interspeech 2005*, pages 797–800, (Lisbon, Portugal).

[Valbret et al., 1992] Valbret, H., Moulines, E., and Tubach, J. P. (1992). Voice transformation using PSOLA technique. *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, 1:145–148.

[Vosnidis and Digalakis, 2001] Vosnidis, C. and Digalakis, V. (2001). Use of clustering information for coarticulation compensation in speech synthesis by word concatenation. In *Proc. of Eurospeech.*

[Websphere,] Websphere, I. Ibm tts system. Accessible at: `http://www-01.ibm.com/software/pervasive/voice_server/` Last accessed: December 2015.

[Wells, 1997] Wells, J. (1997). Sampa computer readable phonetic alphabet. In *Handbook of Standards and Resources for Spoken Language Systems.*, volume Part 4, section B. Gibbon, D., Moore, R. and Winski, R. (eds.).

[Weruaga and Képesi, 2007] Weruaga, L. and Képesi, M. (2007). The fan-chirp transform for non-stationary harmonic signals. *Signal Processing*, 87(6):1504–1522.

[Wu et al., 2015] Wu, Z., Valentini-Botinhao, C., Watts, O., and King, S. (2015). Deep neural networks employing multi-task learning and stacked bottleneck features for speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4460–4464.

[Yamagishi and Kobayashi, 2007] Yamagishi, J. and Kobayashi, T. (2007). Average-voice-based speech synthesis using HSMM-based speaker adaptation and adaptive training. In *IEICE Trans. Inf. & Syst.*, volume E-90-D.

[Yamagishi et al., 2012] Yamagishi, J., Veaux, C., King, S., and Renals, S. (2012). Speech synthesis technologies for individuals with vocal disabilities: Voice banking and reconstruction. *Acoustical Science and Technology*, 33(1):1–5.

[Yoshimura et al., 2000] Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., and Kitamura, T. (2000). Speaker interpolation for hmm-based speech synthesis system. In *Acoustical Science and Technology*, volume 21, pages 199–206.

[Zen and Senior, 2014] Zen, H. and Senior, A. (2014). Deep mixture density networks for acoustic modeling in statistical parametric speech synthesis. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 3872–3876.

[Zen et al., 2013] Zen, H., Senior, A., and Schuster, M. (2013). Statistical parametric speech synthesis using deep neural networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7962–7966.

[Zen et al., 2007] Zen, H., Toda, T., Nakamura, M., and Tokuda, K. (2007). Details of the Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. In *IEICE - Trans. Inf. Syst.*, volume E90-D, pages 325–333, Oxford, UK. Oxford University Press.

[Zen et al., 2009] Zen, H., Tokuda, K., and Black, A. W. (2009). Statistical parametric speech synthesis. *Speech Communication*, 51(11):1039–1064.

[Öhlin and Carlson, 2004] Öhlin, D. and Carlson, R. (2004). Data-driven formant synthesis. In *Proc. FONETIK Dept. of Linguistics, Stockholm University.*

Universitat Ramon Llull

Aquesta Tesi Doctoral ha estat defensada el dia ____ d________________ de 201__

al Centre____

de la Universitat Ramon Llull, davant el Tribunal format pels Doctors i Doctores

sotasignants, havent obtingut la qualificació:

President/a

____

Vocal

____

Vocal *

____

Vocal *

____

Secretari/ària

____

Doctorand/a

____

(*): Només en el cas de tenir un tribunal de 5 membres