TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

**GERARD SANCHEZ GASULLA**

# PROSODY AND WAVELETS: TOWARDS A NATURAL SPEAKING STYLE CONVERSION

Master of Science Thesis

Examiners: HANNA SILEN
Examiners: JANI NURMINEN
Examiners: MONCEF GABBOUJ
Examiners and topic approved by the Council of the Faculty of Computing and Electrical Engineering on 4.12.2013

# ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
**GERARD SANCHEZ GASULLA : Prosody and Wavelets: Towards a natural speaking style conversion**
Master of Science Thesis, 68 pages
March 11th, 2014
Major: Signal Processing
Examiner: Hanna Silén, Jani Nurminen, Moncef Gabbouj
Keywords: Speech processing, voice conversion, prosody conversion, wavelet analysis

Speech is the basis of human communication: in everyday life we automatically decode speech into language regardless of who speaks. In a similar way, we have the ability to recognize different speakers, despite the linguistic content of the speech. Additionally to the voice individuality of the speaker, the particular prosody of speech involves relevant information concerning the identity, age, social group or economical status of the speaker, helping us identify the person to whom we are talking without seeing the speaker.

Voice conversion systems deal with the conversion of a speech signal to sound as if it was uttered by another speaker. It has been an important amount of work in the conversion of the timber of the voice, the spectral features, meanwhile the conversion of pitch and the way it temporarily evolves, modeling the speaker dependent prosody, is mostly achieved by just controlling the level and range.

This thesis focuses on prosody conversion, proposing an approach based on a wavelet transformation of the pitch contours. It has been performed a study of the wavelet domain, discerning among the different timing of the prosodic events, thus allowing an improved modeling of them. Consequently, the prosody conversion is achieved in the wavelet domain, using regression techniques originally developed for the spectral features conversion, in voice conversion systems.

# Preface

This work has been conducted at the Audio Research Team, within the Department of Signal Processing of Tampere University of Technology.

First and foremost, I would like to express my gratitude to my supervisors, MSc. Hanna Silén and Dr. Jani Nurminen for all the help, guidance and support through the development of this thesis, which would not have been the same without their constant advices and recommendations. I also would like to thank Dr. Tuomas Virtanen, for allowing me to develop this dissertation in the Audio Research Team, and helping me when I was looking for supervisors.

Next, I would like to thank Dr. Moncef Gabbouj for accepting me in the multimedia group and participating as an examiner of this thesis. I also would like to thank MSc. Toni Heittola for the help and tips provided in the development of the listening test.

Moreover, I owe special thanks to my officemate, Samuel, whom we shared the introductory work in the thesis, and has been always open to have technical chats (and casual too), sharing a great time together. Also, I want to thank all the people which I met in my Erasmus stay in Tampere, transforming this time in a huge experience, with lots of moments and memories shared, but also for "suffering" my samples in the listening test.

Finally, I would like to thank my family, specially my parents, for their continued support throughout my life, raising me as the person who I am, and encouraging me and pushing me forward when things went wrong. I will always be in debt with you.

# Table of contents

## Terms and definitions

| | |
|---|---|
| AR | Auto-Regressive |
| A/S | Analysis/Synthesis |
| CWT | Continuous Wavelet Transform |
| DCT | Discrete Cosine Transform |
| DFT | Discrete Fourier Transform |
| DKPLS | Dynamic Kernel Partial Least Squares |
| $F_0$ | Fundamental frequency |
| HNM | Harmonic plus Noise Model |
| LP | Linear Predicton |
| LPC | Linear Predictive Coding |
| LSF | Line Spectral Frequencies |
| LSP | Line Spectral Pair |
| MCC | Mel-general Cepstrum Coefficients |
| MFCC | Mel-Frequency Cepstrum Coefficients |
| PLS | Partial Least Squares |
| RMSE | Root Mean Squared Error |
| VC | Voice Conversion |

# 1. Introduction

Voice Conversion algorithms aim to modify the utterance of a first speaker to sound as if it was uttered by a second speaker. The speaker identity is composed by different factors, including short-term spectral characteristics, prosody and linguistic style, thus in Voice Conversion it is important to take into account all of them, in order to transform the global identity of the speaker. A significant amount of work has focused on the conversion of spectral parameters, which reflect the timbre of the voice (how the voice itself sounds). However, the modeling and conversion of the prosody is still one of the most challenging areas within the Voice Conversion framework, and the employed methods are still slightly simplistic.

The prosody of speech, the speaking style of the speaker, is an idiosyncratic feature of the speaker. Nevertheless, depending on what the speaker wants to reflect on the utterance, e.g. irony, sarcasm, question utterances or commands; the emotional state of the speaker, or the target audience of the speech, the prosody of the exact same utterance may change notoriously. Jointly with the difficulties of evaluating the prosody, this is one of the main difficulties of the prosody modeling.

## 1.1 Motivation

Prosody is created by several factors, such as the phone duration, loudness and pause location, however, the main manifestation and the most expressive one is the **pitch**. Different speakers have different pitch ranges, which can be represented by calculating the mean pitch and pitch variance for each speaker.

Nevertheless, the shapes and contour of the pitch, not just the mean and variance, contain speaker specific information that needs to be extracted and transformed in a voice conversion framework, since changing pitch throughout an utterance is usually the most powerful way of expressing emotion or emphasis based on the meaning of the message.

Different speakers may utter the same sentence with different intonation patterns and each speaker may have specific habits of expressing a message, or a particular emotion. In an ideal voice conversion system, it is important to capture these global habits and manipulate the entire pitch contour accordingly while converting from one speaker to another. Therefore, implanting an appropriate pitch contour, modeling the identity and characteristics of the speaker, is crucial to retain the perceived naturalness of converted speech.

## 1.2 Objective, scope and main results

The objective of this thesis consists of developing a new prosody conversion method, based on the transformation of the pitch, the most expressive manifestation of the speaking style. The proposed approach uses a well-known technique in signal processing, the wavelet transformation, to extract further information from the initial pitch. The wavelet analysis allows discerning the different prosodic phenomena present in the speech, and modeling them separately in the posterior conversion.

The hypothesis is studied and evaluated for English intra-gender and cross-gender conversions. The conversion of the spectral parameters of the speech is carried out in order to evaluate the final synthesized speech, however they are out of the scope of this thesis, thus, no attempts are made in order to improve the performance. Other prosodic features, such as syllables and words durations, or the power of the signal, are neither converted. The performance of the method is assessed in an objective framework as well as in a perceptual listening test, comparing the proposed method with the simplest conversion scheme that has been largely used in the literature [Sty98].

The results show a clear preference in the cross-gender conversion, for the proposed method in improving the naturalness of the generated sample when compared against the typical pitch conversion in literature. Moreover, the statistical results of the proposed method also show an improvement of the similarity with the target speaker pitch contour.

In addition to the evaluation of the conversion method, there are presented other interesting results concerning just the wavelet transformation: the prosodic morphology contains different temporal levels, from the intonation pattern along the whole utterance, to the microprosody events present on the phonemes, going through different levels, all of them with significant prosodic information.

## 1.3 Outline

This dissertation is organized as follows:

**Chapter 2** provides the theoretical background related to the human speech production system, and introduces the main speech parametrization and representation methods in speech processing systems.

**Chapter 3** is focused on the linguistic side of prosody, introducing the principal reasons of different prosodic styles, and the major prosodic events in the speech.

**Chapter 4** introduces the theoretical background related to the Voice Conversion task and summarizes the most influencing works of Voice Conversion.

**Chapter 5** describes the new approach presented in this thesis: the conversion of the prosody with wavelets. It is presented the characteristics of the wavelet transformation applied to speech signals, and the relation between the wavelet domain and the prosodic hierarchical model. Finally, the characteristics of the prosody conversion are introduced.

**Chapter 6** presents an evaluation of the proposed prosody conversion method, and a comparison of the performance with the main method in the literature.

**Chapter 7** gives the main conclusions of the thesis, together with the future work that can be considered as extension of this dissertation.

# 2. Speech Feature and Representation

Speech has a central role in human interaction, thus, a lot of research attention has focused in different aspects of the human speech production, speech perception and different features of spoken language understanding. Moreover, it exists an extensive work in the last years modeling the features of speech and developing new speech processing techniques.

This chapter presents a brief introduction of the human speech production process, in 2.1, and introduces the main speech representation techniques used in the literature in 2.2.

## 2.1 Human speech production

The human speech production process is divided in four main steps: the language processing, where the contents of an utterance are divided first into words and then in phonemic symbols in the brain language center; the generation of motor commands to the vocal organs in the brain's motor center; the movement of the organs involved in speech production (Fig. 2.1) based on the sent commands; and the emission of the air sent from the lungs in the final speech form.

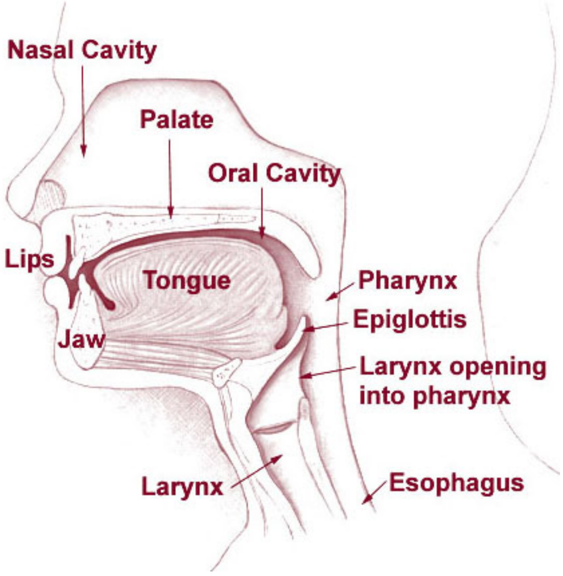

Figure 2.1: Humans organs involved in speech production. (From [Wik])

This final step is the one which generates the main characteristics of speech: when air is released from lungs, it flows through the glottis between the vocals cords which vibrate at regular intervals to produce **voiced** sounds, such as vowels, or remains open on the unvoiced sounds, in **unvoiced** consonants (Fig. 2.2).

Therefore, the V-shaped opening between the vocal cords is the most important sound source in the vocal system. Each person has a different natural length on the vocal cords and the particular mass and tension applied affects the way the vocal cords vibrate (opening and closing) on the voiced portions, generating the characteristic frequency of the voice, the fundamental frequency of voice ($F_0$). Thus, this $F_0$ is different for each person, although men and women can generally be grouped in two separate ranges of $F_0$: between 40 Hz and 180 Hz for men and between 100 Hz and 300 Hz for women.

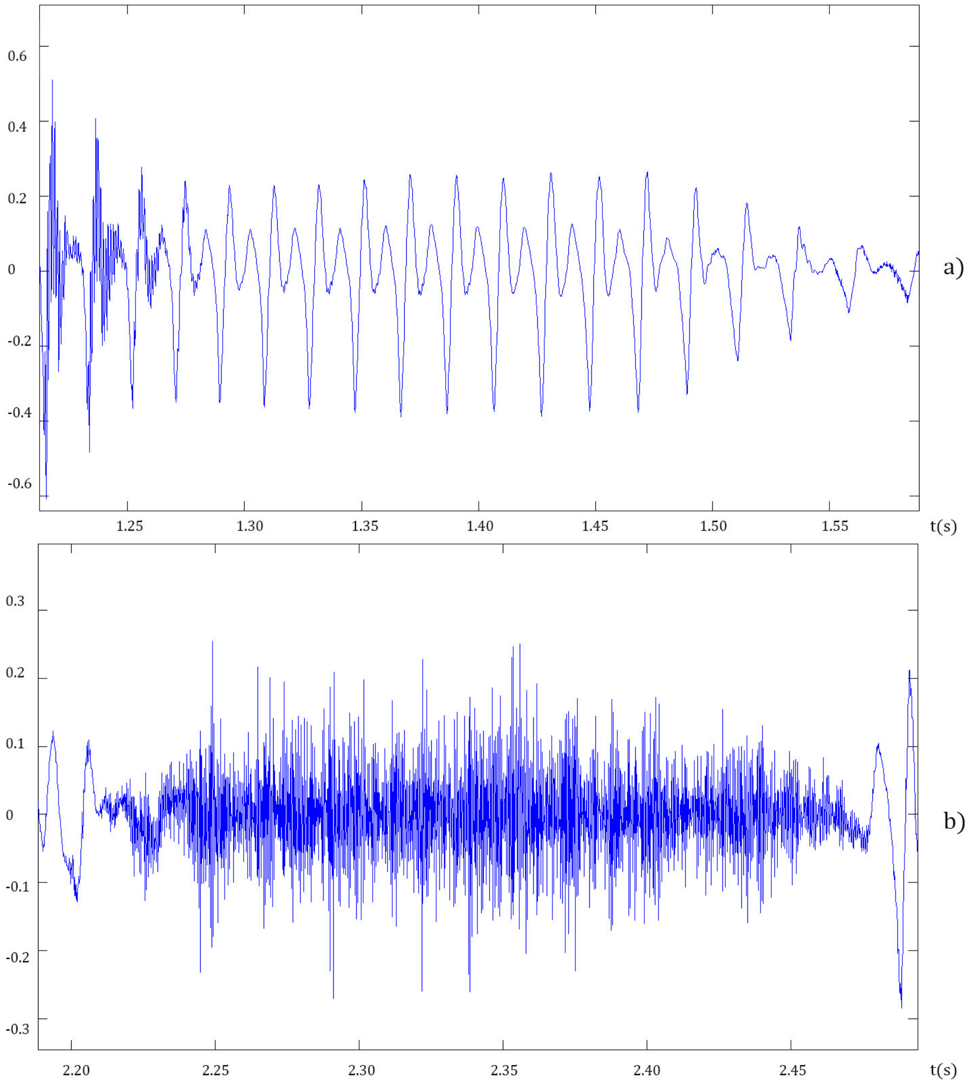

Figure 2.2: a) Voiced speech for the vowel 'i', and b) unvoiced speech for the consonant 's'.

Finally air reaches the oral cavity (and nasal cavity) where the velum, palate, teeth and lips modify the first excitation of vocal cords. Along with the position of the mouth, the position of the tongue and the way the tongue is placed, allows the addition of harmonics, called formants, and the production of the final **phoneme**[1].

## 2.2 Speech Representation

One of the basic tools for analyzing speech is the short-time Fourier analysis: by selecting overlapping segments among 10-30 ms[2], displaced by e.g. 5-10 ms steps and computing a discrete Fourier transform (DFT), the speech signal is decomposed into several frames, each one showing its corresponding spectral characteristics (Fig. 2.3).

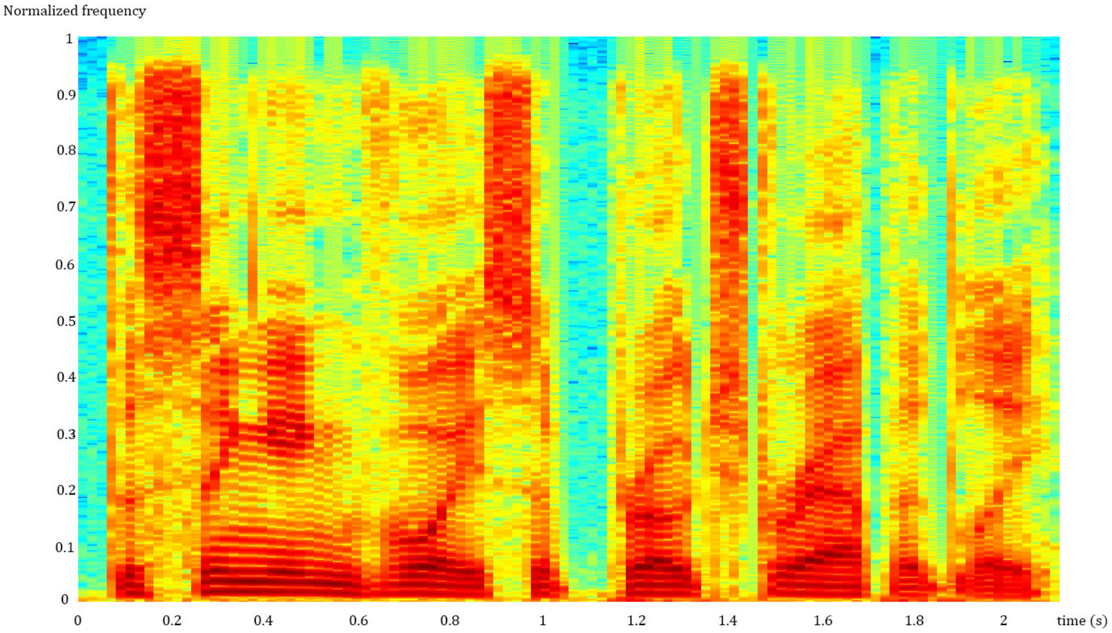

Figure 2.3: Spectrogram of the sentence "The singing voice approached rapidly.", uttered by a male speaker. The speech sample was windowed in 50 ms blocks with a 5 ms displacement.

However using a source-filter model, see Fig. 2.4, the effect of $F_0$ and vocal tract (formants) can be separated, and consequently used in a more accurate way: the source represents the air flow coming, the excitation signal which allows producing the **voiced** sounds, when it consists of the $F_0$ periodic impulses, and the **unvoiced** or fricated sounds, when it is working as a random noise source. On the other hand,

[1] A phoneme is defined as any of the perceptually distinct units of sound in a specified language that distinguish one word from another, for example p, b, d, and t in the English words pad, pat, bad, and bat.

[2] Speech signals are assumed to be approximately stationary in blocks of 10-30 ms, allowing the correct assumption of stationarity in multiple signal processing algorithms.

the filter models the resonances in the vocal tract.

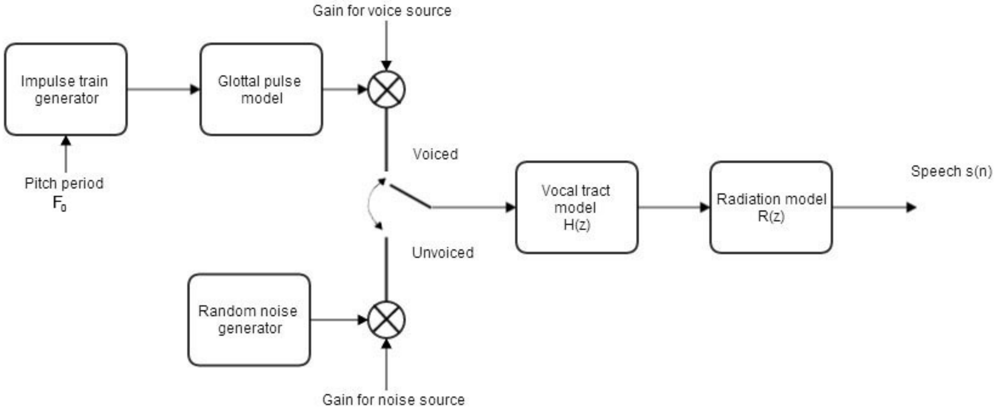

Figure 2.4: Source-filter model of the production of speech.

### 2.2.1 Linear prediction model

Linear prediction (LP) is widely used in speech applications, due to the fact that speech production process is well modeled with LP. The LP model of a speech signal can be written in the following way:

$$x(m) = \sum_{k=1}^{K} a_k x(m-k) + Gu(k) \tag{2.1}$$

where $m$ is the time index, $K$ represents the number of coefficient in the model, $a_k$, $k = 1, \ldots, K$, are defined as the linear prediction coefficients (LPC), $G$ is the gain of the system, and $u(k)$ is the excitation signal. The equation 2.1 can be written, using the z-transform, in the frequency domain:

$$H(z) = \frac{G}{1 - \sum_{k=1}^{K} a_k z^{-k}}, \tag{2.2}$$

which corresponds to an all-pole transfer function.

The LPC can be estimated in various ways, e.g. using the autocorrelation method with Levinson-Durbin algorithm [Lev46; Dur60], or the covariance method with the Cholesky decomposition method [Bel87].

### 2.2.2 Line spectral frequencies

Line spectral frequencies (LSFs) are a more robust representation of the coefficients of linear predictive models. LSFs are obtained from LPCs, computing the roots of two polynomials, called line spectral pair (LSP), $P(z)$ and $Q(z)$:

$$\begin{aligned} P(z) &= A(z) + z^{-(p+1)}A(z^{-1}) \\ Q(z) &= A(z) - z^{-(p+1)}A(z^{-1}) \end{aligned} \tag{2.3}$$

The reverse conversion is also easily computed as follows:

$$A(z) = \frac{1}{2}(P(z) + Q(z)) \tag{2.4}$$

LSF offers a robust representation, good interpolation properties, and a close relationship to the formants, however they just model spectral peaks, corresponding with the formants, but not the valleys.

### 2.2.3 Cepstral features

Another way of parameterizing the speech features is using cepstral analysis. The **cepstrum** is an homomorphic transformation [Hua92] defined as

$$\hat{x}(n) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \ln X(e^{j\omega})e^{j\omega n}d\omega. \tag{2.5}$$

It allows an easy, but not perfect, separation of the source and filter of the produced speech signal: if the interest is in the glottal excitation, the high-quefrency components of the cepstrum are taken, and if the interest is on the vocal tract, it must be kept the low-quefrency components.

The **Mel-Frequency Cepstrum Coefficients** (MFCC) is a representation defined as the real cepstrum of a windowed short-time signal derived from the DFT of that signal. The difference remains mainly in the frequency scale used; MFCC uses the Mel scale, a nonlinear frequency scale which gives a better approximation of the behavior of the auditory system, instead of the linear frequency scale:

$$f^{mel} = 2595 \log_{10}(1 + f/700) \tag{2.6}$$

This scale conversion is achieved using a bank of triangular filters located according the Mel-frequency scale. Finally is applied a discrete cosine transformation (DCT) to the logarithmic energy of the filterbank:

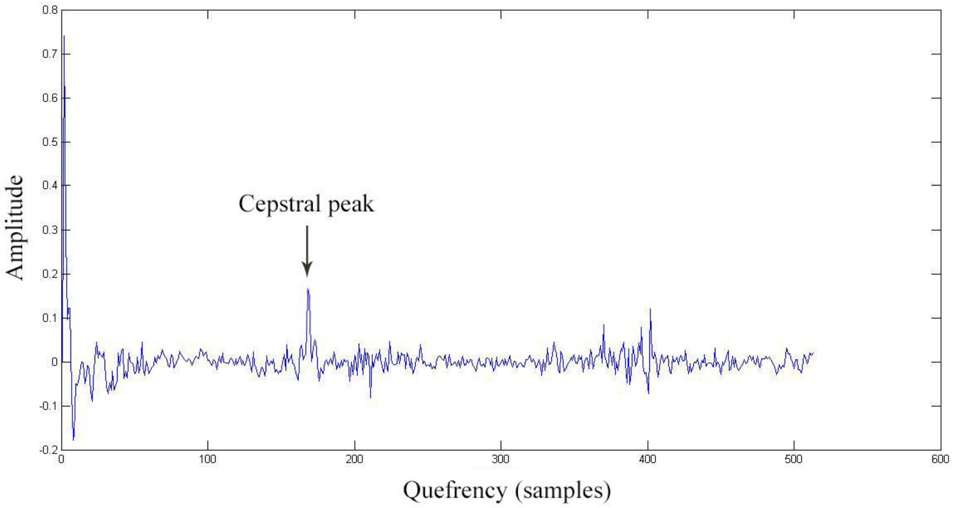

Figure 2.5: Cepstral representation of the source, within a speech sample.

$$c[n] = \sum_{m=0}^{M-1} S[m] cos(\pi n(m + 1/2)/M) \tag{2.7}$$

where $S[m]$ is the logarithmic energy of the $m^{th}$ filter, and $M$ is the number of filters. It varies for different implementations from 24 to 40, but usually is set to $M = 24$ in a 16 kHz sampling rate.

MFCCs are capable of modeling both spectral peaks and valleys, but more important, are reliable for measuring acoustic distances, therefore, they are especially useful for alignment on parallel data.

### 2.2.4 The generalized Mel cepstral analysis

LSF (and also LP) is a good method for obtaining all-pole representation of speech, modeling the spectral peaks, but is not capable of giving information of the valleys, the spectral zeros. On the other hand, cepstral modeling can represent poles and zeros with equal weights, but if a small number of cepstral coefficients are taken, it overestimates the band widths of the formants.

The Generalized Mel cepstral analysis method [Tok94] allows varying the model spectrum continuously from the all-pole spectrum to that represented by the cepstrum according to two control parameters: $\alpha$ and $\gamma$.

The parameter $\alpha$ controls the frequency resolution of the spectrum $\Psi_a(z)$ (Eq. 2.8), from $\alpha = 0$ for linear scale to $\alpha = 0.42$ which approximates the Mel-frequency scale, when the sampling frequency is 16 kHz. The parameter $\gamma$ adjusts the generalized logarithmic function $s_\gamma(w)$ (Eq. 2.9), from $\gamma = 0$ for cepstral modeling to

$\gamma = -1$ for LP representation.

$$\Psi_a(z) = \frac{z^{-1} - \alpha}{1 - \alpha z^{-1}}, \ \ |\alpha| < 1 \tag{2.8}$$

$$s_\gamma(w) = \begin{cases} \frac{w^\gamma - 1}{\gamma}, & 0 < |\gamma| \leq 1 \\ \log w, & \gamma = 0 \end{cases} \tag{2.9}$$

The final Mel-generalized coefficients (MGCs) are generated with the minimization of a cost function (generally the mean square linear prediction error) from an unbiased estimation of log spectrum.

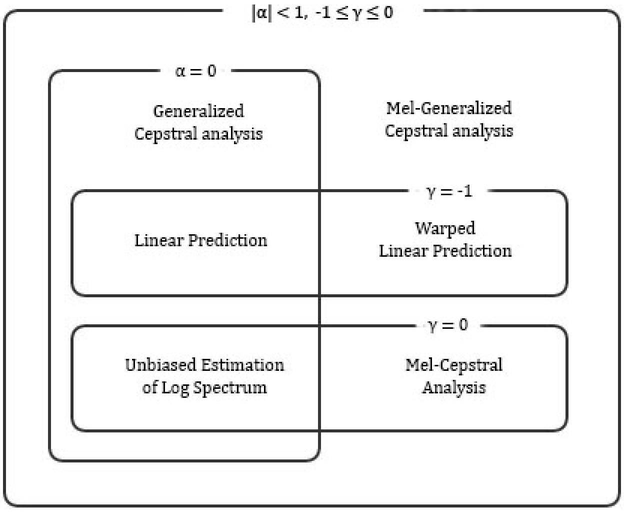
|α| < 1, -1 ≤ γ ≤ 0
α = 0
Generalized Cepstral analysis
Mel-Generalized Cepstral analysis
γ = -1
Linear Prediction
Warped Linear Prediction
γ = 0
Unbiased Estimation of Log Spectrum
Mel-Cepstral Analysis

Figure 2.6: General framework of the cepstral analysis [Tok94]

One interesting case of the MGC, due to its recent popularity in Voice Conversion ([Hel10; Tod07; Hel12], is the **Mel-cepstral coefficients** (MCC), where $\gamma = 0$. The spectrum is modeled as:

$$H(z) = \exp \sum_{m=0}^{M-1} c_\alpha(m) \Psi_\alpha^m(z) \tag{2.10}$$

being $M$ the number (order) of MCC, and $\alpha$ is set to 0.42 for 16 kHz of sampling frequency. The benefits of these parameters are quite similar to MFCCs, allowing the accurate modeling of both peaks and valleys of the spectrum and thus especially useful for the alignment.

# 3. Prosody of Speech

Prosody is a complex weave of physical, phonetic effects, appreciated on the rhythm, stress, tone and intonation patterns of an utterance, which is employed to express attitude, assumptions and attention as a parallel channel in the speech communication. It can reflect many features of the speaker or the utterance, such as the voice characteristics of the speaker, the emotional state of the speaker, the form of the utterance (statement, question or command) or the presence of irony or sarcasm.

In terms of speech processing, modeling the prosody is a problematic issue, since every person involves a particular and different way of speaking, and even the same person can also utter the same sentence in quite different prosodies. Thus, it is assumed the goal in prosody conversion is not a perfect match with some speaker target sentence, but creating a credible prosody, an utterance which can be spoken by the speaker in some situation.

This chapter reviews why the generated prosody is unique for every person in section 3.1, the main prosody phenomena in section 3.2, and finally the morphology of the prosody in section 3.3.

## 3.1 Speaking uniqueness

Prosody does not just depends on the linguistic content of the sentence, but on many other factors such as the speaker's voice and the emotional state of the speaker. Therefore, different people generate different prosody for the same sentence, and a same person could generate different prosody depending on the mood or the emotion which is feeling by the time the sentence is uttered. Another important factor affecting the prosody is obviously the language in which the sentence is uttered, since in some languages the different pronunciation of a word may affect the final meaning.

### 3.1.1 Character

In terms of prosody, it is referred as the long-term, stable and extralinguistic properties of the speaker. Many things can affect the character of a voice, from the individual personality to the membership of a group. Some idiosyncratic features, such as gender, age, physical state, or speech defects affect the prosodic character, and some social characteristics, such as speaker's region, economic status may also affect the prosody of the generated speech.

In Fig. 3.1 it appears the representation of two different speakers pitch profiles saying the same sentence, showing important differences among them: the level, range, duration of the utterance or the shapes of the contours can be clearly distinguished from one speaker to another.

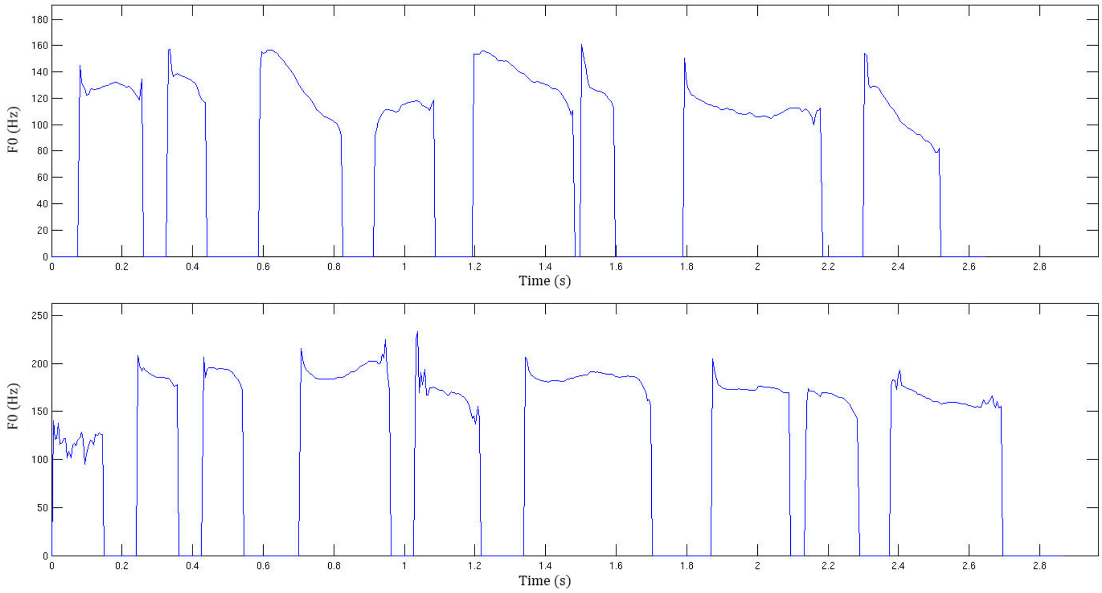

Figure 3.1: Two different speakers, a male and a female, uttering the same sentence "Don't you see, I'm chewing this thing in two.".

### 3.1.2 Emotion

Temporary emotions in the speaker produce an effect on the prosody of the uttered sentence, such as anger, happiness, disgust, fear or surprise. These emotions are generally independent from the speaker's character, since you can imagine any speaker with some social/dialect/gender/age characteristics being in several emotional situations, generating numerous emotional prosodic phenomena (Fig. 3.2).

A large number of high level factors affect the emotional features of speech: spontaneous or acted emotions, culture-depending or universal emotions, the correct interpretation of the emotion by the listener, the strength of the emotion; which produces remarkable differences between the same emotions uttered by different speakers.

Some examples can be clearly understood in the main studied emotions in literature: a controlled pitch sentence, with a low range and close to the monotonicity could be either interpreted as anger or sadness, while another kind of anger, more overtly expressive and with a wide, raised pitch range, is closer to the happiness. Although each emotion has its own characteristics, the difficulties of creating and recognizing them from speakers and listeners, respectively, are one of the challenging parts in emotion recognition and conversion.

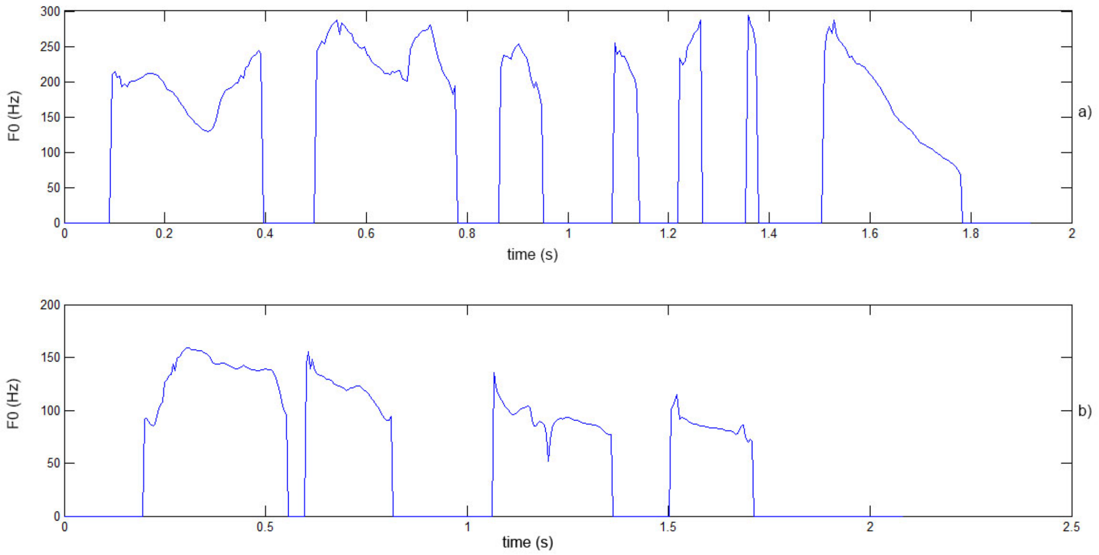

Figure 3.2: The german sentence 'Der lappen liegt auf dem Eisschrank.' for the same male speaker. Sentence a) is uttered in a anger way, meanwhile the sentence b) is uttered simulating boredom. The differences in the pitch range, maximum and duration of the sentence are remarkable.

## 3.2 Prosodic phenomena: stress, tone and intonation

The suprasegmental features of speech, producing changes in the loudness or the pitch of sounds, are commonly classified in stress, tone and intonation. Both of them are used in languages to add information at the text-alone sentence, and even modifying the meaning of the uttered sentence.

### 3.2.1 Stress

Comparing the word 'protest' in the following sentences:

(a) *It started as a student protest against rising tuition fees.*

(b) *The students organized a march to protest against these changes.*

it appears a remarkable difference of pronunciation. In 'student protest', the first syllable, 'PROtest', gets greater emphasis. This emphasis is called **stress**, and it is named that the first syllable is a *stressed* syllable, while the second syllable remains *unstressed*. Stressed syllables tend to be louder and somewhat longer than the unstressed ones. On the other sentence, 'protest' is used as a verb, generating a new stress pattern; the stressed syllable is the second one, 'proTEST'.

Moreover, in polysyllabic words, some syllables appear to be an intermediate degree of stress between stressed and unstressed syllables. Consider the word 'gymnast'; the main stress falls into the first syllable, 'gym-', but the second syllable,'-nast', has also some stress, known as secondary stress.

This type of stress, distinguishing words such as 'PROtest' and 'proTEST' is known as **lexical stress** or **word stress**. Another type of stress, known as **phrasal stress**, allows disambiguating sentences which the purely written form cannot:

*Mike repairs motorcycles.*

The neutral pronunciation of the sentence would provide an amount of stress to each syllable, although 'motorCYcles' would get slightly more stress. However, if the stress is put on the word 'MIKE repairs motorcycles', the sentence provides some extra information by centering the attention on the person who realizes the action, being the natural answer to the question 'Who repairs motorcycles?'. Finally, if the question is 'What does Mike do with motorcycles?', the logical stress of the answer would be 'Mike REPAIRS motorcycles'.

### 3.2.2 Tone

The tone or pitch of the voice is very important in language, since all language make use of it for some purpose. In some languages, such as Mandarin Chinese, Serbo-Croatian or Swahili, different words are even distinguished from each other by means of pitch [Lie67]. These different pitch profiles depending on the meaning of the word, are called **tones**. An example, in Fig. 3.3, is the word 'ma' in Mandarin Chinese:

| word | chinese character | meaning | tone |
|---|---|---|---|
| mā | 媽 | "mom" | — |
| má | 麻 | "hemp" | ／ |
| mǎ | 馬 | "horse" | ∨ |
| mà | 罵 | "scold" | ＼ |

Figure 3.3: Different meanings of the word "ma" in Mandarin, depending on the pronunciation tone.

Not only different levels of tones are used (high or low), even the evolution of the pitch during the course of the syllable affect the meaning of the word, qualifying Mandarin as a **tone language**.

English is not a tonal language, however the variations on the pitch are largely used, not for changing the meaning of the words, but generating intonation, the last prosodic phenomena.

### 3.2.3 Intonation

The variations in the pitches in non tonal languages produce the **intonation** pattern of an utterance, providing information about the attitude expressed by the speaker without, unlike tonal languages, modifying the meaning of the word.

Consider now the instances of the word 'me', where the pitch is represented graphically:

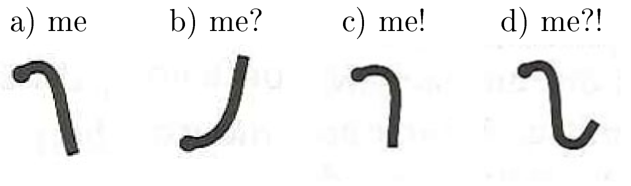

Figure 3.4: Different intonation patterns on the same word, generating different types of sentences.

The pitches has significant variations, however the meaning of the word remains unaltered. From a normal statement (a), a question (b), a strong assertation (c), to an expression of disbelief (d), the information provided due to the intonational pattern of the word has changed.

Unlike tonal languages, the tones generated cannot be assumed as part of a single word; in normal utterances, consisting of more than one syllable, the tone is generated over the whole utterance creating the suprasegmental intonation pattern.

All languages make use of **intonation**, including the tonal ones, however the exact use differs widely from one language to another and from one dialect to another. For instance, the intonational pattern of British English is completely different from the American English, giving the impression of pretentiousness or flattery, while American English sounds rude and pushy for British people.

## 3.3 Prosodic morphology

The prosodic constituent structure of an utterance has generally been proposed [Sel80; Nes86] to be derived rather directly from the morphosyntactic constituent structure of the sentence: syllable, word, phrase and utterance. In many cases the correspondence between these levels and the prosodic levels is quite accurate, however it is not exact:

(1) *He sat down at the table ][ with the vase of flowers.*

(2) *He ][ sat down at the table with the vase of flowers.*

On the first sentence, the prosodic boundary, represented by ’][’, is located on the approximately middle, generating a normal intonation pattern, whereas the second intonation pattern is close to the syntactic boundary (subject-verb) but has no sense in terms of intonation.

With this trivial example, can be noticed that syntactic structure influences the prosodic structure, but other factors, e.g. speaking rate, number of words and syllables, semantic focus or discourse structure, also affect the final prosodic structure.

Current prosodic theories have postulated prosody as a hierarchical system of constituents and the existing relation between pairs of constituting elements on the same level [Lib77; Sel80; Sel86; Bec86], in order to model all these factors which affect the final suprasegmental realization of speech. The constituents of the prosodic hierarchy are normally classified as:

- Intonational phrase
- Phonological phrase
- Prosodic word
- Stress foot
- Syllable
- Mora

A **mora** is a timing unit; each mora takes approximately the same length of time to say. Moreover, it is also the unit which allows the measurement of the weight of the syllable [Pri83; Hym85]. Usually heavy syllables, the ones which contain more than one **mora**, are those that consist of a long vowel or diphthong, such as ’rain’ or ’see’, meanwhile light syllables usually consist of a short vowel, such as the second syllable of ’father’.

The **syllable** constituent corresponds to the grammatical syllable, as well as the **prosodic word** corresponds respectively to the grammatical word.

The **stress foot** is a prosodic level constituted by at least one stressed syllable and usually an unstressed syllable. Consider the words ’modest and ’gymnast’(Fig.

3.5). Both first syllables are stressed, 'mo-' and 'gym', but it exists a remarkable difference on the second syllable: meanwhile the syllable '-dest' is an unstressed one, the syllable '-nast' has a secondary stress, generating a new stress foot.

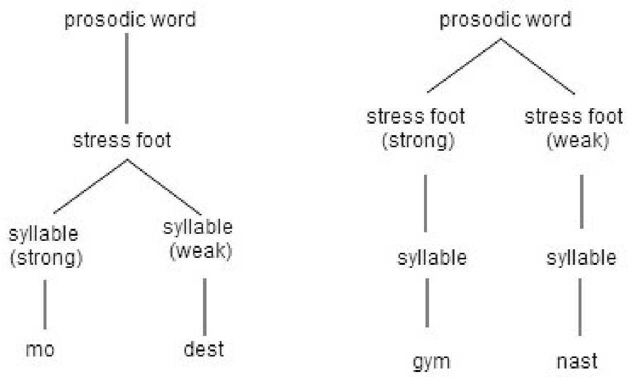

Figure 3.5: Prosodic hierarchy of words 'modest' and 'gymnast'.

However, the two **stress feet** do not have the same amount of weight on the word, considering 'gym-' as a strong stress foot and '-nast' as a weak foot.

The **phonological phrase** is conformed by at least two prosodic words. Contemplating again the sentence 'He sat down at the table with the vase of flowers.', there exist many possible separation into **phonological phrases** but the more logical one might be:

[*He sat down*] [*at the table*] [*with the vase*] [*of flowers.*]

The **intonational phrase** conforms the major prosodic level. It is normally referred as the union of one or more phonological phrases, and normally appear a couple of intonational phrases within a syntactic sentence. In the previous example, the sentence can be uttered separating the speech in a couple of **intonational phrases**, with a small pause in between, or can be uttered in a joint biggest intonational phrase:

[*He sat down at the table*] [*with the vase of flowers.*]

[*He sat down at the table with the vase of flowers.*]

Upper and lower levels can be defined, contemplating i.e. an utterance level or a phoneme level, describing the whole intonation of the speech or the microprosody, respectively. However, the main discussion of linguistics regarding the prosodic levels, is the appearance of middle levels between the prosodic word and the phonological phrase, such as the **clitic group** or the **prosodic list unit**, modeling intermediate prosody phenomena.

# 4. Voice Conversion

**Voice Conversion** (VC) is an area in speech processing, which deals with the conversion of the speaker identity: the original speech produced by a first speaker is transformed to sound as if it was uttered by a second speaker. VC systems have several potential applications, such as hiding the identity of the speaker, voice restoration and vocal pathology, or dubbing games or movies, however the main use is creating new voices in text-to-speech synthesis in a cost-efficient manner.

This chapter reviews the state of the art of VC framework, giving special emphasis to the used methods in the proposed experiments. It is organized as follows: section 4.1 describes the general architecture of the VC systems, section 4.2 reviews the analysis/synthesis framework, section 4.3 deals with the main alignment techniques for parallel data and section 4.4 describes the main mapping functions used to model the conversion.

## 4.1 Architecture of Voice Conversion systems

Currently, all VC systems are based on two main steps: the training phase and the conversion phase (Fig.4.1).

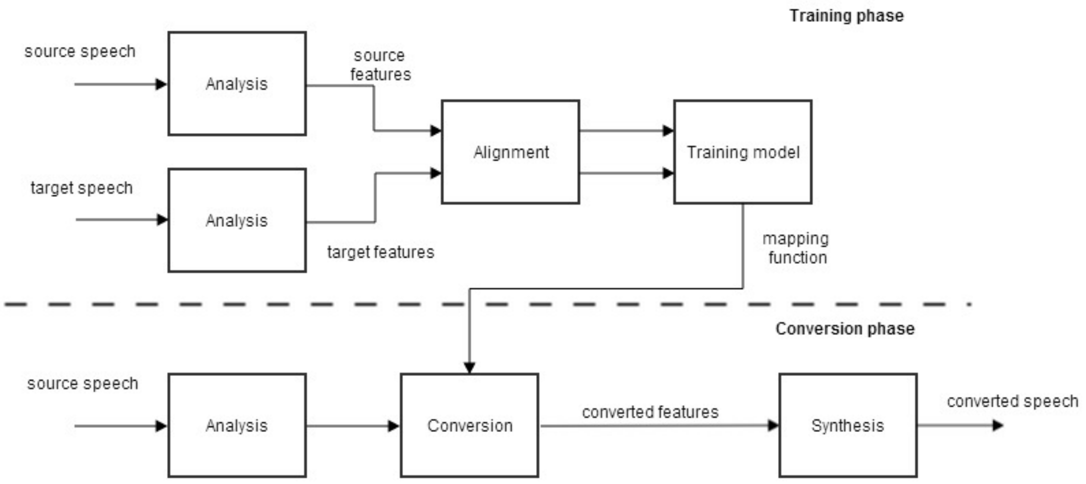

Figure 4.1: Blog diagram of VC architecture

In the training phase, speech samples of both source and target speakers are analyzed, extracting its characteristic speech features. In some VC approaches [Hel07] [Dux04], it is used additional information, such as syllable boundaries, intonation marks or phonetical information. After the features are extracted, and are correctly aligned in case of parallel corpus database[1], the model is trained and the mapping function between source and target is created. In the conversion phase, the source features are extracted as well, and the trained mapping function is applied to them in order to convert the original speech to the target one, with the voice features of the new speaker.

In the next subsections there is a deeper description for the analysis and synthesis phase, the alignment phase, and the mapping functions to create the model.

## 4.2 Analysis/Synthesis framework

The first step and one of the most important ones of any VC is the analysis of the voice signals. It is crucial using an analysis system which provides a high quality analysis of the speech signals and thereby a high quality on the extracted features.

The simplest system used for extracting the speech features is the LPC vocoder. As other LP-based coders, the representation of the spectral envelope is carried out using an all-pole filter and the excitation is modeled as a two-category decision: white noise for unvoiced signals or a sequence of impulses for voiced signals, with a spacing of the pitch period. Consequently, for the frames decided as voiced frames, it is needed an estimation of $F_0$.

Another group of analysis systems are based on a sinusoidal representation: in [Nur06] they assume the excitation signal as a sum of sine waves. The sinusoids can be classified as continuous or random-phased; the first ones represent voiced speech and are modeled using $F_0$ harmonics as frequency and linearly evolving phase while the random-phased are used to model the unvoiced speech, using a fixed value of $F_0$ as a frequency and a random phase.

A different approach on the sinusoidal representation-basis is as harmonic plus noise model (HNM)[Sty05]. It is based on an harmonic modeling for the periodic frames (voiced), using multiples of $F_0$ as frequencies of the sinusoids, and a noise model for the non-periodic (unvoiced), obtained by subtracting the periodic-part from the original speech signal.

[1] A parallel corpus database consists of speech recordings produced by different speakers uttering the exact same sentences.

### STRAIGHT

One of the most used analysis/synthesis (A/S) systems is the one proposed by [Kaw99], known as STRAIGHT. It is a high quality analysis/synthesis method, which uses pitch-adaptative spectral analysis combined with a surface reconstruction method in the time-frequency region. The speech waveform is decomposed into $F_0$ contour and spectrum, but also an aperiodicity map is extracted.

The $F_0$ estimation is performed, under the assumption of a nearly harmonic structure of the speech signal, by a series analyzing continuous wavelet transform, and the final pitch is selected as the one having higher signal to noise ratio of the sinusoidal component and background noise. On the unvoiced frames, no fundamental frequency is detected, returning the expected 0 pitch value (Fig. 4.2).

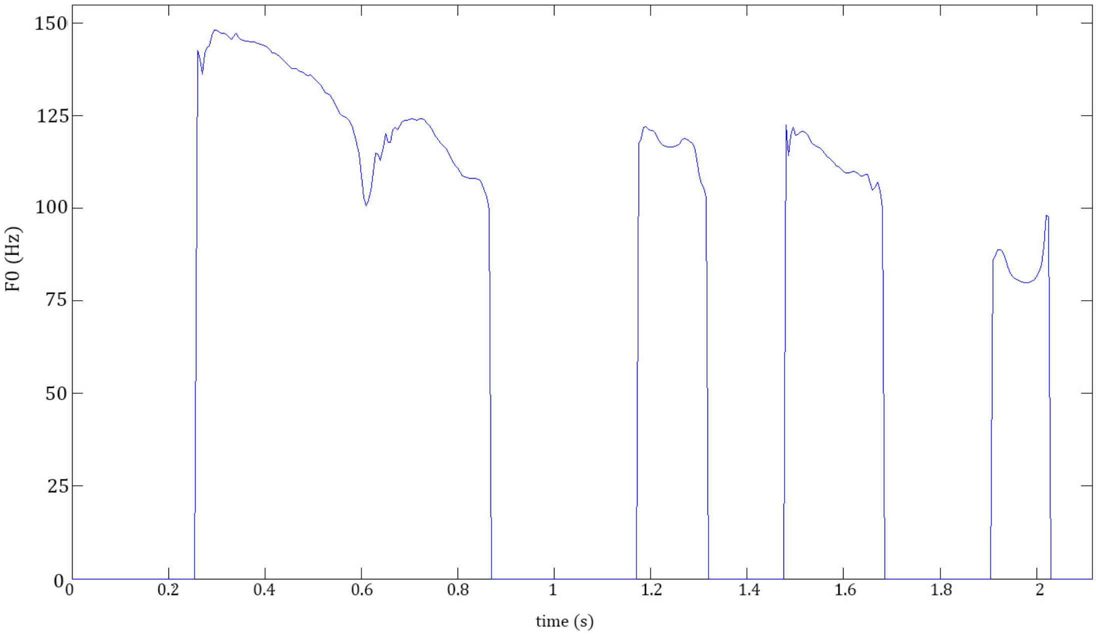

Figure 4.2: STRAIGHT pitch estimation from the sentence "The singing voice approached rapidly", uttered by a male speaker.

The aperiodicity map (Fig. 4.3) represents the deviations from periodicity, which introduce additional components on inharmonic frequencies. It is extracted jointly with the pitch estimation, providing this additional information concerning the source of the speaker that is not represented in the $F_0$ extraction.

Finally, the spectral analysis shows the spectral features of speech with no trace of the periodicity due to the fundamental frequency. The filters used for the extraction of the spectrogram benefits of the previously estimated fundamental frequency to set its bandwidth and, by using a smoothing function to deal with small $F_0$ variations, extract the spectral properties of speech (Fig. 4.4).

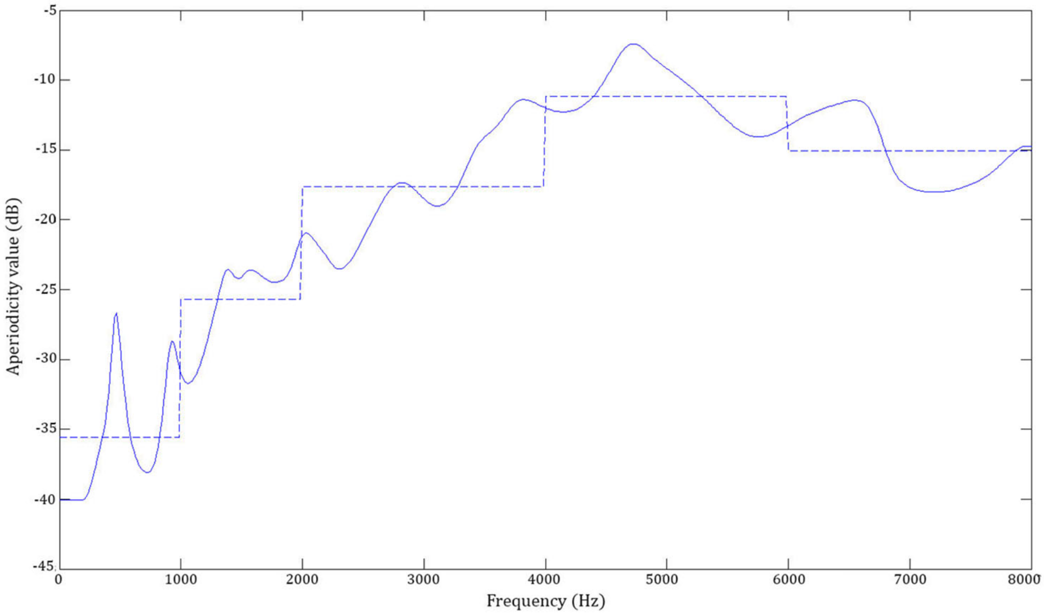

Figure 4.3: STRAIGHT aperiodicity map and the corresponding aperiodicity bands.

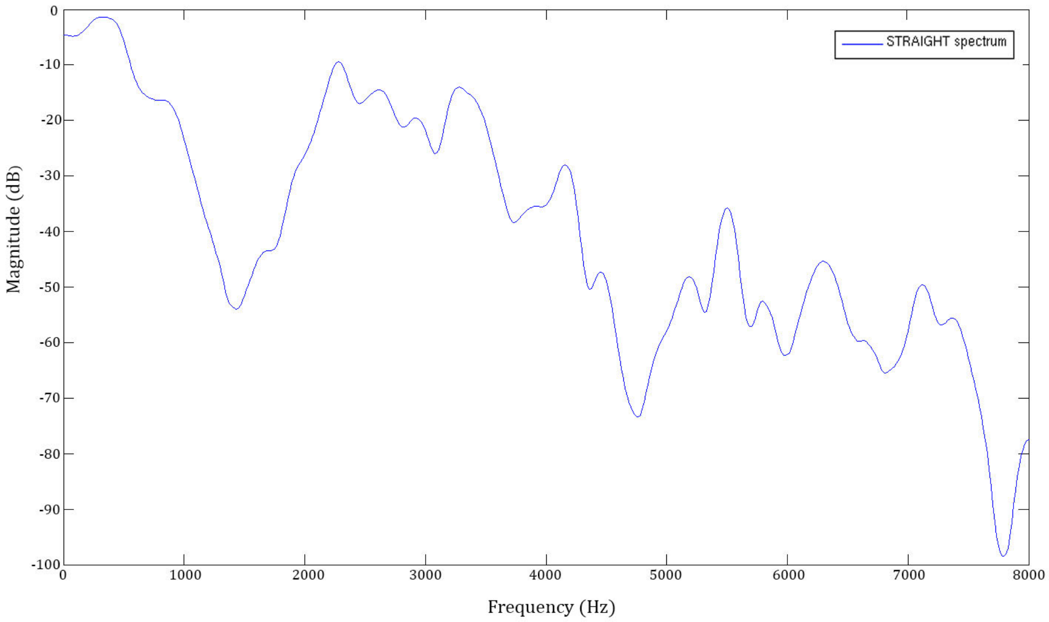

Figure 4.4: STRAIGHT spectrogram estimation from a voiced 5 ms. frame.

## 4.3 Alignment in parallel corpus

Generally, one same sentence uttered by two different speakers is never produced at the same speak rate: neither loudness nor speed are the same for the same word when it is uttered by two speakers. Therefore, when working with parallel data, equal sentences uttered by different speakers, the utterances must be aligned in time so they can preserve linguistic correspondence.

The simplest alignment method is a linear time scaling: the main assumption is the speaking rate is proportional to the duration of the sentence and independent of individual sounds, thus, the two utterances are stretched or compressed linearly so that they become the same length. This way of alignment works reasonable well for monosyllabic utterances but when multisyllabic sentences are included the performance decreases.

The main approach of alignment, in order to work properly with multisyllabic utterances, is dynamic time warping (DTW), which is discussed in the following subsection. Other approaches for aligning parallel data are HMM-based alignments [Err10b], forced-alignment speech recognition or even manual alignment, when phoneme boundaries are available.

For non-parallel data, either intra-lingual VC or cross-lingual VC, it is still necessary an alignment step in order to train the conversion functions. The main approach in literature is the INCA algorithm [Err10a], based on an iterative combination of a Nearest Neighbor search combined with a voice conversion, which in the next iteration is used as the source voice.

### 4.3.1 Dynamic Time Warping

The purpose of Dynamic Time Warping (DTW) [Sak78] is finding an optimal alignment between two given (time-dependent) sequences under terms of a distance function. Given the two sequences $X = (x_1, x_2, \dots, x_N)$ and $Y = (y_1, y_2, \dots, y_M)$, of respective length $N$ and $M \in \mathbb{N}$, is built a cost matrix $C = \mathbb{R}^{NxM}$. Then the goal is having an alignment between $X$ and $Y$, by having minimal overall cost. Intuitively, the optimal alignment runs through the cost matrix $C$, along a low cost path.

Although several ways of computing cost matrices exist, Eq. 4.1 shows the main approach to built the cost matrix:

$$C(i,j) = d(x_i, y_j) + \min\left(C(i-1, j-1), C(i, j-1), C(i-1, j)\right), \tag{4.1}$$

where $C(i,j)$ is the cumulative distance of point $(i,j)$ and $d(x_i, y_j)$ is the distance at the current point. The total cost of the optimal alignment is returned by $C(N,M)$, and the optimal alignment can be obtained by path backtracking.

This path requires some constrictions in the final warping function: endpoint constrains, monotonicity constrains and step size constrains need to be accomplished in order to get a significant warping path function to compare both sequences:

- Boundary condition: $p_1 = (1,1)$ and $p_L = (N, M)$.
- Monotonicity condition: $n_1 \leq n_2 \leq \ldots \leq n_L$ and $m_1 \leq m_2 \leq \ldots \leq m_L$.
- Step size condition: $p_{l+1} - p_l \in \{(1,0),(0,1),(1,1)\}$ for $l \in [l : L-1]$.

A common choice in speech processing is using Euclidean distance as distance measure, and MCCs or MFCCs as alignment features.

## 4.4 Mapping functions

The model which allows converting the speech characteristics from source data to target data is constructed during the training phase, creating a conversion function to map the source feature vector $x_n$ into the target feature vector $y_n$ for each frame $n$.

The first approaches for forming the conversion function were based on a codebook mapping. The main idea in the codebook approaches for VC systems is generating a mapping codebook describing a function between the vector spaces of two speakers. In [Abe88] the source and target feature vectors are quantized frame by frame, and its correspondence is determined using DTW. The mapping function is a linear combination of the target vectors, based on an histogram as a weighting function.

In the recent VC systems are using statistical techniques to find a conversion function $\mathcal{F}(\Delta)$, minimizing the prediction error $\epsilon$:

$$\epsilon = \sum_{n=1}^{N} \|\mathbf{y}_n - \mathcal{F}(\mathbf{x}_n)\|^2 \tag{4.2}$$

One of the main techniques in VC [Sty98; Kai01; Tod07; Che03] to model the distribution of the source and target features vectors is using a Gaussian mixture model (GMM) applied to the spectral features of speech, which is described in Sec. 4.4.1.

An alternative method recently proposed by [Hel12; Sil13], to model the spectral features of speech is the Dynamic Kernel Partial Least Squares (DKPLS). It is an statistical mapping that allows non-linear conversion and improves temporal continuity, since it allows handling the dynamics of speech. It is described in Sec. 4.4.2.

### 4.4.1 Gaussian mixture model

A GMM is a probability density function built as a weighted sum of $M$ Gaussian components:

$$p(\mathbf{x}) = \sum_{m=0}^{M-1} \boldsymbol{\alpha}_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) \tag{4.3}$$

where $\boldsymbol{\alpha}_m$ is the prior probability of the $m^{th}$ Gaussian component and $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)$ is the $p$-dimensional Gaussian function:

$$\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m) = \frac{1}{(2\pi^{p/2}\sqrt{|\boldsymbol{\Sigma}_m|})} \exp\left[-1/2(\mathbf{x} - \boldsymbol{\mu}_m)^T \boldsymbol{\Sigma}_m (\mathbf{x} - \boldsymbol{\mu}_m)\right] \tag{4.4}$$

with $\mathbf{x} = [x_0, x_1, \ldots, x_{p-1}]^{-1}$ a $p$-dimensional random vector, $\boldsymbol{\mu}_m$ the $p$-dimensional mean vector and $\boldsymbol{\Sigma}_m$ the $p \times p$ covariance matrix of the Gaussian distribution. In order to define properly GMM as a probability density function, two extra restrictions must be taken into account: the scalar mixture weights of the GMM must be non-negative, $\alpha_m \geq 0$, $\forall m = 0, \ldots, M-1$, and normalized to 1, $\sum_{m=0}^{M-1} \alpha_m = 1$.

The most popular way to estimate the GMM parameters from the set of source vectors is the expectation maximization (EM) algorithm. The EM algorithm [Dem77] iteratively increases the likelihood of the model parameters by successive maximizations of auxiliary functions.

There are two main approaches in the literature in order to create the mapping function with GMM: just using source features or using both source and target feature vectors to be fit in a GMM .

### GMM with source data

This first approach, proposed in [Sty98], uses the minimum mean square error to estimate the source-target mapping function, assumed to be linear for each Gaussian:

$$\mathbf{x}' = \mathcal{F}(x) = \sum_{m=0}^{M-1} \mathcal{P}(\mathcal{C}_m|\mathbf{x})[\boldsymbol{\nu}_m + \boldsymbol{\Gamma}_m \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_m)] \tag{4.5}$$

where $\boldsymbol{\nu}_m$ denotes the mean target vector belonging to the $m^{th}$ Gaussian, $\boldsymbol{\Gamma}_m$ is the cross-covariance matrix of the source and target vectors:

$$\boldsymbol{\Gamma}_m = E[(\mathbf{y} - \boldsymbol{\nu})(\mathbf{x} - \boldsymbol{\mu})^T] \tag{4.6}$$

and $\mathcal{P}(\mathcal{C}_m|\mathbf{x})$ is the posterior probability of vector x:

$$\mathcal{P}(\mathcal{C}_m|\mathbf{x} = \frac{\boldsymbol{\alpha}_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m, \boldsymbol{\Sigma}_m)}{\sum_{j=0}^{M-1} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} \tag{4.7}$$

The unknown parameters $\boldsymbol{\nu}_m$ and $\boldsymbol{\Gamma}_m$ are computed using least squares approach.

### Joint density GMM

The other mapping model based on GMMs, proposed by [Kai01], considers modeling the joint density of the target and source data with a GMM:

$$p(\mathbf{z}_n) = \sum_{m=0}^{M-1} \boldsymbol{\alpha}_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}) \tag{4.8}$$

where $\mathbf{z}_n = [\mathbf{x}_n^T, \mathbf{y}_n^T]^T$ is the source vector augmented with the target vector, and

$$\begin{aligned} \boldsymbol{\mu}_m^{(z)} &= \begin{pmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{pmatrix} \\ \boldsymbol{\Sigma}_m^{(z)} &= \begin{pmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{pmatrix} \end{aligned} \tag{4.9}$$

are the augmented mean and covariance.

The conversion function is also computed as the one which minimizes the mean squared error between converted source and target vectors:

$$\mathbf{x}' = \mathcal{F}(x) = \sum_{m=0}^{M-1} \mathcal{P}(\mathcal{C}_m|\mathbf{x})[\boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}\boldsymbol{\Sigma}_m^{(xx)}(\mathbf{x} - \boldsymbol{\mu}_m^{(x)})] \tag{4.10}$$

$$\mathcal{P}(\mathcal{C}_m|\mathbf{x}) = \frac{\boldsymbol{\alpha}_m \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)})}{\sum_{j=0}^{M-1} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_j^{(x)}, \boldsymbol{\Sigma}_j^{(xx)})} \tag{4.11}$$

## 4.4.2 Dynamic kernel partial least squares

DKPLS is a statistical mapping technique based on two steps: a pre-processing step consisting of a kernel transformation of the source data and a converting step where partial least squares (PLS) regression is used to estimate the new features. The kernel data is augmented with the previous and following frame, being able to characterize the dependencies between consecutive frames.

### Kernel transformation

The concept of a kernel is a data matrix where exists similarity measures for a specific dataset. In this method the kernel matrix $K$ is built based on a Gaussian transformation of speech features:

$$k_{jn} = e^{-\frac{\left\|x_n - c_j\right\|^2}{s\sigma^2}}, \tag{4.12}$$

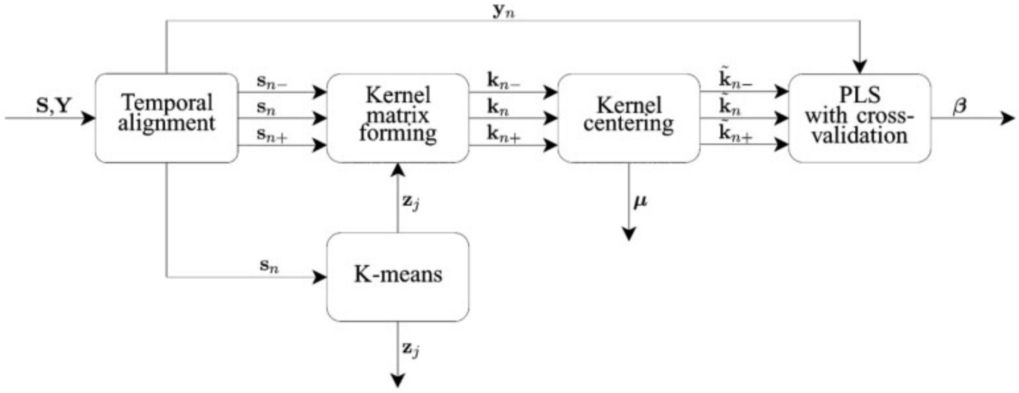

Figure 4.5: Overview of the training procedure using DKPLS

leading to the final kernel matrix

$$\mathbf{K} = \begin{bmatrix} k_{11} & k_{12} & \dots & k_{1N} \\ k_{21} & k_{22} & \dots & k_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ k_{C1} & k_{C2} & \dots & k_{1N} \end{bmatrix} \tag{4.13}$$

where $k_{jn}$ are the final entries of the kernel matrix, $x_n$ are the source features, $c_j$ are the reference vectors, and $\sigma$ is a scaling parameter. The $C$ reference vectors are obtained by k-means algorithm, where the found cluster centers act as reference vectors.

Finally, before carrying out the PLS regression, the kernel has to be centered to force the bias term of the conversion to zero. Since in the kernel space the mean cannot be computed automatically, it is necessary some accuracy in this step [Emb04] : first it is calculated the average of each row and stored for later use, then the average of each column is computed and subtracted from the kernel matrix resulting of row-centering. The new centered kernel is denoted by $\tilde{K} = [\tilde{k}_1, \tilde{k}_2, \dots, \tilde{k}_n, \dots, \tilde{k}_N]$ .

The kernel of the test data needs to be centered similarly: the saved row averages are subtracted from the testing kernel, and the column average is computed and subtracted from the obtained row-centered testing kernel.

### Dynamic modeling

In order to model the time continuity of the speech features, a redundant problem in VC [Hel10; Tod07], the source data is augmented with its previous and next frame data. This allows modeling the dynamic relations of the data, smoothening transitions from frame to frame, and consequently building higher quality models. The predictor variables become:

$$\mathbf{X}_n = [\tilde{\mathbf{k}}_{n-}, \tilde{\mathbf{k}}_n, \tilde{\mathbf{k}}_{n+}] \tag{4.14}$$

being $\tilde{\mathbf{k}}_n$ the centered kernel vector and $\tilde{\mathbf{k}}_{n-}$ and $\tilde{\mathbf{k}}_{n+}$ the centered kernel vectors for the previous and next frames.

### Partial Least Squares Regression

The prediction of **Y** variables from the observed variables **X** is done by a regression model defined as:

$$\mathbf{y}_n = \boldsymbol{\beta}\mathbf{x}_n + \boldsymbol{\epsilon} \tag{4.15}$$

where $\mathbf{x}_n$ are the source observation vectors, $\mathbf{y}_n$ are the target observation vectors, $\epsilon$ denotes the regression residual, and $\beta$ is the regression matrix.

In the DKPLS method, due to the kernel transformation, $\tilde{\mathbf{k}}_n$ becomes linearly dependent, and the addition of the dynamics at the final source observation vector $\mathbf{x}_n$ introduces collinearity. Partial least squares (PLS) regression is a technique for predictive modeling of the relationships between predictor matrix **X** and response matrix **Y**, which can deal with the collinearity of the observation vectors and cases where the number of observations is less than the number of variables.

To perform the regression task, PLS constructs new explanatory variables, called latent vectors, which are a linear combination of the original $x_1, x_2, \ldots, x_N$ vectors. The aim of these vectors is explaining the most relevant information in the **X** variables that is also useful for predicting **Y**. This is a similar way of working to principal component analysis (PCA), but the difference remains in meanwhile PCA just uses **X** to determine the principal components, PLS uses both **X** and **Y** for extracting the latent vectors.

## 4.5 Prosody modeling and conversion in speech processing

The modeling of the pure linguistic prosody phenomena is problematic, since there are no quantitative measures available, and hence they need to be re-estated in other parameters able to be treated and modified. In speech processing are normally parametrized as:

- Pauses: the gaps between words and separating phrases, normally related to punctuation marks.

- Pitch: the rate of vocal-fold cycling (fundamental frequency or $F_0$) as a function of time.

- Rate/relative duration: the phoneme and syllable durations, timing and rhythm of the sentence.
- Loudness: the relative amplitude of the sentence.

Since pitch is the most expressive of the prosodic phenomena, most of the prosody conversion methods in literature are based on a frame-level conversion of the pitch. However, in the last years, new approaches have been working based on a local contour of $F_0$, normally at syllable level, also adding linguistic information and timing references, being able to model the timing parameters as well.

### Mean and standard deviation scaling

The first approach to transform $F_0$ is a simple scaling, referred to as mean and scaling (MS) method. In order to obtain the $F_0$ of the target speaker $f_t$, the following transformation is computed to the $F_0$ of the source speaker $f_s$:

$$f_t = \frac{\sigma_{f_t}}{\sigma_{f_s}}(f_s - \mu_{f_s}) + \mu_{f_t} \tag{4.16}$$

where $\mu_{f_t}$, $\sigma_{f_s}$, $\mu_{f_t}$, $\sigma_{f_s}$ represent the mean and standard deviation of the $F_0$ values for the target and source, respectively. This mapping function is computed based on the assumption that each speaker's $F_0$ values belong to a Gaussian distribution with a specific mean and variance [Ric95]. Although this first transformation method, does not really convert the prosody appearing in $F_0$, is the most common way of converting $F_0$ since results are good enough, as has been shown in [Ina03; Hel07]. This MS method is also performed in logarithmic domain, since the human perception of sound is in a logarithmic scale:

$$\log f_t = \frac{\sigma_{\log f_t}}{\sigma_{\log f_s}}(\log f_s - \mu_{\log f_s}) + \mu_{\log f_t} \tag{4.17}$$

A benefit of MS method is that parallel data is not required, but in terms of prosody conversion, it keeps the shapes and contour of source $F_0$ and is unable to model small changes depending on the target speaker prosody. Thus, technically prosody is not converted.

### Polynomial conversion

This method was proposed by is an improvement of the MS: a higher-order mapping function is estimated without the assumption of a Gaussian distribution. Based on

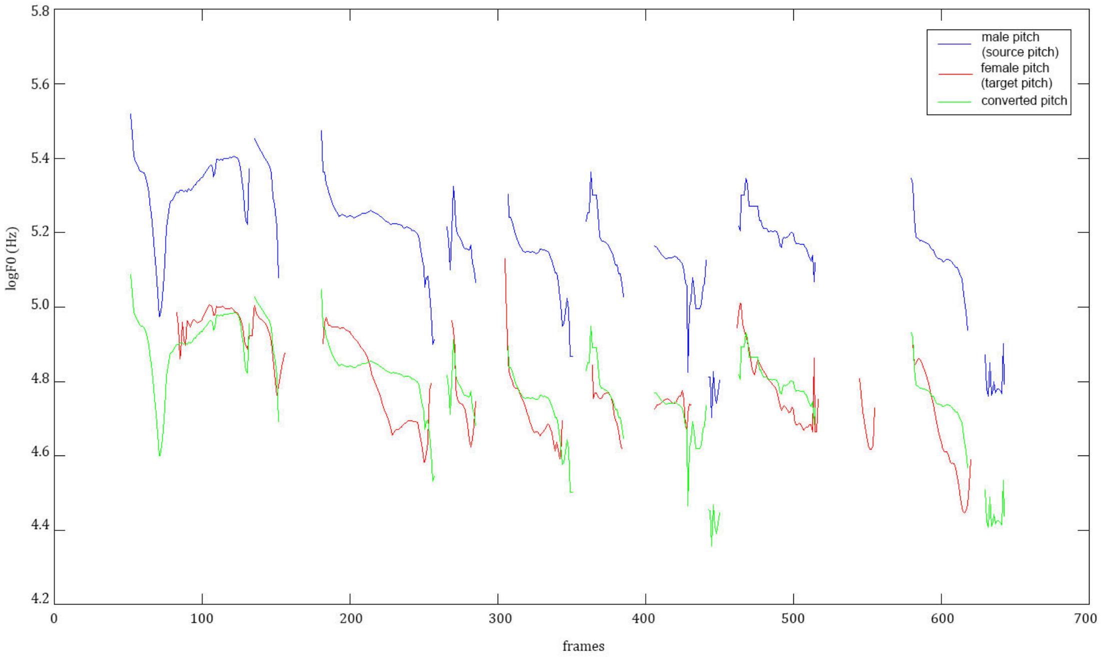

Figure 4.6: Representation of the pitch of two different speakers, male and female, and the converted pitch using the MS method.

the scatterplot model of mean pitch for each speaker least squares method is used to compute the Nth order polynomial function.

### Contour codebook and dynamic time warping

Another approach in prosody conversion is the one presented in [Ina03]. In contrast of the previous approaches, this codebook method is working with the utterance contour instead of at a frame-level, trying to impart an entire pitch contour. Here it is used DTW algorithm to select the closest pitch contour from an available training sentence database. This helps minimize lexical stress while maintaining the large-scale intonation differences between the speakers' utterances. Finally the target pitch contour is warped in order to generate the new pitch, helping maintaining some characteristics of the intonation pattern but also adding those new characteristics of the new speaker.

### Piecewise linear mapping using intonation marks

In [Gil03] is proposed using intonational marks for creating the mapping function able to model the prosody. The $F_0$ contours are parametrized based on four selected points: sentence-initial high, sentence-final low, non-initial accent peaks and post-accent valleys. For each sentence there is one sentence-initial high and sentence-final low, appearing several number of peaks and valleys. The piecewise mapping function is constructed based on the different union of these source and target pair points: post-accent valleys and sentence-final low, post-accent valleys and non-initial peaks, and non-initial peaks and sentence-initial high.

### Syllable codebook with regression trees

The prosody conversion method proposed by [Hel07], introduces a syllable-level codebook containing paired source and target $F_0$ contours. These contours are compressed using discrete cosine transform (DCT), which allows fast comparison and avoids using DTW techniques. A first selection is done between the source speaker $F_0$ and the codebook, selecting possible target candidates. The final decision is based on a classification and regression tree (CART) trained with linguistic and durational information.

### Duration conversion based on regression trees

Another approach using extra information to model the prosody, is the one proposed in [Ina07]. The $F_0$ contours are modeled based on the global intonation contour and the duration of phones. For the modeling of the intonation, they use three-state left-to-right HMMs for each syllable, while the duration is modeled by CARTs using information such as phone identity, previous phone, next phone, lexical stress, word position or word length.

# 5. Prosody modeling and conversion with wavelets

The assumption of a hierarchic model of prosody by phonologists and phoneticians is largely accepted. Recently, some approaches working with a hierarchic prosody model [Wan08] [Lei10], not just in a frame wise of $F_0$, have put in relevance that there is important information, in terms of speech processing, in every linguistic level of the utterance, from microprosody information on phonemes, to the whole prosody of the sentence and utterance.

A function or signal, and certainly speech signals, can be understood as compositions of smooth backgrounds and details on top of it, and if we speak in terms of frequency, the high and low frequencies of the signal. The **wavelet transform** is a tool that splits data into different frequency components, and allows studying each component with a frequency resolution matched to its scale [Mal98].

Due to this multiresolution analysis properties of the wavelets, which allows the study of multiple frequency levels of the input sequence keeping the temporal localization, the wavelet transformation has been lately a trend in many fields, such as physics and chemistry, but specially in signal processing: it has been frequently used in image processing for compressing and denoising images, but also in speech recognition and speech synthesis. [Sun13] has used the wavelet transform to model the prosody of speech and represent the hierarchic model. Moreover, [Vai13] has also used the wavelet transform as an analysis tool, developing a system for an automatic detection of word prominence applicable to a high quality speech synthesis system.

This chapter introduces a review of the theoretical basis of the wavelets in 5.1, the capability of the wavelet transformation to model the hierarchic prosody model in 5.2, by decomposing $F_0$ contours into different wavelet levels, close related to linguistic prosody levels and displaying its containing prosodic information. Finally, section 5.3 presents a new prosody conversion method, based on the properties of the wavelet analysis.

## 5.1 Theoretical approach to wavelets

The standard Fourier transform also gives a representation of the frequency content of the signal, but information concerning time-localization of the interesting frequency is lost. The short-time Fourier transform, allows getting time-frequency localization, by windowing the signal in small frames, and then computing its Fourier Transform.

However, the time-frequency localization properties are inherent in the wavelet transformation, so that the difference between the windowed Fourier transform and the wavelet transform remains in the transformation function: while Fourier transform always consists of the same envelope, translated to the proper time location and "filled in" with higher frequency oscillations from the data, the wavelet transform has time-width adapted to their frequency: high frequency transformation functions are very narrow, while low frequency functions are much broader. The result is the wavelet transform is better able to capture both low frequency and very brief high frequency phenomena.

The **Continuous Wavelet Transform** (CWT) is the basis of the wavelet analysis [She96], and it is written as:

$$T^{wav}_{s,\tau} = |s|^{-1/2} \int dt f(t) \psi \left( \frac{t-\tau}{s} \right) \tag{5.1}$$

This equation produces the decomposition of the function $f(t)$ into the different $\psi_{t,\tau}$ wavelets, all of them generated from the mother wavelet (Eq.5.2) by adjusting $s, \tau$, the scaling and translation parameters.

$$\psi_{s,\tau}(t) = \frac{1}{\sqrt{s}} \psi \left( \frac{t-\tau}{s} \right) \tag{5.2}$$

### 5.1.1 Basis of multiresolution analysis

Considering a function $f(t)$ and labeling the resolution level by $j$ the scale below which all fluctuations on that resolution are ignored is $1/2^j$. The function that approximates $f(t)$ is $f_j(t)$. At the next resolution level $j+1$, the details, denoted by $d_j(t)$, are included in function $f_{j+1}(t) = f_j(t) + d_j(t)$. This procedure can be repeated several times. The function $f(t)$ can be viewed as

$$f(t) = f_j + \sum_{k=j}^{k=\infty} d_k \tag{5.3}$$

Similarly, the space of square integrable functions $\mathbf{L}^2(\mathbb{R})$ can be viewed as compositions of subspaces $W_k$ and subspace $V_j$. $W_k$ contains details $d_k(t)$. The subspace $V_j$, contains $f_{j(t)}$ approximation of function $f(t)$ on resolution level $j$.

The requirements of multiresolution analysis are [Dau92]:

1. Subspace $V$, must be contained in all subspaces on higher resolutions

$$\ldots \subset V_0 \subset V_1 \subset \ldots \subset \mathbf{L}^2(\mathbb{R}) \tag{5.4}$$

2. All square integrable functions must be included at the finest resolution level (5.5) and only zero function on the coarsest level

$$\overline{\cup_j V_j} = \mathbf{L}^2(\mathbb{R}) \tag{5.5}$$

$$\cap_j V_j = \{0\} \tag{5.6}$$

3. All the spaces $\{V_j\}$ are scaled versions of the central space $V_0$. If $f(t)$ is in space $V_j$ and it contains no details on scales smaller than $1/2^j$, then function $f(2t)$ contains no details on scales smaller than $1/2^{j+1}$ and it is from space $V_{j+1}$.

$$f(t) \in V_j \Leftrightarrow f(2t) \in V_{j+1} \tag{5.7}$$

4. If $f(t) \in V_0$, so do its translates by integer $k$, $\{f(t-k)\}$.

$$f(t) \in V_0 \Leftrightarrow f(t-k) \in V_0 \tag{5.8}$$

Once these properties are satisfied by a ladder of spaces $V_j$ there exists a function $\psi(t) \in V_0$ so that $\{\psi(t-k)\}$ constitute an orthonormal basis for $V_0$.

### 5.1.2 Discretization of the wavelets

Since the CWT has infinite wavelets representations, depending on the values of $s$ and $\tau$, it is not practical using it as an analysis tool. Moreover, depending on the selected values of the translation and scaling parameters, the mother wavelet could constitute an orthogonal basis [Dau92], facilitating the posterior reconstruction.

The Discrete Wavelet Transform (DWT) is defined by the discretization of $s$ and $\tau$:

$$T^{wav}_{m,n} = |a_0|^{-m/2} \int dt f(t) \psi\left(\frac{t - nb_0 a_0^m}{a_0^m}\right) = |a_0|^{-m/2} \int dt f(t) \psi\left(a_0^{-m} t - nb_0\right), \tag{5.9}$$

where $a_0 > 1$, $b_0 > 0$, and $m$ and $n$ range over $\mathbb{Z}$. The selection of the values of $a_0$ and $b_0$ is important for the correct representation of the information contained in $f(t)$, and the fact that, for some special values, the $\psi_{m,n}$ constitute an orthonormal basis for $\mathbf{L}^2(\mathbb{R})$. In particular, if $a_0 = 2, b_0 = 1$, there exist $\psi$ with good time-frequency localization properties, such that

$$\psi_{m,n}(t) = 2^{-m/2}\psi\left(2^m t - n\right) \tag{5.10}$$

constitute an orthonormal basis for $\mathbf{L}^2(\mathbb{R})$ as long as $\psi_{m,n}$ are orthonormal.

The first known function which satisfies the orthonormal requirements was the Haar function [Haa10]:

$$\psi(x) = \begin{cases} 1 & 0 \leq x < \frac{1}{2} \\ -1 & \frac{1}{2} \leq x < 1 \\ 0 & otherwise \end{cases} \tag{5.11}$$

but in the recent years, several functions have proven its properties as orthonormal basis, such as the Meyer wavelet, the Daubechies family of wavelets or the Mexican Hat wavelet (Fig. 5.1).

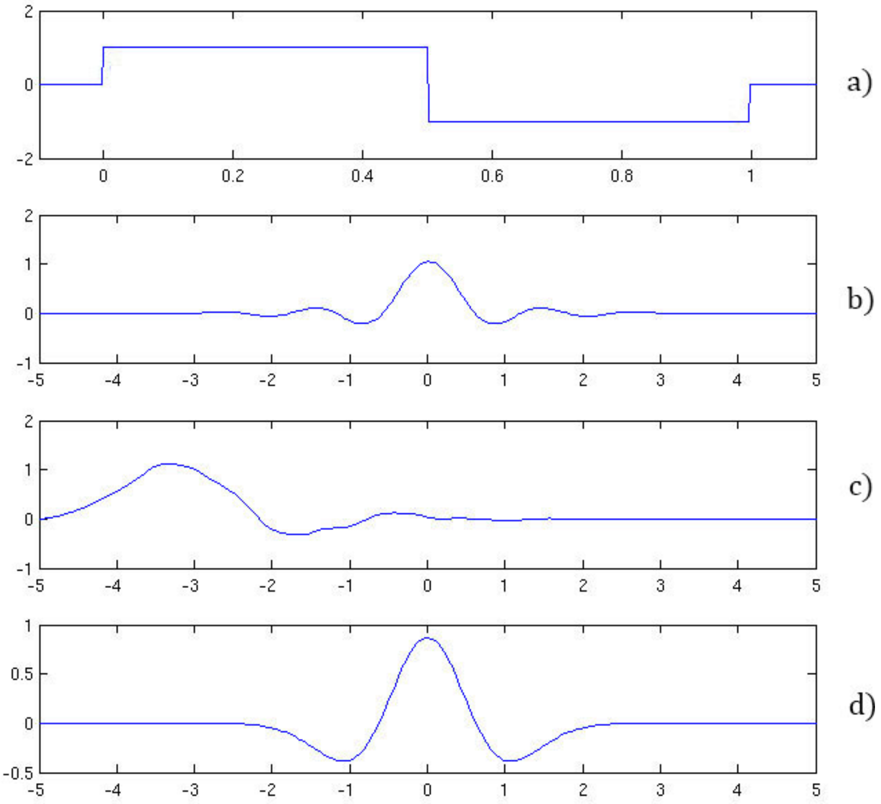

Figure 5.1: Several mother wavelet functions, its progressive dilation constitutes an orthonormal basis. a) Haar wavelet b) Meyer wavelet c) Daubechies1 wavelet d) Mexican Hat Wavelet

### 5.1.3 The problem of reconstruction

When the discretization of the wavelets is used on a transformed continuous signal the result is understood as a series of wavelet coefficients, but it appears a couple of important questions:

> Do the discrete wavelets coefficient completely characterize the original function? Can the original function be reconstructed from the discrete wavelets coefficients?

It has been proven that is possible getting a stable reconstruction as long as the wavelets coefficients satisfy the following condition [Dau92]:

$$A \left\|f\right\|^2 \leq \sum_{j,k} \mid \langle f, \psi_{n,m} \rangle \mid^2 \leq B \left\|f\right\|^2 \tag{5.12}$$

being $\left\|f\right\|^2$ the energy of the original function, $A > 0$, $B < \infty$ and $A, B$ independent from $f$. If $\psi_{m,,}(t); m, n \in \mathbb{Z}$ satisfies this condition, also can be understood as an stability requirement, it is referred as a **frame**. The connection between frames and numerically stable reconstruction from discretized wavelets has been proved by several authors [You80; Gro85].

[Dau92] and [She96] have also proven that if the mother wavelet, joint with the dilation and scaling parameters, constitutes an orthonormal basis, the reconstruction of an arbitrarily signal $f(t)$ can be accomplished by summing the orthogonal wavelet basis function, weighted by the wavelet transform coefficients:

$$f(t) = \sum_{m,n} T^{wav}_{m,n}(t)\psi_{m,n}(t) \tag{5.13}$$

In conclusion, as long as these conditions are achieved, the signal is fully characterized by the discretized wavelet levels and what is more, the original signal can be reconstructed.

### 5.1.4 The equivalence between wavelets and filters

The principal property of the wavelets transformation is the capability of analyzing a time-variant signal on different frequency scales. The details shown on each wavelet level, will depend on the scaling parameter $s$ , or its discrete equivalent $m$ (See Eq. 5.1 and Eq. 5.9), filtering the interesting frequencies on each level.

Furthermore, due to the band-pass like spectrum of the mother wavelet functions

[Dau92], the progressive dilations of wavelet can be understood as a band-pass filter bank (5.2).

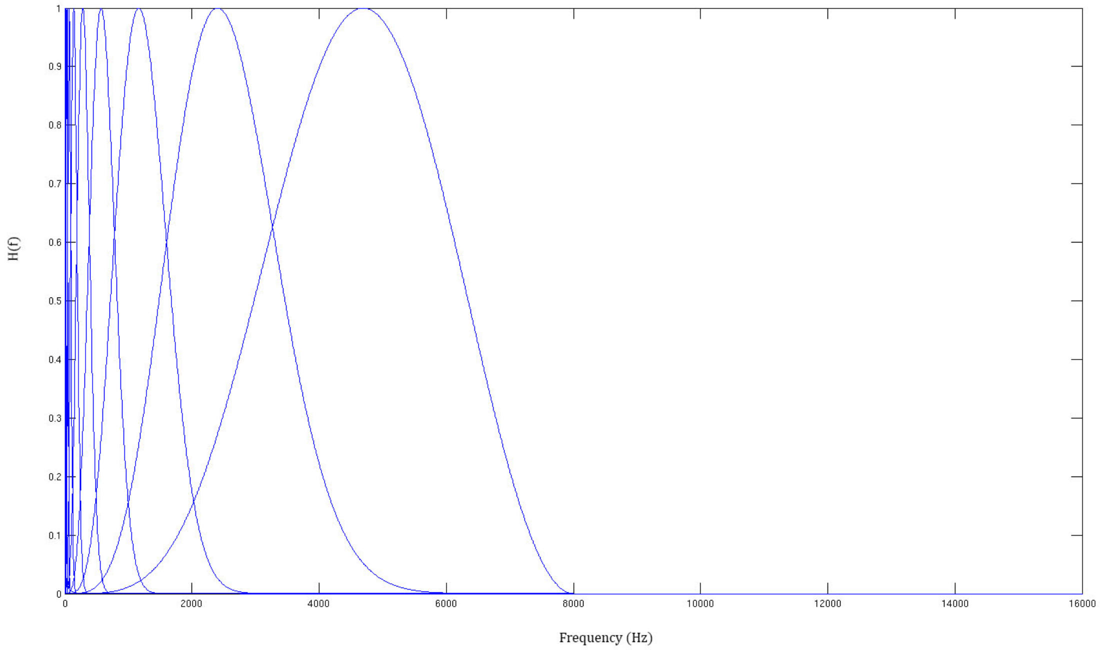

Figure 5.2: Filter bank produced by the progressive dilation of Mexican Hat wavelet, starting from 40 ms and taking ten representation scales.

Adjusting correctly the dilation parameter, the whole spectrum of the signal is covered by the spectra of dilated wavelets. Once again, if $a_0$ is set to 2 and $m \in \mathbb{Z}$, it ensures the full representation of the signal's spectrum and good time-frequency localization properties.

## 5.2 Prosody of speech in the wavelet domain

Using the multiresolution properties of the wavelet analysis, the $F_0$ contours are transformed to the wavelet domain in order to achieve a better capability of understanding the prosodic phenomena present on it. The pitch contours are obtained from STRAIGHT (See Sec. 4.2), using a 5 ms frame update interval. The spectral features, modeled as MCC using the SPTK toolkit [SPT], and the aperiodicity map, are both also obtained from STRAIGHT.

### 5.2.1 Preprocessing $F_0$

The wavelet analysis is sensitive to the gaps of the signal, besides mean and variance, therefore a couple of preprocessing steps, proposed originally by [Sun13], are required to the precise conversion of the signal to the wavelet space.

The first step applied to the $F_0$ contour is a transformation to the logarithmic scale, since the relevant information in the pitch signal is closely related to the logarithmic perceptual scale. To fill the gaps, produced by the unvoiced frames, a simple linear interpolation is applied to an smooth version of the $F_0$ contour, created using a 3-point mean filter. The interpolated gaps are added to the original logarithmic $F_0$, and a 3-point median filter is applied to the final interpolated signal to reduce continuities.

In order to reduce the effect of the edges, constant $F_0$ is added prior and after the utterance. The pre-utterance $F_0$ is set to the mean of the first half $F_0$, while the post-utterance is set to the minimum of the second half $F_0$. Finally, the interpolated $F_0$ contour is normalized to zero mean and unit variance, required by the wavelet analysis, leading to the final $F_0$ preprocessed contour depicted in Fig. 5.3b).

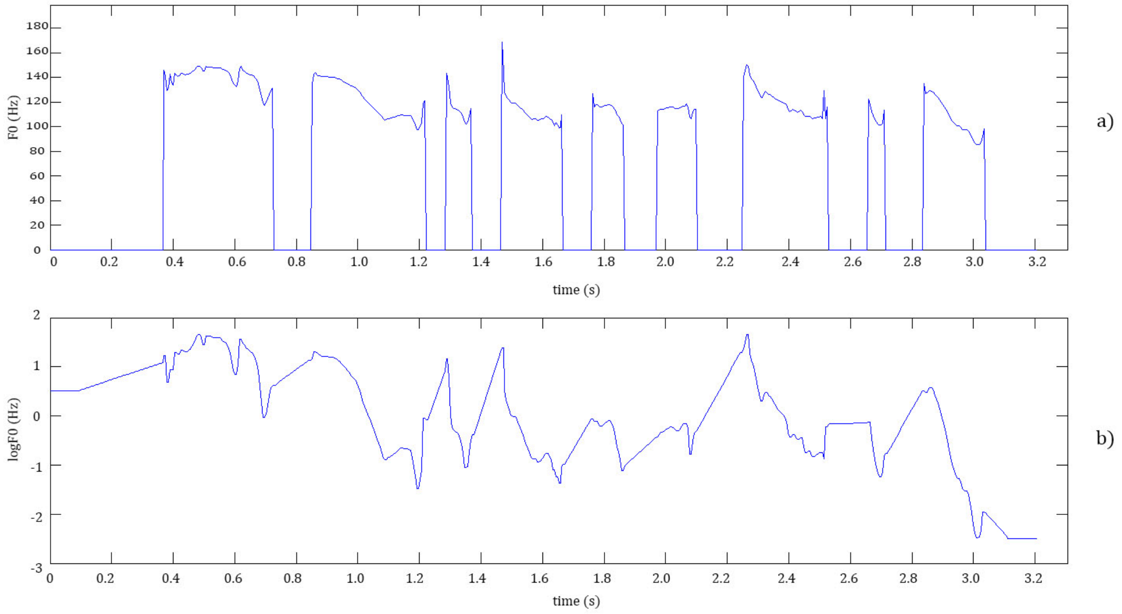

Figure 5.3: $F_0$ contours before (a) and after (b) applying the preprocessing step.

### 5.2.2 Wavelet transformation of pitch contours

The $F_0$ contour is decomposed using the DWT (see Eq. 5.9), with the Mexican Hat wavelet as a mother wavelet (Eq. 5.14), with a standard duration of 5 ms. (for convenience the same interval STRAIGHT uses for the extraction of one sample). The parameters for the wavelet transformation are set to $a_0 = 2$ and $b_0 = 1$, and ten scales has been chosen, one octave apart, for the modeling of the different levels of the wavelet $m = 2, 3, \ldots, 11$.

$$\psi(t) = \frac{2}{\sqrt{3}\pi^{1/4}}(1-t^2)e^{\frac{-t^2}{2}} \tag{5.14}$$

$$T^{wav}_{m,n} = |2|^{-m/2} \int f(t)\psi(2^{-m}t-n)dt, \tag{5.15}$$

These timing scales are the same ones ones proposed by [Sun13], due to its proven relation with the prosodic formants of the hierarchic prosody model (Sec. 3.3)[Vai13], leading to the corresponding frequencies in each level:

| Scale | Duration | Frequency |
|---|---|---|
| 1 | 20 ms | 50 Hz |
| 2 | 40 ms | 25 Hz |
| 3 | 80 ms | 13 Hz |
| 4 | 160 ms | 6 Hz |
| 5 | 320 ms | 3 Hz |
| 6 | 0.64 s | 1.6 Hz |
| 7 | 1.28 s | 0.8 Hz |
| 8 | 2.56 s | 0.4 Hz |
| 9 | 5.12 s | 0.2 Hz |
| 10 | 10.24 s | 0.1 Hz |

Table 5.1: Duration of the mother wavelet and frequencies corresponding on each decomposed wavelet level.

The final wavelet transform of the $F_0$ contour is depicted in Fig. 5.4.

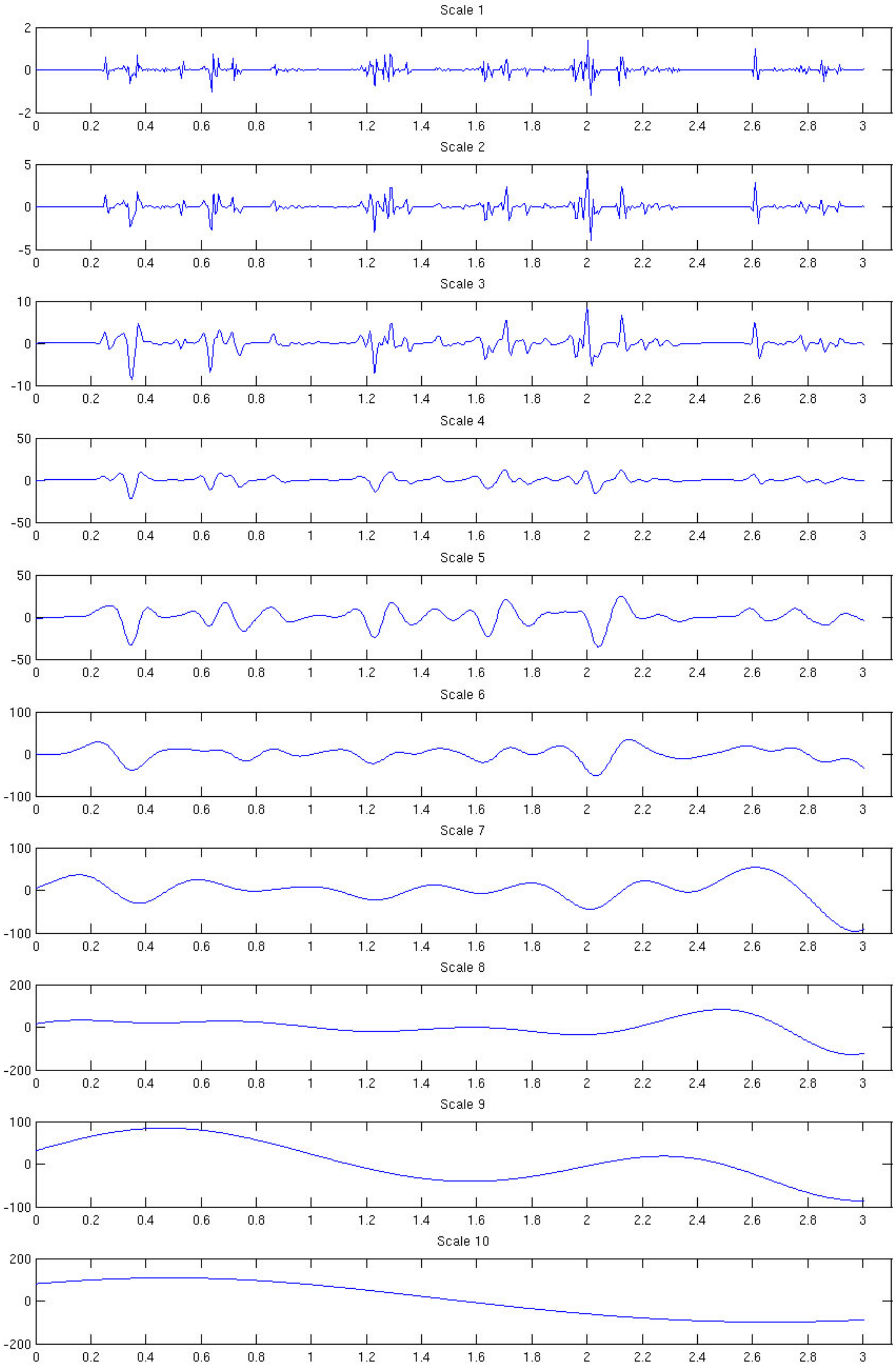

Figure 5.4: Wavelet transform of the $F_0$ signal with the ten choosen scales, from the sentence "A maddening joy pounded in his brain" uttered by a male speaker.

### 5.2.3 Study of the prosody of speech in the wavelet domain

Once the contour has been decomposed in the ten different levels, interesting differences between the speakers uttering the sentence can be perceived.

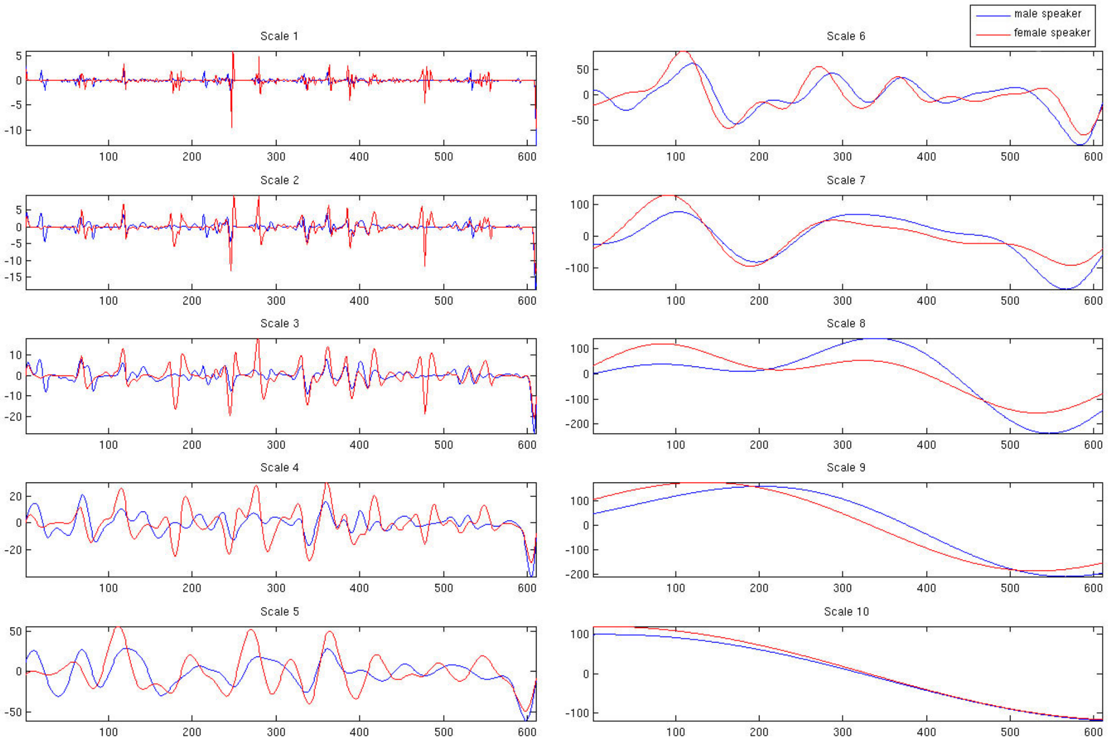

Figure 5.5: Wavelet transform of two $F_0$ contours corresponding to the sentence "He cried, and swung the club wildly." , from a male speaker and a female speaker.

In Fig. 5.5 every wavelet level shows different contours for both speakers, allowing a better and easy modeling of the particular prosody of each speaker, instead of just using the $F_0$ contour. However, if a deep analysis of the wavelet domain is realized for each speaker, by adding the boundaries of phonemes, syllables and words, can be depicted abundant prosodic information on each level, highlighting most of the prosodic phenomena introduced in Sec. 3.2.

The first two levels of the wavelet analysis, corresponding to the phoneme levels, are shown in Fig. 5.6. It can be noticed that almost every "phoneme slot" corresponds to one peak of the signal, normally aligned close to the boundary, showing a correlation between the phoneme level of the wavelet analysis and the real phoneme boundaries. However it is difficult to analyze the suprasegmental prosodic events in these high frequency levels (the duration of the phonemes is estimated to be among 50-100 ms.).

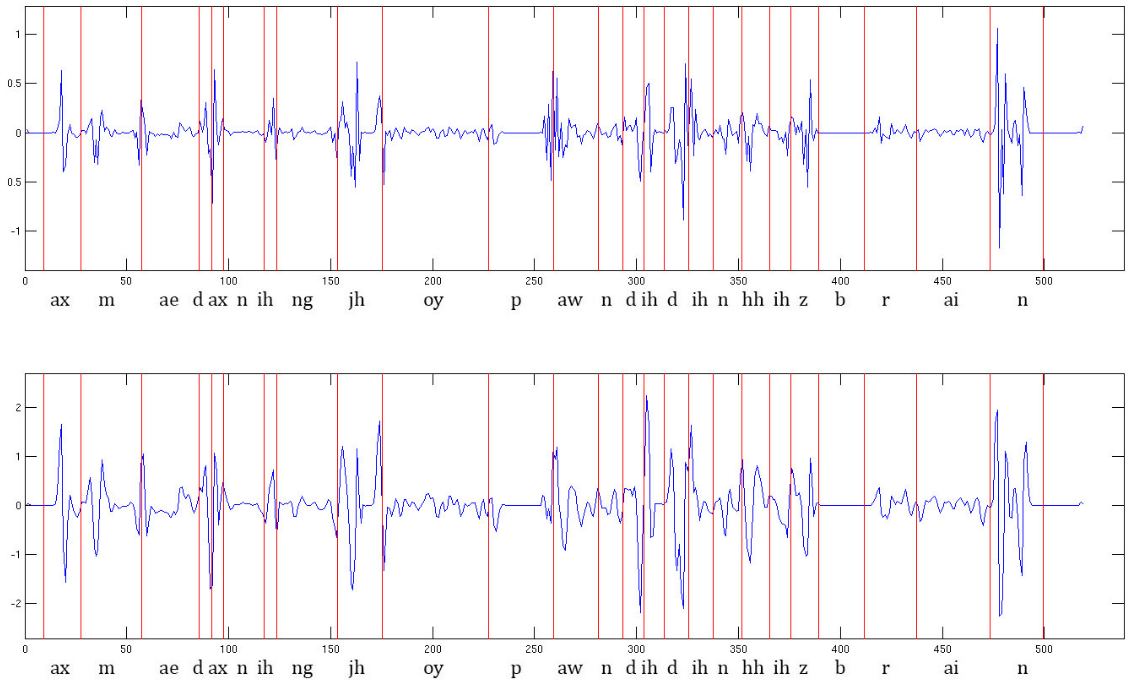

Figure 5.6: Phoneme levels of a $F_0$ contour corresponding to the sentence "A maddening joy pounded in his brain.", from a male speaker. Phoneme boundaries are represented by red vertical lines.

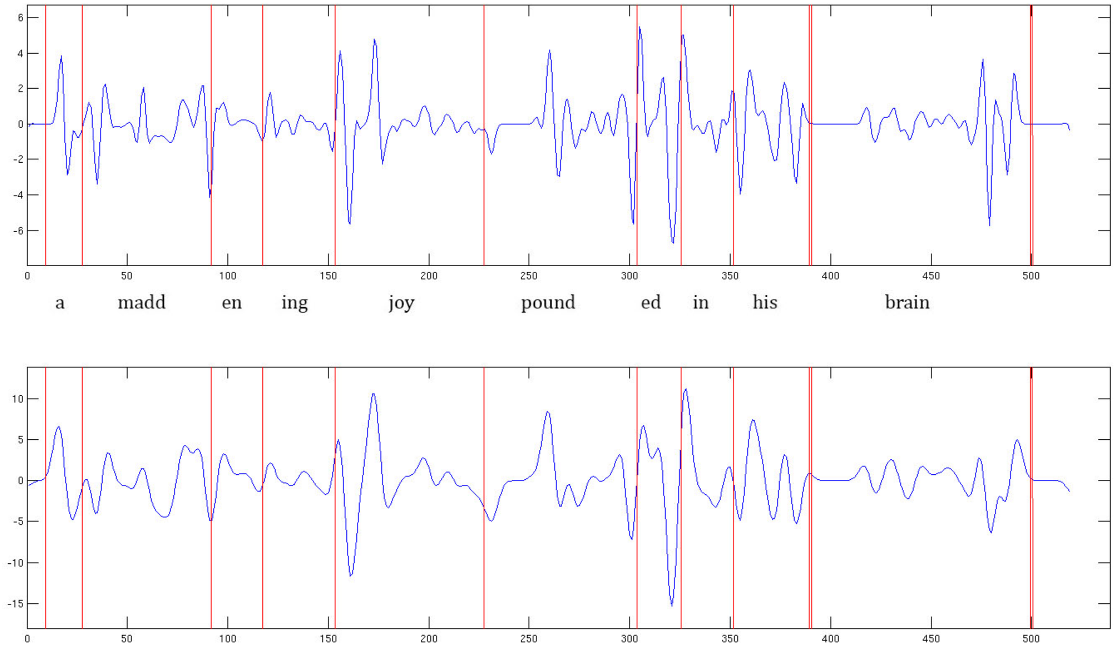

Figure 5.7: Syllable levels of a $F_0$ contour corresponding to the sentence "A maddening joy pounded in his brain.", from a male speaker. Syllable boundaries are represented by red vertical lines.

In the third and fourth levels, corresponding to the syllable levels, it appears several peaks in every "syllable slot", produced by the different phonemes. These different peaks might correspond to the mora level in the hierarchic model, but since there are no boundaries estimated for moras or any estimated timing, it cannot be affirmed properly.

In the example sentence, it is reflected in the words 'maddening', where the first peak corresponds to the 'm' sound and the second and third correspond to 'add' and 'en'; 'joy', with two peaks corresponding to 'j' and 'oy'; or the word 'pounded', where the first peak corresponds to 'pou', the second to 'nd', and third one to 'ed' (Fig. 5.7).

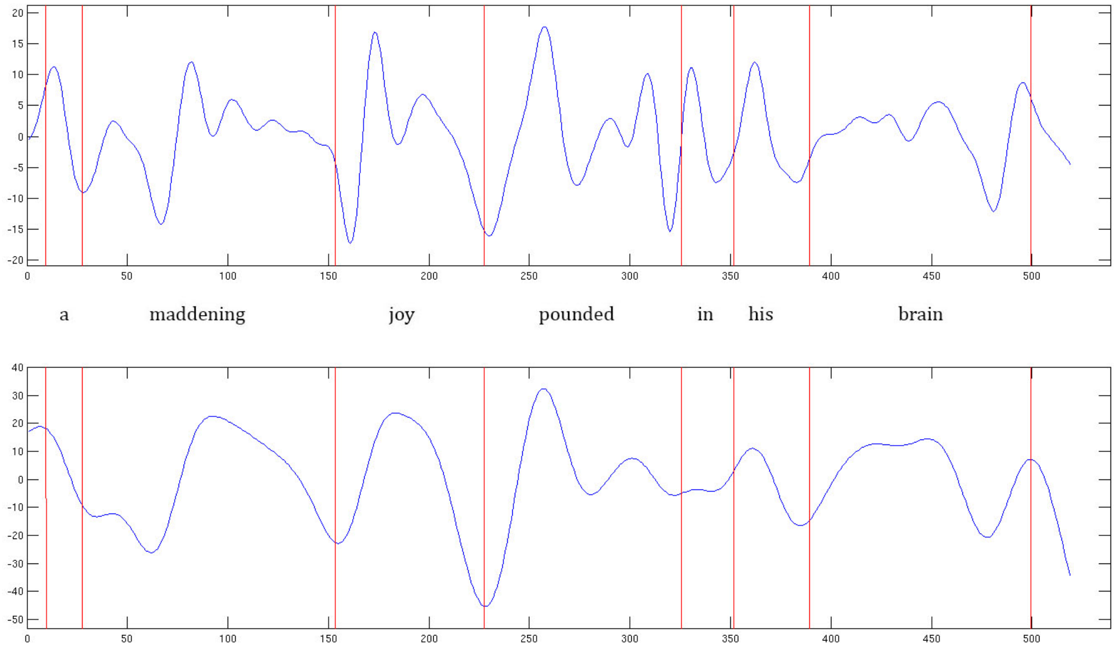

Figure 5.8: Words levels of a $F_0$ contour corresponding to the sentence "A maddening joy pounded in his brain.", from a male speaker. Words boundaries are represented by red vertical lines.

Looking at the word levels, it is recognized one of the main prosodic phenomena: the **stress** of the syllable.

The sentence 'A maddening joy pounded in his brain', its word levels are depicted in Fig. 5.7, the two stressed syllables of the sentence are placed in 'MADDening' and 'POUNDed', a fact that can be appreciated in the wavelet contours: compared to the other peaks in the corresponding words, the stressed syllables are clearly relevant. However the syllables 'joy' and 'in' have also a high peak, showing that this speaker also puts some relevant stress in these syllables, remarking its significance in the sentence.

Another interesting fact related with the stress of the syllables that can be discerned in these levels is where it exactly appears: it is not in the whole syllable but in the start of the vowel in the syllable.

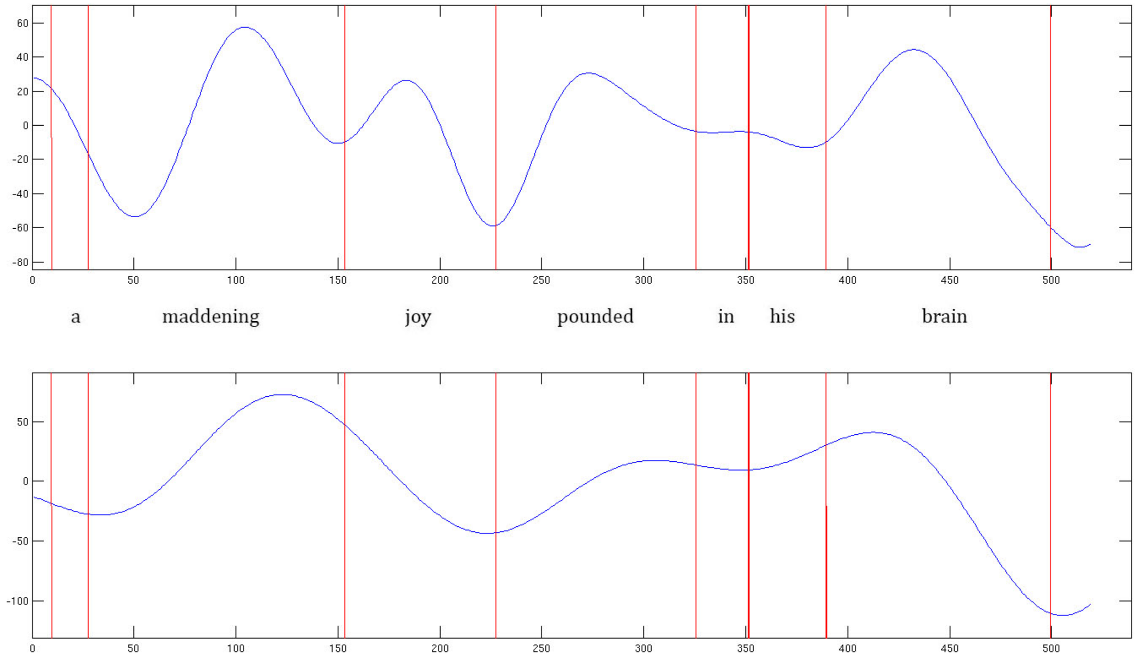

Figure 5.9: Sentence levels of a $F_0$ contour corresponding to the sentence "A maddening joy pounded in his brain.", from a male speaker. Words boundaries are represented by red vertical lines.

Moving forward to the sentence levels, seven and eight of the wavelet transformation, and keep representing the word boundaries, can be perceived one interesting fact related to the word categories: in nouns, verbs, adverbs and adjectives, known as *content words*, words to which an independent meaning can be given, content a peak on it. On the other hand, words such as prepositions, conjunctions, or articles, known as *function words*, that have little semantic contain of its own and chiefly indicates a grammatical relationship, they do not content any peak or just a slight peak.

Both contrasting behavior can be clearly observed in the higher sentence level, as seen in the example sentence (Fig. 5.9): 'maddening', 'joy', 'pounded' and 'brain' are content words and it appears its corresponding peak, while 'a', 'in' and 'his' are function words thus they do not contain any clear peak.

Furthermore, in the second sentence level, high peaks involving several words can be recognized, which may be understood as either phonological or intonational phrases (see Sec. 3.3). However, the separation among these prosodic levels it is not clear thus it is difficult to place the edges of phonological and intonational sentences in a written sentence. In the example, they are suggested two groups of intonation: 'a maddening joy' and 'pounded in his brain', which corresponds to the intonation produced in the audio sample.

This effect produced by phonological and intonation phrases can be easily understood in a sentence with a comma, since it provides a natural division into several intonational sentences. In Fig. 5.10 both commas are clearly identified in the wavelet representation, producing the corresponding valleys in the contour. Therefore can be confirmed that three phonological phrases are present in the utterance.

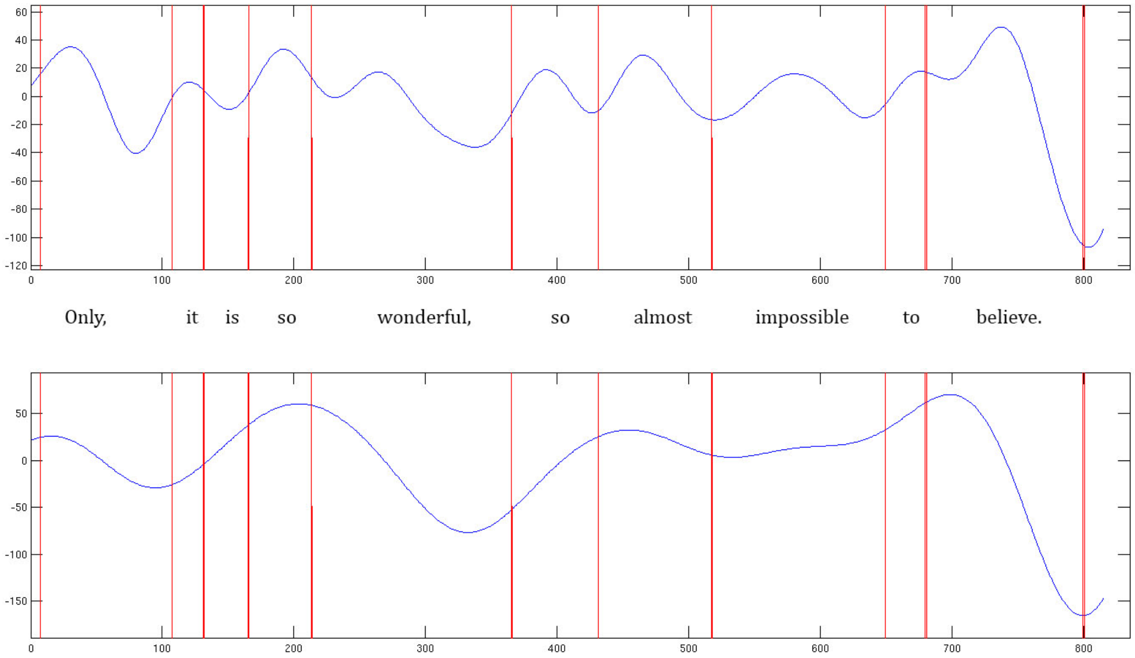

Figure 5.10: Sentence levels of a $F_0$ contour corresponding to the sentence "Only, it is so wonderful, so almost impossible to believe.", from a male speaker. Words boundaries are represented by red vertical lines.

The final levels of the wavelet analysis, corresponding to the utterance levels, show the general trend of the sentence, the general **intonation** pattern. In addition, when two sentences are concatenated to produce a real utterance, it appears the extra information concerning both sentences, showing the separation in between (Fig. 5.11).

In conclusion, every level has its own information and show the main prosodic events represented in a more understandable way. However the levels are not perfectly fit to what it is supposed to be represented on it: in the syllable levels can be appreciated some characteristics from the phonemes, in the word levels can be observed the stress of the syllable, and in the sentence levels can be recognized some characteristics of the words. Consequently some adjustments are necessary in order to achieve a better analysis of the prosody.

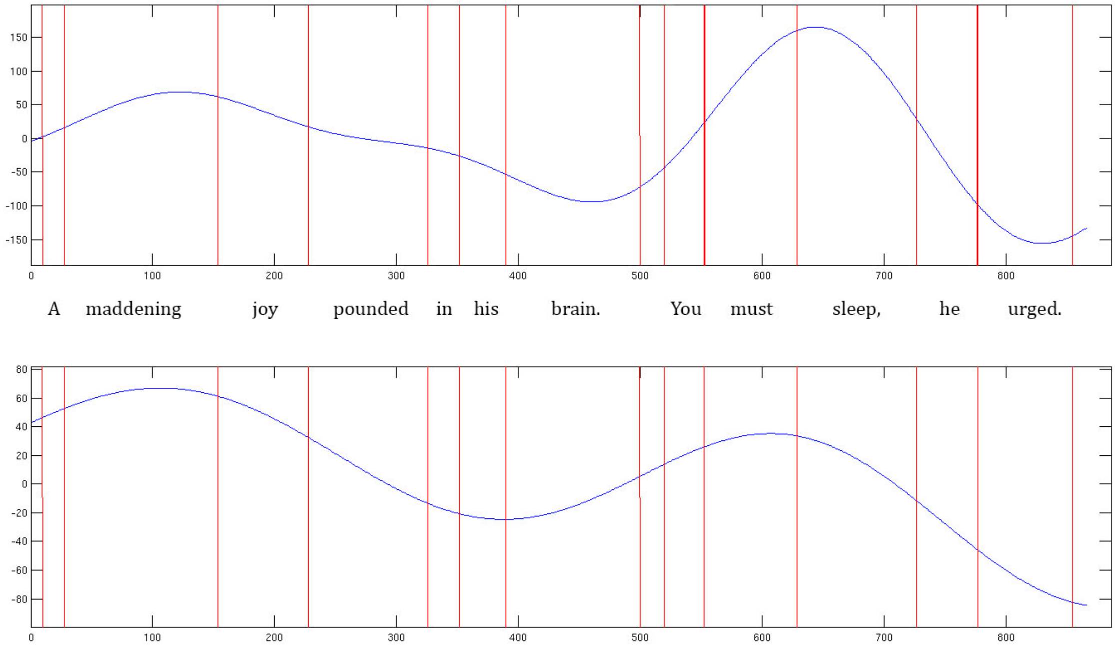

Figure 5.11: Utterance levels of two concatenated $F_0$ contours corresponding to the sentences "A maddening joy pounded in his brain." and "You must sleep, he urged.", from a male speaker. Words boundaries are represented by red vertical lines.

## 5.3 Prosody conversion

The wavelet transformation provides a tool to get a better analysis about the prosody of speech. Comparing the wavelet levels of several speakers can be noticed that the more different ones, thus the ones which carry the identity and personality of the speech, are the syllable, word and sentence levels. Based on this hypothesis it is proposed a new method for the conversion of the prosody in speech (Fig. 5.12):

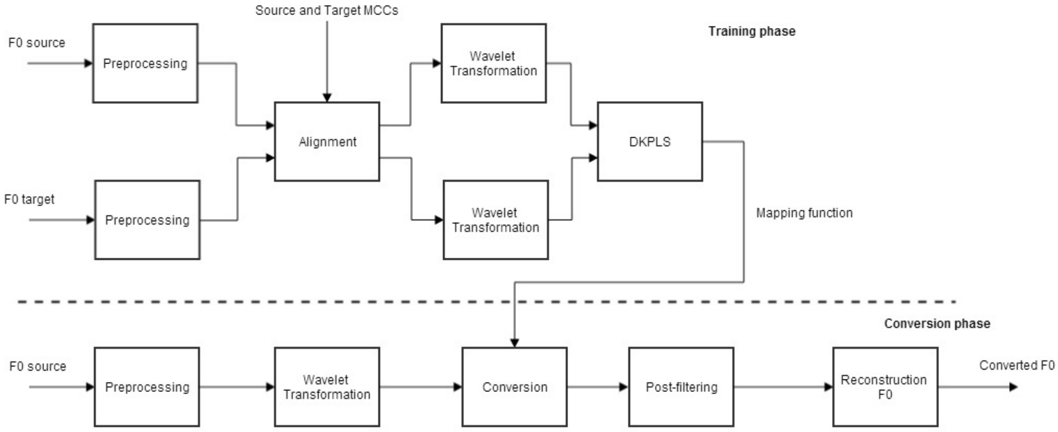

Figure 5.12: Framework of the proposed prosody conversion method.

### 5.3.1 Adjustment and selection of wavelet levels

The scales proposed by [Sun13] which generate the wavelet levels are based in a study of the Finnish word prominence [Vai13]. Although the scales applied to English language are quite accurate and generate reasonable wavelet levels, the differences among the languages. Since Finnish is a predominantly polysyllabic language, word level and syllable level are quite different, thus its corresponding frequencies are clearly differentiated. On the other hand, English is closer to a monosyllabic language, hence syllable frequencies are closer to word frequencies.

This fact can be observed in the wavelet levels: syllable levels show information majorly related with smaller units than the syllables, meanwhile word levels contain information related to the syllables and sentence levels shows information concerning both words and sentences. Consequently, it is proposed increasing the scales by one octave, thus $m = 3, 4, \ldots, 12$ (See Sec. 5.2.2). With this new timing, the wavelet levels have been displaced fitting the properties observed in every level to its corresponding frequency.

A study of the new wavelet levels of 100 $F_0$ contours from the four used speakers in the dataset (See 6.1), two male and two female, is performed, computing for every energy-normalized wavelet level, the RMSE among the same wavelet level of the other speakers. It can be concluded that the levels which have major differences between speakers are the syllable, word and sentence, which reflects the differences present on the main prosodic phenomena (syllable stress, word stress and intonation) uttered by the different speakers.

| | RMSE |
|---|---|
| Scale 1 | 0.6347 |
| Scale 2 | 0.6718 |
| Scale 3 | 0.8622 |
| Scale 4 | 0.8327 |
| Scale 5 | 0.8107 |
| Scale 6 | 0.7422 |
| Scale 7 | 0.6119 |
| Scale 8 | 0.5203 |
| Scale 9 | 0.4162 |
| Scale 10 | 0.5023 |

Table 5.2: Root MSE for 100 $F_0$ contours from four different speakers.

Notice that the first two levels are, comparatively, more different than the levels 7 and 8 (the sentence levels), however they are not selected since the goal is modeling the suprasegmental features of prosody. Thus, levels 3, 4, 5, 6, 7 and 8 are chosen to be transformed.

### 5.3.2 Conversion of wavelet levels and post-filtering

The conversion of the selected wavelet levels is realized using DKPLS (See 4.4.2), and the algorithm used to solve the PLS regression problem is SIMPLS, proposed by [de 93].

In the training phase, 40 sentences are used to create the mapping function, properly aligned using DTW (Sec. 4.3.1), with the extracted MCC as spectral features (Eq. 2.10). The $F_0$ contours are transformed to the wavelet domain, selecting the syllable, word and sentence levels, and normalizing each level by its own energy.

Once the conversion is achieved, transforming the syllable, word and sentence levels and copying the phoneme and utterance levels from the source, a final filtering stage is applied to the converted wavelet levels to avoid the appearance of frequencies not corresponding to the desired wavelet level. A simple low-pass filter is used with the cut-off frequencies from each level, shown in Table 5.3.

| | Cut frequency |
|---|---|
| Scale 3 | 25 Hz |
| Scale 4 | 12.5 Hz |
| Scale 5 | 6 Hz |
| Scale 6 | 3 Hz |
| Scale 7 | 1.5 Hz |
| Scale 8 | 0.8 Hz |

Table 5.3: Cut frequencies applied in the low-past post-filtering stage.

After the filtering, the original energy of the levels is denormalized, leading to the final converted levels, represented in Fig. 5.13.

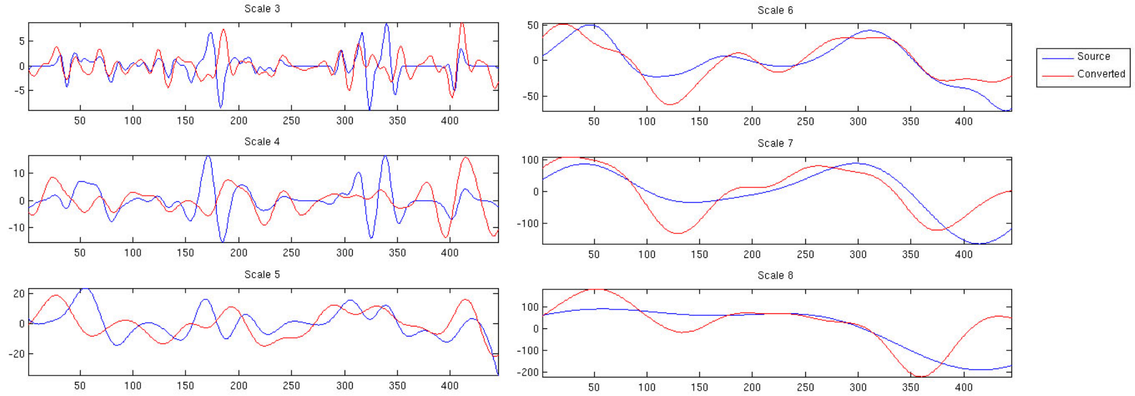

Figure 5.13: Converted wavelet levels from the sentence "A maddening joy pounded in his brain", uttered by a male speaker and converted to a female speaker.

### 5.3.3 Reconstruction of the pitch

The reconstruction of the signal after the wavelet decomposition can be achieved since, all the conditions concerning the mother wavelet equation and the different dilation and scaling parameters are accomplished. However, the final reconstruction method differs from the proposed ones in Sec.5.1.4; it is based on the formula proposed by [Sun13]:

$$f_0(n) = \sum_{m=1}^{10} T_{m,n}^{wav} (m + 2.5)^{-5/2} \tag{5.16}$$

Since this formula was computed *ad hoc*, its accuracy and efficiency must be evaluated: the reconstruction RMSE and the correlation between the original signal and the reconstructed was tested in 20 sentences from male and female speakers. For males, the reconstruction error was 1.67Hz with a 99.95% of correlation, meanwhile in females the reconstruction error was 2.59Hz with a 99.93% of correlation.

Finally, the reconstructed contour is weighted by the mean and variance of the target speaker, and retransformed to the linear scale (Fig. 5.14), in order to resynthesize the final converted speech.

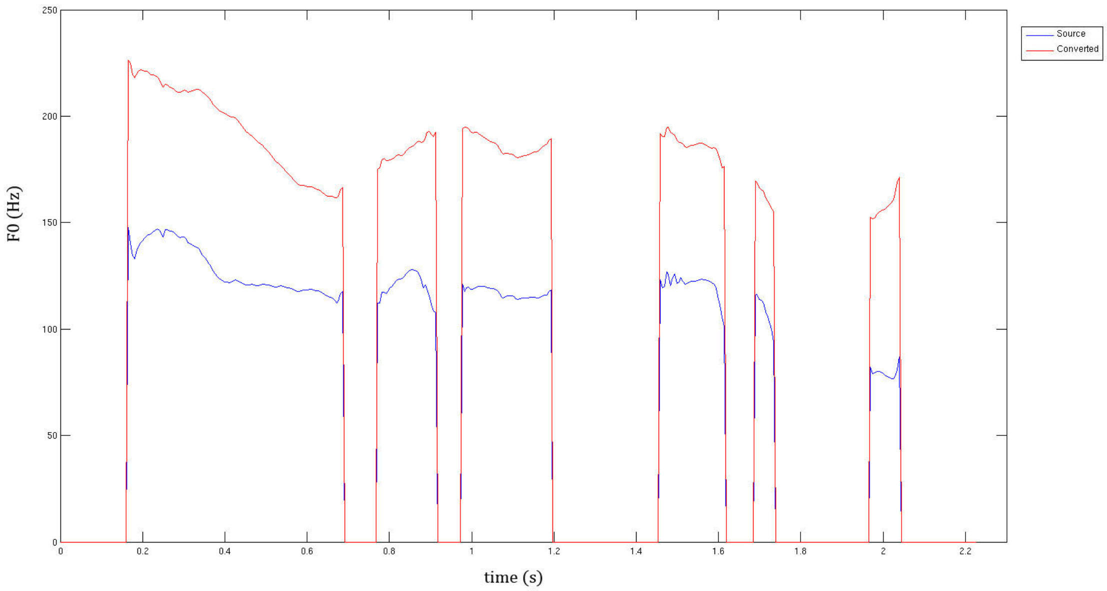

Figure 5.14: Reconstructed pitch from the sentence "A maddening joy pounded in his brain", uttered by a male speaker and converted to a female speaker.

# 6. Evaluation of the prosody conversion method

One of the difficulties about VC systems is the real possibilities to evaluate the quality of the system in terms of identity conversion. The perception of the identity in speech is largely subjective, and it is affected by both speaker and listener: when a sentence is uttered by a speaker different times, each repetition shows different details about the identity. On the other side, the listener is closer to recognize the speaking style if comes from a known person.

Due to these reasons, no unique correct way of converting the speech, and in this case the prosody it is assumed. The objective is achieve a converted speech, which might be uttered by the speaker in some situation, and showing characteristics of the identity and prosody of the target speaker. Based in this goal, some objective measures can be extracted from the converted speech, however the best way of evaluation are the listening tests.

This chapter introduces the dataset used in the experiments in 6.1, and presents both ways of evaluation: section 6.2 introduces an study of the RMSE and the correlation between the converted and target samples, meanwhile section 6.3 presents a listening test, asking by the evaluation of the naturalness of the converted samples.

## 6.1 Speech corpus

The dataset used in this thesis comes from the CMU ARCTIC databases, originally created for speech synthesis by J. Kominek and A. W. Black [CMU]. It consists of nearly 1150 phonetically balanced English utterances, recorded under studio conditions, and packaged with phonetic labels and pitchmark files.

The Arctic corpus consists of four primary sets of recordings, 3 male speakers and 1 female speaker, and several ancillary databases. In this thesis, the used speakers were BDL and RMS, for male speakers, and SLT and CBL, for female speakers. All of them are experienced voice talent, with an US English accent.

## 6.2 Objective evaluation of $F_0$ contours

In order to evaluate the quality and effectiveness of the method, it is proposed studying the RMSE and the correlation of the converted speech using the wavelet approach and target speech, and compare it with the RMSE and correlation achieved with the main proposed method in the literature: the MS method (Sec. 4.5). 25 speech samples for each speaker, two male and two female, were converted and reconstructed, leading to the following results in RMSE:

| | BDL | RMS | SLT | CLB |
|---|---|---|---|---|
| BDL | | 10,35 Hz. | 22,63 Hz. | 16,93 Hz. |
| RMS | 8,07 Hz. | | 21,13 Hz. | 15,54 Hz. |
| SLT | 12,08 Hz. | 14,96 Hz. | | 20,27 Hz. |
| CLB | 9,56 Hz. | 11,66 Hz. | 19,56 Hz. | |

Table 6.1: RMSE between the converted speech with wavelets and sample speech for 25 samples for every speaker. RMS and BDL correspond to the male speakers, meanwhile SLT and CLB are female speakers.

| | BDL | RMS | SLT | CLB |
|---|---|---|---|---|
| BDL | | 11,14 Hz. | 30,68 Hz. | 14,43 Hz. |
| RMS | 9,85 Hz. | | 32,47 Hz. | 16,16 Hz. |
| SLT | 11,36 Hz. | 15,07 Hz. | | 15,80 Hz. |
| CLB | 11,31 Hz. | 13,53 Hz. | 32,58 Hz. | |

Table 6.2: RMSE between the converted speech with MS method and sample speech for 25 samples for every speaker. RMS and BDL correspond to the male speakers, meanwhile SLT and CLB are female speakers.

In the case of the correlation between the signals, the results are:

| | BDL | RMS | SLT | CLB |
|---|---|---|---|---|
| BDL | | 73,62% | 56,62% | 70,82% |
| RMS | 76,10% | | 54,97% | 67,90% |
| SLT | 69,81% | 62,18% | | 69,44% |
| CLB | 64,02% | 62,57% | 60,87% | |

Table 6.3: Correlation between the converted speech with wavelets and sample speech for 25 samples for every speaker. RMS and BDL correspond to the male speakers, meanwhile SLT and CLB are female speakers.

| Involved genders in conversion | RMSE for wavelets | RMSE for MS method | Correlation for wavelets | Correlation for MS method |
|---|---|---|---|---|
| female-to-female | 19,92 Hz. | 22,53 Hz. | 61,93% | 44,83% |
| female-to-male | 12,07 Hz. | 11,34 Hz. | 59,12% | 61,23% |
| male-to-male | 9,22 Hz. | 10,35 Hz. | 74,86% | 71,85% |
| male-to-female | 19,09 Hz. | 22,62 Hz. | 59,02% | 45,46% |

Table 6.5: RMSE and correlation comparison between prosody conversion with wavelets and MS method. The different measurements are computed depending on the genders involved in the conversion.

| | BDL | RMS | SLT | CLB |
|---|---|---|---|---|
| BDL | | 69,52% | 31,73% | 67,25% |
| RMS | 69,22% | | 21,13% | 59,69% |
| SLT | 57,90% | 51,77% | | 59,42% |
| CLB | 66,16% | 56,63% | 24,12% | |

Table 6.4: Correlation between the converted speech with MS method and sample speech for 25 samples for every speaker. RMS and BDL correspond to the male speakers, meanwhile SLT and CLB are female speakers.

Discerning the conversion among genders, the global results of the RMSE and correlation are shown in Table 6.5. Although the RMSE for the proposed system is lower than the MS method, except for the female-to-male conversion, the difference between both methods is minimal (less than 4 Hz). On the other hand, the correlation parameter increases remarkably for the female-to-female and the male-to-female, showing a better approach to the shape of the target $F_0$ contour. In the case of male-to-male conversion, it has also increased, however the source and target $F_0$ contours were already clearly similar. Finally the female-to-male conversion is the only case where the correlation of the signals decreases.

In conclusion, the proposed method shows a better approximation to the target prosody of speech, however it is difficult to evaluate the real prosody performance based in just two objective parameters. In the next subsection it is proposed a listening test in order to evaluate the perception of the converted prosody.

## 6.3 Perceptual evaluation

One of the major difficulties in the evaluation of the prosody of speech is the fact that the person who evaluates the produced speech does not normally know the speakers involved in the original samples, consequently the singular prosody of every speaker is unknown and thus difficult to evaluate.

The proposed listening test was conducted under the assumption of creating a credible sentence, which might be uttered by the target speaker in some situations, therefore the goal was evaluating the naturalness of the converted prosody, instead of the particular prosody of the speaker.

There were evaluated four different scenarios of conversion: male-to-female conversion, male-to-male conversion, female-to-male conversion and female-to-female conversion, consisting of five randomly selected sentences for each speaker pair in every test. The recordings used come from the CMU ARCTIC databases [CMU], originally created for speech synthesis. The samples are recorded under studio conditions, and packaged with phonetic labels and pitchmark files. 16 listeners participated in the test. Nativeness was not required as the test was designed in such a way that also non-native listeners with good English skills can easily judge the relevant issues from the speech samples.

The listeners heard an initial sample, uttered by the target speaker, and two versions of the speech uttered by the source speaker, in which the prosody was converted using the two different techniques, the wavelet approach and the MS method. They were asked to choose the sample that presents best naturalness. The subjects could also choose "equal" and it was possible to listen to the samples as many times as necessary. The spectral features were also converted, however they were asked not to care about quality of the spectral conversion.

### 6.3.1 Results of listening test

The percentages of preference votes that the two methods received as well as the total number of votes are shown in Table 6.6 for the four possible scenarios.

| Method | CW | MS | equal |
|---|---|---|---|
| male-to-female | 63.75% (51) | 20% (16) | 16.25% (13) |
| male-to-male | 30% (24) | 21.25% (17) | 48.75% (39) |
| female-to-male | 72.5% (58) | 13.75% (11) | 13.75% (11) |
| female-to-female | 31.25% (25) | 37.5% (30) | 31.25% (25) |

Table 6.6: Preference votes given to the proposed approach (CW) and to the MS based approach (MS), and the "no preference" votes (equal).

The percentage results show a clear preference for the proposed approach in the cases where speech is converted between different genders, either female-to-male or male-to-female. In the intra-gender conversion, the results do not present a clear preference for any of the methods proposed: in male-to-male conversion the predominating choice is the "no preference", meanwhile in the female-to-female conversion the votes are almost equally distributed.

According to a two-tailed t-test, the confidence intervals are computed for every possible conversion, represented in Fig. 6.1 jointly with the total percentage. It is shown a significant difference between the performance of both proposed methods: the results for both cross-gender conversion scenarios are statistically significant, meanwhile in the intra-gender conversion the performance is highly similar.

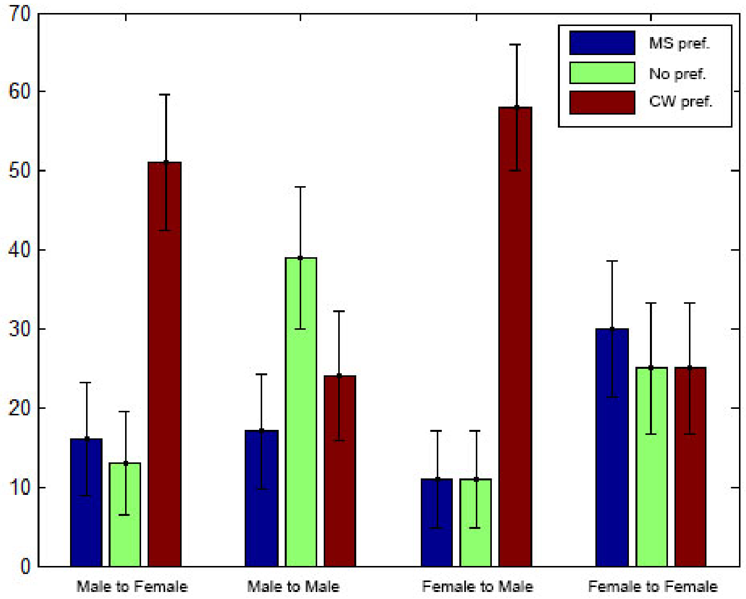

Figure 6.1: Preference percentage, with 95% confidence interval for the MS method and the proposed wavelet method.

Since the database used is originally created for speech synthesis, thus, its utterances are spoken in a controlled prosody style, the differences between speakers from the same gender are minimal, meanwhile for speakers from different gender, the speaking style presents more differences. This is supposed to be one of the reasons why the intra-gender conversion does not present a clear dominance for any of the proposed methods.

On the other hand, the listeners were asked to only rate the naturalness, not how well the prosody matches the expected prosody of the target speaker. Thus, it is an expected result that the two methods performed at a similar level in the intra-gender conversion: in this kind of scenario, it is most likely very hard to get statistically meaningful differences, since much less modification is needed than in the case of inter-gender conversion and, depending a bit on the speaker pair, the simple reference method may already do quite a good job in terms of naturalness. And even more importantly, since the listeners didn't know the speakers, the detailed prosody does not matter that much, as long as the $F_0$ level and scale are fit in the correct range.

In the case of cross-gender conversion, the situation is different. The reference method performs adequately rather rarely because the differences in prosodies are much bigger. It is common that the simple shifting and scaling reduces the naturalness rather much in inter-gender conversion. Thus, there is a lot of room for improvement, being easier to get statistically significant improvements with a more adequate prosody, and the fact that the listeners do not know the speakers does not affect the results that much.

# 7. Conclusions and Future work

This thesis has addressed the study and conversion of the prosody of speech, using a well-known technique in signal processing: the wavelet transform. It is proposed a new prosody conversion system using the wavelet transform to model the different prosodic phenomena, taking into account the different timing of each one with the inherent properties of multiresolution analysis of the wavelets. In contrast with most of the conversion systems present in the literature, the wavelet transformation allows working with an extended set of data for every temporal frame, enlightening extra information in the pitch contour.

Section 7.1 presents the conclusions concerning the study of the prosody phenomena in the wavelet domain, meanwhile section 7.2 discusses the prosody conversion method. Finally, section 7.3 presents lines for future research that can be considered as extensions of the work developed in this dissertation.

## 7.1 The wavelet domain

The prosody modeling system using the wavelet transformation has shown a considerable potential to analyze the prosodic phenomena of the speech. Due to the separation in different streams, the diverse prosodic events emerge and thus, can be easily studied and modeled. Moreover, since the prosodic phenomena are extracted directly from the pitch, it is not required explicit information of the e.g. stressed syllables or the content words.

The stress of the syllables and words, and the intonation pattern are the main prosodic phenomena appreciable in the wavelet domain, nevertheless, it can also be observed other prosodic phenomena, such as the distinction among content and function words or the appearance of the intonational phrases. Other prosodic events, differentiating the morphological levels of prosody, such as the mora or the stress foot, are also perceptible, however since the boundaries of these levels are imprecise and difficult to place, the study of the prosodic events present in these levels is not strictly accurate. On the other hand, the syllable and prosodic word levels are clearly related to the boundaries of syllables and words, respectively: essentially every syllable or word boundary coincide to a valley on the corresponding wavelet level.

However, the entire prosodic phenomena appearing in speech have remarkable differences depending on the language which the sentence is uttered, therefore the timings of the wavelet transformation have to be adjusted and fit according the language the sentence is uttered. In this thesis, the timings proposed by [Sun13] for Finnish language were tested in an English corpus, showing a dissimilarity between the supposed wavelet levels and the prosodic events present on it. Therefore, an adjustment of the wavelet transformation timings was required, in order to represent the entire prosodic morphology properly.

This fact suggests that, for every language, the timings of the wavelet transformation must be reviewed, according to the intrinsic characteristics of the language. Moreover, for different speaking styles in the same language, or for modeling emotions, which also affects and alters the prosody, the timings should also be reconsidered.

## 7.2 A new prosody conversion system

The proposed conversion system, based on a wavelet transformation of $F_0$ contours and the posterior conversion in the wavelet domain, presents better objective measures, in this thesis RMSE and correlation, than the usual method in literature for three of the four tested conversion scenarios. The female-to-male conversion is the only one which presents a worst performance.

Otherwise, the perceptual evaluation shows the clear preference for the new proposed method in the cross-gender conversion (male-to-female and female-to-male), where the differences of source and target pitches are noteworthy. On the other hand, for intra-gender conversion, where the pitch profiles are closer in range and mean level, there is no clear preference for any of the evaluated methods, leading to an equal perceptual performance.

The differences according the objective and perceptual measurements are interesting, since the only case where the RMSE and correlation have decreased in comparison with the MS method, it is precisely the case where better results have been obtained in the listening test. Thus, it is suggested that the proposed method allows creating more natural profiles of pitch than the MS method, in the cases where the differences of speaking style are remarkable among the speakers, although the differences with the actual target pitch are substantial.

## 7.3 Future work

There are several lines for future research that can be considered as extensions of the work developed in this dissertation. Three major areas, which will benefit greatly from further research, are briefly discussed in the following paragraphs.

**Automatic adjustment of the timings for the wavelet analysis.** The selection of the timings of the wavelet is a critical point, since in order to generate an accurate representation of the prosodic events of speech. A suggested system, already used by [Vai13], to the establishment of the correct timings, is studying the peak prominence in the word and syllables levels and relating it with the syllable and word boundaries. The level showing major relation between the peaks and syllables/word slots will be selected as syllable/word level, allowing the construction of the complete wavelet domain.

**Testing other wavelets transforms** Even though the Mexican Hat wavelet has proved good properties in order to represent the prosodic events, other wavelets with different properties can be proven. The Morlet wavelet (or Gabor wavelet), which is highly related with the auditive perception scale of the humans, or wavelets allowing a full reconstruction of the original signal without depending on the dilation and scaling parameters, such as the Daubechies wavelets, can be tested.

**Improving the statistical mapping technique** DKPLS has shown its capabilities to model the prosody using the wavelet domain, however, several improvements can enhance the performance of the system. It is suggested to treat each prosodic unit separately: using the phonemes/syllables/words boundaries on the corresponding wavelet level to model the prosodic unit, for instance with CARTs, based on the position and amplitude of the peak present on the slot.

**Testing the system in diverse databases** The proposed method has shown good results in speakers where the speaking style is clearly different, consequently, the system could also be tested in emotional databases, where the prosody is clearly different for every emotion. Moreover, a complete prosody and emotion conversion system requires a detailed conversion of the speaking rate and the duration of the syllables. The approach proposed by [Nav14], modeling the syllable duration with CARTs, would be an appropriate alternative.

# References

[Abe88] M. Abe, S. Nakamura, K. Shikano, and H. Kuwabara, Voice conversion through vector quantization. *Proc. of ICASSP*, pp. 565–568, 1988.

[Bec86] M. Beckman and J. Pierrehumbert, Intonational structure in japanese and english. *Phonology Yearbook 3*, pp. 255–309, 1986.

[Bel87] M. Bellanger, *Adaptative Digital Filters and Signal analysis.* Marcel Dekker, 1987.

[Che03] Y. Chen, M. Chu, E. Chang, J. Liu, and R. Liu, Voice conversion with gmm and map adaptation. *Proc. of Interspeech*, pp. 2413–2416, 2003.

[CMU] CMU ARCTIC databases for speech synthesis, J. Kominek and A. Black. http://festvox.org/cmu_arctic/index.html, last access, 10-2-2014.

[Dau92] I. Daubechies, *Ten Lectures on Wavelets.* SIAM, 1992.

[de 93] S. de Jong, Simpls: An alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, pp. 251–263, 1993.

[Dem77] A. Dempster, N. Laird, and D. Rubin, Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 39, no. 1:pp. 1–38, 1977.

[Dur60] J. Durbin, The fitting of time-series models. *Revue de l'Institut International de Statistique 28*, pp. 233–244, 1960.

[Dux04] H. Duxans, A. Bonafonte, A. Kain, and J. v. Santen, Including dynamic and phonetic in voice conversion systems. *Proc. of ICSLP*, pp. 5–8, 2004.

[Emb04] M. Embrechts, B. Szymanski, and K. Sternickel, Ch 10: Introduction to scientific data mining: Direct kernel methods and applications. In *Computationally Intelligent Hybrid Systems: The Fusion of Soft Computing and Hard Computing*, pp. 317–363, Wiley Interscience, 2004.

[Err10a] D. Erro, A. Moreno, and A. Bonafonte, Inca algorithm for training vc systems from nonparallel corpora. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[Err10b] D. Erro, A. Moreno, and A. Bonafonte, Voice conversion based on weighted frequency warping. *IEEE Transactions on Audio, Speech, and Language Processing*, 2010.

[Gil03] B. Gillet and S. King, Transforming f0 contours. *Eurospeech*, pp. 101–104, 2003.

[Gro85] A. Grossman, J. Morlet, and T. Paul, *Transforms associated to square integrable group representations.* Journ. Math. Phys., 1985.

[Haa10] A. Haar, Zur theorie der orthogonalen funktionensysteme. *Math. Annal, 69*, pp. 331–371, 1910.

[Hel07] E. Helander and J. Nurminen, A novel method for prosody prediction in voice conversion. *Proc. of ICASSP*, 2007.

[Hel10] E. Helander, J. Nurminen, J. MÃguez, and M. Gabbouj, Maximum a posterior voice conversion using sequential monte carlo methods. *Proc. of Interspeech*, pp. 1716–1719, 2010.

[Hel12] E. Helander, H. Silen, T. Virtanen, and M. Gabbouj, Voice conversion using dynamic kernel partial least squares regression. *IEEE Transactions on audio, speech and language processing*, pp. 806–817, 2012.

[Hua92] X. Huang, A. Acero, and H. Hon, *Spoken Language Processing.* Prentice Hall PTR, 1992.

[Hym85] L. Hyman, *A theory of phonological weight.* Foris Publications, 1985.

[Ina03] Z. Inanoglu, *Transforming Pitch in a Voice Conversion Framework.* Master's thesis, St. Edmund's College, University of Cambridge, 2003.

[Ina07] Z. Inanoglu and S. Young, A system for transforming the emotion in speech: combining data-driven conversion techniques for prosody and voice quality. *Proc. of Interspeech*, 2007.

[Kai01] A. Kain and M. Macon, Spectral voice conversion for text-to-speech synthesis. *Proc. of ICASSP*, pp. 813–816, 2001.

[Kaw99] H. Kawahara, I. Masuda-Katsuse, and A. deChevignÃ¨, Reestructuring speech representations using a pitch-adaptative time-frequency smoothing and a instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, 27:pp. 187–207, 1999.

[Lei10] M. Lei, Y. Wu, F. K. Soong, Z. Ling, and L. Dai, A hierarchical f0 modeling method for hmm-based speech synthesis. *Proc. of Interspeech*, 2010.

[Lev46] N. Levinson, The wiener root-mean-square error criterion in filter design and prediction. *Journal of Mathematics and Physics 25*, pp. 261–278, 1946.

[Lib77] M. Liberman and A. Prince, On stress and linguistic rhythm. *Linguistic Inquiry 8*, pp. 249–336, 1977.

[Lie67] P. Lieberman, *Intonation, perception and language.* MIT Press, Cambridge, Mass., 1967.

[Mal98] S. Mallat, *A wavelet tour of signal processing.* Academic Press, 1998.

[Nav14] S. Navarro, *Automatic conversion of emotion within a speaker independent framework.* Master's thesis, Tampere University of Technology, 2014.

[Nes86] M. Nespor and I. Vogel, *Prosodic Phonology.* Dordrecht: Foris, 1986.

[Nur06] J. Nurminen, V.Popa, J. Tian, Y. Tang, and I. Kiss, A parametric approach for voice conversion. *Proc. of the TC-STAR Workshop on Speech-to Speech Translation*, 2006.

[Pri83] A. Prince, Relating to the grid. *Linguistic Inquiry 14*, pp. 19–100, 1983.

[Ric95] J. Rice, *Mathematical Statistics and Data Analysis.* Duxbury Press, 1995.

[Sak78] H. Sakoe and S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. *IEEE Trans. on Acoustics, Speech and Signal Processing*, pp. 43–49, 1978.

[Sel80] E. Selkirk, The role of prosodic categories in english word stress. *Linguistic Inquiry 11*, pp. 563–605, 1980.

[Sel86] E. Selkirk, *Phonology and Syntax.* Cambridge MIT: Press, 1986.

[She96] Y. Sheng, Wavelet transform. In *The transforms and applications handbook*, pp. 747–827, The Electrical Engineering Handbook Series, 1996.

[Sil13] H. Silen, J. Nurminen, E. Helander, and M. Gabbouj, Voice conversion for non-parallel datasets using dynamic kernel partial least squares regression. *Proc. of Interspeech*, 2013.

[SPT] SPTK toolkit, Speech signal processing toolkit (SPTK) version 3.7. http://sp-tk.sourceforge.net/, last access, 5-2-2014.

[Sty98] Y. Stylianou, O. CappÃ©, and E. Moulines, Continuous probabilistic transform for voice conversion. *IEEE transactions on Speech, Audio and language processing*, 6(2):pp. 131–142, 1998.

[Sty05] Y. Stylianou, Modeling speech based on harmonic plus noise models. In *Nonlinear Speech Modeling and Applications*, Springer Berlin / Heidelber, 2005.

[Sun13] A. Suni, D. Aalto, T. Raitio, P. Alku, and M. Vainio, Wavelets for intonation modeling in hmm speech synthesis. *8th ISCA Workshop on Speech Synthesis*, pp. 285–290, 2013.

[Tod07] T. Toda, A. Black, and K. Tokuda, Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on audio, speech and language processing*, pp. 2222–2235, 2007.

[Tok94] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai, Mel-generalized cepstral analysis - a unified approach to speech spectral estimation. *Proc. of the ICSLP*, 1994.

[Vai13] M. Vainio, A. Suni, and D. Aalto, Continuous wavelet transform for analysis of speech prosody. *TRASP*, pp. 78–81, 2013.

[Wan08] C. Wang, Z. Ling, B. Zhang, and L. Dai, Multi-layer f0 modeling for hmm-based speech synthesis. *Proc. ISCSLP*, pp. 129–132, 2008.

[Wik] Wikipedia, Speech production. Online, accessed Feb. 10, 2014, available: http://en.wikipedia.org/wiki/Speech_production.

[You80] R. M. Young, *An introduction to Nonharmonic fourier Series.* Academic Press, New York, 1980.