8

Improvement of Shimmer Parameter of Oesophageal Voices Using Wavelet Transform

Ibon Ruiz and Begoña García Zapirain
Deusto Institute of Technology, Deustotech-LIFE Unit, University of Deusto, Bilbao Spain

1. Introduction

This chapter presents an oesophageal speech enhancement algorithm. Such an exceptionally special type of voice is due to the laryngectomy undergone by those persons with larynx cancer. An oesophageal voice has extremely low intelligibility. The parameter values characterising the voice go beyond normal levels. This chapter proposes a method to improve its quality, which consists in improving Shimmer parameter using Wavelet transform and stabilizing the transfer function poles of the vocal tract model so as to improve a signal's formants. With this aim, the joint use of two techniques has been applied: on the one hand, Digital Wavelet Transform technique to normalise Shimmer and, on the other hand, an algorithm that transforms the modulus and phase of vocal tract's poles technique. The final speech improvement has been measured with the help of Multidemensional Voice Program (MDVP) (Deliyski, 1993) tools and the Shimmer and Harmonic to Noise Ratio (HNR) parameters.

Communication ability of human beings can be extremely influenced by voice disorders. When any problem in the larynx or changes in the voice pitch appear, it could be important to go to the specialist's office to examine the vocal folds movements.

Specialists use computational tools in the objective diagnosis of vocal folds pathologies by means of a set of acoustic parameters among others. There are some patients with severe degradations of speech, as they are the oesophageal voice of laryngectomees.

Patients who have undergone a laryngectomy as a result of larynx cancer have exceptionally low intelligibility. This is due to the removal of their vocal folds, which forces them to use the air flowing through the oesophagus: this is known as oesophageal speech. The characterization parameters for these kinds of oesophageal voices go beyond normal ranges, due to the low quality of the sound itself and its intelligibility.

The cancer of the vocal folds needs to pay special attention in its diagnosis, treatment, rehabilitation and monitoring mainly because it can cause death. Once the cancer has been detected, the otolaryngology (ORL) arranges the vocal folds removal. This implies that patients in such situation will not be able to produce laryngeal voice and hence, they lose the speaking ability. The second most common type of cancer is larynx cancer with a rate of

95%. Every year approximately 136,000 new cases of larynx cancer are diagnosed in the world, with an average survival rate of 5 years in 68% of the cases.

After the operation and during the rehabilitation, the patient will begin the learning stage of oesophageal speech: the voice produced due to the modulation of the air by means of the oesophagus. This will allow the patient to use oesophageal speech which has a degraded quality but it makes possible to maintain a fluid oral communication.

Low intelligibility is the main problem in both oral and telephone communications with other people. In addition, the noise of this kind of speech signal is especially high. This fact has an extremely negative effect on the objective voice parameters, such as pitch, jitter, shimmer and HNR (Harmonic to Noise Ratio). Thus, it is necessary to process voice signal in order to increase intelligibility. The voice enhancement will be measured by those objective parameters. Therefore, the main aim of this work is to recover the normal range of those parameters, to facilitate the laryngectomized collective communication.

Thus, it is necessary to process voice signals in order to increase intelligibility. The voice enhancement will be measured by the objective parameters. Therefore, the main aim of this work is to recover the normal range of the parameters, to facilitate the laryngectomized people communication.

The general objective of this work is to develop an algorithm to enhance and the voice for people who have voice disorders.

2. Methods and system design

2.1 Acoustic parameters

The voice enhancement will be measured by the objective parameters. Therefore, the main aim of this work is to recover the normal range of the parameters, to facilitate the laryngectomized collective communication.

The pitch (Baken & Orlikoff, 2000) is the property of a sound or musical tone measured by its perceived frequency. Due to de pseudo-periodic nature of the voiced speech, there are variations in the instantaneous frequency f_i so the pitch can be defined as

$$Pitch(Hz) = \frac{\sum_{i=1}^{N} f_i}{N} \tag{1}$$

being N the number of extracted pitch periods.

Fundamental frequency estimation has consistently been a difficult topic in audio signal processing because is so difficult to define the time instants which define the voice cycles used to obtain their related instantaneous frequency, f_i.

Furthermore, in acoustical parameterization it is of capital importance to calculate those instants because they are basic features used in this kind of characterization.

Jitter (Baken & Orlikoff, 2000) is a parameter that represents the variation of the fundamental frequency:

Name	Notation	Definition	Units	id
Absolute Jitter	Jita	$Jitter(Hz) = \frac{\sum_{i=1}^{N-1} \left\| T^{(i)} - T^{(i+1)} \right\|}{N-1}$ T.- time period N.- number of extracted pitch periods	Hz	(2)
Jitter Percent	Jit	$Jitter(\%) = \frac{\frac{\sum_{i=1}^{N-1} \left\| T^{(i)} - T^{(i+1)} \right\|}{N-1}}{\sum_{i=1}^{N} \frac{T^{(i)}}{N}}$	%	(3)
Relative Average Perturbation	RAP	$RAP(\%) = 100 \times \frac{\frac{\sum_{i=2}^{i=N-1} \left\| \frac{T^{(i-1)} + T^{(i)} + T^{(i+1)}}{3} - T^{(i)} \right\|}{N-2}}{\sum_{i=1}^{N} \frac{T^{(i)}}{N}}$	%	(4)
Pitch Perturbation Quotient	PPQ	$PPQ(\%) = \frac{\frac{1}{N-4} \sum_{i=1}^{N-4} \left\| \frac{1}{5} \sum_{r=0}^{4} T^{(i+r)} - T^{(i+2)} \right\|}{\frac{1}{N} \sum_{i=1}^{N} T^{(i)}}$	%	(5)
Smoothed Pitch Perturbation Quotient	sPPQ	$sPPQ(\%) = \frac{\frac{1}{N-sf+1} \sum_{i=1}^{N-sf+1} \left\| \frac{1}{sf} \sum_{r=0}^{sf-1} T^{(i+r)} - T^{(i+m)} \right\|}{\frac{1}{N} \sum_{i=1}^{N} T^{(i)}}$ sf.- smoothing factor (typically odd) m.- ½*(sf-1)	%	(6)

Table 1. Jitter definition formulae

In the other hand, specialists also use the reference of shimmer (Baken & Orlikoff, 2000) which is the parameter that represents the amplitude perturbation of the voice signal. The voice produced in vocal folds is supposed to have the ability to maintain its amplitude almost constant, thus an increased value of shimmer may imply a symptom of a voice disorder.

The possible mathematical definitions of shimmer are the following:

Name	Notation	Definition	Units	id
Shimmer Percentage	Shim	$Shim = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left\|A^{(i)} - A^{(i+1)}\right\|}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}}$ $A^{(i)}$ – Extracted peak-to-peak amplitude data, N – Number of extracted impulses	%	(7)
Shimmer	ShdB	$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left\|20\log\left(\frac{A^{(i+1)}}{A^{(i)}}\right)\right\|$	dB	(8)
Smoothed Amplitude Perturbation Quotient	sAPQ	$SAPQ = \frac{\frac{1}{N-sf+1}\sum_{i=1}^{N-sf+1}\left\|\frac{1}{sf}\sum_{r=0}^{sf-1}A^{(i+r)} - A^{(i+m)}\right\|}{\frac{1}{N}\sum_{i=1}^{N}A^{(i)}}$ sf – smoothing factor (typically odd) m.- ½*(sf-1)	%	(9)

Table 2. Shimmer definition formulae

Several authors have reported decent results using voice cycle detection (Chen & Kao, 2001) (Hagmüller & Kubina, 2006) and there are many techniques widely detailed in literature: time domain estimators (e.g. Zero Crossing Rate (Kedem, 1986)), fundamental frequency estimators (Dorken & Nawab, 1994) (Piszczalski & Galler, 1979), Autocorrelation methods (Yin Estimator, (Cheveigné & Kawahara, 2002)), Phase Space representation (Gibiat, 1988), Cepstrum (Flanagan, 1965) and Statistical Methods (Sano & Jenkins, 1989) (Doval & Rodet, Estimation of fundamental frequency of musical sound signals, 1991) (Doval & Rodet, 1993). Some of them define directly the voice cycles (Chen & Kao, 2001) while others are used to calculate a numerical approximation (Cheveigné & Kawahara, 2002) to the fundamental frequency value. In these ones, a further step is necessary in order to identify clearly which instants define the voice cycles.

The HNR is a general evaluation of noise present in the analyzed signal. It is defined as (10), $r_p(0)$ and $r_{ap}(0)$ being the respective energies of the periodic and aperiodic components:

$$HNR = \frac{r_p(0)}{r_{ap}(0)} \tag{10}$$

The measures have been made with the help of MDVP from Kay Electronics important software that gives good estimations of a signal's parameters. However, this software is not specialized in pathological speech. To fix this problem, the voiced period marks are needed to calculate the pitch and then the HNR have been manually introduced, one by one, on each signal.

The MDVP calculates the HNR as the average ratio of the harmonic spectral energy in the frequency range 70-4500 Hz and the enharmonic spectral energy in the frequency range 1500-4500 Hz (Deliyski, 1993) (Yumoto & Gould, 1982).

2.2 Wavelets

It is well known from Fourier theory that a signal can be expressed as the sum of a, possibly infinite, series of sines and cosines. This sum is also referred to as a Fourier expansion. The big disadvantage of a Fourier expansion however is that it has only frequency resolution and no time resolution. This means that although we might be able to determine all the frequencies present in a signal, we do not know when they are present. To overcome this problem in the past decades several solutions have been developed which are more or less able to represent a signal in the time and frequency domain at the same time.

The idea behind these time-frequency joint representations is to cut the signal of interest into several parts and then analyze the parts separately. It is clear that analyzing a signal this way will give more information about the when and where of different frequency components, but it leads to a fundamental problem as well: how to cut the signal? Suppose that we want to know exactly all the frequency components present at a certain moment in time. We cut out only this very short time window using a Dirac pulse1, transform it to the frequency domain and … something is very wrong.

The problem here is that cutting the signal corresponds to a convolution between the signal and the cutting window. Since convolution in the time domain is identical to multiplication in the frequency domain and since the Fourier transform of a Dirac pulse contains all possible frequencies the frequency components of the signal will be smeared out all over the frequency axis. In fact this situation is the opposite of the standard Fourier transform since we now have time resolution but no frequency resolution whatsoever.

The underlying principle of the phenomena just described is Heisenberg's uncertainty principle, which, in signal processing terms, states that it is impossible to know the exact frequency and the exact time of occurrence of this frequency in a signal. In other words, a signal can simply not be represented as a point in the time-frequency space. The uncertainty principle shows that it is very important how one cuts the signal.

The *wavelet transform* or *wavelet analysis* is probably the most recent solution to overcome the shortcomings of the Fourier transform. In wavelet analysis the use of a fully scalable modulated window solves the signal-cutting problem. The window is shifted along the signal and for every position the spectrum is calculated. Then this process is repeated many times with a slightly shorter (or longer) window for every new cycle. In the end the result will be a collection of time-frequency representations of the signal, all with different resolutions. Because of this collection of representations we can speak of a multiresolution analysis. In the case of wavelets we normally do not speak about time-frequency representations but about time-scale representations, scale being in a way the opposite of frequency, because the term frequency is reserved for the Fourier transform.

2.2.1 The Continuous Wavelet Transform (CWT)

The wavelet analysis described in the introduction is known as the *continuous wavelet transform* or *CWT*. More formally it is written as:

$$\gamma(s,\tau) = \int f(t)\,\Psi^*_{s,\tau}(t)dt \tag{11}$$

where * denotes complex conjugation. This equation shows how a function $f(t)$ is decomposed into a set of basis functions $\Psi^*_{s,\tau}(t)$, called the wavelets. The variables s and - are the new dimensions, scale and translation, after the wavelet transform. For completeness sake equation (11) gives the inverse wavelet transform. I will not expand on this since we are not going to use it:

$$f(t) = \iint \gamma(s,\tau)\,\Psi^*_{s,\tau}(t)d\tau ds \tag{12}$$

The wavelets are generated from a single basic wavelet $\Psi(t)$, the so-called *mother wavelet*, by scaling and translation:

$$\Psi^*_{s,\tau}(t) = \frac{1}{\sqrt{s}}\Psi\left(\frac{t-\tau}{s}\right) \tag{13}$$

In (13) s is the scale factor, τ is the translation factor and the factor $s^{-1/2}$ is for energy normalization across the different scales (Lió, 2003) (Ortolan, Mori, Pereira, Cabral, Pereira, & Cliquet, 2003).

It is important to note that in (11), (12) and (13) the wavelet basis functions are not specified. This is a difference between the wavelet transform and the Fourier transform, or other transforms. The theory of wavelet transforms deals with the general properties of the wavelets and wavelet transforms only. It defines a framework within one can design wavelets to taste and wishes.

2.2.2 Discrete wavelet

Now that we know what the wavelet transform is, we would like to make it practical. However, the wavelet transform as described so far still has three properties that make it difficult to use directly in the form of (11). The first is the redundancy of the CWT. In (11) the wavelet transform is calculated by continuously shifting a continuously scalable function over a signal and calculating the correlation between the two. It will be clear that these scaled functions will be nowhere near an orthogonal basis5 and the obtained wavelet coefficients will therefore be highly redundant. For most practical applications we would like to remove this redundancy.

Even without the redundancy of the CWT we still have an infinite number of wavelets in the wavelet transform and we would like to see this number reduced to a more manageable count. This is the second problem we have. The third problem is that for most functions the wavelet transforms have no analytical solutions and they can be calculated only numerically or by an optical analog computer. Fast algorithms are needed to be able to exploit the power of the wavelet transform and it is in fact the existence of these fast algorithms that have put wavelet transforms where they are today.

As mentioned before the CWT maps a one-dimensional signal to a two-dimensional time-scale joint representation that is highly redundant. The time-bandwidth product of the CWT is the square of that of the signal and for most applications, which seek a signal description with as few components as possible, this is not efficient. To overcome this problem *discrete wavelets* have been introduced. Discrete wavelets are not continuously scalable and translatable but can only be scaled and translated in discrete steps. This is achieved by

modifying the wavelet representation (13) to create (Tohidypour, Seyyedsalehi, & Behbood, 2010) (Daubechies, 1992).

$$\Psi_{j,k}(t) = \frac{1}{\sqrt{s_0^j}} \Psi\left(\frac{t - k\tau_0 s_0^j}{s_0^j}\right) \tag{14}$$

Although it is called a discrete wavelet, it normally is a (piecewise) continuous function. In (14) j and k are integers and $s_0 > 1$ is a fixed dilation step. The translation factor τ_0 depends on the dilation step. The effect of discretizing the wavelet is that the time-scale space is now sampled at discrete intervals. We usually choose $s_0 = 2$ so that the sampling of the frequency axis corresponds to *dyadic sampling*. This is a very natural choice for computers, the human ear and music for instance. For the translation factor we usually choose $\tau_0 = 1$ so that we also have dyadic sampling of the time axis.

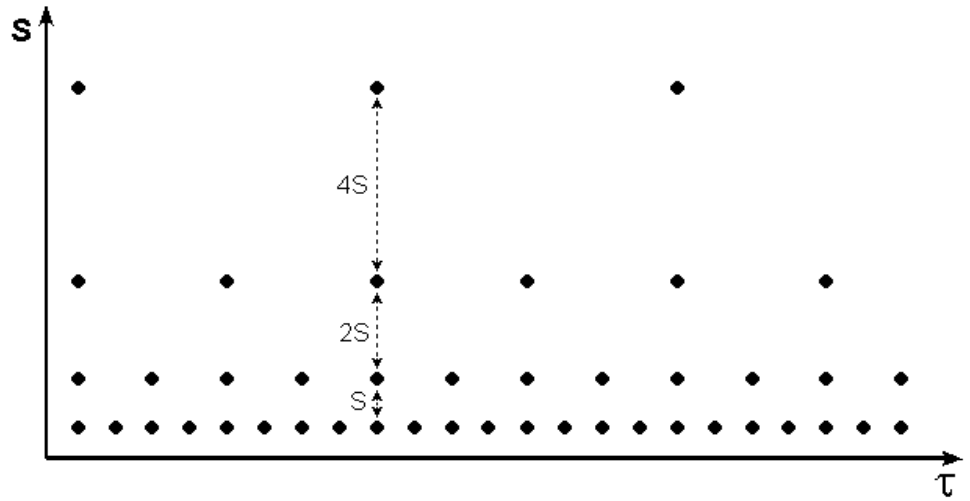

Fig. 1. Localization of the discrete wavelets in the time-scale space on a dyadic grid.

When discrete wavelets are used to transform a continuous signal the result will be a series of wavelet coefficients, and it is referred to as the *wavelet series decomposition*. An important issue in such a decomposition scheme is of course the question of reconstruction. It is all very well to sample the time-scale joint representation on a dyadic grid, but if it will not be possible to reconstruct the signal it will not be of great use. As it turns out, it is indeed possible to reconstruct a signal from its wavelet series decomposition. In (Daubechies, 1992) it is proven that the necessary and sufficient condition for stable reconstruction is that the energy of the wavelet coefficients must lie between two positive bounds, i.e.

$$A\|f\|^2 \leq \sum_{j,k} \left|\langle f, \Psi_{j,k}\rangle\right|^2 \leq B\,\|f\|^2 \tag{15}$$

where $\|f\|^2$ is the energy of $f(t)$, $A > 0$, $B < \infty$ and A, B are independent of $f(t)$. When equation $\left(A\|f\|^2 \leq \sum_{j,k} \left|\langle f, \Psi_{j,k}\rangle\right|^2 \leq B\,\|f\|^2\right.$ (15) is satisfied, the family of basis functions

$\Psi_{j,k}(t)$ with $j, k \in \mathbb{Z}$ is referred to as a *frame* with frame bounds A and B. When $A = B$ the frame is *tight* and the discrete wavelets behave exactly like an orthonormal basis. When $A \neq B$ exact reconstruction is still possible at the expense of a *dual frame*. In a dual frame discrete wavelet transform the decomposition wavelet is different from the reconstruction wavelet.

We will now immediately forget the frames and continue with the removal of all redundancy from the wavelet transform. The last step we have to take is making the discrete wavelets orthonormal. This can be done only with discrete wavelets. The discrete wavelets can be made orthogonal to their own dilations and translations by special choices of the mother wavelet, which means:

$$\int \Psi_{j,k}(t)\Psi^*_{m,n}(t)\,dt = \begin{cases} 1 & if\ j = m\ and\ k = n \\ 0 & otherwise \end{cases} \tag{16}$$

An arbitrary signal can be reconstructed by summing the orthogonal wavelet basis functions, weighted by the wavelet transform coefficients (Sheng, 1996):

$$f(t) = \sum_{j,k} \gamma(j,k)\Psi_{j,k}(t) \tag{17}$$

Equation (17) shows the inverse wavelet transform for discrete wavelets, which we had not yet seen.

Orthogonality is not essential in the representation of signals. The wavelets need not be orthogonal and in some applications the redundancy can help to reduce the sensitivity to noise (Sheng, 1996) or improve the *shift invariance* of the transform (Burrus, Goinath, & Guo, 1998). This is a disadvantage of discrete wavelets: the resulting wavelet transform is no longer shift invariant, which means that the wavelet transforms of a signal and of a time-shifted version of the same signal are not simply shifted versions of each other.

2.2.3 A band-pass filter

With the redundancy removed, we still have two hurdles to take before we have the wavelet transform in a practical form. We continue by trying to reduce the number of wavelets needed in the wavelet transform and save the problem of the difficult analytical solutions for the end.

Even with discrete wavelets we still need an infinite number of scalings and translations to calculate the wavelet transform. The easiest way to tackle this problem is simply not to use an infinite number of discrete wavelets. Of course this poses the question of the quality of the transform. Is it possible to reduce the number of wavelets to analyze a signal and still have a useful result?

The translations of the wavelets are of course limited by the duration of the signal under investigation so that we have an upper boundary for the wavelets. This leaves us with the question of dilation: how many scales do we need to analyze our signal? How do we get a lower bound? It turns out that we can answer this question by looking at the wavelet transform in a different way.

If we look wavelets proprieties we see that the wavelet has a band-pass like spectrum. From Fourier theory we know that compression in time is equivalent to stretching the spectrum and shifting it upwards:

$$F\{f(at)\} = \frac{1}{a} F\left(\frac{\omega}{a}\right) \tag{18}$$

This means that a time compression of the wavelet by a factor of 2 will stretch the frequency spectrum of the wavelet by a factor of 2 and also shift all frequency components up by a factor of 2. Using this insight we can cover the finite spectrum of our signal with the spectra of dilated wavelets in the same way as that we covered our signal in the time domain with translated wavelets. To get a good coverage of the signal spectrum the stretched wavelet spectra should touch each other, as if they were standing hand in hand (see Fig. 2). This can be arranged by correctly designing the wavelets.

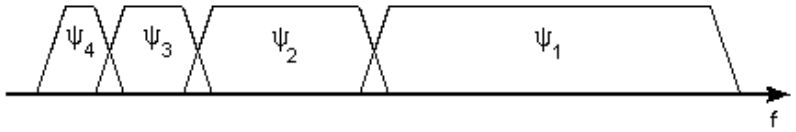

Fig. 2. Touching wavelet spectra resulting from scaling of the mother wavelet in the time domain

Summarizing, if one wavelet can be seen as a band-pass filter, then a series of dilated wavelets can be seen as a band-pass filter bank. If we look at the ratio between the centre frequency of a wavelet spectrum and the width of this spectrum we will see that it is the same for all wavelets. This ratio is normally referred to as the fidelity factor Q of a filter and in the case of wavelets one speaks therefore of a *constant-Q* filter bank.

2.2.4 The scaling function

The careful reader will now ask him- or herself the question how to cover the spectrum all the way down to zero? Because every time you stretch the wavelet in the time domain with a factor of 2, its bandwidth is halved. In other words, with every wavelet stretch you cover only half of the remaining spectrum, which means that you will need an infinite number of wavelets to get the job done.

The solution to this problem is simply not to try to cover the spectrum all the way down to zero with wavelet spectra, but to use a cork to plug the hole when it is small enough. This cork then is a low-pass spectrum and it belongs to the so-called *scaling function*. The scaling function was introduced by Mallat (Mallat, 1989). Because of the low-pass nature of the scaling function spectrum it is sometimes referred to as the *averaging filter*.

If we look at the scaling function as being just a signal with a low-pass spectrum, then we can decompose it in wavelet components and express it like (17):

$$\varphi(t) = \sum_{j,k} \gamma(j,k) \Psi_{j,k}(t) \tag{19}$$

Since we selected the scaling function $\varphi(t)$ in such a way that its spectrum neatly fitted in the space left open by the wavelets, the expression $(\varphi(t) = \sum_{j,k} \gamma(j,k) \Psi_{j,k}(t)$ (19) uses an infinite number of wavelets up to a certain scale j (see Fig. 3). This means that if we analyze a signal using the combination of scaling function and wavelets, the scaling function by itself takes care of the spectrum otherwise covered by all the wavelets up to scale j, while the rest is done by the wavelets. In this way we have limited the number of wavelets from an infinite number to a finite number.

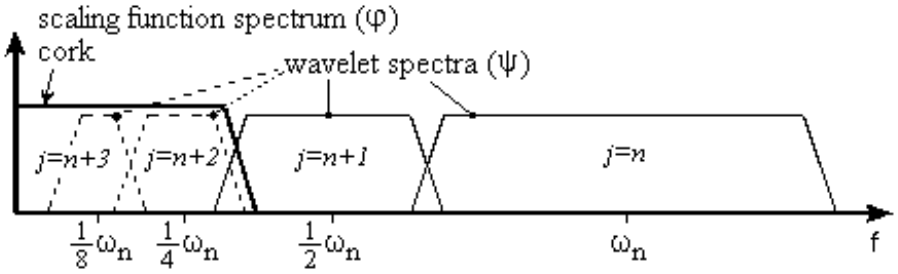

Fig. 3. How an infinite set of wavelets is replaced by one scaling function.

By introducing the scaling function we have circumvented the problem of the infinite number of wavelets and set a lower bound for the wavelets. Of course when we use a scaling function instead of wavelets we lose information. That is to say, from a signal representation view we do not lose any information, since it will still be possible to reconstruct the original signal, but from a wavelet-analysis point of view we discard possible valuable scale information. The width of the scaling function spectrum is therefore an important parameter in the wavelet transform design. The shorter its spectrum the more wavelet coefficients you will have and the more scale information. But, as always, there will be practical limitations on the number of wavelet coefficients you can handle. As we will see later on, in the discrete wavelet transform this problem is more or less automatically solved.

Summarizing once more, if one wavelet can be seen as a band-pass filter and a scaling function is a low-pass filter, then a series of dilated wavelets together with a scaling function can be seen as a filter bank.

2.2.5 Subband coding

Two of the three problems mentioned in section 4 have now been resolved, but we still do not know how to calculate the wavelet transform. Therefore we will continue our journey through multiresolution land.

If we regard the wavelet transform as a filter bank, then we can consider wavelet transforming a signal as passing the signal through this filter bank. The outputs of the different filter stages are the wavelet- and scaling function transform coefficients. Analyzing a signal by passing it through a filter bank is not a new idea and has been around for many years under the name *subband coding*. It is used for instance in computer vision applications.

The filter bank needed in subband coding can be built in several ways. One way is to build many band-pass filters to split the spectrum into frequency bands. The advantage is that the width of every band can be chosen freely, in such a way that the spectrum of the signal to analyze is covered in the places where it might be interesting. The disadvantage is that we will have to design every filter separately and this can be a time consuming process. Another way is to split the signal spectrum in two (equal) parts, a low-pass and a high-pass part. The high-pass part contains the smallest details we are interested in and we could stop here. We now have two bands. However, the low-pass part still contains some details and therefore we can split it again. And again, until we are satisfied with the number of bands

we have created. In this way we have created an *iterated filter bank*. Usually the number of bands is limited by for instance the amount of data or computation power available. The process of splitting the spectrum is graphically displayed in figure 4. The advantage of this scheme is that we have to design only two filters; the disadvantage is that the signal spectrum coverage is fixed.

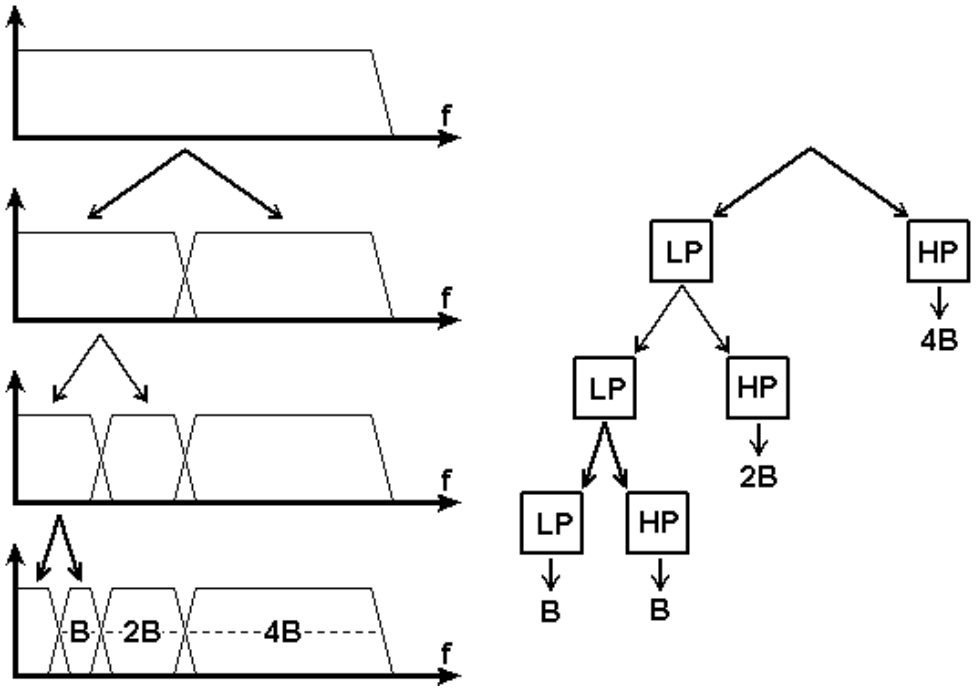

Fig. 4. Splitting the signal spectrum with an iterated filter bank.

Looking at figure 4 we see that what we are left with after the repeated spectrum splitting is a series of band-pass bands with doubling bandwidth and one low-pass band. (Although in theory the first split gave us a high-pass band and a low-pass band, in reality the high-pass band is a band-pass band due to the limited bandwidth of the signal.) In other words, we can perform the same subband analysis by feeding the signal into a bank of band-pass filters of which each filter has a bandwidth twice as wide as his left neighbour (the frequency axis runs to the right here) and a low-pass filter. At the beginning of this section we stated that this is the same as applying a wavelet transform to the signal. The wavelets give us the band-pass bands with doubling bandwidth and the scaling function provides us with the low-pass band. From this we can conclude that a wavelet transform is the same thing as a subband coding scheme using a constant-Q filter bank (Mallat, 1989). In general we will refer to this kind of analysis as a multiresolution analysis.

Summarizing, if we implement the wavelet transform as an iterated filter bank, we do not have to specify the wavelets explicitly! This sure is a remarkable result.

2.2.6 The Discrete Wavelet Transform (DWT)

In many practical applications and especially in the application described in this report the signal of interest is sampled. In order to use the results we have achieved so far with a discrete signal we have to make our wavelet transform discrete too. Remember that our discrete wavelets are not time-discrete, only the translation- and the scale step are discrete. Simply implementing the wavelet filter bank as a digital filter bank intuitively seems to do the job. But intuitively is not good enough, we have to be sure.

In (19) we stated that the scaling function could be expressed in wavelets from minus infinity up to a certain scale *j*. If we add a wavelet spectrum to the scaling function spectrum we will get a new scaling function, with a spectrum twice as wide as the first. The effect of this addition is that we can express the first scaling function in terms of the second, because all the information we need to do this is contained in the second scaling function. We can express this formally in the so-called multiresolution formulation (Burrus, Goinath, & Guo, 1998) or *two-scale relation* (Sheng, 1996):

$$\varphi(2^j t) = \sum_k h_{j+1}(k)\varphi(2^{j+1}t - k) \tag{20}$$

The two-scale relation states that the scaling function at a certain scale can be expressed in terms of translated scaling functions at the next smaller scale. Do not get confused here: smaller scale means more detail.

The first scaling function replaced a set of wavelets and therefore we can also express the wavelets in this set in terms of translated scaling functions at the next scale. More specifically we can write for the wavelet at level *j*:

$$\Psi(2^j t) = \sum_k g_{j+1}(k)\varphi(2^{j+1}t - k) \tag{21}$$

which is the two-scale relation between the scaling function and the wavelet.

Since our signal $f(t)$ could be expressed in terms of dilated and translated wavelets up to a scale *j*-1, this leads to the result that $f(t)$ can also be expressed in terms of dilated and translated scaling functions at a scale *j*:

$$f(t) = \sum_k \lambda_j(k)\varphi(2^j t - k) \tag{22}$$

To be consistent in our notation we should in this case speak of discrete scaling functions since only discrete dilations and translations are allowed. If in this equation we step up a scale to *j*-1, we have to add wavelets in order to keep the same level of detail. We can then express the signal $f(t)$ as

$$f(t) = \sum_k \lambda_{j-1}(k)\varphi(2^{j-1}t - k) + \sum_k \gamma_{j-1}(k)\Psi(2^{j-1}t - k) \tag{23}$$

If the scaling function $\varphi_{j,k}(t)$ and the wavelets $\Psi_{j,k}(t)$ are orthonormal or a tight frame, then the coefficients $\lambda_{j-1}(k)$ and $\gamma_{j-1}(k)$ are found by taking the inner products

If we now replace $\varphi_{j,k}(t)$ and $\Psi_{j,k}(t)$ in the inner products by suitably scaled and translated versions of (20) and (21) and manipulate a bit, keeping in mind that the inner product can also be written as an integration, we arrive at the important result (Burrus, Goinath, & Guo, 1998):

$$\lambda_{j-1}(k) = \sum_m h(m - 2k)\,\lambda_j(m) \quad (24)$$

$$\gamma_{j-1}(k) = \sum_m g(m - 2k)\,\gamma_j(m) \quad (25)$$

These two equations state that the wavelet- and scaling function coefficients on a certain scale can be found by calculating a weighted sum of the scaling function coefficients from the previous scale. Now recall from the section on the scaling function that the scaling function coefficients came from a low-pass filter and recall from the section on subband coding how we iterated a filter bank by repeatedly splitting the low-pass spectrum into a low-pass and a high-pass part. The filter bank iteration started with the signal spectrum, so if we imagine that the signal spectrum is the output of a low-pass filter at the previous (imaginary) scale, then we can regard our sampled signal as the scaling function coefficients from the previous (imaginary) scale. In other words, our sampled signal $f(k)$ is simply equal to $\lambda(k)$ at the largest scale!

As we know from signal processing theory a discrete weighted sum like the ones in (24) and (25) is the same as a digital filter and since we know that the coefficients λj(k) come from the low-pass part of the splitted signal spectrum, the weighting factors h(k) in ($\lambda j - 1(k) = \sum_m h(m - 2k)\,\lambda_j(m)$ (24)) must form a low-pass filter. And since we know that the coefficients γj(k) come from the high-pass part of the splitted signal spectrum, the weighting factors g(k) in (24) must form a high-pass filter. This means that (24) and (25) together form one stage of an iterated digital filter bank and from now on we will refer to the coefficients h(k) as the scaling filter and the coefficients g(k) as the wavelet filter.

By now we have made certain that implementing the wavelet transform as an iterated digital filter bank is possible and from now on we can speak of the discrete wavelet transform or DWT. Our intuition turned out to be correct. Because of this we are rewarded with a useful bonus property of (24) and (25), the subsampling property. If we take one last look at these two equations we see that the scaling and wavelet filters have a step-size of 2 in the variable k. The effect of this is that only every other λj(k) is used in the convolution, with the result that the output data rate is equal to the input data rate. Although this is not a new idea, it has always been exploited in subband coding schemes, it is kind of nice to see it pop up here as part of the deal.

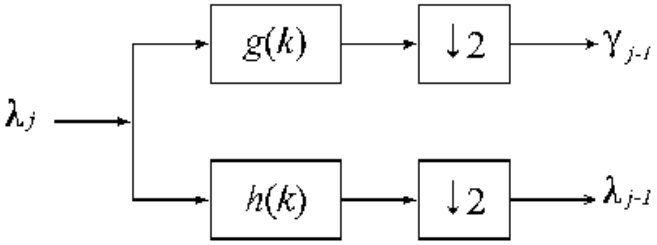

Fig. 5. Implementation of (23) and (24) as one stage of an iterated filter bank.

The subsampling property also solves our problem, which had come up at the end of the section on the scaling function, of how to choose the width of the scaling function spectrum. Because, every time we iterate the filter bank the number of samples for the next stage is halved so that in the end we are left with just one sample (in the extreme case). It will be

clear that this is where the iteration definitely has to stop and this determines the width of the spectrum of the scaling function. Normally the iteration will stop at the point where the number of samples has become smaller than the length of the scaling filter or the wavelet filter, whichever is the longest, so the length of the longest filter determines the width of the spectrum of the scaling function.

2.3 Wavelets algorithm

The general goal of this investigation, and so of all the previous researches (García, Vicente, Ruiz, Angulo, & Aramendi, 2002) is the improvement of the oesophageal voices' quality (García, Vicente, Ruiz, Alonso, & Loyo, 2005). Certainly the specific aim of this research is the spectral and temporal correction of the shimmer and parameter of these voices by the Wavelet Transform.

One of the most important techniques applied in the spectral analysis is the Fourier Transform (STFT), which will allow recognizing the spectral components of speech signal, so it makes possible to distinguish pathological voices and process them. That transform has a resolution problem which is explained by Heisenberg Uncertainty Principle. The Wavelet Transform (WT) was developed to overcome some resolutions related problems of the STFT. It is possible to analyze any signal by using an alternative approach called the Multiresolution Analysis (MRA). MRA, as implied by its name, analyzes the signal at different frequencies with different resolutions. MRA is designed to give good time resolution and poor frequency resolution at high frequencies and good frequency resolution and poor time resolution at low frequencies.

As the signals used are digital, it is more useful to use Discrete Wavelet Transform (DWT) (Mallat, 1999). The DWT analyzes the signal at different frequency bands with different resolutions by decomposing the signal into a coarse approximation and detail information. The decomposition of the signal into different frequency bands is simply obtained by successive high pass and low pass filtering of the time domain signal. The original signal x[n] is first passed through a half band high pass filter g[n] and a low pass filter h[n]. This constitutes one level of decomposition and can mathematically be expressed as follows (Kadambe & Bourdreaux-Bartels, 1991):

$$y_{high}[k] = \sum_{n} x[n] \cdot g[2k - n] \tag{26}$$

$$y_{low}[k] = \sum_{n} x[n] \cdot h[2k - n] \tag{27}$$

Before applying the DWT, the signal is processed. That is, a resample of the original signal, x[n], at a sampling frequency of 12800 Hz. This is so done, as when applying the transformed DWT, the detail signals remain between the frequency bands that are suitable for pitch detection (Kadambe & Bourdreaux-Bartels, A Comparison of a Wavelet Functions for Pitch Detection of Speech Signals, 1991) (Kadambe & Bourdreaux-Bartels, 1992) (Wing-kei, Kwong-sak, & Kin-hong, 1995) (Nadeu, Pascual, & Herdondo, 1991). More specifically, the oesophageal voices have a pitch nearing 60 Hz. On doing the above-mentioned resample and the following transformed DWT, one of the details is found in the frequency band level of 50 Hz – 100 Hz. This means that the original pitch signal's information is located within this detail. Low-frequency noise present in oesophageal voices are found in the 0 Hz – 50 Hz level. We should eliminate this noise before modifying the pitch's peak amplitude.

In short, so as to control the high rates of the shimmer parameter in oesophageal voices, the following steps should be taken: carry out a resample of the original signal at Fs = 12800 Hz; after this the transformed DWT should be done, for which we have used "bior 6.8" as the mother wavelet. Trials with other mother wavelets were done and the results are quite similar to as regards shimmer measurements. Once the DWT transform has been done, the low-frequency noise in the 0 Hz – 50 Hz frequency band is eliminated. After this pre-processing, the amplitude of the maximums in the 50 Hz – 100 Hz frequency band are modified, as this is where the information on oesophageal voices is to be found.

Fig. 6 shows the frequency band tree when DWT is applied.

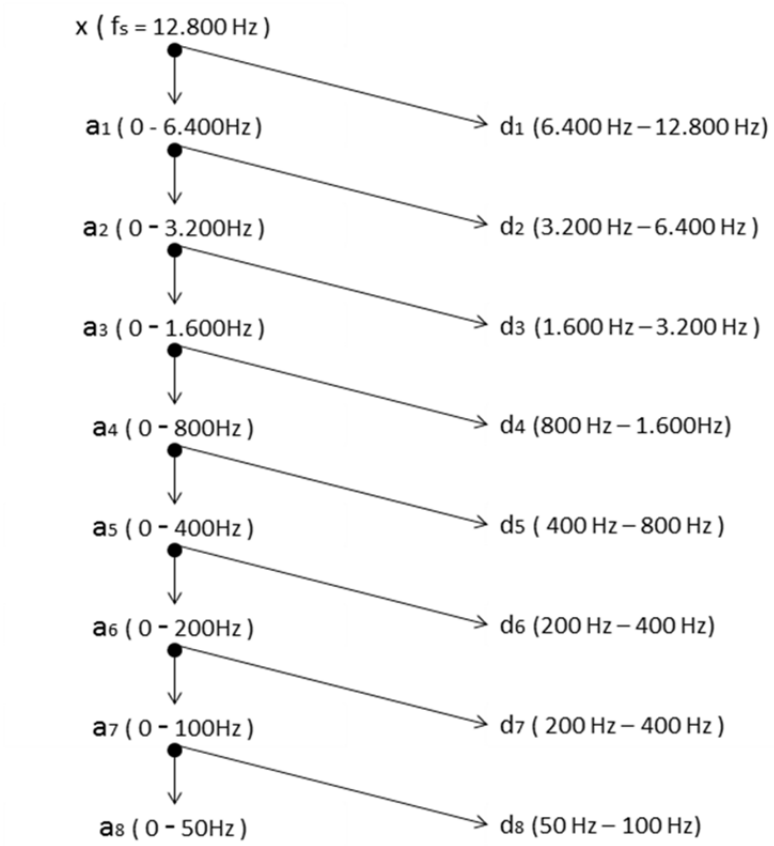

Fig. 6. Frequency band diagram

2.4 Poles stabilization algorithm

The stabilizations of poles, the second algorithm is responsible for analyzing and modifying the poles of the system modelized by the vocal tract. It works with an oesophageal voice signal from which the excitation has been separated from the tract, and it calculates the evolution of modulus and phase of each formant of the vowel modifying such poles.

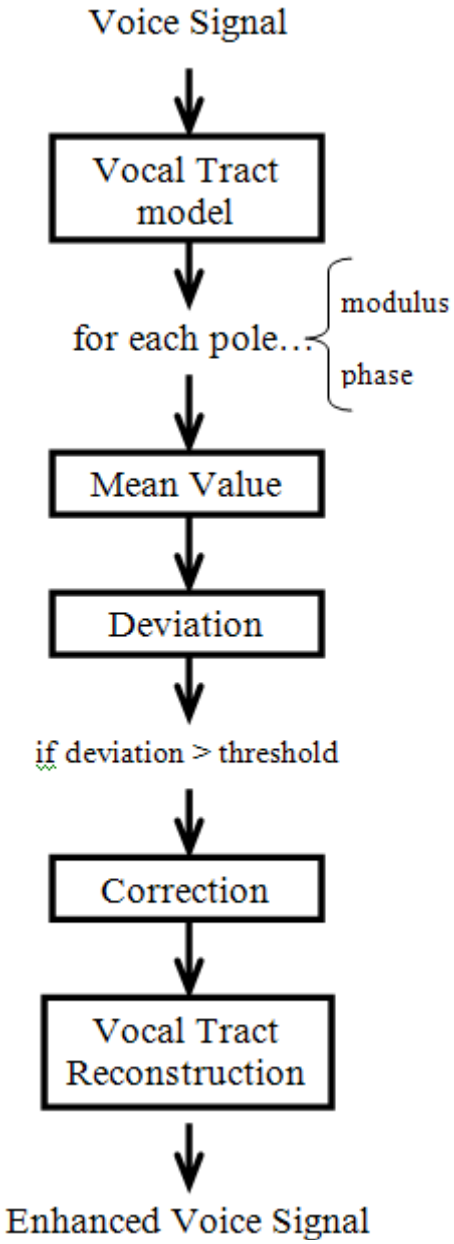

Fig. 7. Pole Stabilization Block Diagram

The stabilization of the first three formants is applied in those values of the vowel which is being enhanced by means of the modification of the first three poles, following these steps:

1. Calculation of the mean value of modulus and phase of each pole through the vocal signal.
2. Calculation of the maximum deviations relative to the mean modulus and phase of the first three poles.
3. Whether the deviations exceed a certain threshold is analyzed, if so the modulus correction is applied:
   a. ModulusModif=modulus+((1-modulus)*ConstMod);
   b. and the phase correction:
      AngModif=Angle-(ContPhase*(Angle+MeanPhase));
      being the correction implanted by means of "ConstMod" and "ConstPhase" parameters which can be adjusted for each voice.
4. Reconstruction of the filter that modelizes the vocal tract with the new poles of the system corrected and stabilized.

3. Results

On the one hand, in the DWT algorithm, the inputs of the developed algorithm are the samples of the oesophageal voice, which shimmer parameter have been previously evaluated. In the 100% of the studied cases obtained value for these parameters is out of the range of the normality. So it could be improved in order to increase the quality of the voice between the normality ranges specified by the scientific community. After the application of the algorithm based on the analysis and processing by Wavelet, the speech signal has been reconstructed. When measuring the shimmer in this reconstructed signal, the obtained results are the following:

Oesophageal Phoneme	Shimmer Real (%)	Transformed Shimmer (%)
a1	18,43	10,27
a2	17,27	10,82
a3	16,05	9,98
a4	19,62	12,79
a5	16,76	8,43
a6	18,58	12,99
a7	12,53	10,11
a8	16,98	9,99
a9	20,31	12,83
a10	18,22	10,61

Table 3. Table of shimmer values

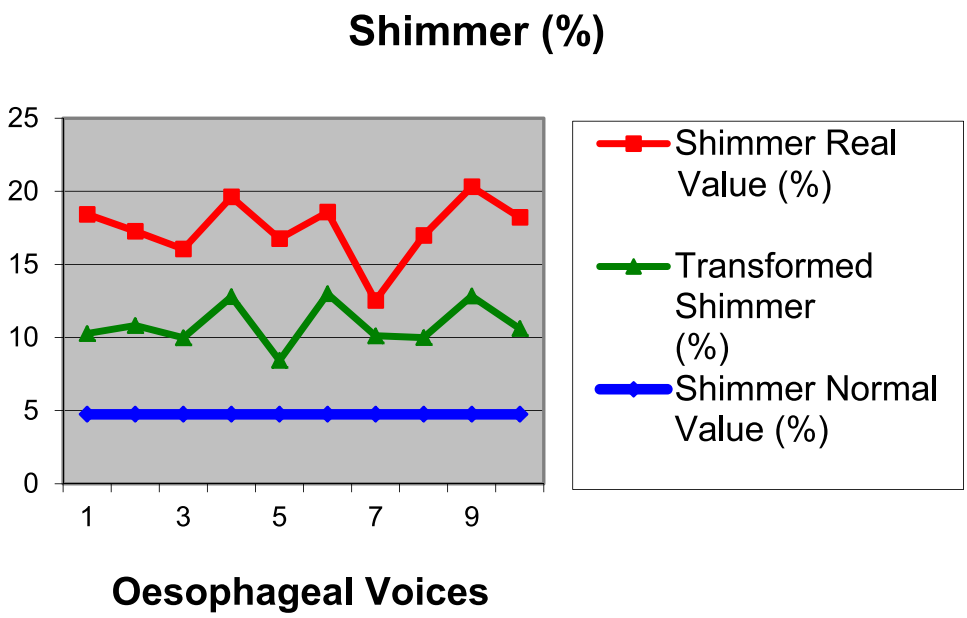

Fig. 8. Shimmer of different voices

Phoneme	Original HNR (dB)	Stabilization HNR (dB)	HNR (dB) increase
a1	-5.001	-1.701	3.300
a2	0.549	1.656	1.107
a3	-3.684	-2.219	1.465
a4	-4.901	-0.668	4.462
a5	-6.375	-2.631	3.744
a6	-6.803	-3.159	3.644
a7	-6.389	-4.451	1.938
a8	-8.724	-5.615	3.109
a9	-3.737	-0.040	3.697
a10	0.930	1.846	0.916
Average			2.941

Table 4. HNR measures with the /a/ phonemes.

As is shown in the table 3 the shimmer has improved in 9 of 10 cases. In four out of the ten voices researched a great goal has been reached. They are not only improved in terms of quality, moreover their values are situated nearest of the limits of normality stipulated in Fig. 8. On the other hand, in the pole's stabilization algorithm, in all cases an increase in harmonics to noise ratio has been achieved. For example, the value of "a1" signal is 5.001dB before processing and 1.701dB afterwards. The increase in improved oesophageal signal, in this case, has been 3.3dB. The fourth column in table 1 shows the enhancement of HNR (dB) before and after processing. It can be appreciated that the improvement in HNR (dB) ranges from 0.916dB, for "a10" signal, to 4.462dB, for "a4". Taking into account all the database, the average HNR improvement (dB) is 2.941dB.

As can be seen in Fig. 9 the increase of the HNR occurs in all voices of the database.

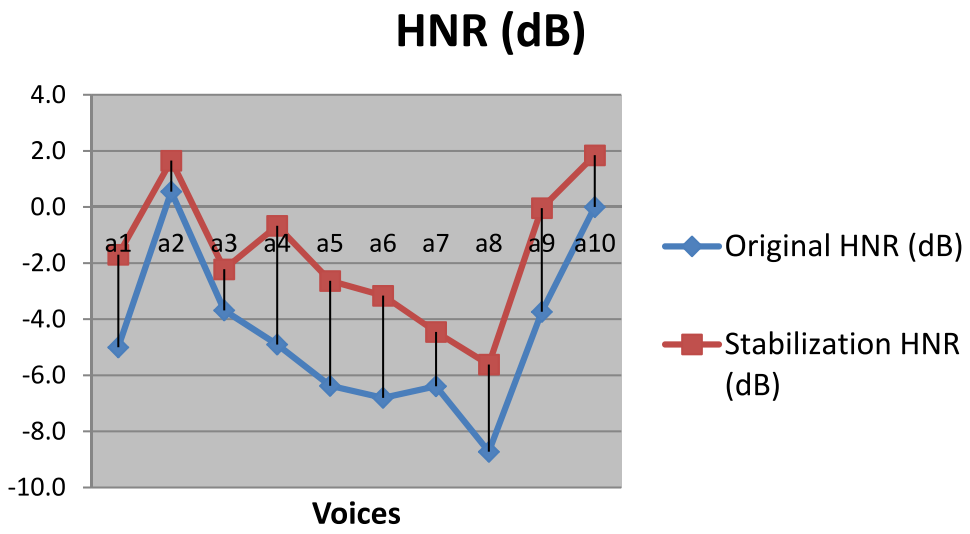

Fig. 9. HNR before and after algorithm

4. Conclusion

It can be concluded that the aimed objectives have been achieved because of the fact that the algorithms are very suitable.

The usage of the Wavelet Transform for the analysis and processing of oesophageal voices is successful in the improvement of the shimmer, which is the aim of the paper. In a extensive analysis its appreciable that it is also good for the improvement of other parameters such as the harmonics to noise ratio. Being a single wavelet detail, optimisation of the computational calculation when processing a simpler signal favours the application of the proposed algorithm to prototypes that process oesophageal signals in real time, in order to improve their quality. On the other hand, the close relationship between characterisation parameters, such as shimmer or jitter and the values of the signal situated in frequency

intervals below 100Hz reinforces the suitability of working with bands inferior to the Wavelet Transform, which distinguish spectral components and enable the focusing on particular components.

Therefore, DWT and both pole stabilization improvement are suitable techniques in the speech enhancement context.

5. Acknowledgment

The authors wish to acknowledge the Deusto University which kindly lend infrastructures and material for this investigation. They would also like to thank to all the scholarships that so enthusiastically have collaborated with this project.

Especially it cannot be forgotten the help of the "Asociación Vizcaína de Laringectomizados" whose members, voluntarily lend his voices for this investigation, without their help it would not be possible to carry out this project.

This work was supported in part by the Basque Country Department of Education, Universities and Research.

6. References

Bagshaw, P., Hiller, S., & Jack, M. (1993). Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. Eurospeech, (págs. 22–25). Berlin.

Baken, P. J., & Orlikoff, R. F. (2000). Clinical Measurement of Speech and Voice. San Diego: Singular Publishing Group.

Burrus, C. S., Goinath, R. A., & Guo, H. (1998). Introduction to wavelets and wavelet transforms, a primer. Upper Saddle River NJ (USA): Prentice Hal.

Chen, J., & Kao, Y. (2001). Pitch marking based on an adaptable filter and a peakvalley estimation method. Computational Linguistics and Chinese Language Processing , 6, 1-112.

Cheveigné, A., & Kawahara, H. (2002). Yin, a fundamental frequency estimator for speech and music, 111 (4) (2002). Journal of the Acoustical Society of America, 11 (4).

Daubechies, I. (1992). Ten lectures on wavelets. Philadelphia: 2nd ed. Philadelphia: SIAM.

Deliyski, D. D. (1993). MDVP Acoustic Model and Evaluation of Pathological Voice Production. Eurospeech. Berlin.

Dorken, E., & Nawab, S. H. (1994). Improved musical pitch tracking using principal decomposition analysis. ICASSP, (págs. 217-220).

Doval, B., & Rodet, X. (1991). Estimation of fundamental frequency of musical sound signals. ICASSP, (págs. 3657–3660).

Doval, B., & Rodet, X. (1993). Fundamental frequency estimation and tracking using maximum likelihood harmonic matching and HMMs. ICASSP, (págs. 221–224).

Flanagan, J. L. (1965). Speech Analysis, Synthesis and Perception. Springer .

García, B., Vicente, J., Ruiz, I., Alonso, A., & Loyo, E. (2005). Esophageal Voices: Glottal Flow Restoration. ICASSP, (págs. 141-144).

García, B., Vicente, J., Ruiz, I., Angulo, J. M., & Aramendi, E. (2002). Esoimprove: Esophageal Voices Characterization and Transformation', :. BIOSIGNAL 2002, (págs. 142-144).

Gibiat, V. (1988). Phase space representations of acoustical musical signals, Journal of Sound and Vibration. Journal of Sound and Vibration , 123 (3), 537–572.

Hagmüller, M., & Kubina, G. (2006). Poincaré pitch marks, 48 (12). Speech Communication , 48 (12), 1650-1665.

Kadambe, S., & Bourdreaux-Bartels, G. F. (1991). A Comparison of a Wavelet Functions for Pitch Detection of Speech Signals. ICASSP , (págs. 449-452).

Kadambe, S., & Bourdreaux-Bartels, G. F. (1992). Application Of The Wavelet Transform For Pitch Detection Of Speech Signals. IEEE Transaction On Information Theory (38), 917-924.

Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. Proceedings of the IEEE , 74 (11), 1477-1493.

Lió, P. (2003). Wavelets in bioinformatics and computational biology: state of art and perspectives. Bioinformatics Review , 19 (1), 2-9.

Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. IEEE Transactions on Pattern Analysis and Machine Intelligence , 11 (7), 674- 693.

Mallat, S. (1999). A Wavelet Tour of Signal Processing. A. Press.

Nadeu, C., Pascual, J., & Herdondo, J. (1991). Pitch Determination Using The Cepstrum Of The One-Sided Autocorrelation Sequence. ICASSP.

Ortolan, R. L., Mori, R. N., Pereira, R. R., Cabral, C. M., Pereira, J. C., & Cliquet, A. (2003). Evaluation of Adaptive/Nonadaptive Filtering and Wavelet Transform Techniques for Noise Reduction in EMG Mobile Acquisition Equipment. IEEE transactions on neural systems and rehabilitation engineering, 11 (1), 60-69.

Piszczalski, M., & Galler, B. A. (1979). Predicting musical pitch from component frequency ratios. Journal of the Acoustical Society of America , 66 (3), 710–720.

Sano, H., & Jenkins, B. K. (1989). A neural network model for pitch perception. Computer Music Journal , 13 (3), 41–48.

Sheng, Y. (1996). Wavelet transform. En The transforms and applications handbook Series (págs. 747-827). Boca Raton, Fl (USA): CRC Press.

Tohidypour, H. R., Seyyedsalehi, S. A., & Behbood, H. (2010). Comparison between Wavelet Packet Transform, Bark Wavelet & MFCC for Robust Speech Recongnition tasks. International Conference on Industrial Mechatronics and Automation (ICIMA). Wuhan.

Wing-kei, Y., Kwong-sak, & Kin-hong, W. (1995). Pitch Detection Of Speech Signal In Noisy Environment By Wavelet. SPIE , 2491, 604-614.

Yumoto, E., & Gould, W. J. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. Journal of the Acoustical Society of America , 71 (6), 1544–1550.

Wavelet Transforms and Their Recent Applications in Biology and Geoscience

Edited by Dr. Dumitru Baleanu

ISBN 978-953-51-0212-0
Hard cover, 298 pages
Publisher InTech
Published online 02, March, 2012
Published in print edition March, 2012

This book reports on recent applications in biology and geoscience. Among them we mention the application of wavelet transforms in the treatment of EEG signals, the dimensionality reduction of the gait recognition framework, the biometric identification and verification. The book also contains applications of the wavelet transforms in the analysis of data collected from sport and breast cancer. The denoting procedure is analyzed within wavelet transform and applied on data coming from real world applications. The book ends with two important applications of the wavelet transforms in geoscience.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821