ENHANCEMENT OF SHIMMER AND HNR IN OESOPHAGEAL SPEECH

Ibon Oleagordia-Ruiz, Begonya García-Zapirain
DeustoTech-Life Unit. DeustoTech Institute of Technology. University of Deusto
Av. Universidades 24, 48007, Bilbao (Spain)
{ibruiz, mbgarciazapi}@deusto.es

ABSTRACT
This research work presents an oesophageal speech improvement algorithm using Wavelet and Kalman Filtering approach. In both techniques, it has been used different mother wavelet for wavelet approach and different measurement noises for Kalman filtering. People who have suffered from larynx cancer have enormously low intelligibility due to the surgery. A new algorithm has been developed to improve the speech quality. The algorithm consists of enhancing the Shimmer and HNR parameters using Discrete Wavelet Transform (DWT) and Kalman Filtering. By taking advantage of the Wavelet Transform's special time-frequency properties, a corrective algorithm in the form of a wave is applied for the signal intervals in which the shimmer measurement goes beyond normal levels. Therefore, an atomized control of the signal peaks is carried out, having an effect on the normalization of the shimmer. Regarding Kalman filtering, it is proved that the noise obtained from an oesophageal voice during periods of silence is the most suitable. The speech enhancement has been measured using Multidimensional Voice Program tool (MDVP) [1].

KEY WORDS
Speech processing, Oesophageal voice, Discrete wavelet transform, Kalman filtering, Shimmer and HNR.

1. Introduction

The laryngectomy is an operation that is performed in people with laryngeal cancer. This paper is presented with the purpose of helping people who suffered a laryngectomy, that is, laryngectomees. This research work intends to provide a solution to the problem suffered by these people with respect to communication. The removal of the larynx, or laryngectomy, has proven over the years to be a very effective treatment for larynx tumours. This operation is performed with serious illness such as laryngeal cancer. After surgery, patients must learn to speak again. They use the oesophagus to blow air into the vocal tract. Due to the removal of larynx, the intelligibility of oesophageal speech is very low. The main aim is the enhancement of oesophageal speech quality in order to make it more understandable.

The latest incidence statistics for laryngeal cancer in Spain are 2013. In this year, there were 3401 new cases of this kind of cancer [2]. It is estimated that there are 136,000 new cases diagnosed in the world.

The algorithm improves the intelligibility of laryngectomees in communications. The main parameters transformed in the voice improvement algorithm were shimmer [3], and Harmonic to Noise Ratio (HNR) [4]. It is well known that these parameters are related to intelligibility [5]. In this process, wavelet transform and Kalman filter techniques were used. The wavelet transform is used to reduce the breathing noise, very typical in oesophageal voice. At the same time, the shimmer of the speech is improved. The Kalman filter is used to reduce the noise of speech.

This work focuses in the improvement of Spanish /a/ phoneme like in other researches [5, 6]. The speech enhancement has been measured using Multidimensional Voice Program tool (MDVP) [1].

2. Methods

2.1 Acoustic Parameters Used

The improvement of voice will be measured by the analysis of shimmer and HNR. In order to obtain these two parameters it is necessary to measure the pitch.

$$Mean\ F_0 = \frac{\sum_{i=1}^{N} F_i}{N} \qquad (1)$$

where N is the number of pitch instants and F_i are pitch periods [7, 8, 9].

Shimmer is a parameter which represents the amplitude perturbation of speech at the instants of pitch [3]. The voice that is produced in the vocal cords is nearly constant in amplitude, thus increasing the value of shimmer may involve a symptom of a voice disorder. The equation in (2) is used in this paper to measure the shimmer:

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1}\left|20\log\left(\frac{A_{i+1}}{A_i}\right)\right| \qquad (2)$$

being A_i the amplitude in the instant *i*. There are many measurements of shimmer but in this work the parameter has been calculated in dB [12].

DOI: 10.2316/J.2014.216.818-0012

The HNR is defined as the ratio between periodic (r_p) and aperiodic energy (r_{ap}) components (3) [4]:

$$HNR\,(dB) = \frac{r_p}{r_{ap}} \qquad (3)$$

As said in the previous section, these two measurements are carried out with the help of the MDVP software. This software gives automotive good estimations of laryngeal voice signal; however, it does not work properly with oesophageal speech. In order to overcome this problem the pitch period marks have been introduced manually. By so doing, the pitch is calculated and then Shimmer and HNR are obtained.

The MDVP is set to measure HNR as the harmonic spectral energy in 70 - 4500 Hz frequency band and enharmonic spectral energy in 1500 - 4500 Hz band [5, 6].

2.2 Wavelet Transform Theory

The wavelet transform is a special type of Fourier transform which represents a signal in terms of translated and scaled versions of a finite wave. This wave is called mother wavelet [10, 11 and 12]. Wavelet transform can consider as a forms of time-frequency representation. This is called the multi-resolution analysis [13, 14]. There are many basis functions or mother wavelets families. Among of them the most important are: Haar, Symlet, Coiflet, Daubechies, Meyer, Biorthogonals etc. [12, 14].

In this research work Discrete Wavelet Transform (DWT) is used, as in other many applications [16]. DWT uses two discrete-time filter banks: finite impulse response (FIR) and infinite impulse response (IIR). In summary, the DWT of a $x[n]$ signal, speech in our case, is calculated by passing two filters: low pass filter and high pass filter:

$$y_{low}[n] = \sum_{k=-\infty}^{\infty} x[n]\, g[2n-k] \qquad (4)$$

$$y_{high}[n] = \sum_{k=-\infty}^{\infty} x[n]\, h[2n-k] \qquad (5)$$

where the $g[n]$ is the low pass filter and $h[n]$ is the high pass filter. One of the filters represents the mother wavelet, the highest level of the bandwidth, and the other one, the scaling function cover the lowest part [13, 14].

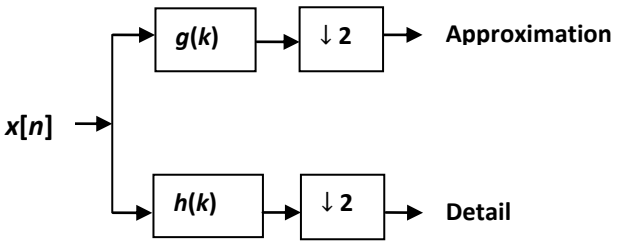

Figure 1: Implementation of (4) and (5) as one stage of an iterated filter bank.

The decomposition has halved the time resolution but it has half frequency band, therefore, the output has double resolution.

This decomposition can be repeated as many times as wanted, resulting new decomposition levels. The approximation coefficients are decomposed with high and low filter in order to obtain a new level. The detail coefficients are maintained.

2.3 Kalman Filter

The Kalman filter is a set of mathematical equations that provide an efficient recursive computational solution of the mean square error method. The filter is very powerful in several aspects: it states compatible with the past, present and future estimations and can do so even when the precise nature of the modelled system is unknown. Therefore, we could say that the Kalman filter is a linear estimator that solves the problem of state estimation of dynamic systems with different noises.

Kalman filtering has been used many times in speech and it was introduced first in [15]. In many contributions white noise is used but coloured noise is introduced in [16]. Some Kalman Filter developments have been proposed for speech improvement focusing in speech modelling [17, 18], others on parameter estimation [19, 20].

The speech is characterised by clean speech and additional zero mean coloured noise. This is performed using Autoregresive (AR) model:

$$x(n) = \sum_{k=1}^{p} a_k\, x(n-k) + \omega(n) \qquad (6)$$

where a_k are Linear Prediction Coefficients (LPC). The state transition matrix, $A(x)$, consists of Linear Prediction Coefficients (LPC). In the equation (6), p represents the system order and is chosen 14 in this work.

3. Design

The general goal of this investigation, and so of all the previous researches [5] is the improvement of the oesophageal voices' quality [6]. Certainly the specific aim of this research is the spectral and temporal correction of the shimmer and HNR parameters of these voices by the Wavelet Transform and Kalman Filtering. For that purpose two different stages have been concatenated: firstly, wavelet stage, and secondly, Kalman filter stage.

The first stage, Wavelet transform, is implemented using DWT. Multiresolution analysis says that DWT gives good frequency resolution and poor time resolution at low frequencies. The general idea of this stage is to apply DWT in order to decompose signal in frequency bands for denoising 0 - 50 Hz frequency range.

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 [21, 22, 23, 24, 25 and 26]. 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. We have used several mother wavelets.

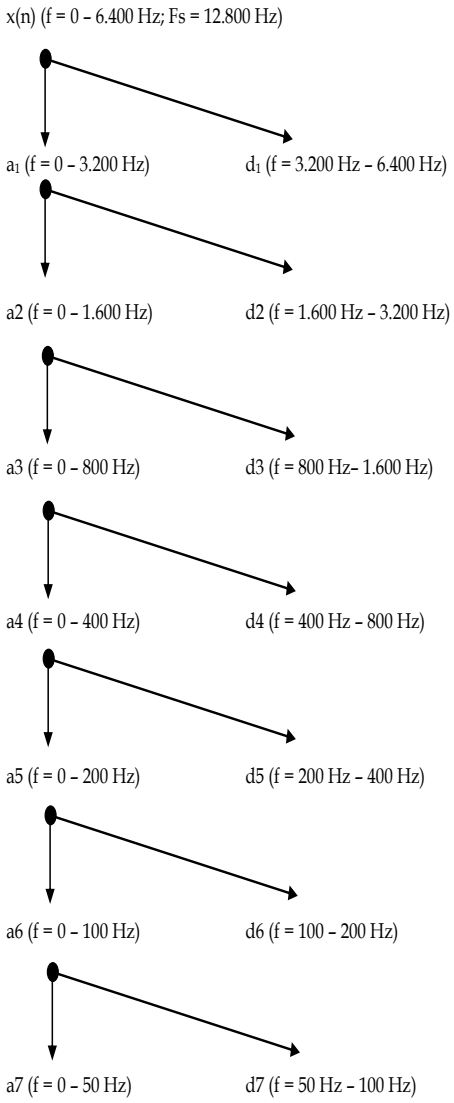

Figure 2: Frequency band diagram

To be precise, we use: Symlet, "bior 6.8", Coiflet, "db6", Haar and Meyer. 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 is modified, as this is where the information on oesophageal voices is to be found. Figure 2 shows the frequency band tree when DWT is applied.

The second important step of the algorithm is Kalman Filtering (KF). The state transition matrix is performed as:

$$A = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \\ a_p & a_{p-1} & a_{p-2} & \cdots & a_1 \end{bmatrix} \quad (7)$$

In the KF, the incoming real signals are the speech and noise. The noise could be randomly created, but in this research work different noises have been used: white, brown, violet, pink and noise obtained from an oesophageal voice during periods of silence. Quantification noise is an unwanted but also unavoidable problem of every digital system. In this case, as both signals are quantified separately, this effect can be minimized without much computational load.

4. Results and Conclusion

On the one hand, in the DWT + Kalman Filtering algorithm, the inputs of the developed algorithm are the samples of the oesophageal voice, which shimmer and HNR parameters have been previously evaluated.

After the application of the DWT algorithm based on the analysis and processing by different mother wavelet, the speech signal has been reconstructed. When measuring the shimmer in this reconstructed signal, the obtained results are shown in table 1. The differences among the processed voices with respect to the original ones are shown in figure 3.

As is shown in the table 1 the shimmer has improved in 29 of 30 cases for Bior 6.8 mother wavelet. Other cases are not significant. This fact can be appreciated in figure 3; the red line corresponds to Biot 6.8. It shows that this wavelet mother is reducing more shimmer. Focusing on statistics results, Wilcoxon test has been developed due to the originals samples are not normally distributed. Regarding to test, $P<0.0001$, which means that reject the null hypothesis (the average of differences of originals and bior 6.8 are equal to zero) and there is significances between two groups of samples. The test shows that using other wavelets results are: DB6 (P=0.116), Meyer (P=0.118), Haar (P=0.139), Coiflet (P=0.181) and Symlet (P=0.484).

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 one of the paper. Therefore, DWT improvement is suitable techniques in the speech enhancement context using a suitable mother wavelet.

Regarding to the Kalman filtering and HNR parameter, the table 2 shows that the most effective measurement noise is the oesophageal voice during periods of silence. It reduces the HNR parameter in 3.354 dB in average. This fact is appreciated in figure 4 in which green line shows the greatest improvement of HNR parameter. Focusing on statistics results, T-test has been developed due to the originals samples are normally distributed. Regarding to test, $P<0.0001$, which means that reject the null hypothesis (the average of processed data with noise during the period of silence are equal to originals) and there is significances between two groups of samples. The test shows that using other noises results are: white ($P=0.142$), pink ($P=0.298$), brown ($P=0.035$) and violet ($P=0.005$). As can be seen brown and violet measurement noises are significant with respect to originals. Nevertheless, the noise during silence gives the best results.

The global improvement in average taking to account two stages for shimmer has been 0.5 dB and 2.605 dB for HNR.

Table 1: Shimmer (dB) before and after algorithm for different mother wavelets

	Originals	**Symlet**	**Bior 6.8**	**Coiflet**	**db6**	**Haar**	**Meyer**
A1	0,594	0,51	0,106	0,53	0,372	0,0549	0,389
A2	0,468	0,363	0,227	0,524	0,369	0,565	0,374
A3	2,734	0,433	0,254	0,563	0,365	0,874	0,932
A4	0,573	0,831	0,372	0,751	0,809	0,919	0,823
A5	1,323	1,059	0,959	1,273	1,058	1,062	1,059
A6	0,796	0,677	0,521	0,642	0,624	0,73	0,819
A7	0,409	0,382	0,175	0,373	0,381	0,39	0,387
A8	0,342	0,391	0,186	0,384	0,391	0,38	0,383
A9	0,673	0,56	0,556	0,567	0,564	0,571	0,586
A10	0,339	0,537	0,171	0,676	0,528	0,669	0,629
A11	1,5	1,256	0,816	1,058	1,369	1,232	1,221
A12	0,412	0,491	0,289	0,484	0,415	0,492	0,474
A13	0,909	0,965	0,519	0,934	0,955	1,011	0,972
A14	0,868	0,564	0,346	0,56	0,55	0,587	0,573
A15	0,416	0,66	0,123	0,622	0,651	0,651	0,634
A16	0,23	0,603	0,264	0,654	0,582	0,694	0,704
A17	0,359	0,632	0,289	0,58	0,546	0,554	0,604
A18	0,71	0,803	0,413	0,772	0,823	0,861	0,746
A19	2,01	2,504	1,97	2,051	2,436	2,534	2,632
A20	0,343	0,538	0,124	0,501	0,925	0,483	0,493
A21	0,863	0,618	0,477	0,663	0,737	0,59	0,665
A22	0,997	1,318	0,272	0,272	0,287	0,335	0,398
A23	1,98	1,377	0,712	1,165	1,37	1,121	1,156
A24	0,494	0,759	0,275	0,519	0,794	0,874	0,533
A25	1,164	1,706	0,541	1,871	1,698	1,571	1,807
A26	2,069	0,812	0,599	1,543	0,813	0,755	0,808
A27	2,002	3,045	2,073	3,737	3,044	3,377	3,036
A28	2,133	2,361	1,151	1,048	1,151	1,177	1,141
A29	1,461	0,394	0,208	1,019	0,407	0,446	0,433
A30	2,772	1,956	1,566	1,989	1,657	1,702	1,636

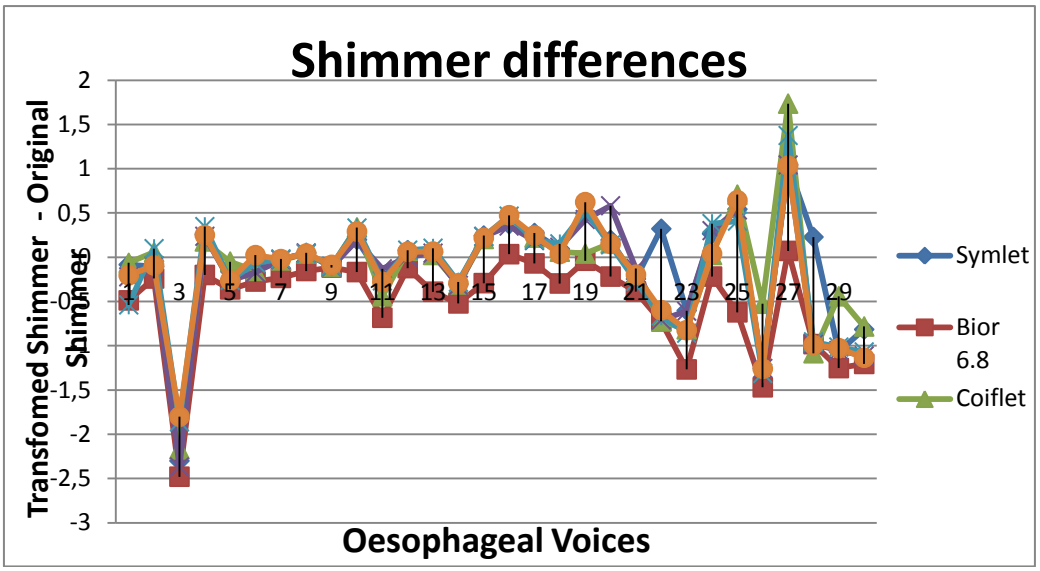

Figure 3: Shimmer differences with respect the originals

Table 2: HNR (dB) before and after algorithm for different measurement noises

	Originals	White	Brown	Oesophageal Noise	Pink	Violet
A1	-2,098	-0,452	0,409	2,037	0,288	-0,404
A2	-6,959	0,265	1,062	1,265	0,536	0,781
A3	-7,070	-6,185	-6,238	-4,954	-5,976	-5,805
A4	-7,979	-7,637	-6,678	-6,35	-6,721	-7,795
A5	-6,641	-5,836	-5,418	-4,393	-5,026	-7,978
A6	-0,802	3,658	3,205	3,938	2,591	2,648
A7	-9,191	-7,28	-6,779	-5,857	-6,067	-6,47
A8	-8,481	-6,783	-6,953	-5,202	-5,824	-7,013
A9	-2,526	-3,259	-2,026	-1,512	-1,856	-2,675
A10	-7,851	-4,589	-4,345	-3,057	-3,598	-5,237
A11	-7,298	-4,025	-5,954	-3,356	-4,561	-5,438
A12	-5,425	-4,176	-3,418	-2,782	-3,579	-3,929
A13	-5,700	-3,899	-3,659	-2,82	-3,762	-4,213
A14	-2,292	0,264	0,269	1,922	0,936	0,276
A15	-6,480	-5,462	-5,916	-4,349	-4,671	-4,627
A16	-5,687	-3,061	-3,733	-2,954	-3,498	-3,751
A17	-8,557	-4,851	-5,601	-3,525	-4,978	-5,131
A18	-7,735	-2,491	-2,045	-2,01	-2,802	-2,248
A19	-7,370	-5,976	-5,136	-4,862	-5,721	-6,657
A20	-6,570	-5,915	-6,754	-4,914	-5,614	-5,482
A21	-5,070	-2,887	-2,324	-1,645	-2,741	-4,053
A22	-5,040	-3,162	-2,871	-2,615	-2,936	-3,22
A23	-3,644	-1,951	-1,844	-1,012	-1,462	-1,486
A24	-5,010	-4,812	-4,802	-4,289	-4,524	-4,863
A25	-6,419	-5,852	-5,763	-5,096	-5,731	-5,591
A26	-9,309	-1,082	-1,466	-0,944	-1,618	-1,375
A27	-9,191	-6,504	-6,827	-5,781	-6,615	-6,106
A28	-6,638	-3,183	-5,045	-2,252	-4,051	-6,248
A29	-3,772	-2,615	-1,509	-1,427	-1,849	-2,674
A30	-7,796	-6,562	-5,672	-5,180	-5,561	-5,418

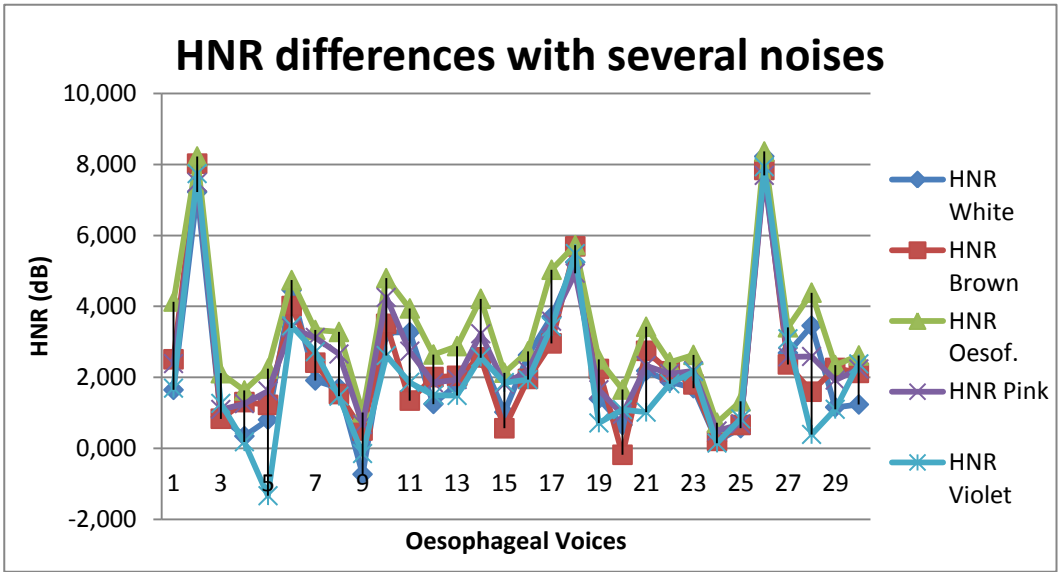

Figure 4: HNR differences with respect the originals

Acknowledgements

The authors wish to acknowledge 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. It must be pointed out the Education, University, and Research Basque Department supports the project.

References

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

[2] SEOM. Sociedad Española de Oncología Médica. http://www.seom.org/es/prensa/el-cancer-en-espanyacom/104018-el-cancer-en-espana-2013

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

[4] Severin, F., Bozkurt, B., Dutoit, T., "HNR extraction in voiced speech, oriented towards voice quality analysis", Proc. EUSIPCO'05, Antalya, Turkey.

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

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

[7] 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.

[8] 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).

[9] 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.

[10] Burrus, C. S., Goinath, R. A., & Guo, H. (1998). Introduction to Wavelets and Wavelet Transforms, A Primer. Upper Saddle River NJ (USA): Prentice Hall.

[11] Sheng, Y. (1996). WAVELET TRANSFORM. The transforms and applications handbook Series (pp. 747-827). Boca Raton, Fl (USA): CRC Press.

[12] Daubechies, I. (1992). Ten Lectures on Wavelets. Philadelphia: 2nd ed. Philadelphia: SIAM.

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

[14] Mallat, S. (1989). A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11 (7), 674- 693.

[15] K. K. Paliwal and A. Basu, "A Speech Enhancement Method Based on Kalman Filtering," in Proc. ICASSP'87, pp. 177–180

[16] J. Gibson, B. Koo, and S. Gray, "Filtering of coloured noise for speech enhancement and coding," IEEE Trans. Signal Process., vol. 39, no. 8, pp. 1732–1742, Aug. 1991.

[17] Z. Goh, K.-C. Tan, and B. Tan, "Kalman-filtering speech enhancement method based on a voiced-unvoiced speech model," IEEE Trans. Speech, Audio Process., vol. 7, no. 5, pp. 510–524, Sep. 1999.
[18] S. Gannot, D. Burshtein, and E. Weinstein, "Iterative and sequential Kalman filter-based speech enhancement algorithms," IEEE Trans. Speech, Audio Process., vol. 6, no. 4, pp. 373–385, Jul. 1998.
[19] P. Sorqvist, P. Handel, and B. Ottersten, "Kalman filtering for low distortion speech enhancement in mobile communication," in Proc. IEEE Int. Conf. Acoustics, Speech, Signal Processing, vol. 2, Munich, Germany, Apr. 1997, pp. 1219–1222.
[20] Biddle A, Watson L, Hooper C, et al. "Criteria for Determining Disability in Speech-Language Disorders". Evidence Report/Technology Assessment No. 52 AHRQ Publication No. 02-E010. Rockville, MD: Agency for Healthcare Research and Quality. January 2002.
[21] Hagmüller, M., & Kubina, G. (2006). Poincaré pitch marks, 48 (12). Speech Communication, 48 (12), 1650-1665.
[22] Kadambe, S., & Bourdreaux-Bartels, G. F. (1991). A Comparison of a Wavelet Functions for Pitch Detection of Speech Signals. ICASSP , (págs. 449-452).
[23] 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.
[24] Kedem, B. (1986). Spectral analysis and discrimination by zero-crossings. Proceedings of the IEEE, 74 (11), 1477-1493.
[25] Nadeu, C., Pascual, J., & Herdondo, J. (1991). Pitch Determination Using The Cepstrum Of The One-Sided Autocorrelation Sequence. ICASSP.
[26] Sano, H., & Jenkins, B. K. (1989). A neural network model for pitch perception. Computer Music Journal, 13 (3), 41–48.