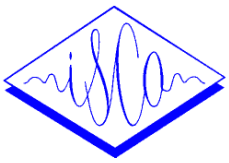

4th European Conference on
Speech Communication and Technology
EUROSPEECH '95
Madrid, Spain, September 18-21, 1995

MODULATED GAUSSIAN WAVELET TRANSFORM BASED SPEECH ANALYSER (MGWTSA) PITCH DETECTION ALGORITHM (PDA)

Léonard Janer
e-mail: leonard@tsc.upc.es
Dept. TSC Universitat Politècnica de Catalunya
08034 - Barcelona
SPAIN

ABSTRACT

In this paper, a new Pitch Detection Algorithm (PDA) based on a Wavelet Transform (WT) Analyser of speech signals is proposed. It provides a value for the fundamental frequency at a pitch period rate. The method is described and evaluated in this paper. The algorithm uses both time and frequency information from the front-end analyser to detect relevant events for the estimation of the fundamental frequency.

1. INTRODUCTION

In the speech signal processing field, pitch period estimation is one of the most intensively studied problems. The accuracy of this estimation is crucial for most analysis and synthesis speech processing systems.

In the last five years, some Pitch Detection Algorithms (PDAs) based on the Wavelet Transform (WT) have been presented([1], [2], [3], [4], [5], [6]). The basic idea, in all these systems is to use the good time and frequency resolution of this Transform to detect the location of the abrupt changes that occur in the glottal closure. The algorithm this paper proposes has the same fundamentals.

In most cases ([1], [2], [4], [5], [6]) the Dyadic Wavelet Transform (DyWT) is used as front-end for the PDAs. Instead of using DyWT a Wavelet Transform Analyser based on Modulated Gaussian Wavelet Functions is implemented, because they show a better resolution (in the frequency domain that is far clear as Figure 1 reveals: In DyWT the scales change always by a factor of 2, instead of that in MGWT this factor is smaller than 2 and done adaptively). Therefore MGWTs can determine with less errors the glottal closure position and the pitch period.

Almost all classical algorithms used to estimate the pitch period prefilter the input signal and take only the information contained in the lowest part of the spectrum, as they suppose the value fluctuating between 50 Hz. and 500 Hz.. For the Wavelet Transform Based systems, only some bands of the Transform are considered. In section 2 our New PDA is described in detail and its differences with other WT based PDAs are enumerated.

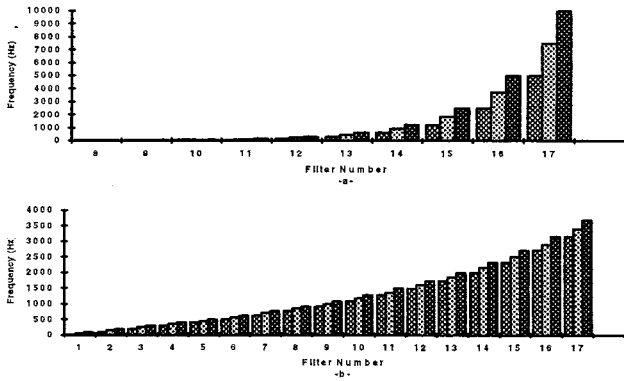

Figure 1. Central frequencies and Bandwidths for the resulting filter bank in both solutions (DyWT -a- and MGWT-b-) In case -a- the central frequencies decrease faster than in case -b- and then we can use less bands in the PDA.

In section 3, some results of the algorithm are presented according to classical performance criteria, and finally in section 4, the main conclusions are outlined.

2. THE MODULATED GAUSSIAN WAVELET TRANSFORM BASED SPEECH ANALYSER (MGWTSA) PITCH DETECTION ALGORITHM (PDA)

2.1. Description of the Modulated Gaussian Wavelet Transform Based Speech Analyser

The DyWT of a signal f(t) using a mother wavelet function $\psi(t)$ is given by (1)([7]):

$$\mathbf{DyWT}f(2^j,\tau) = \frac{1}{2^j}\int_{-\infty}^{\infty} f(t)\psi^*\left(\frac{t-\tau}{2^j}\right)dt \qquad (1)$$

In this equation, the scale parameter is discretised

across the dyadic sequence (2^i) and the translation parameter (τ) is maintained to its maximum resolution (with direct relationship with the sampling frequency of the input speech signal).

In the case of a Modulated Gaussian Wavelet Transform the mother wavelet function is a Gaussian function which is dilated or contracted up to the desired scale adjusting the ξ_o (2). In our MGWTSA this parameter is adapted to have 17 bands/scales with their central frequencies and bandwidths representing a Bark scale filter bank, at the sampling rate of the analysing speech signal. This class of WT is more appropriate than the DyWT in speech signals analysis to take into account the spectral information of the speech (Gaussian mother functions have been chosen because of their good time-frequency resolution).

2.2. Pitch Detection Algorithm

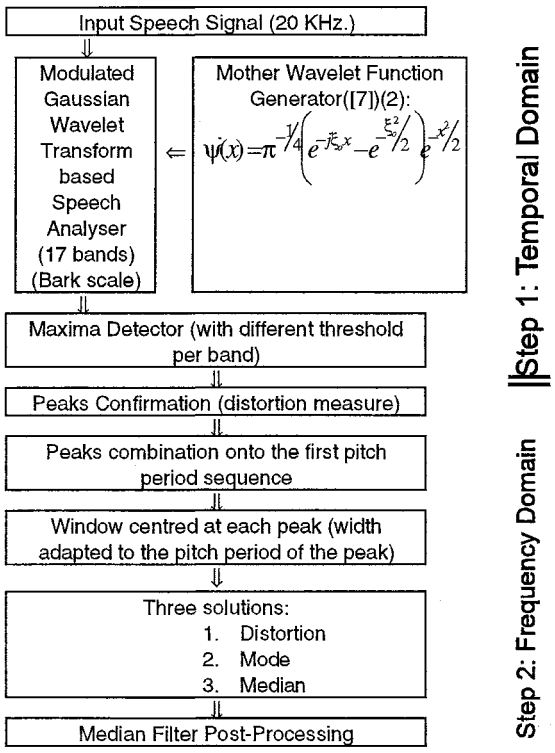

Figure 2: Block Diagram for the complete system. There are two main steps: in the time and in the frequency domains.

The current event detection Pitch Determination Algorithm works both in the time and in the frequency domains to estimate the pitch period. In fact the algorithm works in two steps(Figure 2).

In the first step, the system implements a maxima detector for all the bands we want to be taken into account in the analysis. All 17 bands are included because sometimes, high frequency bands represent better than the lower ones the periodicity of the signal. Anyhow to take into consideration the fact that the fundamental frequency should be between 50 and 500 Hz., a different threshold is used for the bands in this maxima detector., that should be adaptive because depending on its value we consider some voiced values as unvoiced (when the level is minimum) [6].

Once the maxima are detected for all the bands, a pitch period confirmation step is introduced to consider the previous marks as potential fundamental frequency labels. In doing so, the system applies a tracking algorithm to the time interval between two consecutive marks: the interval between all two consecutive peaks is computed and then these values are subtracted two at a time; this difference is what we call the "distortion measure" of the pitch period estimates.: it represents the correctness and the smoothness for all the estimated pitch period values. All the marks (maxima ones) that pass the tracking test are an estimation of the pitch period of the speech signal for every scale of the system (distortion lower than a threshold)

This first part of the algorithm is done in the time domain. We analyse with 17 different scales the temporal signal, and obtain the position of some of their peaks. These peaks, their interval and their position seems to be related with the Glottal Closure Instant (GCI).

In the temporal domain step we work with all the bands due to the fact that higher frequency bands can detect voiced frames when lower frequencies lose them: this is because our particular mother wavelet functions have different bandwidths at every scale. These bands have also the opposite effect: we consider as voiced some unvoiced segments (see Figure 3); this problem will be avoided with an energy distribution decision as the first block in the frequency domain step.

In the second step, the algorithm works in the frequency domain; it combines the information contained on the peaks detected with the 17 bands of the system. The idea is to harmonise the results of the temporal domain block, projecting all the maxima confirmed marks onto a single sequence, which we call the "first estimation of the pitch period values". This sequence plotted in Figure 3, has a high variability due to some irregular marks detected in some of the bands of the analysis filters bank.

To eliminate some of these irregularities the algorithm introduces some necessary post-processing tools with a smoothing purpose. We

propose three different solutions (and therefore their statistical results); for all these solutions we work with temporal windows centred at each peak value detected, and with a width associated with its pitch period estimated value:

Solution 1: For every window we take the value associated with the lowest above mentioned distortion measure.

Solution 2: For every window we take the mode of pitch period values.

Solution 3: For all the values, we calculate their median.

As a last post-processing block for all these three schemes, we filter the output with a median filter of size 3. All the outputs are presented in Figure 3.

3. EVALUATION OF PERFORMANCE WITH CONTINUOUS SPEECH

To evaluate the MGWTSA PDA a continuous speech Database composed by five male and five female speakers, with English sentences at a sampling rate of 20KHz. was used.[1] The average duration per speaker is about 40 seconds.

A semi-automatically segmented pitch contour using a laryngogram signal was used for objective measurements of comparison[8]. Pitch period values for the reference signal and the automatically segmented output were compared at each pitch period ("not frame by frame"). For an objective measure of correctness eight classic parameters were used: i) GPER(%) Gross Pitch Error Rate: errors greater than 1 msec., ii) PDER(%) Pitch Doubling Error Rate, iii) PHER(%) Pitch Halving Error Rate, iv) VuVER(%) Voiced-Unvoiced Error Rate, v) FPER(%) Fine Pitch Error Rate, vi) MFPE Mean of Fine Pitch Errors, vii) StdFPE Standard Deviation of Fine Pitch Errors and viii) PR(%) Performance Rate(1-GPER-PDER-PHER-VuVER)

Table 1 represents results, both for male and female speakers, for the solution based on the distortion measure; table 2 is for Solution 2 and table 3 for the third Solution.

All three solutions work better for female speaker than for male speaker; mainly because we have a fixed Gross Pitch Error decision (1 msec.) instead of an adaptive level (lying on the pitch period value, that would be another Gross Error decision criterium). For female speakers all the solutions work more or less the same, but for male speakers median filter increases GPER due to a false decision in the median computation (with male speakers the maxima detection block detects more near half pitch period values than with female speakers and most of the times these values are chosen on the median filter decision).

	GPER(%)	*PDER(%)*	*PHER(%)*	*VuVER(%)*
Male	5,50	0,00	4,22	6,63
Female	2,90	0,25	0,63	4,27
	FPER(%)	*MFPE*	*StdFPE*	*PR(%)*
Male	73,87	2,27	0,27	**83,65**
Female	72,38	0,93	0,17	**91,95**

Table 1: Results for Solution 1 (distortion measure)

	GPER(%)	*PDER(%)*	*PHER(%)*	*VuVER(%)*
Male	3,76	0,00	4,87	6,63
Female	3,15	0,25	0,38	4,27
	FPER(%)	*MFPE*	*StdFPE*	*PR(%)*
Male	73,65	2,37	0,28	**84,74**
Female	72,51	1,07	0,17	**91,95**

Table 1: Results for Solution 2 (Mode)

	GPER(%)	*PDER(%)*	*PHER(%)*	*VuVER(%)*
Male	10,72	0,00	5,30	6,85
Female	4,67	0,38	0,25	3,89
	FPER(%)	*MFPE*	*StdFPE*	*PR(%)*
Male	72,78	3,94	0,28	**77,13**
Female	80,46	2,81	0,17	**90,81**

Table 3: Results for Solution 3 (Median)

4. CONCLUSIONS

In this paper a new PDA has been proposed and some statistical results show its performance capabilities. The system works both in the time and in the frequency domains, and uses the information contained in 17 bands to extract precise information about pitch period values. This PDA works with MGWT to take profit of their good time-frequency resolution, and to have a great number of bands to analyse in the temporal domain and to combine in the frequency domain step (the more information we have the better the system works).

5. ACKNOWLEDGEMENTS

The author gratefully acknowledges fruitful discussions with Dr. Asunción Moreno Bilbao and Dr. Eduardo Lleida Solano; also he would like to thank Ignasi Esquerra, Dr. Eric Mousset and Juan Luis Navarro for their anonymous collaboration, and Dr. Fabrice Plante and Dr. Georg Meyer from Keele University for the Database, all of them in the SPHERE Human Capital and Mobility Project.

[1] Keele University UK

6. REFERENCES

[1] S.Kadambe, G.F.Boudreaux-Bartels, *Application of the Wavelet Transform for Pitch Detection of Speech Signals*, IEEE Trans. on Information Theory, March 1992, Vol.38, No.2, pp. 917-924.

[2] Nuria González, Domingo Docampo, *Application of singularity detection with wavelets for pitch estimation of speech signals*, procs. EUSIPCO'94, pp. 1657-1660, September 1994, Edinburgh.

[3] Silvio Montrésor, Marc Baudry, *Pitch estimation of speech signal with the wavelet Transform*, Procs. EUROSPEECH'93, pp. 2017-2020, September 1993.

[4] Ewa Lukasik, Stefan Grocholewski, *Two Pass Robust Pitch Extraction Algorithm Using the Dyadic Wavelet Transform*, Procs. EUSIPCO'94, pp. 1681-1684, September 1994, Edinburgh.

[5] Shubha Kadambe, G.F.Boudreaux-Bartels, *A comparison of Wavelet Functions for Pitch Detection of Speech Signals*, Procs. ICASSP'91, pp. 449-452, May 1991, Toronto.

[6] Francisco J. Ancin, Brian L. Burrows, Rolando A. Carrasco, *A novel DyWTVT approach for continuous speech pitch estimation*, Procs. EUSIPCO'94, pp. 1677-1680, September 1994, Edinburgh.

[7] Ingrid Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conference Series in Applied Mathematics.

[8] Juan Luis Navarro, Ignasi Esquerra, *A time-frequency approach to epoch detection*, Procs. EUROSPEECH'95, September 1995, Madrid.

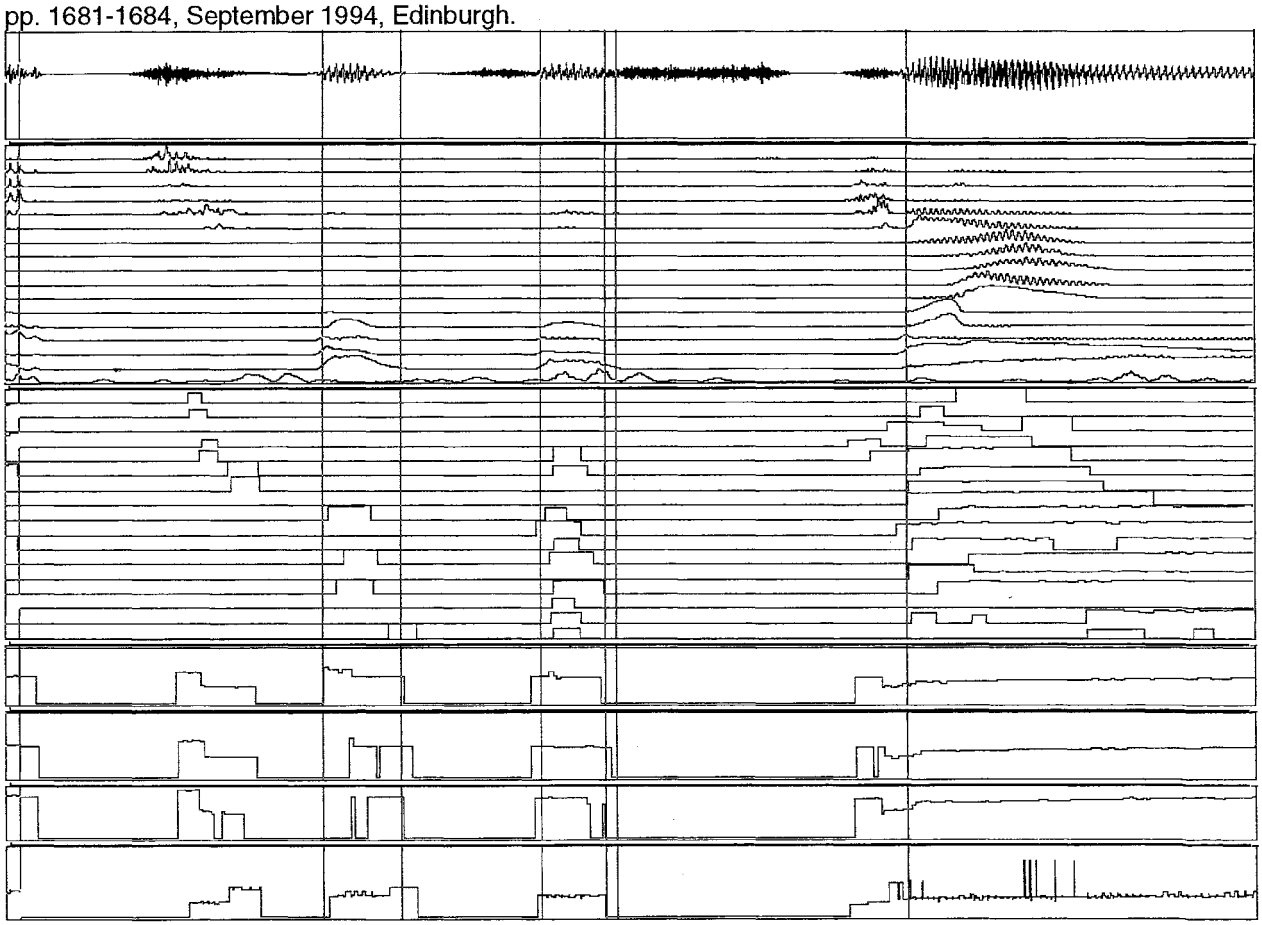

Figure 2. Results of the three solutions of the MGWTSA PDA, for the segment /(wh)ich was the strong(er)/ of the sentence /The northwind and the sun were disputing which was the stronger when a traveller came along, wrapped in a warm cloak/, for a female speaker. At the top of the figure it is represented the temporal signal. Then it is showed the output for all the 17 bands of the Modulated Gaussian Wavelet Transform based Speech Analyser (from low to high frequency). The third plot represents the maxima (as pitch period values) detected after the confirmation step, for all the bands. Then finally it is plotted the output for the three solutions: Statistical Median, Distortion Measure and Statistical Mode). The last plot is the projection of all the bands in the frequency domain analysis of the MGWTSA PDA. The vertical marks represent Voiced/Unvoiced or Unvoiced/Voiced limits segmented at hand.