Pitch Detection in Vocal Music Monophonic Signals

António Sá Pinto, Ana Maria Tomé

***Resumo* – Este trabalho é acerca da determinação de pitch em sinais musicais monofónicos produzidos por uma fonte vocal. O Algoritmo de Detecção de Pitch (PDA) baseia-se na função de autocorrelação, um dos mais explorados métodos de detecção da frequência fundamental. Contudo, é desenvolvida uma nova aproximação ao processo de estimação de pitch. Esta nova estratégia consiste na introdução de uma unidade interactiva de processamento lógico que aumenta a cooperação entre o extractor central e o bloco de pós-processamento (dois dos três blocos que caracterizam a maior parte dos PDAs), de forma a evitar estimativas erradas de pitch.**

***Abstract* - This work is concerned with pitch[*] determination in vocal music monophonic signals. The proposed Pitch Detection Algorithm (PDA) is based on the autocorrelation function, one of the most explored fundamental frequency detection methods. However, a new approach to the estimation process is developed. This new strategy consists in the introduction of a new logic processing interaction unit that enhances the co-operation between the central extractor and postprocessor blocks (two of the three blocks that characterise most PDAs), in order to avoid erroneous pitch estimates.**

I. INTRODUCTION

The problem of estimating the fundamental frequency of speech signals occupies a key position in the signal processing research area. It has many potential applications in different areas such as transmission, synthesis and recognition of speech, and plays the leading role in systems for helping to correct speech impediments of the handicapped [1]. Pitch determination of speech signals is not a simple task. The arduousness of this operation arises from the non-stationary nature of the speech waveform.

This paper will however focus on the application of pitch estimation to vocal music processing, an area that has been relatively little explored in comparison to the massive investigation carried out by the speech community. However, an automatic pitch detector, capable of extracting the fundamental frequency from singing voices would have many interesting applications, such as: systems for computer-assisted singing teaching or ear-training, automatic score transcription, analysis of microtonal non-Western music, real-time control of MIDI devices, etc. [2].

The main differences of vocal music signals, in comparison to speech signals, are related to the wider range of fundamental frequency (from ≈82.4Hz to ≈987.7Hz)[6], and the enormous variations in timbre (and therefore in spectral content) that a singer can produce in a single piece of music. These aspects should be carefully regarded in order to develop a reliable method of extracting the fundamental frequency of vocal music signals.

In this work we present a Pitch Detection Algorithm based on the autocorrelation function[3]. Some modifications are introduced in the PDA structure in order to increase its performance, reliability and accuracy. The software implementation of this PDA was developed in MATLAB.

We begin with a brief description of the structure of general PDAs, and the problems associated with this kind of approach. Then, the basic characteristics of the pitch estimation method of this PDA, the autocorrelation function, are described. Afterwards, we approach the implementation of the proposed PDA and finally, we present and analyse the results obtained with the analysis of synthesized, samples and real signals.

II. PITCH DETECTION ALGORITHMS

Most of the PDAs are characterised by the following blocks: the pre-processor, the central extractor and the postprocessor [1]. The central extractor performs the main task: it converts the input signal into a series of pitch estimates. The task of the pre-processor is data reduction and enhancement in order to facilitate the operation of the central extractor. The postprocessor operates in a more application-oriented way. Some of its typical tasks are error correction, smoothing the pitch contour and refining the pitch estimation.

The main problems with this structure occur when in presence of more complex situations, such as voicing transitions. In our point of view, this is due to a lack of an effective interaction between the central extractor and the postprocessor, since in this model, the pitch estimation and its correction are made separately.

[*] Although there is a psychoacoustical distinction between "pitch" as a perceived quantity and "fundamental frequency" as a physical quantity, in this paper, these terms are used indistinctly in reference to the fundamental frequency of voice and the measurement unity used is Hz.

III. THE AUTOCORRELATION FUNCTION

There are several reasons why autocorrelation methods have generally met with good success. The autocorrelation computation is made directly on the waveform, is a fairly straightforward (albeit time consuming) computation, and, above all, is phase insensitive [3]. The use of a zero phase method is particularly promising for the study of musical signals, since this means that contributions from all of the harmonics occur at the period of the fundamental, and any problem of a non-existent or weak fundamental is thus circumvented [4]. However, there are several problems associated with the use of this method.

Although the autocorrelation function of a voiced section of a vocal piece generally displays a prominent and isolated peak at the pitch period, there are also often present peaks due to the detailed formant structure of the waveform [1]. Another problem is the required use of a window for computing the short-time autocorrelation function. This exigency comprises three difficulties. First there is the problem of choosing an appropriate window. Second, no matter which window is selected, it will taper the autocorrelation function smoothly to 0, an effect known as linear tapering [5]. This effect tends to compound the difficulty mentioned above in which formant peaks in the autocorrelation function (which occur at lower indices than the period peak) tend to be of greater amplitude than those due to the fundamental. A final difficulty is the problem of choosing an appropriate analysis window size. The ideal analysis frame should contain from two (necessary) to three (preferred) complete pitch periods. Thus, for male voices (low pitch), the analysis frame should be long, whereas for female voices (high pitch) it should be kept short.

IV. THE PROPOSED ALGORITHM

The proposed method includes a logic processing interaction unit, an intermediary processing step between the central extractor and the postprocessor. The goal of this interaction unit is to prevent the erroneous estimation of pitch, in opposition to typical PDA analysis, for which the burden of correcting estimation errors produced by the central extractor is exclusively imputed to the postprocessor. This new interaction processing step implements a logic based on a four-state model, capable of dealing with the characteristics of different-nature segments of this type of signals.

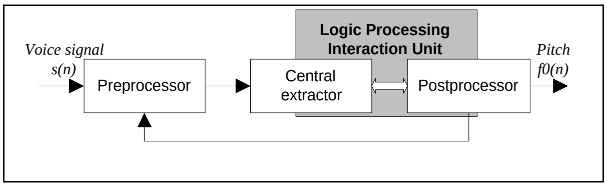

Figure 1. Block diagram of the proposed PDA

A. Pre-processor

The goal of the pre-processor is to eliminate, or at least reduce the problems of the autocorrelation method as a voice pitch estimator.

Our pre-processor is composed of four blocks. The first one implements the adaptive segmentation of the voice signal, which allows the use of appropriate values for the analysis frame size and reduces the computational cost of the autocorrelation computation. The default analysis frame size is 30.8 ms, the maximum size required to cope with pitch values corresponding to the low end of the fundamental frequency voice range (≈65.4Hz). After five consecutive voiced segments, the window size is altered to the triple of the average of the pitch periods of these five segments. The factor of 3 allows up to a 50 percent variation in pitch period from the estimated average pitch period, and still ensures that at least two complete pitch periods are contained within each analysis frame. The second block distinguishes between silent or final transient segments and other type of segments. The silence level threshold is set to 1/15 of the magnitude of the maximum peak in the whole voice signal. A final transient segment is defined as one for which its maximum peak is below ½ of the peak magnitude of the previous segment. The third block function is to whiten or spectrally flatten the signal, with a time domain non-linear distortion method, based on previous works [2]. The analysed frame is centre and peak clipped, resulting in a signal which can assume one of three possible values: -1, 0 or 1. The last block deals with the computation of the autocorrelation function. The autocorrelation function (which is equivalent to the inverse Fourier Transform of the power spectrum [5]) is calculated with the use of FFT techniques, in order to decrease its processing time. Then, it is normalised to unity at origin (lag 0). Finally, the effect of linear tapering is corrected with its inverse transformation.

B. Central Extractor and Postprocessor

As mentioned above, these two blocks collaborate within the logic processing interaction unit, which is based in a four-state model, depicted in figure 2.

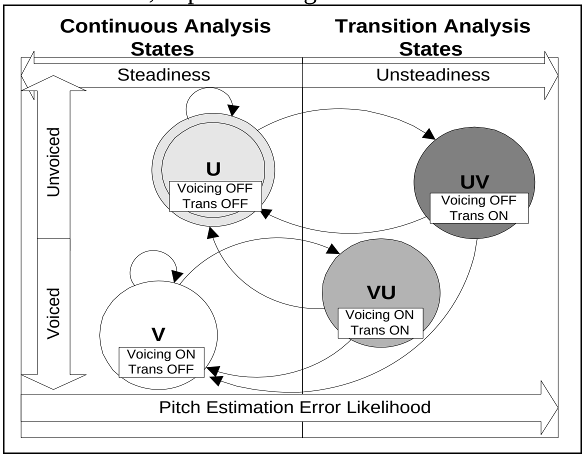

Figure 2. Four–state analysis model

There are four separate logic paths (or states), each of which are selected, based on two control variables (*voicing* and *trans*). These two variables can assume two values: on or off. The first one regards the voiced or unvoiced classification of the last segment. The second one indicates if there were detected evidences of a possible voicing onset or offset transition.

The goal of the transition analysis states (**V**oiced-**U**nvoiced or **U**nvoiced-**V**oiced) is to ascertain the veracity of the voicing transition hypothesis, raised by their corresponding continuous analysis state (**V**oiced or **U**nvoiced). These logic paths implement a more cautious processing, given the ambiguous nature of the segments they analyse. This is achieved, for example, with the use of higher constraints for threshold parameters.

The pitch estimation is made basically with an inspection of the maximum of the autocorrelation function (detected by the central extractor), and comparison with predefined thresholds, empirically obtained from the analysis of several musical phrases from different singers. Whenever the last segment (*n-1*) analysis exhibited a prominent peak, its position is used in order to restrict the range of acceptable autocorrelation peaks for the *n*th. segment, since there are obvious physical limitations to voice fundamental frequency variability in adjacent frames. This pitch tracking strategy reduces both computational effort and pitch error estimation likelihood.

One of the major drawbacks of the autocorrelation function, its proneness to harmonic or subharmonic detection, is due to the fact that the function is itself periodic in the true pitch period [1]. The harmonic detection likeliness is reduced by the linear tapering correction routine in the pre-processor. The subharmonic detection impairs drastically the prospect of success for the overall pitch detection, since both the adaptive segmentation and the logic processing unit are very susceptible to gross pitch errors. This problem is circumvented with the introduction of an algorithm that checks the autocorrelation samples corresponding to the sub-multiple positions of its maximum peak, and chooses the lowest order peak whose magnitude exceeds 80% of the original peak magnitude. Since the subharmonic detection probability attains its maximum at the onset of voicing (for high-pitched singers), the correction algorithm mentioned above is activated for the states U and UV.

Finally, a frequency-domain method of interpolation was implemented in order to refine the pitch estimation for high-pitch signals, since the autocorrelation resolution progressively falls for increasing fundamental frequency values.

V. RESULTS

One main aspect for the development of a PDA is its evaluation through standard databases. Such databases exist only for speech signals[2], which constitutes one of the main difficulties in developing a musical PDA. In order to surpass this problem, we developed a database composed of synthesized[13], samples[14] and real signals. The three types of signals allow us to analyse the performance of our PDA on a growing difficulty logic. The use of synthesized voice signals allows the objective measurement of PDA accuracy, but does not faithfully represent the characteristics and peculiarities of the human voice. Since the samples may have vibrato, we don't have exact information on their fundamental frequency. Nevertheless, this kind of signals allow us to evaluate the PDA in a global way, since we can reproduce rather fast and complex melodies(with a great degree of certainty concerning the pitch), representing all kinds of human voices(from the bass to the soprano). Although the human nature of voice impedes an absolute control over the produced fundamental frequency, thus turning the evaluation process in some kind of a subjective measurement, the real signals allow the analysis of the overall performance and robustness of the PDA, for the situations for which it was created.

A. Synthesized Signal

Synthesized Signal				**Results**		
Begin Time (ms)	**Note**	**Frequency (Hz)**	**Obtained* Frequency (Hz)**	**Unvoiced Frames**	**Voiced Frames**	**Voiced Estimates Average**
112,1	E1	82,410	82,645	1	20	82,645
354,1	F1	87,310	86,957	0	23	86,957
607,1	F#1	92,500	92,593	0	23	92,593
855,5	G1	98,000	98,039	0	22	98,039
1100,3	G#1	103,830	104,167	0	23	104,167
1359,5	A1	110,000	109,890	0	23	109,890
1605,2	A#1	116,540	116,279	0	22	116,279
1854,6	B1	123,470	123,457	0	23	123,457
2105,7	C2	130,810	131,579	0	23	131,579
2356,5	C#2	138,590	138,889	0	22	138,889
2601,3	D2	146,830	147,059	0	23	147,059
2852,9	D#2	155,560	156,250	0	23	156,250
3102,5	E2	164,810	163,934	0	23	163,934
3352,6	F2	174,610	175,439	0	23	175,439
3603,4	F#2	185,000	185,185	0	23	185,185
3851,8	G2	196,000	196,078	0	23	196,078
4101,7	G#2	207,650	208,333	0	23	208,333
4351,3	A2	220,000	222,222	0	23	222,222
4603,3	A#2	233,080	232,558	0	23	232,558
4852,7	B2	246,940	250,000	0	23	250,000
5100,7	C3	261,630	263,158	0	23	263,158
5351,5	C#3	277,180	277,778	0	24	277,778
5603,5	D3	293,660	294,118	0	24	294,118
5851,7	D#3	311,130	312,500	0	25	312,500

Table 1. Synthesized Signal Analysis

The results obtained with the analysis of the synthesized signal were very satisfactory, as we can see by Table 1. The only error reported was a voiced-unvoiced decision, made by our PDA at the beginning of the first note. All

* The Obtained Frequency of the signal refers to the real frequency of the synthesized signal, due to the discrete nature of the voice synthesizer.

other results allowed us to verify the accuracy of our PDA, since the pitch estimates coincided (with a precision of 0.0005Hz) with the synthesized signal (obtained) frequency.

For the results shown in Table 1, it were not included the pitch estimates corresponding to different note overlapping frames.

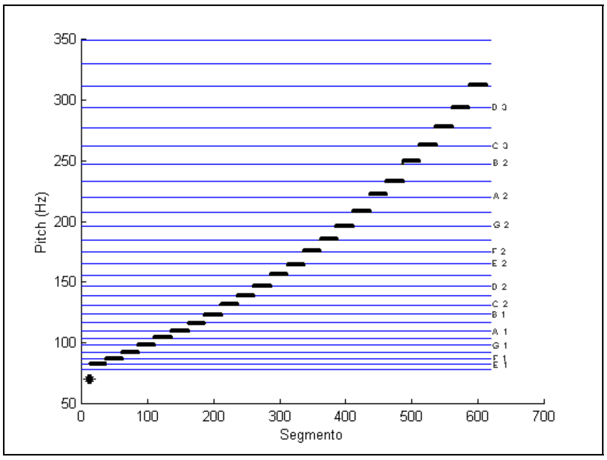

Figure 3. Chromatic Scale (2 octaves)

B. Samples

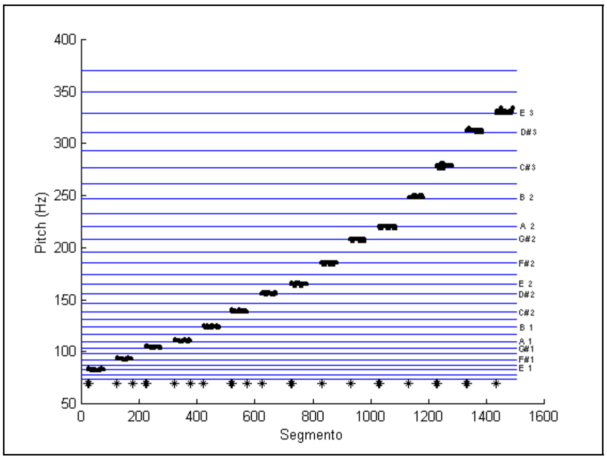

Figure 4. Bass sample, E Maj scale(2 octaves)

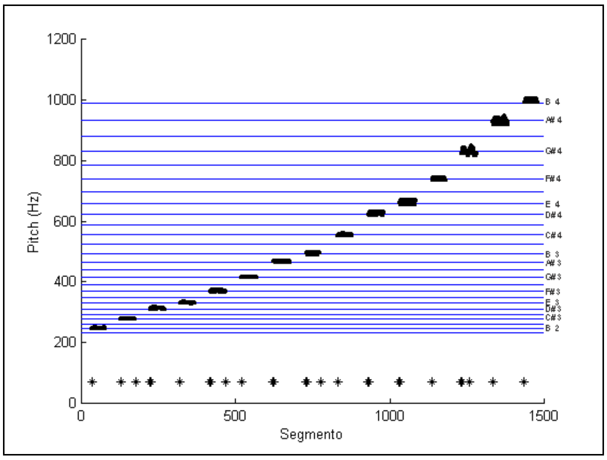

Figure 5. Soprano sample, B Maj scale(2 octaves)

The major contribution of the samples analysis was the ability to evaluate the PDA performance with different kinds of voices. As we can see by figure 4 and figure 5, the PDA proved its ability to analyse the full human voice fundamental frequency range. Once more, it were detected erroneous voiced-unvoiced decisions at the voicing onset.

C. Real Signals

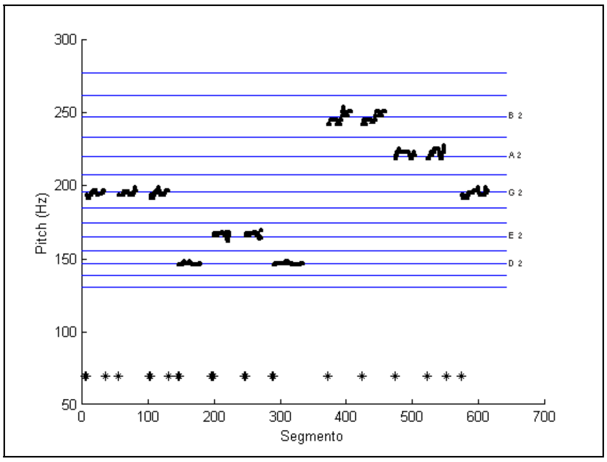

Figure 6. "Old MacDonald had a farm" - Male voice

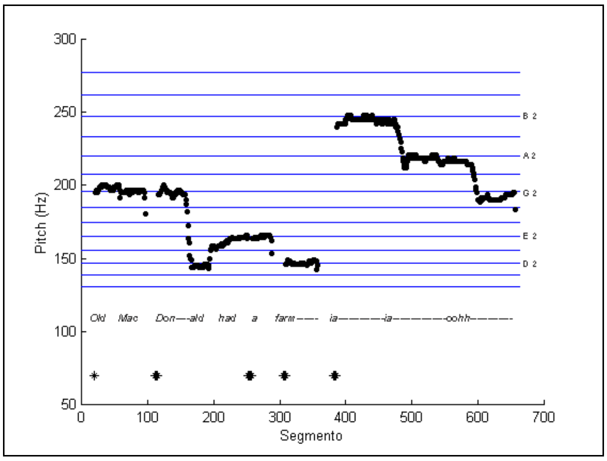

Figure 7. "Old MacDonald had a farm" - Male voice (with lyrics)

The analysis of the real signals demonstrated the robustness of the PDA, namely with the analysis of a vocal excerpt with (figure 6) and without lyrics (figure 7), that introduce uncharacteristic aperiodicities on the signal, thus raising difficulties to the good performance of the PDA.

The fine detail of our PDA was also demonstrated with the detection of some musical occurrences characterised by minor frequency variations: vibrato, portamento and minor untunings.

VI. CONCLUSIONS

We think that the good performance of this PDA is mainly due to its innovative architecture, that reflects the understanding of the human voice production system and the performance of the autocorrelation function as a pitch detector.

The overall performance of our PDA was highly satisfactory, since it could detect the pitch of different kinds of voice signals, covering the full human voice fundamental frequency range. The only inaccuracy detected was its aptness to classify segments of low periodicity (typically the voicing onset) as unvoiced segments. This innacuracy is due to the severe constraints imposed when processing onset voicing segments, which in most of the PDAs originate gross pitch determination errors.

VII. REFERENCES

[1] Hess,W.J. (1983). “ Pitch Determination of Speech Signals-Algorithms and Devices “, Springer-Verlag, Berlin, Germany

[2] F.J.Casajús-Quirós; P.Fernandez-Cid (1994). “Real-time, Loose-Harmonic Matching Fundamental Frequency Estimation for Musical Signals“, ICASSP-IEEE International Conference on Acoustics, Speech, and Signal Processing, 221-224.

[3] Rabiner,L.R. (1977). “On the Use of Autocorrelation Analysis for Pitch Detection“, IEEE Transactions on Acoustics,Speech and Signal Processing 25:24-33

[4] Brown,J.C.; Puckette, M.S. (1991) “Calculation of a Narrowed Autocorrelation Function“ J.Acoust.Soc.Am., vol.85, nº4: 1595-1601

[5] Rabiner,L.R.; Schafer, R.W. (1978) “Digital Processing of Speech Signals“ Prentice-Hall

[6] Doscher, B.M. (1994) “The Functional Unity of the Singing Voice“-2nd ed., The Scarecrow Press, Inc.

[7] W.B.Kuhn. (1990) “A real-time pitch recognition algorithm for music applications“, Comp. Music Journal : 60-71

[8] Carey, Michael J.; Parris, Eluned S.; Tattersall, Graham D. (1997) “Pitch Estimation of singing for re-synthesis and musical transcription“, ESCA, EuroSpeech97 :887-890

[9] Doval, Boris.; Rodet, Xavier (1991) “Estimation of Fundamental Frequency on Musical Sound Signals“, ICASSP, vol.5 :3657-3660

[10] Brown,J.C. (1992) “Musical Fundamental Frequency Tracking using a Pattern Recognition Method“, JASA, vol.92, nº3: 1394-1402

[11] Yavelow, C. (1987) "Personal Computers and Music - The State of the Art ", JAES (Journal of the Audio Engineering Society) 35(3):160-193

[12] Sondhi,M.M. (1968) “New methods of pitch extraction “, IEEE Transactions on Acoustics, Speech and Signal Processing 16:262-266

[13] Rabiner,L.R. (1976) “A comparative performance study of several pitch detection algorithms“, IEEE Transactions on Acoustics,Speech and Signal Processing 24:399-417

[14] Teixeira, A.; Vaz, F.; Príncipe, J. (1997) “ A Software Tool to Study Portuguese Vowels“, ESCA, Eurospeech97, Rhodes, Greece, vol.5, pp.2543-2546

[15] Obtained with Yamaha SY77 AWM Samples