Proceedings of the Institute of Acoustics

PITCH DETECTION : AUDITORY MODEL VERSUS INVERSE FILTERING

F Plante (1), G Meyer (2), W A Ainsworth (1,2)

(1) Communication and Neuroscience Dept. Keele University, Keele
(2) Computer Science Dept., Keele University, Keele

1. INTRODUCTION

Pitch detection is one of the more important problems in speech processing. The required accuracy of the pitch period estimate depends on the task (eg. perception, coding or labelling). Different levels of pitch analysis could be considered: voiced/unvoiced discrimination, pitch frequency estimation, glottal closure instant detection and glottal waveform extraction. The analyses are increasingly complex and require a better knowledge of the glottal waveform. In this paper, analyses up to the extraction of the Glottal Closure Instant (GCI) are discussed. The pitch period gives information about intonation patterns while the GCI allow pitch synchronous analysis of the speech signal, which is important in synthesis or coding.
For all types of analysis a number of algorithms have been proposed over the years but there is no algorithm that is able to deal with all conditions and all voices [1]. These algorithms can be divided into three classes, according to the knowledge required [2]. The first and oldest method is based on pitch extraction from either time or spectral representations, the second class is based on perceptual theories and the third uses auditory models. The first two classes have been extensively studied [1,3] while comparisons with auditory model based pitch extractors are rare [4].
Two algorithms, one an auditory model and the other an inverse filter based approach, were studied and are compared with a reference trace obtained from the derivative of the laryngograph signal. To allow a meaningful test of the two algorithms for both problems, pitch and GCI extraction, the raw data is processed with shared methods, fig. 1.

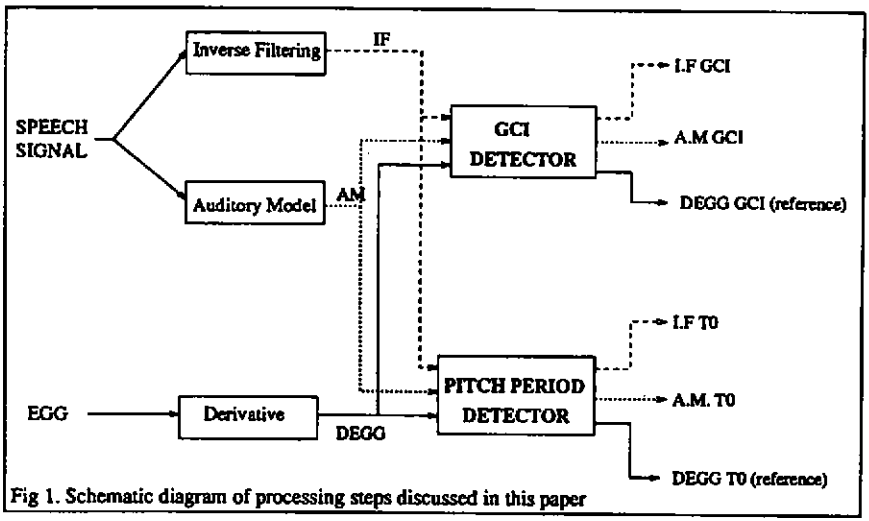

Fig 1. Schematic diagram of processing steps discussed in this paper

2. METHODS

To obtain a quantitative measure of performance both the output of the auditory model and the residual signal from inverse filtering are compared with the laryngograph signal. The generation of these signals is described in the following sections. These signals were processed with same GCI and pitch detector (Fig.1).

2.1 Laryngograph

The laryngograph measures the impedance across the vocal cords [5]. The signal obtained (EGG) is the best estimate of vocal cord activity [6] because it is not modified by the vocal tract. The glottal closure represents a discontinuity in the kinetics of the vocal cord which is extracted by calculating the derivative of EGG [7].

The derivative of EGG (DEGG) is used as a reference in all experiments. Figure 2 shows the DEGG signal with a speech trace. Note that glottal activity can be seen even for near silent speech sections (fig 2 A & B) or vice versa periodicities without glottal activity (fig.2 C & D).

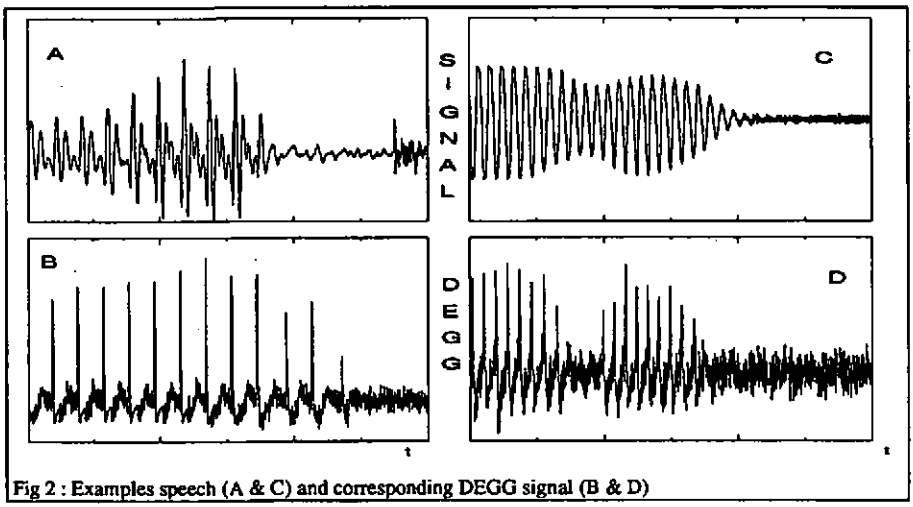

Fig 2 : Examples speech (A & C) and corresponding DEGG signal (B & D)

2.2 The Auditory model

Auditory models are attractive for pitch extraction because they predict pitch perception data well [8] and are robust in noise [4, 9]. The AMPEX pitch extraction algorithm [4] is one of the best pitch extraction algorithms [2] but does not predict pitch perception data because the nerve model driving it generates an envelope representation.

Physiological experiments [11] have shown that a population of neurones in the cochlear nucleus, the first processing stage in the auditory pathway, selectively extract the fundamental frequency of complex stimuli. Models of these neurones predict the pitch of complex tones more robustly than a cochlear nerve model [10].

The model used here contains the key elements of the AMPEX algorithm, but models of cochlear nucleus onset units replace the cochlear nerve envelope representation used as the input to the original algorithm. The onset models selectively enhance pitch periodicities by summing cochlear nerve activity over wide

frequency bands (7 barks) and performing a peak picking operation. In contrast to the original AMPEX algorithm this method is able to predict pitch effects, such as pitch shift of harmonic components [10]. As the auditory model is sensitive to peaks in the signal it is interesting to see how well it performs for GCI detection.

2.3 The Residual signal

Markel and Gray [12] propose that the LPC residual is a good basis for extracting glottal information. In theory linear prediction finds the filter characteristics of the vocal tract, so using inverse filtering the glottal information can be estimated. This is the principle of the SIFT algorithm [13]. To increase the noise robustness of this method we use the instantaneous envelope of the LPC residual. The analysis is shown schematically in fig 3. The signal is first passed through a pre-emphasis module improving the accuracy of the LPC analysis [12]. The analysis is performed on 25.6ms asynchronous windows, overlapping by 12.8ms. The filter corresponding to the vocal tract is calculated from the LPC coefficients and the residual is obtained by inverse filtering. This operation is successful only if the signal is energetic enough. To increase the residual amplitude for voiced frames it is weighted by the energy ratio between the original and the pre-emphasized version. This minimizes the influence of noise on low amplitude residuals [1] and artefacts produced by fricatives.

From a theoretical point of view, the residual signal contains only the glottal information if the LPC filter represents exclusively the vocal tract. In practice this is never achieved because the characteristics of vocal tract are unknown. This means that the residual signal generally contains some noise corresponding to vocal tract characteristics. To remove some of this, the signal is clamped [14] and low pass filtered [12]. This signal is then used as the input to the peak detector. Performance can be improved further by using a quadratic detector [15]. Here the instantaneous envelope of the signal computed with the Hilbert transformation is used.

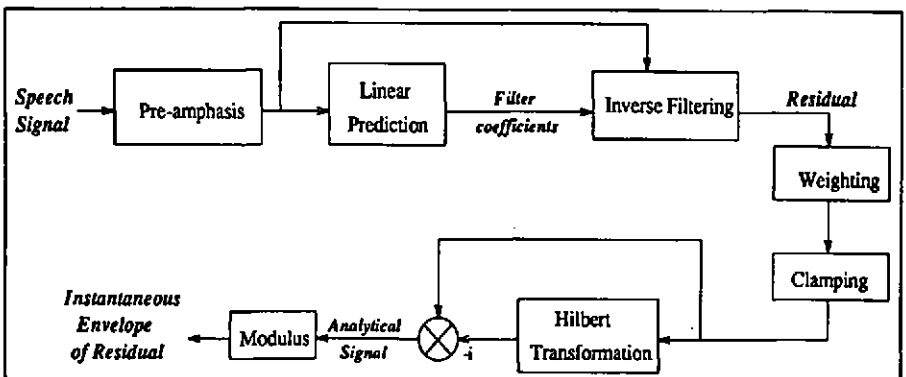

Fig 3: The processing scheme employed to calculate the instantaneous envelope of the residual signal.

2.4. Pitch Extraction

The signals are processed by two different algorithms to obtain

1. a pitch estimate or 'unvoiced flag' for each 10ms frame of the signals,
2. the exact time of glottal closure for each glottal pulse. The analysis is only performed for the voiced sections.

The two experiments are discussed in turn.

2.4.1 Pitch Estimation and Voiced/Unvoiced Decision

The aim of pitch estimation is to obtain a precise measurement of the pitch in each of the frames making up a speech signal. The pitch is estimated by calculating the autocorrelation function ($R_{F,t}$) for the IF and DEGG signals and by using the AMPEX algorithm on the auditory model output. Pitch estimates are obtained every 10ms *(F)* and rectangular windows *(W)* of 30 ms duration are used. The pitch period is taken to be the maximum position in the autocorrelation time lag function. $S[x]$ represents sample x of speech signal, and $\overline{S}$ the mean over the window.

$$R_{F,t} = \sum_{n=0}^{W} (S[F+n] - \overline{S})(S[F+n+t] - \overline{S}).$$

The most problematic step in pitch extraction is the voiced/unvoiced decision. The decision algorithm is based on the value of the maximum autocorrelation lag. If it lies below a set threshold the frame is considered unvoiced, whereas if the maximum value exceeds the threshold the frame is considered voiced.

The AMPEX algorithm is slightly different. It calculates a quasi-autocorrelation using a 'minimum operator' rather than the product. Evidence is also collected over 2 frames preceding and following the frame to be evaluated. The AMPEX algorithm was evaluated for the ENV and DL signals but was found to perform worse than a simple autocorrelation.

The decision thresholds have to change with background noise level. It is assumed that the noise level is known. The threshold is optimised to give the best possible average performance for all speakers but varies with noise level.

2.4.2 Glottal Closure Instant Extraction

In order to compare the different signal representations the same peak detector is used, only the thresholds are adjusted according to the nature of the signal. The GCI detection is performed in two steps, figure 4:

Firstly, a local analysis is achieved. If an harmonic structure is found (using autocorrelation) the potential GCI peaks are detected. Only the largest peaks within each neigbourhood are considered as potential candidates. This step also allows the extraction of voiced segments from the signal.

Secondly, the potential peaks are processed for each voiced period. The best peak to start the process is found. For this, we choose the highest peak with peaks with close amplitudes in the possible pitch range area only. This allows the removal of peaks produced by phoneme transitions in the residual. After the area for the next peak is determined with continuity criteria. In this area the GCI is determined taking the centre of gravity of higher peaks.

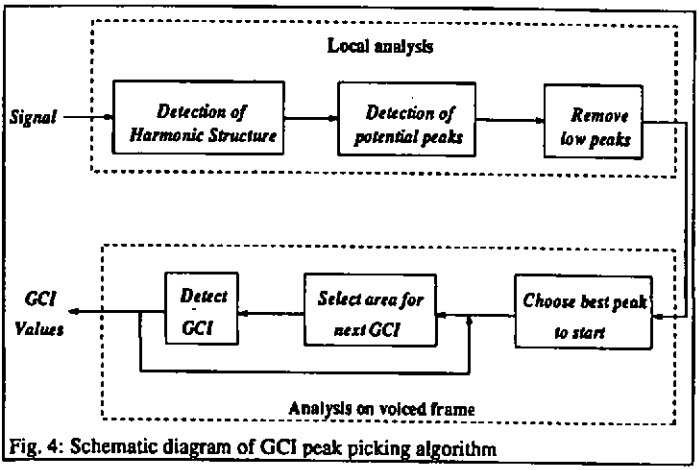

Fig. 4: Schematic diagram of GCI peak picking algorithm

3. RESULTS

3.1 The speech corpus

All experiments are carried out on a total of 110 secs of speech recorded from two female and two male speakers reading "The North Wind and the Sun", a phonetically balanced text. Speech and laryngograph were recorded simultaneously onto a DAT tape and sampled at 20kHz. The robustness in noise of the algorithms is tested using additive white (gaussian) noise. Signal to noise ratios are given a S_{rms}/N_{rms}. Performance measures are calculated against reference data obtained from the ELG recordings.

3.2 Pitch estimation

The speech data were processed by both algorithms using 30ms rectangular windows. Pitch estimates are returned every 10ms. Errors are calculated as in [4]: A pitch estimate is correct if the estimate lies within 20% of the reference for voiced speech, or if the frame is correctly identified as unvoiced. Three types of errors can occur, unvoiced frames can be mislabelled as voiced frames, and vice versa or the pitch estimate can differ by more than 20% from the reference.

3.2.1 Auditory Model versus Inverse Filtering.

Fig 5 demonstrates that the pitch estimation performances of the auditory model (AM) deteriorate with increasing of noise level (81.4% for clean speech, and 61.7% at 0dB S/N). Using inverse filtering (IF), no significant difference is noted between clean and 20dB S/N conditions. Performance decreases from 10dB S/N (5.6% less). 0dB S/N does not cause a further significant deterioration. Inverse filtering seems more robust in noise perturbations than the auditory model.

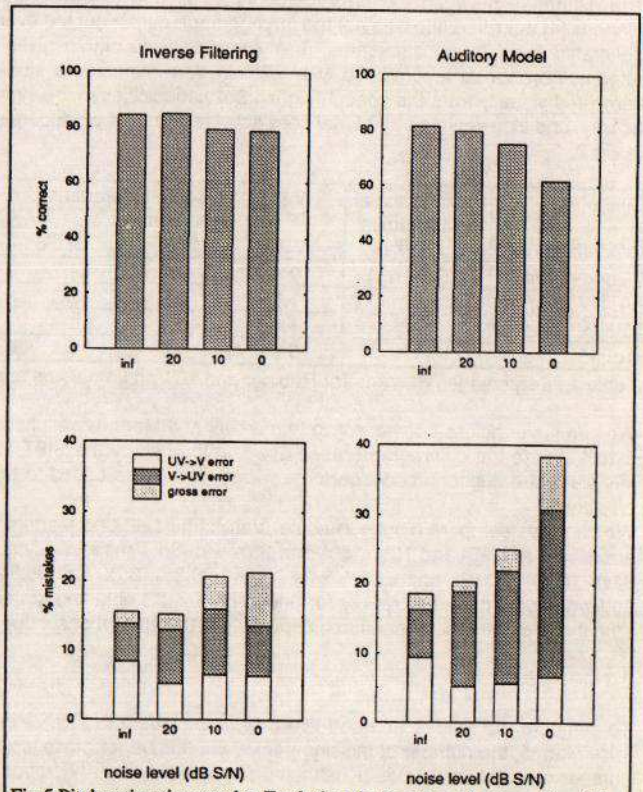

Fig 5 Pitch estimation results. Evolution, in function of noise ratio, of the correct pitch estimation (top) and the errors (bottom) for IF (left) and AM (right)

It is also interesting to compare patterns of error rates between AM and IF. It is apparent that unvoiced/voiced error is similar for the both signals. In contrast, voiced/unvoiced error increase considerably in the case of AM (6.9% for clean speech to 24.4% at 0dB S/N), while this error is constant in the case of IF. Moreover voiced/unvoiced decision performance is constant for IF (except at 10dB S/N

with 2% less) while it deteriorates more for AM (from 83.5% to 69.9%). In fact the decrease of pitch estimation performance for IF is due to the gross errors (from 1.7% to 7.7%). This percentage is stable for AM until 0dB S/N.

3.2.2 Comparison with other studies.
To allow a meaningful comparison of the results with previous data [3, 4] our database was processed using the SHS algorithm [2] which in turn has been compared with a number of other algorithms. The error rates (in %) are given in table 1.
The difference between the error percentages is due to the reference used. In Van Immerseel and Martens [4] the reference was not the laryngograph output but the average of a number of pitch estimates generated by different algorithms. The same SHS algorithm gives 14.5% of error with laryngograph reference on our data compared with 5.3% in Van Immerseel's study. In Rabiner the pitch reference is computed visually from the speech signal. But periodicities in the signal do not always reflect glottal cord activity, and vice versa some glottal cord activitie does not produce periodic signals. Examples are shown in Fig 2.

Authors	Rabiner and al. 3845 frames	Van Immerseel and Martens 5600 frames			This study 11036 frames		
Algorithms	SIFT	AMPEX	SIFT	SHS	AM	IF	SHS
Gross errors %	4.5 (>1ms)	0.9	1.4	1.7	2.3	1.7	1.2
V->UV errors %	3.2	1.6	4.5	2.3	6.9	5.5	1.3
UV->V erros %	3.7	1.4	2.8	1.4	9.4	8.3	11.9
Total %	11.4	3.8	8.6	5.3	18.6	15.7	14.5

Table 1 : Percentages of errors for Rabiner and al. Van Immerseel and Martens and this study.

We can nevertheless compare Van Immerseel and Martens' results with this study assuming an offset of 9-10% due to the different reference used. In their study the SIFT algorithm gives worst results while in this study the auditory model performs worst. The steps added to the computation of the residual seem efficient.
We can also compare results in noise. Van Immerseel and Martens obtained around 5%, 7% and 27% errors with AMPEX and 10%, 15% and 26% with SIFT respectively at 20, 10 and 0dB S/N. In our case we have 15.2%, 20.8% and 21.4% with IF and 20.3%, 25% and 38.3% for AM. The performances of the auditory model decrease rapidly for 0dB S/N. Results obtained at 0dB S/N with IF are significantly better tahn those of the other algorithms, especially if the 10% of offset due to the different references is added.

3.3 GCI RESULTS

To compare the results we used five criteria according to studies on pitch detection [1]. The number of false alarms, the number of missing peaks, the number of gross errors, the number of fine errors and the number of good detections. Krishnamurphy and Childers [7] report that the derivative of laryngograph gives the glottal Closure Instant at an accuracy of two sample points at 10kHz. Here estimates within 5 samples (0.25ms) of the DL signal are considered correct. For the limit between gross error and fine error we take 1ms (20 pts) according to Rabiner et al. [3] The DEGG gives 7360 GCI for the entire database. Results are shown in Fig 6.
Inverse filtering gives significantly better results than the auditory model in all noise conditions.
For IF the performance deteriotes with noise level. In the case of the auditory model, detection is robust for low levels of noise (20dB of S/N), but deteriorates more with increasing noise level.

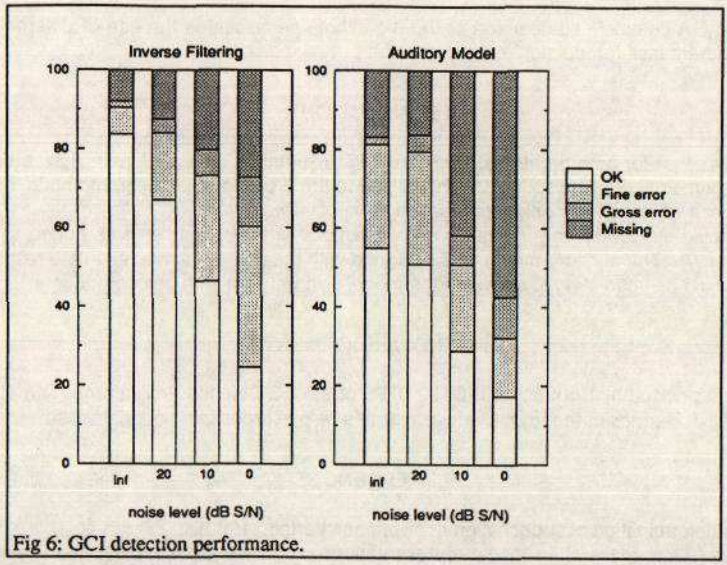

Fig 6: GCI detection performance.

Different patterns in the change of error with noise level are also visible. In the case of the auditory model performance deteriorates due to missing the GCI while for inverse filtering the number of gross and fine errors increases.

The fact that the auditory model performs worse in this type of task is not surprising considering that the processing is a 'peak picking' or coincidence detection operation across a number of cochlear nerve channels, each of these channels receives input from a band-pass filter with different group delays which is only partially compensated (in line with physiological data).

4 DISCUSSION

In both cases, pitch or GCI estimation, the inverse filtering gives better results than the auditory model for all noise levels tested. The performance of the auditory model deteriorates considerably for high levels of noise. Our data does not differ from Van Immerseel and Martens study. The AM performs a peak picking operation. At low level of noise, peaks produced by the closure of glottal chords are still visible. When the noise level increases, peaks are increasingly obscured. This is supported by the fact that the percentage missing or voiced/unvoiced errors increases most in noise.

In the case of inverse filtering, it is not the peak due to the GCI that is detected but modifications of the characteristics of the signal. So it is more robust in noise. Gross errors for pitch estimation and fine and gross errors for GCI increase most strongly. Inverse filtering always detects modifications but has increasing difficulty in returning accurate timing information.

The choice of reference is an important problem in pitch analysis. We saw that results differ significantly according to the reference used. This problem poses the fundamental question of why the pitch analysis is carried out. If laryngograph data is used as reference, models of speech production are evaluated. The main aim is to detect glottal cord activity and not its effect on the signal. It consequently seems natural that a model based on production such as inverse filtering is better than methods based on the perception

or the physiology. A complete comparison of the algorithms necessitates the use of different references according to different task (production or perception).

5 CONCLUSION

This study shows that for both problems, pitch and GCI estimation, inverse filtering gives better results than the auditory model, at all noise levels. This is due to the processing of these methods, and in part to our pitch reference which reflects only a production point of view.
Future work should include testing the relative contributions of the enhancements added to the inverse filtering algorithm. As The auditory model was designed with the primary aim to simulate responses in the brain stem, it would be interesting to correlate the model performance with perceptual data.

6 ACKNOWLEDGEMENTS

This work was supported by Contract SCI-CT92-0786 of the EC Science Programme. The authors thank especially Dr. D. J. Hermes of the IPO, Eindoven for this help in processing our database.

7 REFERENCES

[1] Hess "Pitch determination of speech signals" Springer-Verlag, (1983).
[2] D.J. Hermes "Pitch analysis" In Visual representations of speech Cooke, Beet and Crawford Eds, 1-25, (1993)
[3] L. M. Rabiner, M.J. Cheng, A.E. Rosenberg, C. A. McGonegal "A comparative performance study of several pitch detection algoritms." IEEE ASSP Vol.24 399-418, (1976).
[4] L Van Immerseel and J-P Martens "Pitch and voiced/unvoiced determination with an auditory model" (J Am Soc Acoust 91(6) 3511-3526, (1992)
[5] A.J. Fourcin, E. Abberton "First application of a new laryngograph" Medical and Biological Illustration 21, 172-182, (1971).
[6] G. Childers A.M. Smith G.P. Moore "Relationships between electroglottography speech and vocal cord contact" Folia Phoniatrica Vol.36, 105-118 (1984).
[7] K. Krishnamurphy, D. G. Childers "Two-channel speech analysis" IEEE ASSP Vol.34, 730-743, (1986).
[8] R.Meddis and M.J. Hewitt "Virtual pitch and phase sensitivity of a computer model of the auditory . periphery. 1: Pitch identification" J Acoust Soc Am 89, 2866-2882 (1991).
[9] Ainsworth and G.F. Meyer."Recognition of plosive syllables in noise: Comparion of an auditory model with human performance", J Acoust Soc Am 96/2, 687-695 (1994).
[10] G.F. Meyer and I. Dewar."Comparing pitch extraction in the cochlear nerve and cochlear nucleus", this volume (1994).
[11] A.R. Palmer and I.M. Winter" Coding the fundamental frequency of voiced sounds and harmonic complexes in the cochlear nerve and cochlear nucleus" In The Mammalian Cochlear Nuclei Ed: Merchan et al. Plenum Press , 373-384 (1993) .
[12] D. Markel, A. H. Gray "Linear prediction of speech" Springer-Verlag, (1976).
[13] J.D. Markel "The SIFT algorithm for fundamental frequency estimation" IEEE Trans. Audio Elec. 20, 367-377, (1972).
[14] M. Cheng, D. O'Shaughnessy "Automatic and Reliable Estimation of Glottal Closure Instant and Period" IEEE ASSP Vol.37, 1805-1814, (1989).
[15] L. Atlas, J. Fang "Advantages of general quadratic detectors for speech representations" In Visual representations of speech Cooke, Beet and Crawford Eds, 161-168, (1993).