Acoustics'08
Paris
June 29-July 4, 2008
www.acoustics08-paris.org

Using sets of combs to control pitch estimation errors

Jean-Sylvain Lienard, Claude Barras and Francois Signol

LIMSI-CNRS, BP133, 91403 Orsay Cedex, France
jean-sylvain.lienard@limsi.fr

We analyze the errors of a Pitch Estimation Algorithm using the Pitch Function (response of the periodicity estimator as a function of the frequency parameter F_c). The estimator's maximum response for a single signal of fundamental frequency F_0 is expected to occur for $F_c=F_0$. Actually the pitch function exhibits many secondary peaks which occasionally cause the errors. When several signals are mixed the main peaks do not reliably represent the F_0s of the component signals. By taking as periodicity estimator the correlation of the spectrum module with a uniform infinite spectral comb of fundamental frequency F_c we show that each peak corresponds to a particular value of the ratio $F_c/F_0=p/q$ (p and q positive integers). It follows that some secondary peaks can be cancelled either by augmenting the comb with intermediary negative teeth, or by setting to zero some of its teeth. These modified combs can be viewed as combinations of uniform combs of different F_cs. The present study aims at precisely defining and combining the modified combs so that the main peaks of the new Pitch Function reliably indicate the F_0s of the components. Examples are given on mixtures of voiced segments extracted from natural speech.

1 Introduction

Despite some recent improvements [1,2,3,4], pitch estimation of speech signals remains an error-prone process. The main sources of errors are i) the voiced-unvoiced decision and ii) the selection of harmonics or sub-harmonics instead of the fundamental frequency, yielding what is named the gross errors (i.e. estimation error > 20%). The problem is still more complex if one considers that several voices may be mixed in the signal [5].

In the present study we focus on the gross errors problem in voiced frames of short duration, in the multipitch perspective. Pitch is estimated in the spectral dimension by use of a spectral comb [1]. We define the Pitch Function PF as the response of a given pitch estimator in a given pitch interval. This notion embodies the notions of Period Histogram and Product Spectrum [6], as well as what is called 'pitch strength' by some authors. In the monopitch case, each peak of this function can be labeled by a couple of positive integers p and q, named respectively the harmonic and sub-harmonic indices. For the pitch estimation to be correct one has to select the peak (1,1). However this peak is not always the highest one, even in the monopitch case, and this is the main cause of the gross errors. Our efforts aim at increasing the prominence of the main peak by reducing the magnitude of the secondary peaks. To do so we define some particular combs, with missing or negative teeth, which produce particular pitch functions called Suppression Functions, each one addressing a given peaks family. They are combined into the final PF, in which the main peak gets reinforced, thus reducing the risk of gross errors. We present two ways of combining the Suppression Functions, the Alternate Comb and the Suppression Comb.

2 Infinite Uniform Comb

The Infinite Uniform Comb IUC(F_c) is made of an infinite number of unitary teeth regularly spaced in frequency. We call comb frequency F_c the frequency of its first tooth. The Pitch Function value for F_c is obtained by the scalar product of the comb IUC(F_c) with the module |S(F)| of the sound spectrum.

Although this comb is theoretically infinite, the scalar product is finite because the spectrum is bounded in frequency, as for any physical sound of limited energy. If {Bmin, Bmax} is the frequency band of the sound, and {$F_c$min, $F_c$max} the frequency range in which the estimator computes the PF, a finite comb gives the same PF as an infinite comb if it has at least nteeth:

$$nteeth >= \ Bmax/F_cmin$$

2.1 Functioning

The functioning of the IUC has been described in [7]. It is briefly illustrated in Fig. 1. When the comb frequency F_c is equal to the fundamental frequency F_0 of the sound (here a schematic sound made of a series of spectral peaks) each spectral peak is matched by a tooth of the comb and the scalar product is maximum, yielding a peak of the PF. When F_c equals a multiple p of F_0 it also produces a peak of the PF, but only 1/p peaks are matched; thus the corresponding peak of the PF gets smaller. When F_c equals a sub-multiple q of F_0 all of the spectral peaks are matched by some teeth. Because the comb extends up to Bmax, the scalar product takes the same maximum value as in the $F_c=F_0$ case. More generally, the PF peaks appear when the following relation is observed:

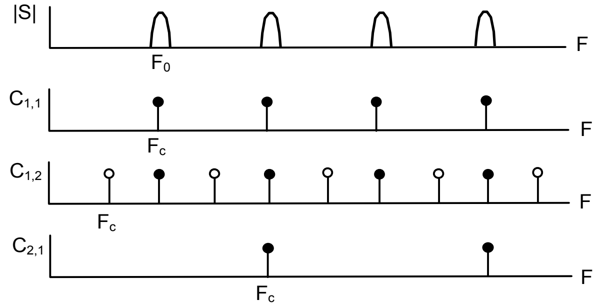

Fig 1. Functioning of the Infinite Uniform Comb on a schematic spectrum F_0. The IUC is represented for 3 values of its frequency F_c: $F_c=F_0$, $F_c=F_0/2$ and $F_c=2F_0$

$$F_c/F_0 = p/q \qquad p,\ q,\ integers > 0$$

2.2. The peaks of the Pitch Function

Let us consider a physical sound made of a series of pulses at F_0=250 Hz, Hanning windowed, of duration 50 ms. Fig. 2 shows the PF obtained by use of an IUC, normalized to the height of the main peak. Some of the numerous peaks obtained are identified in terms of the harmonic and sub-harmonic indices p and q. In the following we shall consider that p and q are the integer terms of the irreducible fraction F_c/F_0. The peak (4,2) does not exist; it is just a redundant notation for the peak (2,1).

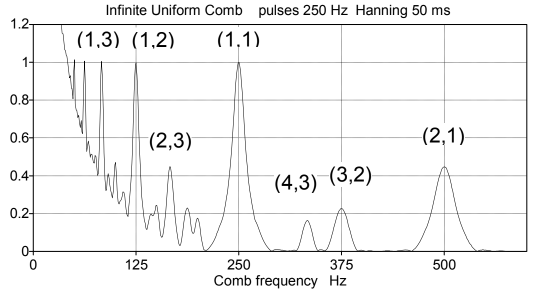

Fig. 2 Pitch Function, Infinite Uniform Comb applied to a 50 ms Hanning-windowed 250 Hz series of pulses

As mentioned above the sub-harmonic peaks (1,q) have the same maximum magnitude $A_{1,1}$ as the main peak (1,1). However, the magnitude of the harmonic peaks (p,1) decreases theoretically according to the inverse of their index p, because only one tooth out of p matches a spectral peak:

$$A_{p,1} = A_{1,1} /p$$

Besides the harmonic (h) and sub-harmonic peaks (sh) one observes secondary (or fractionary) peaks, the frequencies of which are given by:

$$F_{p,q} = (p/q)\ F_{1,1}$$

Their magnitude may vary with the spectral composition of the sound, but theoretically they tend toward the values given by:

$$A_{p,q} = A_{1,q} /p$$

In the case of the IUC the magnitude $A_{1,q}$ of the sub-harmonic peaks (1,q) equals that of the main peak, thus:

$$A_{p,q} = A_{1,1} /p$$

In Fig. 2 for instance, the magnitude of the peak (3,2) at 375 Hz should be close to $A_{1,1}/3$ but the observed value is lower, due to the decreasing spectral envelope of the series of pulses used in our example sound.

The fact that the sub-harmonic peaks have the same value as the main peak makes the use of the IUC problematic for the estimation of F_0, even in the monopitch case. Before examining some possible solutions we have to mention the role of the signal windowing in the PF shape.

2.3. Signal duration and windowing

When the signal duration gets shorter, the spectral peaks get thicker. For a 100 ms Hanning window, for instance, the width of a spectral peak at half height is about 20 Hz and becomes about 40 Hz when the window gets shortened to 50 ms. The peaks of the PF, as they result from the accumulation of several spectral peaks, evolve in the same way. This widening alters the resolution of the pitch estimation, especially in the low F_c range.

Another effect of this widening appears when F_c gets very low, so that several teeth of the comb can take place in the width of a single spectral peak. This produces the decreasing hyperbolic component observed in the low range of the PF (Fig. 2). The resulting masking effect may be tolerated to some extent if one is more interested in the peaks of the PF than in its valleys. In the example illustrated above the PF peaks remain unmasked for F_c>40 Hz, which constitutes the real lower bound of the estimator's frequency range. Several solutions may be implemented to reduce this effect, such as using a better window or modeling the spectral peaks in order to diminish their apparent width.

3 Control of the sub-harmonic decrease

In order for the comb to function as a pitch estimator it is mandatory to reduce the magnitude of the sub-harmonic peaks (1,q). Two techniques, among others, have been used to achieve this task.

3.1. Limiting the number of teeth

Limiting the number of teeth to a fixed small number Nt<nteeth practically ensures the magnitude of the main peak $A_{1,1}$ to be the real maximum of the PF, in the monopitch case. For $F_c=F_0$ each tooth matches a spectral peak, so that $A_{1,1}$ is proportional to Nt. For $F_c=F_0/2$ only the Nt/2 even teeth match spectral peaks. The spectral peaks located above Nt/2 do not contribute any more to $A_{1,2}$, which yields $A_{1,2}<= A_{1,1}$. The 'equal' case may occur if the part of the spectrum matched by the teeth located between Nt/2 and Nt is of zero magnitude. The same line of reasoning holds for the other sub-harmonic peaks. Setting the number of teeth to a value comprised between 5 and 10 yields a progressive decrease of the sub-harmonic peaks, sufficient in practice to promote (1,1) as the main peak of the PF.

This technique implies a loss of information because the spectral peaks lying above the last tooth are not taken into account in the periodicity estimation. This loss is not critically important for speech because i) those high-frequency peaks convey less energy than the low ones and ii) their are somewhat packed down by the jitter.

3.2. Decreasing the teeth magnitudes

Another way to ensure the decrease of the sub-harmonic peaks is to reduce the magnitude of the teeth as a function of their rank ***r***. An exponential decrease in $\boldsymbol{r}^{\boldsymbol{-k}}$ is often chosen. This technique is just a variation of the previous one in that it favors the effect of the lower teeth. However its action is smoother, as the teeth contribution to the PF decreases with their rank. The counterpart is that it may be more sensitive than the previous one to any alteration of the energy carried by the fundamental (telephone speech, or partial masking by a low frequency noise).

The exponent k is often empirically given the value 0.5. This produces a decrease of √2 of the peak (1,2), and a similar increase of the peak (2,1), compared to the values they had by use of the IUC. Thus the two peaks presenting the maximum risk of gross error get the same height.

Both techniques can be used together, as illustrated in fig 3. The comb used has a limited number of teeth (10 teeth), decreasing in √r. The goal of reducing the sub-harmonic peaks is attained, but the rejection of the secondary peaks remains poor and the risk of octave or sub-octave errors is still high, especially in the multipitch case.

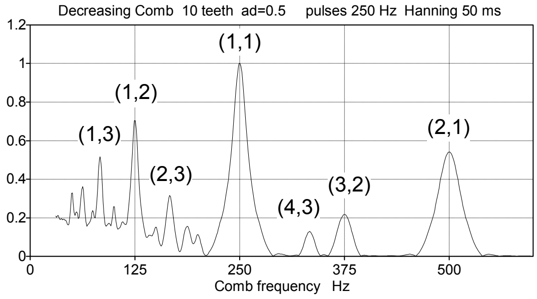

Fig 3. Pitch Function of a 10-teeth comb decreasing in √r. Same sound as in Fig. 2

4 Control of the gross errors by use of irregular combs

In this section we present a novel approach of the control of the gross errors, based on the use of irregular combs. The particular Pitch Functions obtained with those combs are called Suppression Functions.

4.1. Missing Teeth Comb and sub-harmonic suppression

Let us consider an Infinite Uniform Comb, from which the even-numbered teeth of rank 2k have been removed (odd comb, Fig. 4). The PF generated in the presence of a harmonic spectrum of fundamental F_0 exhibits a minimal response (close to zero) for the peaks (p,2k):

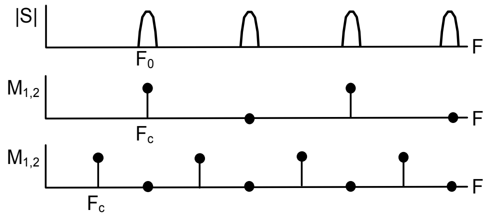

Fig. 4 Missing Teeth Comb of order 2 (odd comb $M_{1,2}$), positioned at F_c=F_0 and F_c=F_0/2. This comb cancels the sub-harmonic peaks (p,2k), p,k integers >0 p>2k

$$F_c/F_0 = p/2k \qquad p,\ k \text{ integers} > 0 \qquad p < 2k$$

Figure 5 shows that the peaks (1,2), (1,4), (3,4), (1,6), (5,6)... have been practically cancelled, to the extent that they were distinct in the initial PF (compare with Fig. 2). Let us denote (sh2) this order 2 sub-harmonic peaks family. The corresponding PF, after processing of the spectrum |S|, is called order 2 sub-harmonic Suppression Function and denoted SH2.

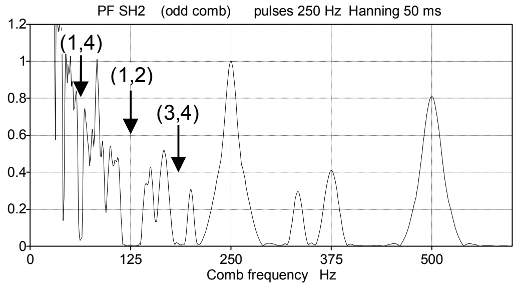

Fig 5. Suppression function SH2 obtained by processing the 250 Hz sound with the odd comb. Peaks (1,2), (1,4), (3,4)... at 125, 62.5, 187.5 Hz have been cancelled

Similarly, a Missing Teeth Comb of order 3 is defined by removing the teeth of rank 3k. This comb cancels the peaks:

$$(p,3k) \qquad p,\ k \text{ integers} > 0 \qquad p < 3k$$

Those peaks (1,3), (1,6), (1,9)... constitute the order 3 sub-harmonic family denoted (sh3). The PF obtained with the spectrum |S| is the Suppression Function SH3.

Continuing with the orders greater than 3, we have to observe that the (sh4) and (sh8) families are totally included in the (sh2) family. Similarly, the (sh6) family is included in the (sh2) and (sh3) families. Thus the corresponding suppression functions are redundant. Only the families of prime order are necessary to cancel all the error-prone sub-harmonic peaks.

If $F_c$max has been set at 600 Hz and $F_c$min at 75 Hz, then it is useless to cancel any peak beyond order 8. As we only need the prime orders, computing SH2, SH3, SH5 and SH7 is sufficient. We note that all those functions let the other peaks, including (1,1), at their proper location, while reducing their magnitude in the proportion of the missing teeth.

4.2. Alternate Teeth Comb and harmonic suppression

In order to control the harmonic peaks (p,1) an approach similar to the previous one consists in placing some negative teeth between the regular teeth of the Infinite Uniform Comb.

Let us consider the order 2 (Fig. 6). The positive teeth are located at frequencies kF_c (k integer >0). When the IUC is positioned on the octave, F_c=$2F_0$, it produces the parasitic peak (2,1) with a magnitude lower than the magnitude $A_{1,1}$ of the main peak. To eliminate it one has to subtract from the

PF a part of $A_{1,1}$. This may produce a negative value in place of the peak (2,1). A thresholding strategy will be necessary to use this new suppression function H2 in a multiplicative combination to get the final PF.

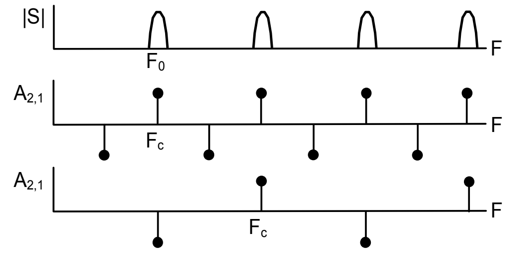

Fig. 6 Alternate Teeth Comb of order 2 $A_{2,1}$, positioned at F_c=F_0 and F_c=2F_0. This comb cancels the harmonic and fractionary peaks (2k,q) with 2k>q

Actually the family (h2) is not limited to the peaks (2k,1), but includes also some fractionary peaks such as (6,5), which is actually the 6th multiple of the 5th sub-multiple of F_0(Fig.7). In symbolic terms the (h2) family may be denoted:

(h2) = (2k,q)   2k >q   k, q integers >0

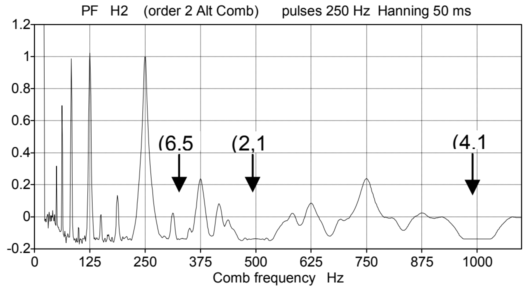

Fig 7. Suppression function H2 obtained by processing a periodic spectrum with the order 2 Alternate Teeth Comb. The function has been normalized to 1

At this point we have to mention that subtracting the middle part of the interpeak parts of the spectrum to improve the pitch estimation has been investigated with success in [3] and [4], by following approaches different from ours.

At order 3 the Alternate Teeth Comb comprises 2 intermediary negative teeth, equally spaced between the regular (positive) teeth. This comb cancels the (h3) peaks family defined by:

(h3) = (3k,q)   k, q integers >0   3k > q

Among the members of this family, one finds peaks (3,1), (6,1)... (3,2), (9,2)... (6,1), (6,5)... (9,1), (9,2), (9,4)...

As before, we remark that some harmonic suppression families are redundant: (h4), (h6) and (h8) are included in (h2), (h6) is included in (h2) and (h3) etc. In other terms the prime order families (h2), (h3), (h5) and (h7) are sufficient in order to cancel all the harmonic and fractionary peaks appearing in a {$F_c$min, $F_c$max} interval less than 11 times $F_c$min.

4.3. Alternate Comb

The Pitch Estimation Algorithm presented in [7] is an implementation of two of the approaches described above. The reduction of the sub-harmonic peaks is done by the limitation of the number of teeth and their exponential decrease, while the reduction of the harmonic peaks is done by using an additive combination of two Alternate Teeth Combs, of orders 2 and 3. Two coefficients a2 and a3 were used to control their part on the harmonic reduction process, and another coefficient ad (the exponent of the teeth decrease) was used to control the decrease of the teeth magnitude. An experimental work yielded the optimal values of a2, a3 and ad. After evaluation on two classical pitch databases and comparison with other PEAs, the performance of the Alternate Comb appeared to be at the level of the best published results.

4.4. Suppression Comb

The Suppression Comb aims at consistently implementing the harmonic and sub-harmonic suppression functions described in the preceding sections. Each one of those functions preserves the main peak and controls some families of parasitic peaks. Thus a multiplicative combination - completed by the use of positive thresholds to avoid the negative parts - looks more appropriate than the additive one implemented in the Alternate Comb.

A first implementation consists in using as many irregular combs as the number of suppression functions needed. If we need to cover a 10 times pitch interval we need to achieve the suppression up to the order 10, and a set of 8 combs is convenient: 4 Missing Teeth Combs to get the functions SH2, SH3, SH5, SH7, and 4 Alternate Teeth Combs to get H2, H3, H5 and H7. This method requests some amount of computing but does not bother - or only marginally - with the problem of the low frequency hyperbolic increase of the PFs.

Another implementation is based on the fact that the suppression functions can be computed from the PFs of the Infinite Uniform Comb, according to the following formulae. PF designates the Pitch Function of the IUC, i designates the order of the reconstructed irregular comb. The coefficients of the H functions ensure that the Alternate Combs are centered (null mean value):

$$SHi(f_c) = PF(f_c) - PF(i\,f_c)$$

$$Hi(f_c) = (i/(i-1))PF(f_c) - (1/(i-1))\,PF(f_c/i)$$

This method is faster than the first one, but it implies the computation of the PF at very low frequencies ($F_c$min/7 for the function H7), and thus encounters the problem mentioned above.

In the following examples the final PF, obtained by multiplying the 8 suppression functions, was weighted in frequency by $\sqrt{f_c}$, in order to get a supplementary gain on the SH peaks such as the one evoked in section 3.2.

Figure 8 represents the final PF of the same sound used thoughout the paper. It is to be compared with figures 2 and

5, to get an idea of the efficiency of the whole suppression process.

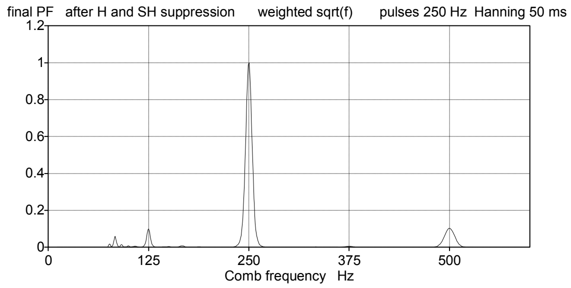

Fig 8. Final PF after H and SH suppression and √f weighting

Figure 9 shows the final PF obtained on the same sound, to which an equally intense white noise has been added.

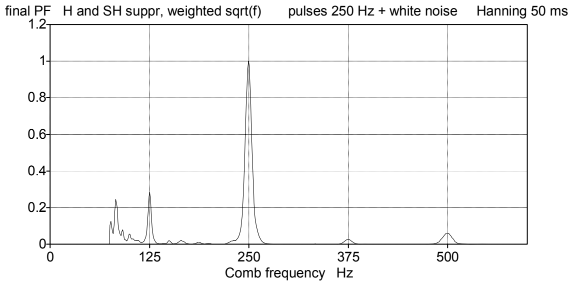

Fig 9. Same as in fig 8, but the sound has been mixed with white noise of same intensity

Figure 10 represents the final PF obtained from the mix of two vocalic sounds uttered by a single male voice. Both sounds are extracted from continuously pitch-varying and spectrum-varying sequences. The main peak of the most acute sound is less prominent, but it appears nevertheless in second position.

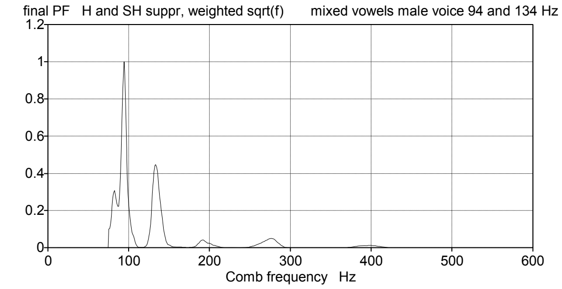

Fig 10. Mix of two vowels at 94 and 134 Hz from the same male speaker

Finally, figure 11 shows a mix of two artificial vowels of same intensity, whose F_0s are exactly one octave apart. The correct peaks emerge clearly, which can be seen as a positive feature for a harmonic suppression algorithm.

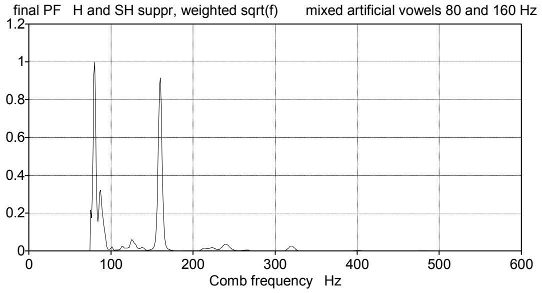

Fig 11. Mix of two artificial vowels of same intensity, with F0s exactly one octave apart (80 and 160 Hz)

5. Conclusion

The main goal of this work was to improve the ability of the spectral comb methods to provide an error-free pitch estimator, in order to be usable in both mono and multipitch estimation. We proposed to use irregular combs, either with some missing teeth or with alternate negative teeth, which have the property of canceling some families of erroneous solutions. An evaluation of ths approach in the mono and multipitch cases is in progress [8].

References

[1] Martin, P., "Comparison of pitch detection by cepstrum and spectral comb analysis", IEEE ICASSP, 180-183, 1982.

[2] De Cheveigné, A., "YIN, a fundamental frequency estimator for speech and music", J. Acoust. Soc. Amer., 111, 1917-1930, 2002.

[3] Sun X., "A pitch determination algorithm based on sub-harmonic-to-harmonic ratio", 6th ICSLP, Beijing, 2000.

[4] Camacho A., "SWIPE: a sawtooth waveform inspired pitch estimator for speech and music", PhD dissertation, Univ. of Florida, 2007.

[5] De Cheveigné, A., "Multiple F0 estimation", in Computational Auditory Scene Analysis, Wang and Brown eds, IEEE Press, Wiley-Interscience, 2006.

[6] Schroeder, M. R., "Period Histogram and Product Spectrum: New Methods for Fundamental-Frequency Measurement", J. Acoust. Soc. Amer., 43, 829-834, 1968.

[7] Liénard J.S., Signol F., Barras C., "Speech Fundamental Frequency estimation using the Alternate Comb", InterSpeech 2007, Antwerpen.

[8] Signol F., Barras C., Liénard J-S., "Evaluation of the Pitch Estimation Algorithms in the Monopitch and Multipitch cases", Acoustics 2008, Paris.