Speech fundamental frequency estimation using the Alternate Comb

Jean-Sylvain Liénard, François Signol and Claude Barras

LIMSI-CNRS 91403 Orsay Cedex, France

{jean-sylvain.lienard, francois.signol, claude.barras}@limsi.fr

Abstract

Reliable estimation of speech fundamental frequency is crucial in the perspective of speech separation. We show that the gross errors on F_0 measurement occur for particular configurations of the periodic structure to be estimated and the other periodic structure used to achieve this estimation. The error families are characterized by a set of two positive integers. The Alternate Comb method uses this knowledge to cancel most of the erroneous solutions. Its efficiency is assessed by an evaluation using a classical pitch database.

Index Terms: F_0 estimation, pitch detection, multipitch, spectral comb, speech separation

1. Introduction

Separating two speech signals mixed in a single channel, although easy for a human listener, proves to be difficult for automatic processing. Fundamental frequency F_0 is considered as the main feature usable for this task. Therefore it is necessary to develop robust Pitch Estimation Algorithms (PEA) that are able to yield satisfactory results even when multiple voiced signals are mixed together. A recent review of this problem can be found in [1].

Our objective is to analyze the nature of the errors produced by PEAs and to design a mechanism able to reduce them. The errors can be classified into 3 categories: voicing decision, gross errors and fine errors.

Voicing decision is ambiguous. The phonological point of view demands a binary decision, namely, Voiced or UnVoiced, although physical reality shows that there is always some gradual transition between the two states. Thus, it is necessary to fix a threshold, above which the frame is declared voiced.

In a voiced frame, F_0 estimation is performed by a particular function (periodicity indicator) computed for any F_c comprised between the arbitrary limits F_0min and F_0max. This estimation can be biased in two ways. First, a wrong extremum may be chosen by the decision algorithm, yielding what is usually called a gross error. Second, when the system chooses the right extremum, it may produce fine errors, due to small voice fluctuations, presence of noise, the window too narrow or too wide, or computational precision. Usually the limit between the two types of errors is fixed at +- 20% of the reference F_0, corresponding approximately to +- 3 semitones. Gross errors can occur with any type of periodicity indicator, be it spectral, temporal or spectro-temporal. In the present study we consider a purely spectral method, in the line of [2], [3], [4], among others.

First we explain the principles that lead to gross errors by way of the spectral structure that we call Simple Comb. We propose a modification which reduces some of those errors. The functioning of the new device called Alternate Comb is illustrated with real signals. Then we describe a monopitch evaluation including comparisons with other PEAs on the same database.

2. Origin and structure of the gross errors

Let us consider a spectral function |S| composed of N F_0-spaced peaks having a unity amplitude, and a spectral comb C unlimited in frequency, i.e. an infinite series of pulses of height unity and fundamental frequency F_c. Let us vary F_c.

When F_c=F_0 all of the spectral peaks are matched by the N first teeth of the comb (Figure 1), the scalar product of both functions is maximum and equals N. When F_c=2F_0 we get another product maximum, equaling the integer part of N/2. Choosing this maximum to represent the fundamental

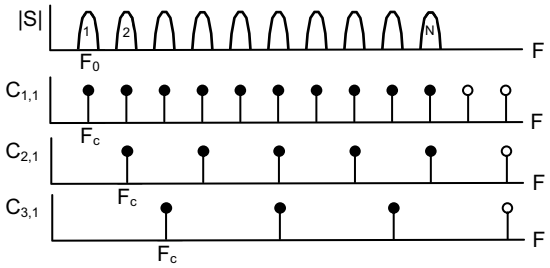

Figure 1: *series of 10 spectral peaks of fundamental frequency F_0 (top) and uniform infinite combs of fundamental frequencies F_c=F_0 , 2F_0 and 3F_0. The matching teeth are painted in dark.*

frequency of |S| yields an octave error. By proceeding upwards, several maxima of decreasing amplitude appear each time that F_c becomes a multiple of F_0. These maxima ("pitch peaks") correspond to harmonic errors of order p = 2, 3 ... etc.

Moving backwards from the starting position we encounter a new peak at F_c=F_0/2, although the first tooth does

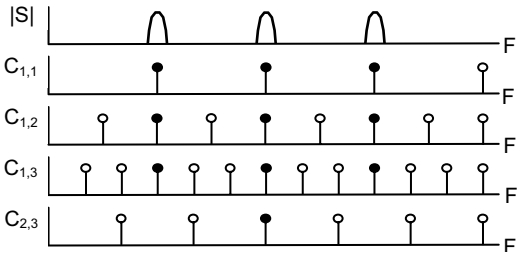

Figure 2: *series of 3 spectral peaks of fundamental frequency F_0 (top) and uniform infinite combs of fundamental frequencies F_c = F_0, F_0/2, F_0/3 and 2F_0/3. The matching teeth are painted in dark.*

not match any spectral peak (Figure 2). This is the order 2 sub-harmonic error, actually the sub-octave error. Because we consider an infinite comb, the scalar product amounts to the same value N as for the main peak at F_c=F_0.

There is a similar peak at F_c=F_0/3, which produces an order 3 sub-harmonic error. Again, the scalar product equals N. There is another related peak at F_c=2F_0/3, which produces

another sub-harmonic error of order 3. Thus we have to use two numbers, the harmonic order p and the sub-harmonic order q, to specify a pitch peak (p,q). The previous peaks are labeled (1,3) and (2,3). It is easy to identify other sub-harmonic peaks such as (1,4), (2,4) and (3,4), (1,5), (2,5). As N is limited, the amplitudes of the peaks (p, q) for which p is greater than 1 do not reach the value of the main peak (1,1). We have to notice that peaks (1,2) and peaks (2,4) are two different labels for the same entity and should preferably be designated by the simplest, irreducible form (1,2).

The subharmonic peaks observed for $F_c<F_0$ have replicas in all of the intervals between successive multiples of F_0. They are characterized by p>q and their magnitudes are globally decreasing.

The above considerations come very close to the basic notions developed by Schroeder in [2]: period histogram, frequency histogram, Harmonic Product Spectrum (HPS). Let us call "pitch function" the generalization of the above scalar product as a function of F_c. It differs from HPS by the fact that the products are not expressed in log units. Figure 3 shows the pitch function of a physical signal (series of pulses at F_0=250 Hz), analyzed by a uniform comb.

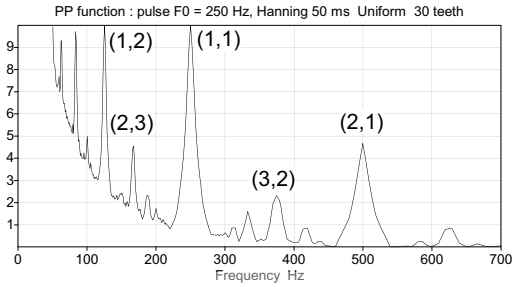

Figure 3: *Uniform Comb applied to a 250 Hz Hanning windowed pulse series. Some of the peaks are labeled with their (p,q) orders.*

3. The Simple Comb

The pitch function presented above is prone to gross errors, as it exhibits many peaks having the same maximum value in the region $F_c<F_0$. In order to make the main peak (1,1) dominate the others there are two solutions. One is to limit the number of teeth (usually 10), so that when decreasing F_c the set of teeth encompasses a smaller part of the spectrum. The other is to apply a decaying shape to the teeth. Both solutions may be implemented simultaneously. We used an exponential decay governed by a parameter (**ad**) chosen in {0,1}. Common values are **ad**=0 (no decay), **ad**=0.5 (decay in 1/sqrt(m), m being the tooth index), or **ad**=1 (decay in 1/m). As **ad** comes close to 1 or goes beyond 1, the pitch function tends to get identical to the part of the spectrum lying between F_0min and F_0max.

Figure 4 shows the same sound as in Figure 3, analysed with a 10-teeth Simple Comb decaying in 1/m : the sub-harmonic peaks (q>1) are somewhat attenuated and become less confusing than the harmonic ones (q=1).

There is a problem regarding the unit in which the spectrum module is best expressed in the pitch function calculation: linear (related to amplitudes), quadratic (related to energy and autocorrelation) or logarithmic (related to the decibel scale). As noticed in [5], as the voiced speech spectrum becomes globally less intense in the high frequencies, the quadratic units exaggerate the importance of the lowest part of the spectrum, and the logarithmic units give

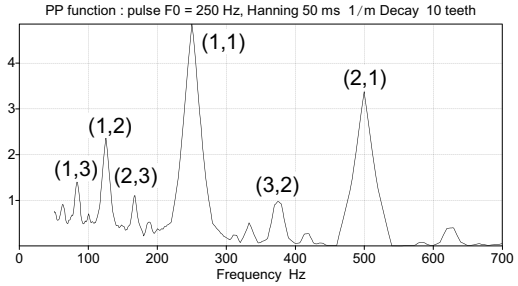

Figure 4: *Simple Comb applied to a 250 Hz Hanning windowed pulse series. Peaks of sub-harmonic order q>1 are attenuated compared to harmonic peaks q=1*

too much weight to the highest part or to the weakest spectral components. In the multipitch perspective the linear units should be preferred, as illustrated in Figure 8 below.

The Simple Comb, as well as the equivalent methods based on the accumulation of spectral shifts (for instance [4]) gives good results, even for telephone voice or in the presence of noise. The implementations differ in several respects: units of spectral magnitude, F_0min and F_0max limits, number of teeth, decaying function, spectrum pre-processing, and accumulation process. These variants aim at reducing the magnitude of the secondary peaks compared to the main one. None eliminates them completely, but it is not a real drawback in the perspective of single pitch estimation.

However, in the perspective of speech separation, reliable multiple pitch estimation is necessary. Mixing two periodic signals of fundamental frequencies $F_0$1 and $F_0$2 produces in the pitch function two peak families interfering in complex ways. Although one can presume that the main peak represents one of the two periodicities, identifying the other or assessing its absence is a difficult task. For this reason the pitch estimator has to produce the smallest possible number of reliable candidates.

4. The Alternate Comb

In order to reduce the magnitude of the harmonic peaks we propose the Alternate Comb. To the positive teeth of the Simple Comb we adjunct some intermediary negative teeth, positioned at the exact frequencies that may produce the harmonic errors (Figure 5).

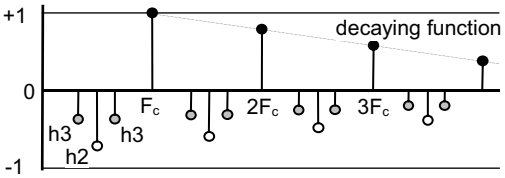

Figure 5: *Alternate Comb. The positive teeth are the same as in the Simple Comb. The negative teeth of magnitudes h2 and h3 contribute to reducing the harmonic errors (2,1) and (3,1).*

In the pitch function calcultation, subtracting the spectral components placed halfway from two successive positive teeth yields a large reduction of the octave error $F_c=2F_0$. The negative teeth placed at 1/3 and 2/3 of the positive teeth intervals reduce the error at $F_c=3F_0$. Weighting coefficients h2, h3 ... hp are attached to each harmonic order. Setting them to 0 transforms it back into a simple comb. By changing them gradually one can evaluate the impact of the proposed strategy. Figure 6 shows the pitch function obtained with the

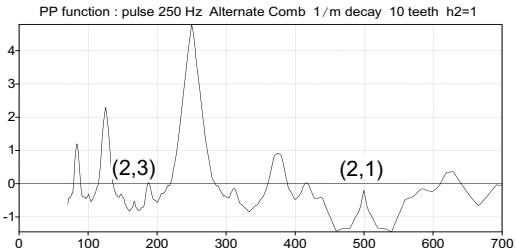

Figure 6: *Alternate Comb applied to the same 250 Hz Hanning windowed pulse series as in Figure 4.*

Alternate Comb on the same signal as above (250 Hz pulse series). Coefficient h2 has been set to 1. As a consequence peak (2,1) gets cancelled out, as well as the other peaks of harmonic order (p=2), for instance (2,3).

The pitch function can now take negative values. In order to ensure the existence of positive peaks the mean value is subtracted. The magnitude of the peak retained as possibly representing F_0 is compared to a threshold depending on the maximum surrounding level, within a +- 1 second interval.

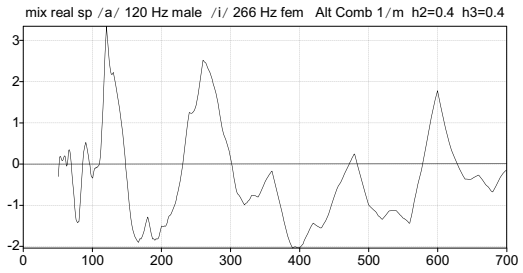

Figure 7: *Alternate Comb applied to a mix of two real vowels (50 ms Hanning windowed): /a/ male voice 120 Hz and /i/ female voice 266 Hz.*

Figure 7 shows the pitch function obtained on a frame extracted from the sum of two speech signals of equal level: /a/ (male voice 120 Hz) and /i/ (female voice 266 Hz). The Alternate Comb was tuned with h2=0.4 and h3=0.4. The octave peak at 240 Hz is practically cancelled, as well as the half-octave peak at 133 Hz. The peak at 600 Hz corresponds to the 5th harmonic of the first vowel (p=5, q=1). It is not cancelled because the coefficient h5 was set to zero in this tuning.

Figure 8 demonstrates the capability of the Alternate Comb to simultaneously process two synthetic speech signals of equal level that have F_0s exactly at one octave interval. One can observe that octave cancellation does not eliminate the peak at 160 Hz of the second signal. This result is important in the perspective of multipitch estimation. It is a direct consequence of the use of linear units in the calculation of the pitch function.

The Alternate Comb method bears some similarities with other published work, particularly [7, 10], devoted to the reduction of the octave error. Our method differs in three respects: i) it is based on the analysis of the different types of gross errors and not on considerations related to voice quality; this analysis is valid for any periodicity estimator ii) we use linear units for the spectral magnitude and pitch function computation, and iii) we place our study in the perspective of multiple pitch estimation.

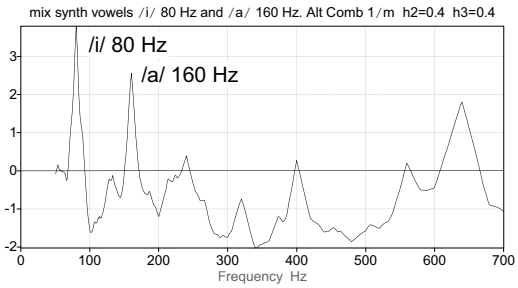

Figure 8: *Alternate Comb applied to a mix of two synthetic vowels, /i/ and /a/ (50 ms Hanning windowed), with their F_0s exactly one octave apart (80 and 160 Hz).*

5. Evaluation

The tests reported below have been conducted with the Keele database [7], totalling 337.1 seconds of speech uttered by 10 speakers (5 males, 5 females), i.e. 33710 frames concatenated into a single file without any level equalization. The voicing and F_0 values taken as references were those provided by the authors from the analysis of the electro-glottographic signal.

The performance of a given algorithm or tuning was estimated by two main indicators. VUV (Voiced-Unvoiced) is the ratio between the number of frames that have been misclassified regarding their voicing state and the total number of frames of the database. GER (Gross Error Rate) is the ratio between the number of gross errors and the number of frames declared voiced by both the reference file and the PEA tested.

The figures reported in Table 1 refer to two algorithms available in the Praat software [5]. The first one is the standard algorithm (called with "To Pitch..."), which gives highly reliable results in most practical situations. It is based on autocorrelation and uses an efficient post-processing. We used it with the standard F_0 range (75-600 Hz). Its results are good, but it does not provide a fair basis for comparison with the other PEAs, which do not implement any post-processing. The other one is also an autocorrelation-based algorithm, called with "To Pitch (ac)...". We used it with a setting mentioned in [10], which removes some of its post-processing capabilities.

In the 3rd and 4th lines we reported the GERs given in [9] and [10] with two PEAs based on different principles. Neither of them used any post-processing and the F_0 range was maintained to the same value for all speakers.

Table 1: *results from other PEAs on the same data*

	VUV %	GER %
Praat *To Pitch... 0.01 75 600*	10.63	1.65
Praat *To Pitch (ac)... 0.01 50 15 1 0 0 0.01 0.35 0 550*	41.15	3.39
YIN (as reported in [9])	n/a	2.40
SHRP (averaged from [10])	n/a	1.91

For the measurements on the Alternate Comb we first had to determine the best value of the decay parameter **ad**. Figure 9 shows the results obtained when **ad** varied between 0 and 1, with a minimum GER around **ad**=0.5. This illustrates the main feature of the Alternate Comb as compared to the Simple Comb: the decaying function is essential to reduce the

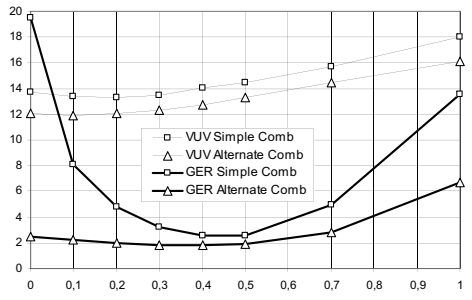

Figure 9: *GER and VUV (in %) as a function of the decay parameter* **ad**, *for the Simple Comb as well as for the Alternate Comb (h2=0.4 h3=0.2).*

prominence of secondary peaks in the Simple Comb (cf Figures 3 and 4), while the Alternate Comb specifically eliminates them at the source. For further measurements we chose 0.3 instead of 0.5 as the optimal **ad** value, because the minimum VUV was observed for the smallest **ad** values.

Table 2 shows the best results obtained with four different settings of the Alternate Comb. Only h2 and h3 were varied. The decay parameter was fixed at 0.3. The pitch function was centered, i.e. its mean value was removed. The window width was fixed at 40 ms and the F_0 range was fixed at 75-600 Hz, which are the default values of the Praat standard algorithm.

Table 2: *Comb results with width=40 ms*

	VUV %	GER %
Simple Comb h2=0 h3=0	13.51	3.23
Alt. Comb h2=0.8 h3=0	12.04	1.99
Alt. Comb h2=0 h3=0.4	12.84	3.00
Alt. Comb h2=0.4 h3=0.2	12.35	1.85

Globally, these results compare favorably with those reported in Table 1. They are not far from the best Praat result, which implements a sophisticated post-processing. They show that h2 alone, which specifically cancels the errors of order (2,q), provides the largest improvement with respect to the Simple Comb case. This confirms the observations reported in [6] and [10]. However they also show that the effect of h3 alone is less important than h2, and almost insignificant when both are optimally combined.

A possible explanation is that, for the lowest F_0 values, the resolution of the spectral peaks becomes poor when one uses a short (40 ms) time window. This loss of resolution affects differently the teeth h2 and h3: as the h3 teeth are closer to each other, they tend to fail more frequently at discriminating the parts of the spectrum module they should separate.

Table 3: *Comb results with width=80 ms*

	VUV %	GER %
Simple Comb h2=0 h3=0	15.91	1.65
Alt. Comb h2=0.8 h3=0	14.72	1.33
Alt. Comb h2=0 h3=0.4	14.97	1.35
Alt. Comb h2=0.4 h3=0.2	14.86	1.06

In order to check this explanation, we doubled the window size to increase the spectral resolution. The results shown in Table 3 confirm that the effects of h2 and h3 are cumulative. Besides, performance has improved for all the comb settings, because all the spectral peaks are better resolved, which benefits all of the comb methods in the lower F_0 range.

With this setting the Alternate Comb outperforms the other PEAs, including the Praat standard algorithm with its post processing. The slight deterioration of VUV is the consequence of the increase of the window duration, which increases the uncertainty of the voicing decision on the frames located near the limits of the voiced segments.

6. Conclusions

We have presented an approach to the problem posed by the gross errors in the F_0 estimation of speech signals. This approach was motivated by the multipitch perspective. Even in the monopitch case, the problem is error-prone, and we tried to understand why.

We enumerated and counted the coincidences occurring when a periodic structure of fundamental frequency F_0 is confronted to a periodic set of pulses of variable fundamental frequency F_c and decaying magnitude (Simple Comb). We found that the confusions were maximally plausible at certain locations, indexed with two positive integers p and q, named respectively the harmonic and sub-harmonic orders. The Alternate Comb method consists in introducing negative teeth at those precise locations in order to reduce the prominence of the corresponding peaks.

Evaluation on a popular database proved the method to give satisfactory results, thus validating our approach in the monopitch framework. A multipitch evaluation is in progress.

7. References

[1] De Cheveigné, A., "Multiple F_0 estimation", in *Computational Auditory Scene Analysis*, Wang and Brown eds, IEEE Press, Wiley-Interscience, 2006.

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

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

[4] Hermes, D. J., "Measurement of pitch by sub-harmonic summation", J. Acoust. Soc. Amer., **83**, 257-263, 1988

[5] Camacho, A. and Harris, J. G., "A spectral-based pitch estimation algorithm and pitch perception model using an integral transform with a truncated decaying cosine kernel", 4th ASA/ASJ joint meeting, Honolulu, 2006.

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

[7] Plante, F., Ainsworth, W.A. and Meyer, G.,"A Pitch Extraction Reference Database", Eurospeech Madrid, 837-840, 1995.

[8] Boersma P. and Weenink, D. "Praat: doing phonetics by computer", http://www.praat.org/

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

[10] Sun X., "Pitch determination and voice quality analysis using sub-harmonic-to-harmonic ratio", IEEE ICASSP, 333-336, Orlando, 2002.