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 F0 measurement occur for particular configurations of the periodic structure to estimate and the other periodic structure used to achieve the 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 on a classical pitch database.

Index Terms: F0 estimation, 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 an automatic processing. Fundamental Frequency F0 is considered as the main usable indices for this task. Thus it is necessary to work out robust Pitch Estimation Algorithms (PEA) able to give satisfactory results even when several voiced signals are mixed (multipitch estimation). However F0 estimation is a difficult, error-prone operation, even when one is certain that there is only one single voice in the signal. A recent review of this problem can be found in [1].

Our objective is to analyze the nature of the errors produced by a PEA 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, either in the production or in the perception perspective. From the Signal Processing point of view one also uses to consider that a given frame should be voiced or not, although physical reality shows that there is always some progressivity in the signal transition between the voiced and unvoiced states. Thus it is necessary to fix some threshold, above which the frame is declared voiced. It is well known that such a threshold cannot be valid for any kind of speech signal and any situation.

In a voiced frame F0 estimation is performed by a particular function (periodicity indicator) which computes a non-dimensional value for any value Fc comprised between the arbitrary limits F0min and F0max. Periodicity is indicated by the position of a given extremum of this function. The estimation can be biased in two ways. First, the extremum decision may choose a wrong one, which is easy because periodicity indicators often happen to be periodic themselves. This produces what is usually called gross errors. Second, when the system effectively chooses the right extremum, it may produce fine errors, which may have multiple causes: small voice fluctuations, presence of noise, window too narrow or too wide, computational precision. Usually the limit between the two types of errors is fixed at ±20% of the reference F0, corresponding approximately to ±3 semitones. Gross errors can happen with any type of periodicity indicator, spectral, temporal or spectro-temporal. In the present study we use a purely spectral method, in the line of [2], [3], [4], among others.

First we explain the principles by which gross errors are formed by the spectral structure that we call Simple Comb. We propose a modification of its structure which reduces some of those errors. The functioning of the new device called Alternate Comb is illustrated with real signals. Then we propose a monopitch evaluation in comparison with a popular autocorrelation-based PEA freely available with the Praat software [5].

2. Origin and structure of the gross errors

Let us consider a spectral function |S| composed of N harmonic peaks, of fundamental frequency F0 and amplitude unity, and a spectral comb C unlimited in frequency, i.e. an infinite series of pulses of height unity and fundamental frequency Fc. There is no spectral component between the peaks. Let us vary Fc.

When Fc = F0 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 Fc = 2*F0 there is still another product maximum, equaling the integer part of N/2. Choosing this peak to represent the fundamental frequency of |S| yields an octave error. By proceeding upwards one can see that peaks of decreasing amplitude appear each time that Fc becomes a multiple of F0. These peaks correspond to the harmonic errors of order p = 2, 3 ...

![Figure 1: series of 10 spectral peaks of fundamental frequency F0 (top) and uniform infinite combs of fundamental frequencies Fc = F0, 2*F0 and 3*F0. The matching teeth are painted in dark.](image)

If we move backwards from the starting position we see easily that we encounter a new peak at Fc=F0/2, although the first tooth does not match any peak (Figure 2). This is the order 2 subharmonic error, actually the sub-octave error. The problem is that, because we use an infinite comb, the scalar product amounts to the same value as for the main peak at Fc=F0.

There is a similar peak at Fc=F0/3, which produces an order 3 subharmonic error. Again, its scalar product equals N. In addition, we see that there is another related peak at Fc=2*F0/3. It produces a second order 3 subharmonic error. Thus we have to use two orders, the harmonic order p and the subharmonic order q, to specify a peak (p, q). The previous peaks are labeled (1, 3) and (2, 3). It is easy to identify other subharmonic peaks such as (1, 4), (2, 4) and (3, 4), (1, 5), (2,
5) etc. 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 form (1, 2).

Finally we observe that 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 \( p > q \). Their amplitudes are globally decreasing, due to two causes: i) the scalar product tends to take \( 1/p \) peaks in the summation when \( N \) tends to infinity and ii) \( N \) is limited.

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

There is a problem concerning the unit in which the spectrum module is best expressed in the PP calculation: linear (related to amplitudes), quadratic (related to energy and autocorrelation) or logarithmic (related to the decibel scale). As noticed in [6], as the voiced speech spectrum is globally less intense in the high frequencies, the quadratic units exaggerate the importance of the lowest part of the spectrum, and the logarithmic units gives too much weight to the highest part or to the weakest spectral components. According to our experience the linear units are better adapted to the problem.

3. The Simple Comb

The PP function presented above is prone to gross errors, as it exhibits many peaks having the same maximum value, especially in the region \( < F_0 \). In order to make the main peak \((1, 1)\) dominate the others there are two solutions. One is to limit the number of teeth, 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 may be implemented together. Common values are 10 for the number of teeth and \( 1/m \) or \( 1/\sqrt{m} \) for the decaying function \((m \) is the tooth index). Figure 4 shows the same sound as in Figure 3, analysed with a 10-teeth Simple Comb decaying in \( 1/m \). Generally, the subharmonic peaks are somewhat attenuated and become less confusing than the harmonic ones.

![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.](image1)

![Figure 3: Uniform Comb applied to a 250 Hz Hanning windowed pulse series. Some of the peaks are labeled with their \((p, q)\) orders.](image2)

3. The Simple Comb

In order to reduce the amplitude 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 contribute to reducing the harmonic errors of orders \((2, 1)\) and \((3, 1)\).](image3)
Subtracting from the PP summation the spectral components placed halfway from two successive positive teeth produces a large reduction of the octave error \( F_c=2\times F_0 \). The negative teeth placed at 1/3 and 2/3 of the positive teeth intervals reduce the error at \( F_c=3\times F_0 \). As the optimal height of the negative teeth cannot be computed a priori, weighting coefficients \( h_2, h_3 \ldots h_p \) are attached to each harmonic order. These coefficients are the main parameters of the Alternate Comb. Fixing 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 PP function obtained with the Alternate Comb on the same signal as above.

The function PP can now take some negative values. In order to ensure the existence of positive peaks the mean value is subtracted. The amplitude of the peak retained as possibly attenuated, with the exception of the one located at 640 Hz. Most of the undesired peaks are strongly cancelled out. Compare to Figure 5. As a consequence peak \((2, 1)\) gets cancelled out. Figure 6 shows the PP function obtained with the Alternate Comb on the same signal as above.

The Alternate Comb method bears some similarities with other published work, particularly \([7]\), where the author implements a processing devoted to the elimination 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; ii) we use linear units in the spectral magnitude computation, and iii) we place our study in the perspective of multiple pitch estimation.

5. Evaluation

For preliminary studies we used speech data extracted from the Speech Separation Challenge \([8]\), in particular 10 sentences (5 males, 5 females) totalling 17 seconds. The tests reported here have been conducted with the Keele database \([9]\), totalling 337.1 seconds of speech uttered by 10 speakers (5 males, 5 females), ie 33710 frames, of which 14936 were considered voiced by the reference algorithm.

We chose to compare several tunings of the Alternate Comb to an algorithm widely used in the speech community. The Praat AC PEA is based on autocorrelation and uses an efficient post processing. Prior to any other measurements, we compared the results given by the same algorithm on the audio signal (reference) and on the egg signal (test). As a result we observed a rather large rate of voicing errors and a rather small rate of gross errors (table 1, first line). This indicates that, as long as the gross error rate remains larger, taking as reference the standard Praat AC algorithm on the audio signal is legitimate.

As indicated above, the results obtained by a given PEA on a given database may differ according to the voicing criterion used. We minimized the corresponding bias by adjusting the voicing threshold so that the undervoicing rate (the PEA tested declares less voiced frames than the reference) is of the same order of magnitude than the overvoicing rate (the PEA tested declares more voiced frames than the reference). We checked that the gross error rates do not vary much if the undervoicing and overvoicing rates are kept within the interval of 2 to 8%.

Our evaluation was not directed towards any rigorous performance comparison with other PEA models, the results of which have been published in several papers such as \([6]\), \([7]\) or \([10]\). Instead, it aims at investigating the parameters of the Alternate Comb when gradually introducing negative teeth of orders \( hp \) \((p=2\) and \(p=3)\) in the Simple Comb. As some parameters are interdependent, the general idea was to seek the best result for each setting of the \(hp\) and voicing threshold parameters from a trial set (a part of the whole database
comprising 6147 frames out of 33710). The values given in table 1 were computed from the whole database with those values. All other settings were kept constant across measurements and algorithms. In particular the window width was fixed at 40 ms and the F0 interval was fixed at 75-600 Hz, which are the default values of the Praat standard algorithm.

Table 1: summary of the evaluation results

<table>
<thead>
<tr>
<th></th>
<th>VUV %</th>
<th>GER%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Praat egg signal vs audio</td>
<td>12.18</td>
<td>1.13</td>
</tr>
<tr>
<td>Simple Comb h2=0 h3=0</td>
<td>8.87</td>
<td>14.37</td>
</tr>
<tr>
<td>Alt Comb h2=-1.0</td>
<td>7.99</td>
<td>1.90</td>
</tr>
<tr>
<td>Alt Comb h3=-0.4</td>
<td>7.74</td>
<td>1.85</td>
</tr>
<tr>
<td>Alt Comb h2=-0.4 h3=-0.4</td>
<td>7.29</td>
<td>1.43</td>
</tr>
</tbody>
</table>

VUV is the ratio between the number of frames that have been misclassified regarding the voicing state, and the total number of frames of the database. GER represents the ratio between the number of gross errors and the number of frames declared voiced by both reference and tested PEs.

We did not report here the mean deviation of the F0 values found in the fine error category. In all the situations examined, the average difference was less than 0.07 semitone, with a standard deviation of less than 0.30 semitone. In other words, when there is no gross error, the value found for F0 is practically exact.

The first line shows the result of the reference algorithm applied to the egg signal band-pass filtered between 50 and 1000 Hz. The result shows large discrepancies concerning the voicing decision. The audio signal is declared less voiced than the egg signal, which casts a doubt on the value of the egg signal as a voicing ground truth: in most cases the vocal folds vibrate but the sound produced is inaudible or too low in frequency to correspond to the perceptive voicing. On the other hand, when both signals are declared voiced, the rate of gross errors is quite low.

The second line corresponds to the Simple Comb. The surprise comes from the rather high rate of gross errors. This could probably be improved by adjusting more precisely the number of teeth and their decaying function. However, there is a very large gap to fill to compete with the next case.

The 3rd line shows the drastic effect of a perfect cancellation of the octave error, with the coefficient h2 equal to 1. This confirms the observations reported in [7] and [11].

The 4th line shows the effect of partially cancelling the p=3 harmonic error. This effect is as strong as the previous one. It may be explained by the fact that for low-pitched voices and short frame durations the spectral peaks tend to merge. Their processing with the order 3 interteeth produces more or less the same effect than the single negative tooth of order 2.

Finally, using both orders yields the best result (line 5). It must be noted that the final gross error rate is still superior to the one obtained on the egg signal, which confirms the statistical validity of our results.

Although our evaluation was not done in order to compete with other PEs, it should be noted that other authors using a very similar setup and the same database obtain results in the same range. For instance, on the Keele database, with width=40 ms, F0min=50 and F0max=550 Sun [11] gets a gross error rate of 2.08% for male speakers and 1.74% for female speakers, i.e. around 1.9% for the whole database, for which we found a best rate of 1.43%. However this difference is to be appreciated with caution, due to the difference in the choice of the reference data, as well as in the many small differences that occur from one experimental setup to another.

6. Conclusions

We have presented an approach to the problem posed by the gross errors in the F0 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 F0 is confronted to a periodic set of pulses of variable fundamental frequency Fc (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 subharmonic orders. Thus, as we knew where the gross errors could happen, we could reduce from the start the noicity of these locations. This was the basis of the Alternate Comb method, in which some negative teeth indicate where the spectral amplitude should be reduced to minimize the danger of confusion.

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

7. References
