A PROBABILISTIC PITCH ESTIMATION METHOD WITH COMBINATION OF SEVERAL TECHNIQUES

G. Bruno, M.G. Di Benedetto, A. Gilli
Università di Roma

Abstract

In this paper, a bayesian approach to pitch estimation is presented. The pitch frequency is assumed to be a random variable. At step k, some detections of the pitch $F_k$, obtained by different classical procedures, are used to modify the a priori pdf of $F_k$. In this way, the estimation of $F_k$ is made by choosing the expected value of the a posteriori pdf of $F_k$. In addition, the previous measurements are used to update the a priori pdf of $F_k$ for the next step. The results obtained by the application of this algorithm, considering the combination of two fast techniques of pitch detection, are very satisfactory and better than those obtained by applying each method separately.

1- Model

With reference to fig.1, let $f_k$ indicate the true value of the pitch frequency corresponding to the $k^{th}$ segment of the speech signal, which is supposed to be voiced; in addition, let $f_k$ be obtained by the relation:

$$f_k = f + e_k$$  (1)

where $f$ is the expected value of the pitch frequency of the speaker and $e_k$ is an error. In expression (1), the quantities $f$, $e_k$, and therefore $f_k$, will be treated as determinations of random variables, which will be indicated by capital letters. In particular, one assumes:

1- that the sequence $\{E_k\}$ be stationary and memoryless;
2- that $F$ and $E_k$ be stochastically independent.

The estimation of the pitch frequency $\hat{F}_k$ is made by combining the results of the two different measurement procedures which will be described in section 4 and indicated by $a_k$ and $b_k$ in fig.1, with the a priori information available concerning $F_k$ in position $k$. In particular, the quantities measured $a_k$ and $b_k$ are also considered to be determinations of the random variables obtained from $f_k$ by means of the relations:

$$a_k = f_k + q_k ; \quad b_k = f_k + r_k$$  (2)

in which $\{q_k\}$ and $\{r_k\}$ are supposed to be stationary, stochastically independent reciprocally and of $\{E_k\}$, $\{F_k\}$ and $F$. In addition, the quantities $a_k$ and $b_k$ are used to update the pdf of $F_k$ necessary for the estimation of $F_{k+1}$. For the sake of simplicity, all the pdf updated on the basis of the measurements made in the previous positions will be indicated by the index $k$, without explicitly mentioning this condition.

As regards the estimation of $f_k$, the Bayes criterion is used with a quadratic cost function for the estimation error. Consequently, one has:

$$\hat{f}_k = \frac{p_k(f_k/a_k,b_k)}{\int p_k(f_k/a_k,b_k) df_k} = \int p_k(f_k/a_k,b_k) df_k$$  (3)

where from (2) one gets:

$$p_k(f_k/a_k,b_k) = \frac{p(a_k,f_k/b_k) p(b_k/f_k) p_k(f_k)}{\int p(a_k,b_k/f_k) p_k(f_k) df_k}$$  (4)

and

$$p_k(f_k) = \int p_k(f_k/t) p(t) df + \int p_k(f_k/t) p_k(t) df_k$$  (5)

as, due to the hypothesis made, the pdf of $E_k$ is not required to be updated and therefore, conditioned to the event $F \neq f$, the random variable $F_k$ is independent of the random variables $A_1, A_2, ..., A_{k-1}, A_k$. As regards updating, from equations (4) and (5) it is evident that the only quantity which must be updated is $p_k(f)$. Using for this purpose the measurements $a_k$ and $b_k$, one gets:
with:

\[ C = v^2.s^2 + u^2.s^2 + u^2.v^2 \]
\[ D_{k-1} = C + (v^2 + u^2).s_{k-1} \]

i.e., by setting \( Q = (u^2 + v^2) / C \), in particular one obtains:

\[ s_{k}^2 = s_{k-1}^2 / [1 + (k-1).Q.s_{k-1}^2] \]

Therefore, the variance \( s_{k}^2 \) tends to zero with increasing \( k \), with the evident physical significance that when \( k \) increases the expected value of the pitch of the speaker becomes known. Relations (14) and (16) constitute the updating relations required.

Finally, as regards the estimation relation, from equations (1) and (3) one gets:

\[ p(f_k/f) = p_k(f_k) = g(f_k; s_{k}^2) \]

and by taking into account equations (5), (13), (17):

\[ p_k(f_k) = g(f_k; m_k, s_{k}^2 + s^2) \]

From the latter and taking into account relations (4), (10) and (11), it is inferred that also \( p_k(f_k; m_k, s_{k}^2 + s^2) \) belongs to a gaussian distribution with parameters given by:

\[ f_k = (s^2 + s_{k}^2) \cdot (v^2.s_k + u^2.b_k)/D_k + u^2.v^2.s_{k}^2/m_k \]

Relation (19) constitutes the relation of estimation required.

4- Experimentation

This algorithm has been experimentated at the "Centre Scientifique IBM France" in Paris, on an IBM Series/1 computer. The speech signal is first filtered at 4.8 kHz by means of a Butterworth 8-pole low-pass filter and then sampled at 10 kHz. The samples so obtained are stacked on a disk and analyzed, considering 128 samples for each frame, which correspond to windows of 12.8 ms duration. The two methods which have been combined have the characteristic of being very simple and implementable in real-time. Both methods make the computation by calculating the mean value of the pitch, window by window (128 samples). Let's examine the principle of these two algorithms.

Maxima: The signal is first filtered by means of a low-pass filter (Chebychev filter). This filtering is made by software. Then, after a test using a threshold, the distance between two significant maxima is computed. This algorithm has been used on an Intel 8086 and works in real-time. A flow chart of this algorithm is shown in Fig.2

Autocorrelation: On windows of 256 samples with an overlap of 128 samples, only the signs of the signal are conserved. The signal is not filtered so that the algorithm can work in real-time on an Intel 8086. After the computation of the pseudo-autocorrelation function, the value for which the autocorrelation is maximum over a threshold is chosen. This algorithm is shown in Fig.3.

One should note that there is no V/U/V classification system before the pitch detection (53). Both methods make the classification. It is obvious why typical errors will be considered: an error for the pitch detection and one for the V/U/V classification.

Experimental results: This algorithm has been tested on 28 words spoken in English by a professional speaker. For these words, the glottal signal has been obtained by means of an accelerometer. During a training phase the values \( u^2 \) and \( v^2 \) have been evaluated and it has been found that: \( u^2=60 \) (maxima), \( v^2=50 \)
(autocorrelation). As mentioned previously, two errors must be considered:

1- the classification error defined as:

\[
CLE = \frac{VUV + UUV}{NTOT}
\]

where VUV is the number of voiced windows classified as unvoiced, UUV the number of unvoiced windows classified as voiced and NTOT the total number of windows considered.

2- the pitch error:

\[
PE = \frac{1}{NTOT} \sum_{i=1}^{NTOT} |FG(i) - P(i)|
\]

where FG(i) is the real value of the pitch and P(i) the estimated value for the \(i\)th voiced window. PE gives the mean value of the absolute error in the evaluation of the fundamental period expressed in number of samples.

The results obtained are the following:

<table>
<thead>
<tr>
<th>Method</th>
<th>CLE</th>
<th>PE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maxima</td>
<td>0.07</td>
<td>1.38</td>
</tr>
<tr>
<td>Autocorrelation</td>
<td>0.02</td>
<td>1.59</td>
</tr>
<tr>
<td>New method</td>
<td>0.04</td>
<td>1.07</td>
</tr>
</tbody>
</table>

For example, if FG=25 (400 Hz at 10 kHz), one has the following variation for the estimated pitch:

\[
25 - PE \leq P \leq 25 + PE
\]

As one can see, the results obtained by the combination of the two methods are better than the ones obtained by applying each method separately.