



PITCH DETERMINATION BASED ON WAVEFORM SUPERPOSITION

C. Nadeu, E. García-Melendo and J. Alsina *

ABSTRACT

Based on the idea of waveform superposition, three algorithms for pitch determination and voiced/unvoiced decision are developed. Although this new approach is intimately related with the autocorrelation analysis method, the obtained results show an improvement with respect to it that is more remarkable in the presence of noise.

INTRODUCTION

Pitch determination and voiced/unvoiced (v/uv) decision are important tasks in speech processing; for this reason, a large number of techniques have been developed (ref. 1). However, none of them is reliable and accurate enough for a wide variety of working conditions (ref. 2), specially when the speech signal is contaminated by noise (ref. 3).

Algorithms based on the autocorrelation function yield good performance for both clean (ref. 4) and noisy (ref. 3) speech and they are a tradeoff between accuracy and computational complexity. In this paper we describe some new methods of fundamental frequency extraction based on the idea of waveform superposition (ref. 5) and we relate and compare them with the autocorrelation method.

WAVEFORM SUPERPOSITION TECHNIQUES

Consider a speech frame $x(n)$ consisting of N samples, from $n=0$ to $n=N-1$, and an integer value m that satisfies $m \leq N/2$. We can build a new sequence $S_m(n)$ according to

$$S_m(n) = \sum_{k=1}^r x(n+(k-1)m) \quad (1)$$

where $n=0, 1, \dots, m-1$ $r = \lceil N/m \rceil$

and $x(n)=0$ if $n \geq N$

In other words, each frame of the speech signal is divided into N/m parts that are added up to obtain a superposition sequence $S_m(n)$ which is m samples long. If m is very close to the actual pitch period M_0 , the addition is performed coherently, whereas if M_0 is far from it, the superposition is not coherent.

* Dept. de Teoria del Senyal i Comunicacions. Universitat Politècnica de Catalunya, Ap.30.002, 08080 Barcelona, Spain.

This work is supported by the CAICYT, grant number 21096/84.

In order to obtain the pitch estimate we should extract from $S_m(n)$ a parameter for every m which measures the coherence degree of the waveform superposition, i.e. the periodicity of the signal. A first method (WS1) uses the following difference of energies

$$D(m) = \frac{1}{2} \left[\sum_{n=0}^{m-1} (S_m(n))^2 - \sum_{n=0}^{N-1} (x(n))^2 \right] \quad (2)$$

This expression has an interesting interpretation in terms of correlation. Since $S_m(n)$ results from the addition of r segments of signal, the m numerator involves the correlation of every segment with itself and all the others. Consequently, if $R_x(n)$ is the autocorrelation function computed from the speech frame $x(n)$ and extended with zeroes from N to ∞ , it follows that

$$D(m) = \sum_{k=1}^{r-1} R_x(km) \quad (3)$$

This average suggests that the corresponding algorithm will be attractive for the determination of fundamental frequency of speech contaminated by noise.

A second way of extracting information about periodicity from the superposition sequence $S_m(n)$ is based on the idea of principle cycles. According to Miller (ref. 5), if an excursion cycle consists of that part of the waveform between consecutive zero crossings, the first excursion cycle that occurs in a pitch period (its principle cycle) tends to have large amplitude and long duration and, consequently, it has considerable energy. It is reasonable to expect that, for $m=M_0$, the principle cycle appears clearly in $S_m(n)$. Therefore, it makes sense to search for the maximum energy excursion cycle and use its energy $E(m)$ to estimate the pitch period.

Two different measures are built with the above value. The first one (method WS2) is the value itself, i.e. $E(m)$, and the second one (method WS3) is like expression (2) but performing the two sums only inside the maximum energy excursion cycle, i.e.

$$DE(m) = \frac{1}{2} \sum_{n=n_1}^{n_2} \left[(S_m(n))^2 - \sum_{k=1}^r (x(n+(k-1)m))^2 \right] \quad (4)$$

being n_1 and n_2 the first and last samples of the maximum energy excursion cycle.

Once $D(m)$, $E(m)$ or $DE(m)$ are computed for all values of m belonging to the interval where the pitch period is searched, a plot of them reveals a significative peak around $m=M_0$ when the current frame is voiced, so that a peak detector would suffice to estimate it. However, from ex-

pression (3) it is clear that the peak around the harmonics M_0/i , $i=2,3,\dots$, may easily be higher than the peak around M_0 ; namely, $D(m)$ favours low values of m in front of high values, because if m decreases, the number r of summands in (1) increases. In order to avoid errors due to second or higher harmonic tracking, specially when the signal is band limited, we can divide $D(m)$ by N/m , which is equivalent to use $mD(m)$ instead of $D(m)$ as input to the peak detector that obtains the fundamental frequency estimate. Another way of avoiding harmonic errors consists of computing the peak areas, since the peak widths are roughly proportional to m ; in fact, a shift $\Delta\omega$ from $m=M_0/i$ produces the summand $R_x(M_0+i\Delta\omega)$ in (3). However, it will be suitable to perform adaptive center clipping before the area computation to isolate peaks. The choice of the clipping level is very important because it controls the relative weight given to harmonics and subharmonics. Indeed, all the above considerations about octave errors can be extended to $E(m)$ and $DE(m)$ functions.

Since all three introduced methods belong to the same family than the autocorrelation pitch determination algorithm they easily allow a v/uv detection. The decision threshold is made equal to $h.R_x(0)$, where h is a constant value.

EXPERIMENTS AND RESULTS

Some preliminar experiments were carried out. The speech material consisted of a sentence M1 uttered by a male speaker and a sentence F2 uttered by a female speaker. The reference pitch contour and v/uv decision were evaluated by visual inspection. Signals were sampled at 8KHz and gaussian random noise was added to them in order to obtain signal-to-noise ratios (SNR) from ∞ to 0 dB.

We carried out tests with the three above mentioned waveform superposition methods and the autocorrelation with adaptive center clipping method (ACC) (ref. 4). Neither low pass filtering nor other kind of preprocessing was employed.

The algorithms used in the methods WS1, WS2 and WS3 compute the areas of peaks after adaptive center clipping of the corresponding $D(m)$, $E(m)$ or $DE(m)$ function. The clipping threshold was not optimized; it was made equal to the average value of the function at each frame. The frame length was 40 ms and a pitch value was extracted every 10 ms exploring the range from 50Hz to 400Hz. To compare the methods, the same approach as Rabiner et al. (ref. 2) was employed.

Table 1 shows a meaningful subset of results corresponding to SNR= ∞ and 0. Gross errors (differences from the reference value greater than 1ms) as well as standard deviations (SD) of fine errors (differences smaller or equal to 1ms) were measured in labeled voiced frames independently from the v/uv decision; voiced errors correspond to voiced frames where the algorithm decision was unvoiced or a gross pitch error took place. Unvoiced errors mean that an v/uv decision error occurred in an unvoiced frame. The number of unvoiced

frames is 33 for M1 and 30 for M2, and the number of voiced frames is, respectively, 186 and 153.

We can observe that all three algorithms have a relatively small standard deviation which does not increase significantly with noise. The method WS1 achieves the least amount of gross pitch errors, whereas if we take into account the global performances, WS2 is preferred, except for F2 and SNR=0dB, where WS3 obtains the best results. It is needed to mention that the constant h of the v/uv decision was temporarily set to 0.7, 0.7 and 0.2, respectively, in WS1, WS2 and WS3 algorithms; when the data base will be enlarged, these values may require a readjustment.

CONCLUSIONS

According to preliminary experiments, waveform superposition techniques obtain better results than the ACC method for speech signals corrupted by noise. This observation agrees with the kind of average that they perform on the autocorrelations or related functions. Furthermore, by using the idea of principle cycle, they also may improve ACC method in clean speech. Nevertheless, the computational complexity of both methods is comparable. Presently, we are carrying out a greater number of tests that will be presented at the conference.

REFERENCES

1. W. Hess, Pitch Determination of Speech Signals (Springer-Verlag, 1983).
2. L.R. Rabiner et al., IEEE Trans. on ASSP, Vol. 24, No. 5, (1976).
3. K.A. Oh and C.K. Un, ICASSP-84, pp.18B.4.1-4, (1984).
4. L.R. Rabiner, IEEE Tran. on ASSP, Vol. 25, No.1, (1977).
5. E. García-Melendo and C. Nadeu, URSI-86, pp. 69-71, (1986).
6. N.J. Miller, IEEE Trans. on ASSP, Vol. 23, No. 1 (1975).

TABLE 1

M1 ∞dB	SD (ms)	Gross	Voiced	Unv.	F2 ∞dB	SD (ms)	Gross	Voiced	Unv.
ACC	0.43	32	53	6	ACC	0.25	1	3	7
WS1	0.49	23	65	7	WS1	0.21	1	3	11
WS2	0.49	24	33	7	WS2	0.21	1	2	6
WS3	0.46	28	49	5	WS3	0.22	2	2	11

M1 0dB	SD (ms)	Gross	Voiced	Unv.	F2 0dB	SD (ms)	Gross	Voiced	Unv.
ACC	0.56	48	137	0	ACC	0.30	36	80	0
WS1	0.44	35	116	0	WS1	0.24	5	32	0
WS2	0.52	42	80	0	WS2	0.21	6	86	1
WS3	0.52	51	87	4	WS3	0.24	6	15	4