A MODIFIED ZERO-CROSSING METHOD FOR PITCH DETECTION IN PRESENCE OF INTERFERING SOURCES

François GAILLARD, Frédéric BERTHOMMIER, Gang FENG, Jean-Luc SCHWARTZ
Institut de la Communication Parlée, UPRESA 5009
46 avenue Félix Viallet 38031 GRENOBLE cedex 01
Tel: +33 04 76 57 47 15 FAX: +33 04 76 57 47 10, E-mail: gaillard@icp.grenet.fr

ABSTRACT

This paper evaluates, in terms of speech signal processing, a non-linear method of pitch detection based on the detection of the zero-crossings of the signals (ZC method), in adverse conditions of interference.
First, F0 identification is evaluated according to the relative level of energy between the components in mixtures of pure tones or pairs of vowels; then, we introduce in the double-vowel paradigm a *confidence measure based on the standard deviation of inter-zero intervals*. Finally, the robustness of this confidence measure is tested in two cases of interference : pure tones + noise, and vowels + noise.
We show that the method allows to detect periodicity without any knowledge about the nature of the interfering sources, and then to identify their fundamental frequency.

1. INTRODUCTION

The human auditory system is able to extract one voice from mixed speech signals. In terms of signal processing, this auditory mechanism of separation can be modelled by extracting from a mixture a primitive of the speech signals that would help characterizing each of the mixed sources. This strategy can be placed into the context of Auditory Scene Analysis [1], in which structural cues like the fundamental frequency (F0) or the interaural delay (ITD) were proposed to separate mixed voices.
In the framework of F0-dependent separation, we have implemented, modified and evaluated a non-linear method of pitch detection based on the extraction of the zero-crossings of the signal (ZC method, often used in speech processing to detect voicing), in different conditions of interference.
After being tested in presence of sine wave mixtures, in clean and noisy conditions, the method is evaluated in the double-vowel paradigm and in conditions of noise interference.

2. THE ZC METHOD

2.1. Basic algorithm

Our ZC method includes three steps: first, all the positions of the zero-crossings with positive slope contained in a 80ms-window of signal are detected; second, these positions allow to build an histogram of the intervals contained into the 80-ms window. Third, this histogram provides a mean interval (μ) and a standard deviation (σ) of the first-order intervals inside the 80-ms window. The estimated F0, that we note F0* (in Hertz), corresponds to the inverse of μ, and the standard deviation (σ, in bins) estimates the variations of the interval lengths around the mean value.

2.2. Evaluation with mixtures of pure tones. Main characterictics

To start the evaluation of the ZC method in interfering conditions, we first mixed two pure tones at different Relative Levels (RL, difference in dB between the energy of one tone and the energy of the other). The evolution of F0* and σ when RL varies between -15dB and 15dB are shown on Figure 1. These variations have been smoothed by an average over 100 random phases, so that estimation is not dependent on phases. For comparison, we show the variations of the estimated F0 by linear autocorrelation function (ACF), in the same experimental conditions.

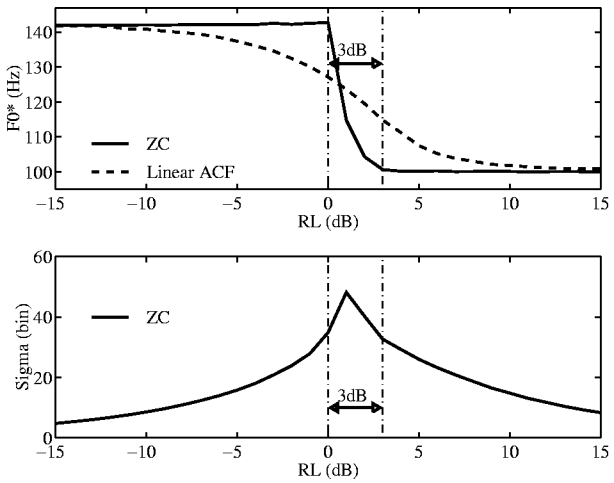

Figure 1. Mixture of two pure tones $F0_1$=100Hz, and $F0_2$=142Hz, at relative levels RL∈ [-15..15]dB.

We show that the transition area is narrow, smaller than 3dB, and corresponds to low values of |RL|. In this area, F0* is intermediate between the two expected frequencies, and σ is high. Notice that the transition is

not exactly symmetric: as soon as RL = 0dB, the zero-crossings of one source appear (or disappear) abruptly inside the temporal signal. This explains the sharpness and the asymmetry of the transition. In comparison, the transition area is broader with the linear ACF method.
The observation of the F0* variations with RL shows that the ZC pitch detection method provides strong dominance effects: outside the transition area, one of the sources masks the other, without being affected by it. In the following we will test this property of the ZC method in the double-vowel paradigm.

3. A FIRST STUDY OF THE DOUBLE-VOWEL PARADIGM

3.1. Experiment

We prepared mixtures of two among six French vowels, [a, e, i, o, u, y] with RL=0dB (equal *rms level* for both vowels in a pair). For each pair of vowels, one F0 is fixed at 100Hz, the other being chosen among 12 semitones-scaled other values, from 106Hz up to 200Hz, hence 360 pairs altogether.
The processing system is composed of a first spectral analysis stage based on a second order gammatone filterbank [2] with 32 channels. The second stage realizes a demodulation in medium- and high-frequency channels, thanks to a Half-Wave Rectification (HWR) followed by a bandpass filtering in the band 50-200Hz. The last stage performs pitch detection on each channel, by using the ZC method, during 80ms. Then, for each pair in each channel, the system provides a mean (μ) and a standard deviation (σ) of the ZC intervals within 80ms.
To characterize the behavior of the system, we computed the relative levels between the two filtered sources at the output of the spectral analysis, inside each channel, namely $RL_{i=1..32}$. For a given channel and a given pair of vowels, we consider that the source with the highest energy in the channel should be *dominant*. Then, we define three situations: F0* corresponds (within a 5Hz resolution) to the dominant source ("Dominant"), to the Non-Dominant source ("Non-Dominant") or to neither of them (« Errors »).

3.2. Results

A first inspection of the results showed a rather unefficient behavior of the ZC method in the low frequency channels, certainly because harmonics are resolved in these channels due to their small bandwidth. Hence, in the following, we consider only channels 10 to 32, corresponding to frequencies higher than 700Hz.
For all the channels and all the pairs, we computed the percentage of cases corresponding to the three situations described previously, namely Dominant, Non-Dominant and Errors. Table 1 provides the obtained rates.
When looking at these results, we observe that, while the percentage of F0 identification of the Non-Dominant source is quite low, the percentage of Errors is unexpectedly high.

Dominant	Non-Dominant	Errors
63.8%	2.2%	34.0%

Table 1. Detection rates in the double-vowel task.

Hence, though this method leads to strong dominance effect in presence of pure tones interferences, the error rate is high with complex sources like pairs of vowels, when interferences are produced by fundamentals *and* harmonics. The ZC method doesn't seem to be as efficient as expected for F0 identification.
However, we have seen that, with pure tones mixtures in the transition area, where F0* corresponds to neither of the two expected frequencies, the standard deviation of the intervals histogram becomes high. This parameter could be used as a confidence measure parameter, first to detect the presence of a dominant periodicity, and then, in this case, for F0 estimation.

4. INTRODUCTION OF A CONFIDENCE MEASURE BASED ON σ

4.1. Definition of a confidence criterion for double vowels.

In the previous section, we exploited only μ values at the output of the filterbank. In Figure 2, we consider both μ and σ values. In this figure, one point corresponds to one estimation for one pair of vowels in one channel.
A striking point emerges from this figure. Indeed, it is obvious that "good" estimations of the dominant F0 produce low σ values, while "errors" produce high ones, and, more importantly, that there seems to exist a clear-cut separation between the two groups of values. The boarder involves a σ threshold decreasing with F0*, but the reason and the way it decreases are not yet completely understood in theory. Hence we have used at present a statistical approach (by the way of a ROC curve) to build a separation line between "correct" and "wrong" estimations. The result, based on the analysis of the 360 pairs of vowels (hence, altogether 8280 (F0*,σ) points since there are 23 useful channels providing one point each), is superimposed to the data in Figure 2; it seems quite in line with the two groups of points.
We provide in Table 2 the percentages of correct estimations of the dominant F0 (within a 5Hz resolution), and the percentage of Errors (including estimations of the Non-Dominant F0, quite marginal in our results) above and below the separation line respectively.

	Dominant	Errors
Above	9%	**91.0%**
Below	**94.4%**	5.6%

Table 2. Detection rates above and below the separating line built in the (F0*,σ) plane, in the double-vowel task.

A closer look at the distributions of points in Figure 2 reveals two interesting points. First, what we called "Errors" below the line are mainly due to our 5Hz-resolution absolute criterion; if we replace it by a criterion relative to the frequency (e.g. 5% of the

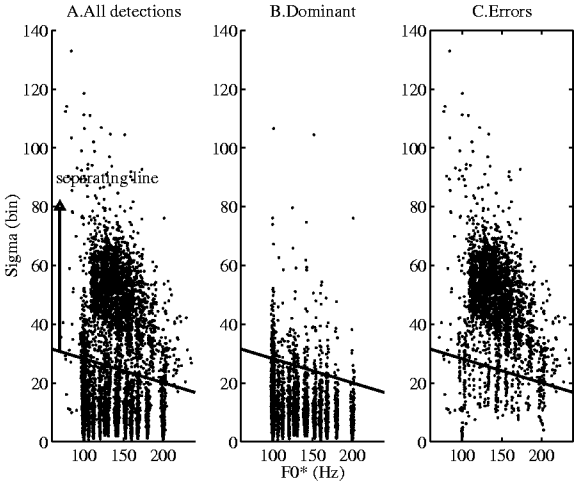

Figure 2. ZC detection in the double-vowel task, represented in the (F0*,σ) plane.

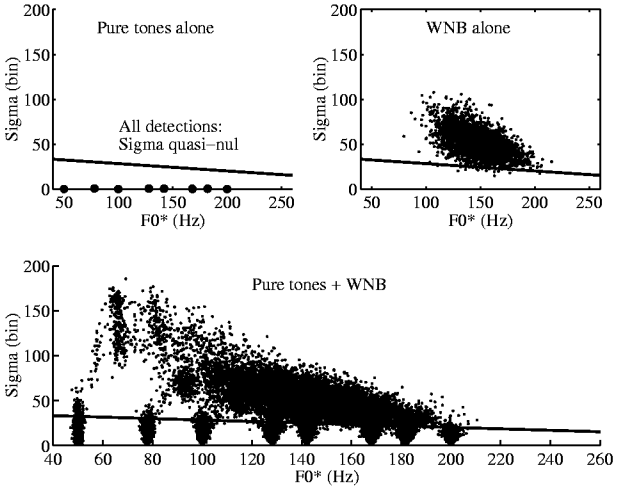

Figure 3. ZC detection with pure tones in noisy conditions, represented in the (F0*,σ) plane.

frequency), we obtain 99% correct estimations of the frequency of the tone below the boarder. Second, when we look at the distribution of $RL_{i=1..32}$ above and below the line, we note that points localized below the line mainly correspond to high positive or high negative values of $RL_{i=1..32}$ (one vowel is dominant in energy), and that points localized above the line mainly correspond to values of $RL_{i=1..32}$ around zero.

This shows that we can define a criterion based on the standard deviation of the estimation: When the (F0*,σ) point is localized below the separating line, we can conclude, with a low probability of errors, the presence of a dominant periodic source, and then identify its F0. When the (F0*,σ) point is localized above the line, the presence of a dominant periodicity is quite unlikely.

This criterion has been defined in the double-vowel paradigm. We will now test its robustness in other interference paradigms.

4.2. Application of the confidence criterion to other interference paradigms.

4.2.1. Pure tone in acoustic noise.

We have mixed a 50-200Hz White Noise Band (WNB) with eight pure tones with frequencies in the 50-200Hz band, at different SNR (difference in dB between the energy of the pure tone and the energy of the WNB) between -15dB and 15dB. For each pure tone, we make 100 estimations with different 50-200Hz WNB. Each estimation is plotted in the (F0*,σ) plane in Figure 3.

We have superimposed on the representation the separating line defined from the double-vowel paradigm. This line seems to realize a good separation of the two clusters obtained in this new experiment. To confirm this observation, we have calculated the rate of detection of the "Pure tone" (detection of the frequency of the pure tone, within a 5Hz resolution) and the rate of "Errors" (other detections) above and below the line. These rates are given in Table 3. As in section 4.1., a look at the distributions of points above and below the line reveals that "Errors" below the line are mainly due to the choice of 5Hz resolution, and that, according to a 5% relative criterion, 98% of the detections are correct. We also note that the points localized below the line mainly correspond to high values of SNR (pure tone dominant).

	Pure tone	Errors
Above	5.7%	**94.3%**
Below	**95.6%**	4.4%

Table 3. Detection rate above and below the separating line in the (F0*,σ) plane, with pure tones in acoustic noise.

4.2.2. Vowels in noisy conditions

In this second experiment, we have mixed a white noise (WN) with six French vowels [a, e, i, o, u, y], at eight F0 values between 100Hz and 200Hz. These mixtures are presented to the detection system described in section 3.1. Taking into account that we exclude the nine first channels of the filterbank, we built the mixture with equal *rms level* between the noise and the vowel present in the mixture at the output of the filterbank inside the channels 10 to 32. For each couple (vowel, F0), we made 10 estimations with 10 different WN. The results in the (F0*,σ) plane, together with the separating line defined in section 4.1 are displayed in Figure 4.

Like in the last case, the separating line seems to provide a good separation between errors and detections of the F0 of the vowel present in the mixture. The same rates as in section 4.2.1. are calculated above and below the line: "Vowels" corresponds to the detection rate of the F0 of the vowels present in the mixtures, "Errors" to other detections. The results are shown in Table 4.

A look at the distributions of points above and below the line shows that, with a 5% relative criteria, 98% of the detections are correct; moreover, the detections of the vowels below the line correspond to high positive SNR values, hence in the vowels formants areas, and Errors above the line to negative SNR values (noise dominant).

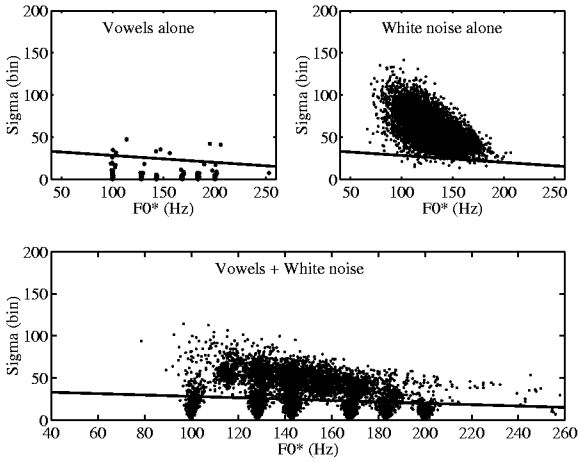

Figure 4. ZC detection with vowels in noisy conditions, represented in the (F0*,σ) plane.

	Vowels	Errors
Above	9.9%	**90.1%**
Below	**90.0%**	10%

Table 4. Detection rate above and below the separating line in the (F0*,σ) plane, with vowels in noisy conditions.

4.3. Synthesis of the two experiments

The two experiments in section 4.2. show a striking correspondence with the boarder line defined in section 4.1. Altogether, in the three experiments, namely the double-vowel paradigm, the pure tone + noise and the vowel + noise paradigms, the separating line defined in section 4.1. allows to detect a dominant periodicity in a mixture, and to estimate the F0 of the dominant source.

The interest of this periodicity detection method is that it has similar characteristics in presence of interferences produced by noise or other signals. In 98% to 99% of the cases, the F0* value below the boarder line corresponds to the F0 value of the dominant source with a relative error less than 5%. Errors above the boarder line mainly come from the cases where there is no real dominance between the components of the mixture.

5. CONCLUSION

These experiments show that, according to a confidence measure based on the standard deviation of the inter-zero intervals, the ZC method allows periodicity detection with small errors, and with no assumptions about the nature of interference.

There exist a number of methods of separation of concurrent sources based on pitch estimation, usually performed in the time domain, by calculating the Autocorrelation Function (ACF) [3] [4], or the Average Magnitude Difference Function (AMDF) [5]. This paper can be placed in the same global strategy, but the temporal method we use highly restricts the information carried by the temporal signal, by focussing on zero-crossing points.

In this paper, we show that this sampling is sufficient for periodicity detection, because of the temporal redundancy of speech signals, and leads to interesting selective dominance effects.

However, even if the efficiency of the modified ZC method for the identification of interfering vowels remains to be proved, a first step could be to identify the two fundamental frequencies across medium and high frequency channels, according to the groups of points localized below the boarder line on Figures 2,3,4. Note that this identification doesn't need any exhaustive representation, like autocorrelogramms [3] [4] or AM-maps [6]. We have also observed in the double-vowel paradigm a rather unefficient behavior of the ZC method after demodulation in the low frequency channels, because harmonics are resolved in these channels. Then, a first improvement for vowels separation will be to find a strategy that deals with these channels, to perform the F0 identification on the 32 channels (instead of only 23 at present).

At last, the redundancy in the frequency domain and the fact that features are statistically detectable in a time-frequency domain even if there are interferences, because of the strong dominance effect, would allow to group channels at the output of a filterbank according to the identified F0. Each group of channels could then feed a partial recognition process [7] to achieve the separation and the identification of each vowel in the scene.

REFERENCES

[1] BREGMAN A.S. (1990), *Auditory Scene Analysis,* MIT Press, London.

[2] PATTERSON R.D. et al. (1992), in *« The auditory processing of speech »*, Schouten, M.E.H. (Ed.), Mouton de Gruyter, Utrecht, 67-83.

[3] MEDDIS R., HEWITT M. (1992), « Modelling the identification of concurrent vowels with different frequencies », *Journal of the Acoustical Society of America, **91**, 233-245.*

[4] ASSMANN P.F., SUMMERFIELD Q. (1990), « Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies », *Journal of the Acoustical Society of America, **88**, 680-697.*

[5] DE CHEVEIGNE A. (1993), « Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing », *Journal of the Acoustical Society of America, **93**, 3271-3290.*

[6] BERTHOMMIER F., MEYER G. (1995), « Source separation by a functional model of amplitude demodulation », *Proc. Eurospeech Madrid, 135-138.*

[7] COOKE M., MORRIS A., GREEN P. (1996), « Recognizing occluded speech », *Proc. of Workshop on the Auditory basis of speech perception, Keele, 297-300.*