AUDITORY MODELING TECHNIQUES FOR ROBUST PITCH EXTRACTION AND NOISE REDUCTION

Piero Cosi, Stefano Pasquin** and Enrico Zovato***

*Institute of Phonetics – C.N.R.
Via G. Anghinoni, 10 - 35121 Padova (ITALY)
e-mail: cosi@csrf.pd.cnr.it www: http://www.csrf.pd.cnr.it

**University of Padova, Department of Electronic Engineering and Computer Science
Via G. Gradenigo, 6 - 35100 Padova (ITALY)

ABSTRACT

A novel method for robust pitch[1] extraction, based on the correlogram output of the Lyon's cochlear model is described. The value of the autocorrelation lag for which the signals of the cochlear channels have the same periodicity can be computed thus tracking how the pitch of the input signal varies in the time domain. In the case of a stationary noise, a sort of '*spectral-subtraction*' technique, built in the correlogram domain named '*correlogram subtraction*', is applied to enhance the signal before computing its fundamental frequency. Finally, a correction algorithm based on an '*island driven*' strategy, working on particular zones of the signal with stable pitch values, is used to refine the pitch estimate. This method of pitch extraction is extremely reliable, even in the case of a signal to noise ratio of 0dB. The same subtraction technique, with some new specific filter-bank energy-based modifications, is considered to re-synthesize, by an inversion strategy, a clean version of an input noisy signal. The quality of the re-synthesized signal is quite promising, leading us to try, in the future, to use this technique as a new signal enhancement scheme.

1. INTRODUCTION

Pitch is a fundamental parameter that conveys most of the prosodic information contained in the speech signal. As such, pitch determination plays a central role in many speech-related fields. Despite the multitude of pitch determination algorithms proposed, an algorithm that is absolutely reliable remains to be found, especially in highly degraded signal conditions. Among the many different methods, one of the oldest and most celebrated is based on short-time autocorrelation (STA) analysis [1]. The method proposed in this paper is again based on STA analysis, but this is now applied to the outputs of an auditory model known as the Lyon's cochlear model [2]. This new algorithm is strongly inspired by the Licklider's duplex theory of pitch perception [3] subsequently adopted by M. Slaney and R.F Lyon in their various works on auditory modeling summarized in [2]. In practice, Licklider originally proposed the '*correlogram*' as a pitch model, but the high computational cost delayed its practical implementation. Pitch is, in fact, an obvious quantity to measure with the correlogram that is a three-dimension function of time, frequency, and periodicity or autocorrelation time delay, or lag, shown on the horizontal axis. If a signal is periodic the value of the autocorrelation lag for which the signals of the cochlear channels have the same periodicity can be computed thus tracking how the pitch of the input signal varies in the time domain. The proposed algorithm follows entirely this idea but before computing the pitch, a specific enhancement technique is applied in the correlogram domain named '*correlogram subtraction*' in order to improve the accuracy of the computation in low SNR situations. The same enhancement technique is then applied together with the above pitch extraction algorithm to a new re-synthesizing procedure, driven by an inversion strategy [4], [5], in order to clean speech signals highly degraded by stationary-like noise.

2. AUDITORY FRONT-END

The front-end processing groups together the Lyon's cochlear model and the correlogram building procedure.

2.1. The Lyon's cochlear Model

Following the work of M. Slaney and R.F. Lyon [2], the Lyon's cochlear model was adopted as a front-end. The outline of the Lyon's cochlear model is described in Figure 1, where the three main characteristic blocks of the model are indicated. The first filtering stage models, by a broadly tuned cascade of low-pass filters, the propagation of energy as waves on the basilar membrane (BM). The second detection stage non-linearly (Half Wave Rectification) converts BM velocity into a representation of Inner Hair Cells (IHC) receptor potential or auditory nerve (AN) firing rate. Finally the third compression stage continuously adapts, by an automatic gain control (AGC), the operating point of the system in response to its level of activity. This stage compress widely varying sound input levels into a limited dynamic range of BM motion, IHC receptor potential, and AN firing rate. In summary, the Lyon's cochlear model, instead of trying to replicate the detailed time structured firing events of each IHCs or AN fibers, aims to globally model the neural firing rate as a function of cochlear place or best frequency (BF) or characteristic frequency (CF) versus time.

[1] 'pitch' and f0 (fundamental frequency) are considered synonyms throughout this paper despite their psychoacoustical and physical different meaning.

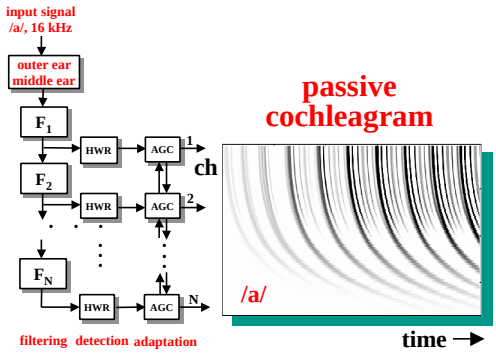

Figure 1: Outline of the Lyon's cochlear model. The passive cocheagram refers to the Italian vowel /a/.

2.2. The Correlogram

Each of the outputs of the Lyon's cochlear model is short-time windowed, by an L-length window, and autocorrelated thus obtaining as many STAs as the number of channels. The correlogram consist of the ensemble of these STAs. In the present implementation a modified Hamming window [4] is utilized to reduce the classical frame border effect and moreover to exploit some computational advantages incorporated in the inversion procedure outlined in the following 'Speech Enhancement' paragraph. For each channel i, of the Lyon's cochleagram, the following STA is computed:

$$r_{x_i}(m,t)=\frac{1}{L}\sum_{n=0}^{L-|t|-1} x_{w_i}(n,m)\,x_{w_i}(n-|t|,m) \qquad (1)$$

where: i=1,…,N channel index
(N=86 in the present implementation)
L window length in samples
τ autocorrelation lag

and $x_{w_i}(n,m) = x_i(n)\cdot w(m-n)$.

Given that the autocorrelation itself is a function of a third variable, the autocorrelation lag, the resulting correlogram is a three-dimensional function of frequency or number of channel of the cochlear filter bank, time, and autocorrelation lag or delay. The correlogram referring to the Italian vowel /a/ sampled at 16 kHz computed with L=512 samples (t=32ms), at a frame shift of S=128 samples (t=8 ms), thus $C_i(mS,t)=r_{x_i}(mS,t)$ is illustrated in Figure 2.

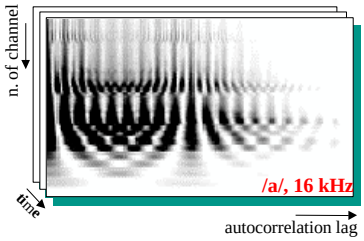

Figure 2: Correlogram relative to the Italian vowel /a/.

3. PITCH EXTRACTION

As illustrated in Figure 2, the correlogram is not a static representation but a dynamic one. Given that the periodicity information remains unchanged along all the transformations occurring in the Lyon's cochlear channels, if a sound is periodic, the autocorrelation functions of a big number channels shows a clear peak at the position t_p corresponding to the period of the input signal. As stated in the original implementation of the algorithm [2], if a sum is executed with the values of the correlogram *cor*(*t*,*i*) along all the channels a function called *summary correlogram SC*(*t*) is obtained

$$SC(t)=\sum_{i=1}^{N} cor(t,i) \qquad (2)$$

where the peak corresponding to the period of the input signal, related to the pitch frequency by the relation $f_p = 1/t_p$ results enhanced. In practice, given that in the computation of the short-time autocorrelation, the input signal is multiplied by the modified Hamming window *w(n)* [4] thus reducing the classical frame border effect, all the STAs considered in the computation of SC(t) are normalized by the term

$$r'_x(m,t)=\frac{r_x(m,t)}{r_w(t)} \qquad (3)$$

The Summary Correlogram is thus modified as illustrated in Figure 3 relative to the Italian vowel /i/.

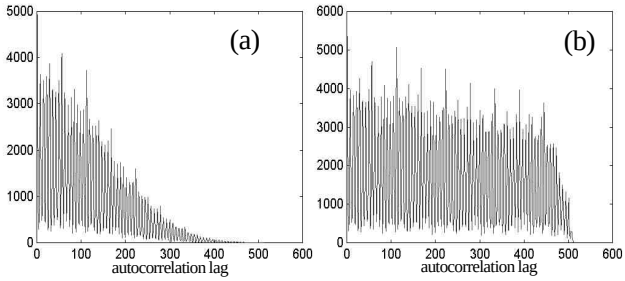

Figure 3: Summary Correlogram for the Italian vowel /i/ without (a) and with (b) the application of the correction.

When the signal is highly degraded by noise the above procedure has to be modified in order to become more robust. In the case of a stationary noise, a sort of *spectral-subtraction* technique [6] in the correlogram domain named '*correlogram subtraction*' is proposed. In fact, a mean correlogram of the noise, computed by

$$rif_{\text{noise}}(t,i)=\frac{1}{nframes}\sum_{m=1}^{nframes} cor_{\text{noise}}(m,t,i) \qquad (4)$$

where: m frame index
τ autocorrelation lag
i channel index

is subtracted to the correlogram of the signal plus the noise. The main hypothesis for this procedure to be applicable is that there must be some parts of the signal where only the noise is present; in this way its statistics can be computed. This is obviously a quite common situation in the case of the speech signal where silence is frequently inter-mixed within speech. The above procedure is shown to result quite effective despite the fact that the non-linearity transformations implied in the auditory front-end should not justify the super-imposition effect which is instead explicitly accepted considering the subtraction

operation. In order to justify this assumption the new quantities computed by the sum along the time lag of cor(t,i) are defined:

$$S(i) = \sum_t cor(t,i) \qquad (i=1,\ldots,N) \tag{5}$$

These quantities represent the '*degree of excitation*' of the input signal x(n) for each channel and for each frame of analysis, and are, even if improperly, indicated as 'energies' because they are strictly related to the true signal energies X_i. In fact, given that the energies of x(n) for each channel of the filter-bank are related to their Fourier transform and to the filter-bank responses by the following relation:

$$X_i = \int_{-p}^{p} |F_i(w)|^2 |X(w)|^2 \frac{dw}{2p} \tag{6}$$

where: X(ω) Fourier transform of X(n)
Fi(ω) filter bank answers
i channel index

with the hypothesis that the following HWR stages almost halves each of these *energies* and that the following non-linear stages behave uniformly across all the channels, it is possible to hypothesize that

$$S_i \cong a X_i \qquad (i=1,\ldots,N) \tag{7}$$

with α proportionality constant due to the uniform transformations of non-linear stages on each channels. The white noise case constitutes an experimental proof of this assumption. In this case, in fact (6) becomes:

$$N(i) = N_0 \int_{-p}^{p} |F_i(w)|^2 \frac{dw}{2p} = N_0 E(i) \tag{8}$$

where: N_0 power spectral density of input noise n(k)
E(i) Energy of the impulse response of the analysis filters

$$E(i) = \int_{-p}^{p} |F_i(w)|^2 \frac{dw}{2p}$$

Considering (4), and computing the sum on the time lag, it can be noted (Figure 4), that $\hat{N}(i) = \sum_t rif_{noise}(t,i) \approx a\, N(i) = a\, N_0 E(i)$

Another experimental finding is that considering the sums on the time lag (5), the superimposition effect is approximately applicable thus leading to write: $SN_i \cong S_i + N_i$. In the case of a stationary noise, this property suggests the idea to clean the correlogram by a '*spectral subtraction*' like strategy [6] such as $\hat{S}(i) = SN(i) - N(i)$ named '*correlogram subtraction*'. The results of the application of the correlogram subtraction on the sum of the time lag (5) to an Italian noisy vowel /e/ is graphically illustrated in Figure 5.

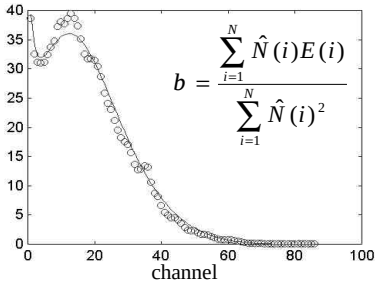

Figure 4: White noise '*energies*' *E(i)* (continuous line) and computed sum on the time lag $\hat{N}(i)$ (circle line). These quantities are multiplied by the b constant that minimizes the quantity $\sum_i \left[E(i) - b\,\hat{N}(i)\right]^2$.

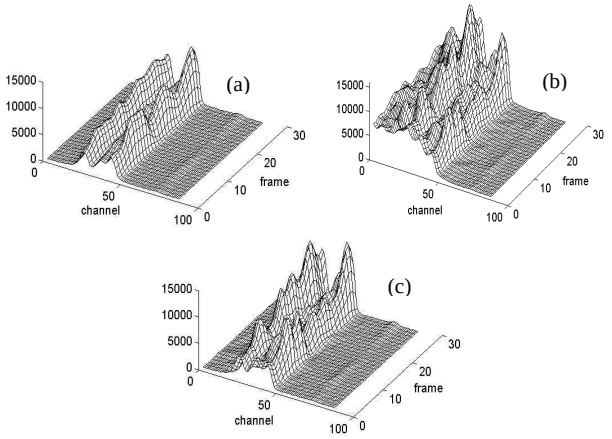

Figure 5: Sums on the time lag for the Italian vowel /e/. (a) refers to the clean case, (b) refers to the nosy case and (c) refers to the subtraction case.

The statistic of the noise computed in (4) is the arithmetic mean of the noise correlogram frames thus it constitutes only a first approximation of the true noise correlogram frame by frame, but if the noise is stationary, this constitutes a good estimate of the noise statistics. The sums on the time lag in (4), given by $\hat{N}(i) = \sum_t rif\_noise(t,i)$, are the estimate of the '*energies*' of the noise frame by frame, thus the correlogram subtraction can be directly derived by the new formula:

$$\hat{S}_c(i) = \max(0, SN(i) - c\hat{N}(i)) + 10^{-3} \tag{9}$$

where: c = 1 or 2

where the estimated quantities $\hat{N}(i)$ substitute the true noise sums on the time lag frame by frame *N(i)*. This form forces the estimated energies to be positive overcoming the problems given by the fluctuations of *N(i)* relative to $\hat{N}(i)$ and by the fact that the application of the superimposition effect to the '*energies*' of signals which pass through various non linear transformations is quite inadequate. The case c=2 grants a better cleaning but has the disadvantage to eliminate also part of the channel where *S(i)* is significant. In the case of a noisy signal the $\hat{S}(i)$ quantities lead to discriminate the channels more degraded from those less degraded by noise. As a result of that, a new version of the Summary Correlogram where the less degraded channels have a bigger influence on the final computation can be defined:

$$SC(t) = \frac{SC_w(t)}{r_w(t)} \quad \text{with} \quad SC_w(t) = \sum_{i=1}^{N} C(i) \frac{cor(t,i)}{\sum_t cor(t,i)} \tag{10}$$

where: $C(i) = \frac{\hat{S}(i)}{E(i)}$

In summary, in Figure 6, the complete pitch extraction algorithm is illustrated. An '*island driven*' error correction procedure is also included. First it searches for I_k stable pitch region, where stable means regions with at least 8 consecutive values of pitch whose differences do not exceed the $\Delta f_p \le 30Hz$ range, then corrects isolated gross errors within those regions, and, finally, joins together those regions by a continuity procedure working frame by frame with the SC and looking for the correct peaks. An example of the application of the proposed algorithm to a very noisy signal (0db SNR) is given in Figure 7

where the improvement of the extracted pitch curve is quite evident.

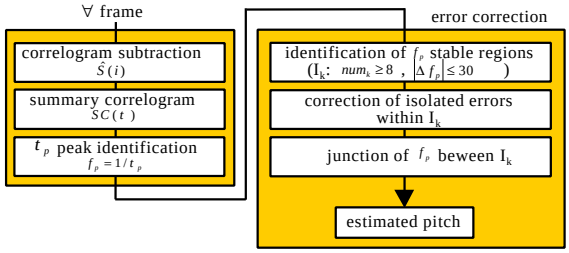

Figure 6: Outline of the pitch extraction algorithm.

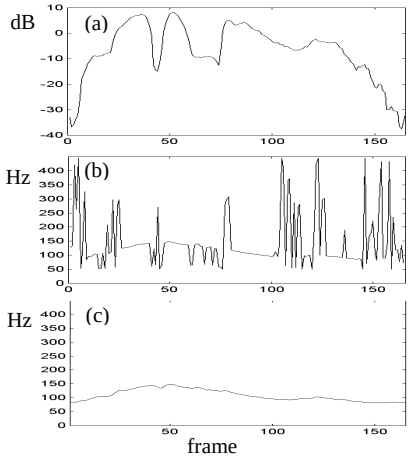

Figure 7: Referring to the the Italian noisy voiced word /lavan'daja/ ('washerwoman'): in plot (a) the frame-by-frame signal to noise ratio corresponding to a global 0dB SNR is indicated; in (b) the pitch estimate from the original Summary Correlogram [2] is given, while in (c) the final estimated pitch computed with the proposed algorithm is illustrated.

4. SIGNAL ENHANCEMENT

The same '*correlogram subtraction*' technique is then applied to enhance noisy speech signal. Before applying the inversion strategy procedure described in [4-5], a clean correlogram *corSNC* is reconstructed by the following formula operating on each frame *m* and on each channel *i*.

$$corSNC(m,t,i) = \hat{S}(m.i)\frac{corSN(m,t,i)}{\sum_t corSN(m,t,i)} \qquad \forall i, \forall m \quad (11)$$

Using (11) the following condition is satisfied:

$$\sum_t corSNC(m,t,i) = \hat{S}(m,i) \qquad (12)$$

therefore noisy channels are strongly attenuated. The clean correlogram is then forced to belong to a true periodic signal with period $t_p = 1/f_p$ corresponding to the previously computed pitch. This is obtained working in the correlogram domain, for each channel, by the use of an anti-aliasing window $F_p(t)$ isolating the period $[-t_p/2, t_p/2]$ and periodically repeating the signal $x_F(t) = F_p(t)\,x(t)$ (extended for parity to negative values for $t \in [-t_{MAX}, t_{MAX}]$) with period t_p obtaining:

$$x_p(t) = [F_p(t)x(t)] * \left[\frac{1}{2n+1}\sum_{i=-n}^{n} d(t - i t_p)\right] \text{ with } n = \left\lfloor \frac{t_{MAX}}{t_p} \right\rfloor \quad (13)$$

Finally the window utilized in (3) is multiplied for each channel and the obtained correlogram is passed to the inversion procedure described in [4-5]. It should be noted that, differently from the *spectral subtraction* technique described in [6], this enhancement procedure works only at the '*energy*' level leaving to the inversion procedure the task of reconstructing all the information on the phases. Therefore the '*musical-noise*'-related problems are inherently avoided and the computation is quite straightforward. An example for the noisy Italian word /lavan'daja/ is given in Figure 8.

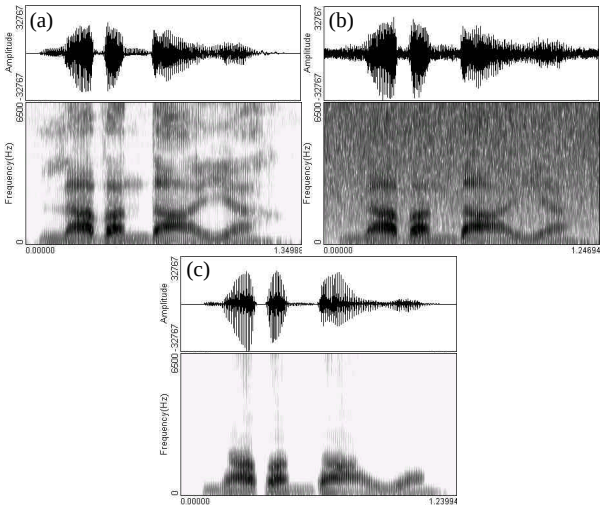

Figure 8: Application of the '*correlogram subtraction*' enhancement technique to the Italian noisy (10dB SNR) word /lavanda'ja/ ('washerwoman'): (a) clean [SOUND 1053_01.wav], (b) noisy [SOUND 1053_02.wav] and (c) enhanced [SOUND 1053_03.wav] .

5. CONCLUSIONS

A new pitch extraction algorithm and a new enhancement procedure for noisy speech signal have been presented. Very promising results have been so far obtained thus leading to hypothesize the future use of these techniques in new front-ends for robust speech applications.

4. REFERENCES

1. L.R. Rabiner, "On the use of autocorrelation analysis for pitch detection", *Proc. IEEE*, Vol. 58, 1970, pp.707-712.
2. M. Slaney and R.F. Lyon, "On the importance of time – a temporal representation of sound", in *Visual Representation of Speech Signals*, M. Cooke, S. Beet and M. Crawford (eds.), John Wiley & Sons Ltd, 1993, pp. 95-116.
3. J.C.R. Licklider, 'A duplex theory of pitch perception', Experentia 7, 128-133. Also reprinted in Pysiological Acoustic (E.D. Shubert ed.), Dowden, Hutchinson and Ross Inc. Stroudsburd, PA, 1979.
4. P. Cosi and E. Zovato, "Lyon's Auditory Model Inversion: a Tool for Sound Separation and Speech Enhancement", *Proc. of ESCA Workshop on 'The Auditory Basis of Speech Perceprion*, Keele University, Keele (UK), 15-19 July, 1996, pp.194-197.
5. E. Zovato, "Sintesi di suoni mediante inversione di un modello uditivo", Dipartimento di Elettronica e Informatica , Università di Padova, Tesi di Laurea, A.A. 1994-95.
6. S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *Trans. Acoust., Speech and Signal Processing.* vol. ASSP- 27, Apr. 1979.