Proceedings of the Institute of Acoustics

CORRELOGRAMS AND AUDITORY IMAGES.

M. Allerhand, R. Patterson.

MRC Applied Psychology Unit,
15 Chaucer Road, Cambridge CB2 2EF.

1. INTRODUCTION.

The correlogram is a multi-channel autocorrelation analysis originally introduced in Licklider's (1951) Duplex model to explain human pitch perception of complex sounds. Several types of correlogram are currently used, and Slaney and Lyon's (1992) is possibly closest to Licklider's original Duplex model. Their correlogram is a time-varying two-dimensional array usually plotted as frequency against autocorrelation lag. The autocorrelation analysis adds a third dimension to the conventional time-frequency analysis, and the correlogram can be viewed as a time-evolving cartoon which is richer in information than the traditional speech spectrogram. Patterson and Holdsworth's (1992) auditory image is functionally very similar, but with some computational differences. The auditory image model has the same general architecture as the Duplex model of pitch in the sense that it is composed of a functional model of the cochlea and a multi-channel neural post-processor. However this model was originally conceived as a method of neural temporal integration to stabilise the timbre information of periodic sounds. Instead of the explicit autocorrelation used in the Duplex model, the stabilised auditory image is constructed in a manner analogous to oscilloscope images by periodically refreshing a decaying image buffer, using a data driven trigger mechanism. Nevertheless the auditory images of complex sounds appear to be very like correlograms. Allerhand (1990) suggested that this was due to an underlying similarity; that the construction of the auditory image was closely related to a recursive implementation of the running autocorrelation used to construct the correlogram.

Licklider (1951) proposed an analog computer to generate correlograms which consisted of a series of integrators linked to taps in a delay line by multiplier units. The signal is multiplied by a delayed version of itself to yield lagged products, and these are averaged by a bank of low-pass filters, one for each autocorrelation coefficient. The process was a multi-channel version of the running autocorrelation introduced by Fano (1950) and Stevens (1950). This is a short-time autocorrelation function in which the weighting function to window the data is provided by the impulse response of a low-pass filter, and the integration is performed as the convolution. Current methods for constructing correlograms (Slaney and Lyon, 1992; Meddis and Hewitt, 1991) are digital, using the FFT to generate autocorrelation coefficients. However it turns out that the running autocorrelation can be generated by a recursive process which is significantly more efficient than conventional block methods for computing a running autocorrelation. It also turns out that the construction of Patterson and Holdsworth's (1992) auditory image is closely

related to this recursive running autocorrelation. This establishes the relationship between the auditory image model (Patterson and Holdsworth, 1992) and the correlogram model (Slaney and Lyon, 1992), which is based upon the original Duplex model.

2. A RECURSIVE RUNNING AUTOCORRELATION.

The autocorrelation function for a particular lag is an average of lagged products of a signal over all time. The average can be computed by an analogue first–order low–pass filter. This simple resistor–capacitor network attenuates higher frequencies, so that the output emphasises the low frequencies, and in particular the d.c. level, which is the average value of a stationary signal. The filter is a simple linear system which is characterised by its impulse response; the output at any time is given by the convolution of the input with the system impulse response. When the input to the system is the lagged products, then the output is the average lagged product, which is the autocorrelation. Writing the convolution integral with the input as lagged products derives the running autocorrelation function (1). The exponential impulse response of the first-order low–pass filter can be seen as a weighting function to window the lagged products into the past. Below we derive a discrete approximation of this analogue model using a recursive digital low–pass filter to compute the running autocorrelation.

Licklider's (1951) "Duplex model of pitch perception" is a multi–channel autocorrelation which used the running autocorrelation introduced by Fano (1950) and Stevens (1950). The running autocorrelation function (ACF) is a function of time as well as lag, and can be defined (eg see Fano, 1950; Schroeder and Atal, 1962):

$$\phi_t(\tau) = \int_{-\infty}^{t} h(t-u).f(u).f(u-\tau)\,du \qquad (1)$$

(where u is a dummy variable). This is the ACF of windowed data, where $h(t)$ is a physically realizable weighting function, $h(t) = 0$ for all $t < 0$, which provides the short-time window over the data. When (1) is seen as a convolution integral, then $\phi_t(\tau)$ is the zero-state response of a linear time-invariant filter with impulse response $h(t)$ and input $f(t).f(t-\tau)$.

We define the running ACF of a discrete signal (to correspond with (1)) as:

$$\psi_n[m] = \sum_{k=-\infty}^{n} h[n-k].f[k].f[k-m] \qquad (2)$$

where $h[n] = 0$ for all $n < 0$. This can be seen as a discrete convolution, so that $\psi_n[m]$ is the zero-state response of a digital filter with impulse response $h[n]$ and input $f[n].f[n-m]$.

As Fano (1950) indicated, an appropriate filter to perform the averaging required by a

running autocorrelation process is the first-order low-pass ("leaky integrator") filter. For convenience and simplified notation we describe one filter with input $x(t) = f(t).f(t-\tau)$ and output $y(t) = \phi_t(\tau)$, and with input-output relationship of the form:

$$T\frac{dy(t)}{dt} + y(t) = x(t) \tag{3}$$

where $T = RC$, (R=resistance, C=capacitance), is the system time constant. The discrete counterpart of the continuous filter (3) is developed as follows. The complete response of the system (3) evolving from time t_0 is (Takahashi et al, 1970, pp12–18):

$$y(t) = y(t_0)e^{-(t-t_0)/T} + \frac{1}{T}\int_{t_0}^{t} e^{-(t-u)/T}x(u)\,du \quad , \quad t \geq t_0 \tag{4}$$

where the first term is the zero-input response which evolves from an initial condition $y(t_0)$, and the second term is the zero-state response to the input $x(t)$. This second term is the convolution of the input, a particular lagged product $x(t) = f(t).f(t-\tau)$, with the impulse response function. The impulse response of the system (3) is $h(t) = \frac{1}{T}e^{-t/T}$, and its form provides an exponential window folded to weight into the past. If the initial condition is $y(t_0) = 0$, and we let $t_0 \to -\infty$, then the response of the system (4) takes the form required to calculate the running ACF (1).

The complete response (4) is generalized for any instant t_0, and so we will apply the complete solution over a single sample interval of T_s seconds from $t_0 = (n-1)T_s$ to $t = nT_s$, so that $t - t_0 = T_s$. Note that continuous $y(t) = y(nT_s)$ corresponds with discrete $y[n]$, and continuous $y(t_0) = y((n-1)T_s)$ corresponds with discrete $y[n-1]$. Substituting we have:

$$y[n] = y[n-1].e^{-T_s/T} + \frac{1}{T}\int_{(n-1)T_s}^{nT_s} e^{(nT_s-u)/T}x[n]\,du \tag{5}$$

Assuming that $x(t)$ is a constant from $x(t_0) = x((n-1)T_s)$ to $x(t) = x(nT_s)$, (a staircase approximation), then $x[n]$ is a constant. The result of integration is then:

$$y[n] = a.y[n-1] + (1-a).x[n] \tag{6a}$$

$$\text{where :} \quad a = e^{-T_s/T} \tag{6b}$$

This difference equation is a discrete approximation of the differential equation (3), and its complete response evolving from time index n_0 is:

$$y[n] = y[n_0].a^{n-n_0} \quad + \quad (1-a)\sum_{k=n_0+1}^{n} a^{n-k}x[k] \quad , \quad n \geq n_0 \tag{7}$$

where, (corresponding with (4)), the first term is the zero-input response which evolves from an initial condition $y[n_0]$, and the second term is the zero-state response to the input $x[n]$. This second term is the discrete convolution of the input, a particular lagged product $x[n] = f(nT_s).f((n-m)T_s) = f[n].f[n-m]$, with the impulse response function. The impulse response of the system (6) is $h[n] = (1-a)a^n$, and again its form provides an exponential window folded to weight into the past. If the initial condition is $y[n_0] = 0$, and we let $n_0 \to -\infty$, then the response of the system (7) takes the form required to calculate the running ACF of a discrete signal (2).

Figure 1 shows vowel correlograms generated using the recursive algorithm (6). To set the system time constant, T, we suggest two criteria. The first is to set $T > 1/\omega$, (where ω is the angular frequency of the input, describing the longest period over which the system may be required to integrate), since the filter begins to operate as an approximate integrator at the transition from passband to stopband, when $\omega T = 1$. The second is to set the length of the exponential window (the fading memory of the system) so that it decays to 0.01 (by 99%) of its initial value in a given period. Considering the unit step response, and the definition of the decay constant (6b), this period is $-T\log_e(0.01) = 4.6T$. As an example, consider a maximum integration interval of 20ms, (roughly the longest period over which the auditory system detects tones rather than resolving them into buzzes or clicks). The first criterion has $T > 20/2\pi \approx 3$ms, and the second has $T > 20/4.6 \approx 4$ms.

3. CONSTRUCTION OF AUDITORY IMAGES.

The stabilised auditory image (Patterson et. al., 1991) is constructed in a manner analogous to oscilloscope images by periodically refreshing a decaying "image buffer", using a data driven trigger mechanism. The n'th sample of the signal is aligned with a "trigger point", which is a datum at the zeroth lag. At each discrete time step, $n = 1, 2, \ldots$, values of the aligned signal are added into the image buffer. At the same time the contents of the buffer are also proportionally scaled down by a decay factor, a. The value added at any point along the buffer is a function of two values of the signal: the value aligned with the buffer at that point and the value aligned at the trigger point. When the function used is a product, then the image consists of a sum of lagged products, constructed as follows:

$$\begin{aligned}
\psi_n[0] &= f[n].f[n] + a.f[n-1].f[n-1] + a^2.f[n-2].f[n-2] + \cdots \\
\psi_n[1] &= f[n-1].f[n] + a.f[n-2].f[n-1] + a^2.f[n-3].f[n-2] + \cdots \\
&\cdots \\
\psi_n[m] &= f[n-m].f[n] + a.f[n-1-m].f[n-1] + a^2.f[n-2-m].f[n-2] + \cdots \\
&= \sum_{k=1}^{n} a^{n-k}.f[k].f[k-m] \qquad (8)
\end{aligned}$$

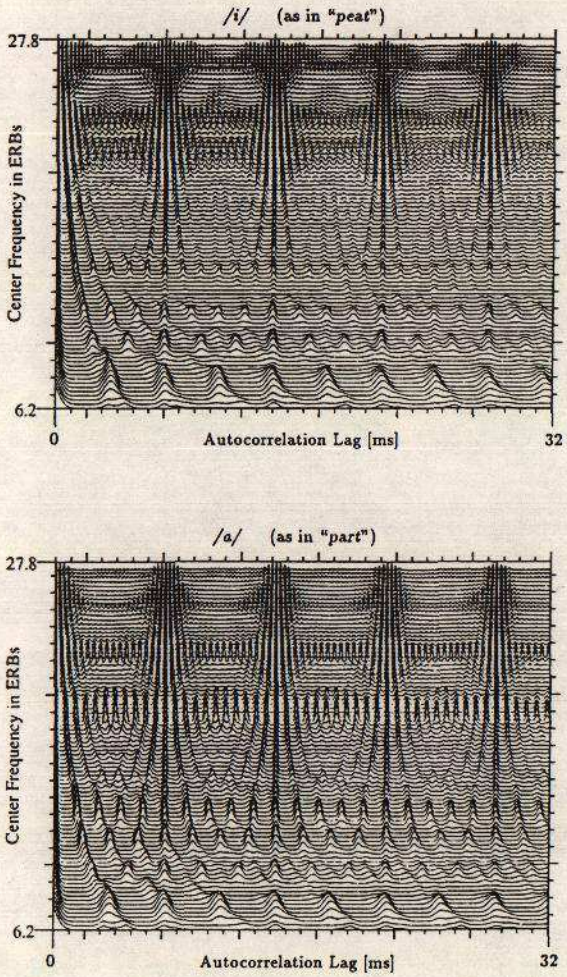

Figure 1. Correlograms of vowels.

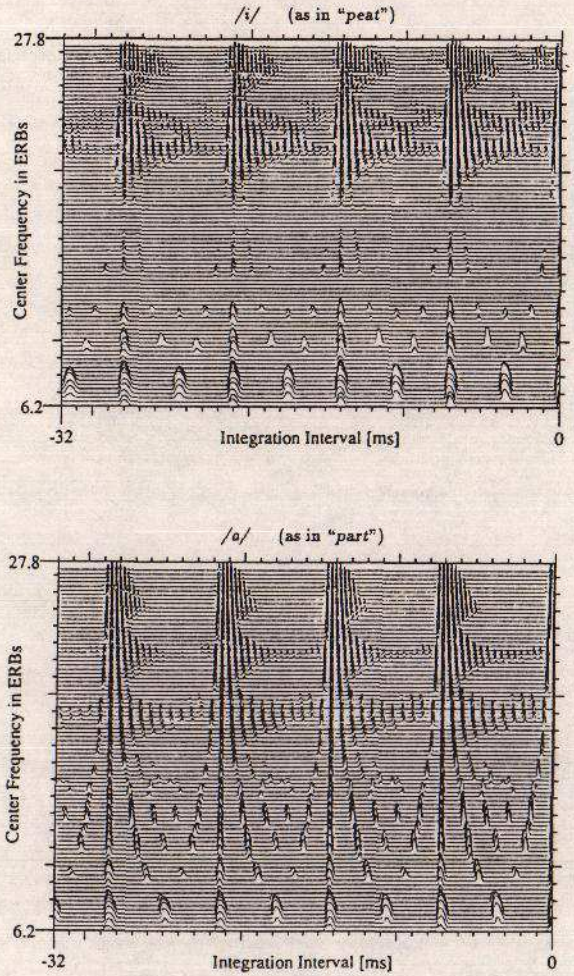

Figure 2. Auditory Images of vowels.

The recursive difference equation which generates this exponentially weighted sum of lagged products (8) for each lag is:

$$y[n] = a.y[n-1] + x[n] \tag{9}$$

where the input $x[n] = f[n].f[n-m]$.

Comparing (8) with (7) it can be seen that the zero-state response of the above construction differs from that of the discrete running autocorrelation by a constant scale factor of $1-a$. In addition, comparing the corresponding difference equations (9) and (6) confirms the underlying similarity between the auditory image and the correlogram model. Both systems compute a short–time autocorrelation in the form of a running average of lagged products, exponentially weighted into the past. The system (6) is based upon an exponentially weighted mean, and the system (9) is based upon an exponentially weighted sum.

The auditory image originally described by Patterson et. al. (1991) is a "sparse" method based upon the above construction. The main difference is that lagged products are not added into the image buffer at every discrete time step, but only when triggered by peaks in the signal. The peak-picker adapts to local periodicity and is designed to trigger the addition process once per fundamental period; for example, once per glottal period during voiced speech. The image is updated at specific points in time, synchronised with signal peaks. This method is equivalent to modeling the output of each channel by a train of impulses occurring at the positions of the maxima; a method which has been used since in other work (Van Immerseel and Martens, 1992) where it is called a "pseudo autocorrelation analysis". Another important difference is that the function used to update the image buffer is a sum instead of a product, so that the resulting image is constructed from lagged sums rather than lagged products.

4. DISCUSSION.

The correlogram and the auditory image are similar in many respects due to the underlying similarity of the processes used to generate them. Figure 1 shows vowel correlograms generated using the recursive algorithm (6). Figure 2 shows auditory images of the same vowels generated using the sparse method based upon the recursion (9). Both figures show a form of multi–channel running autocorrelation with a vertical frequency axis, and horizontal lag axis. The maximum lag is 32ms; (the auditory image is conventionally plotted with the zeroth lag on the right hand side, representing the most recent event).

The sparse construction of the auditory image has some advantages. It is faster by a number of multiplications equivalent to the number of samples between peaks in the signal. However, many of the inter-peak samples in the rectified cochlea output are zero and could be neglected in any case. The use of lagged sums rather than products reduces image contrast, so that global normalization of the image is unnecessary. This auditory image

also preserves the assymetry of the firing patterns in the neural activity pattern, as a result of the highly selective synchronised process which replaces the usual symmetrical shifting process of autocorrelation. This is clearly illustrated by a comparison of Figures 1 and 2. Whether or not the assymetry of the neural activity patterns, preserved in the auditory image of figure 2, confers additional information is still to be decided. However, it can be seen that the selective process used to construct the auditory image has a narrowing effect on the individual peaks, as compared with those in the correlogram.

The main disadvantage with this auditory image is it's reliance upon a peak-picker for triggering the process. Triggering is a discontinuous process capable of introducing artifacts into the information. The sum in the image buffer is, in theory, very sensitive to mis-triggering. The weighted sum is designed to compute a running average, but it turns out that there are only about three significant components to this average. For example, the time constant of the exponential decay should be no greater than 15ms so as not to smear rapid temporal changes. With this time constant, the exponential window decays by 99% in around 60ms. If the fundamental period is 20ms, the peak-picker would find just three peaks inside the window. If it misses one, the effect on the average would be big. In practice, however, it turns out that the sum is less sensitive to changes in the value of its few components because the use of lagged sums rather than products tends to de-emphasize the differences. Furthermore, the decaying threshold used for the trigger mechanism makes it very unlikely to miss a peak when they are 20ms apart.

Acknowledgements.
The work presented in this paper was supported by the UK Medical Research Council and grants from Espirit BRA (3207) and MOD PE (XR 2239 and SLS/42B/663).

REFERENCES.

Allerhand M. (1990) Autocorrelation and the Stabilised Auditory Image. Research Report No.1. Air Applications Model of Human Auditory Processing, MOD PE SLS/42B/663, 1990.

Fano R.M. (1950) Short-Time Autocorrelation Functions and Power Spectra. *J. Acoust. Soc. Am.* 22 5 pp546–550.

Licklider J.C.R. (1951) A Duplex Theory of Pitch Perception. *Experientia* VII/4 128–134.

Patterson R.D., Holdsworth J. (1992). A functional model of neural activity patterns and auditory images. In: *Advances in Speech, Hearing and Language Processing*, W.A. Ainsworth, (ed.), Vol 3. JAI Press, London. (in press).

Patterson R.D., Robinson K, Holdsworth J, McKeown D, Zhang C, and Allerhand M. (1991) Complex sounds and auditory images. In: *Auditory physiology and perception*, eds Y. Cazals, L. Demany, and K. Horner. Pergamon, Oxford, 429-446.

Schroeder M.R., Atal B.S. (1962) Generalized Short-Time Power Spectra and Autocorrelation Functions. *J.Acoust.Soc.Am.* 34 11 pp1679–1683.

Slaney M., Lyon R.F. (1992) On the Importance of Time – A Temporal Representation of Sound. In: *Visual Representations of Speech Sounds*. Eds: M.Cooke, S.Beet. Pubs: J.Wiley 1992.

Stevens K.N. (1950) Autocorrelation Analysis of Speech Sounds. *J.Acoust.Soc.Am.* 22 6 pp769–771.

Takahashi Y., Rabins M.J., Auslander D.M. (1970) Control and Dynamic Systems. Addison-Wesley 1970.

Van Immerseel L., Martens J.P., (1992) Pitch and voiced/unvoiced determination with an auditory model. *J. Acoust. Soc. Am.* 91 3511–3526.