# Correlation Network Model applied to F0-adaptive spectral estimation (CREST CMAP abstract, do not distribute until July 2002)

*Alain de Cheveigné* — Ircam - CNRS, Paris, France, cheveign@ircam.fr

*Hideki Kawahara* — CREST, Wakayama University, Wakayama, Japan, kawahara@sys.wakayama-u.ac.jp

This paper shows how the functionality of F0-adaptive spectral envelope estimation (Kawahara et al. 1999a) might be implemented within the auditory system, using the framework of the Correlation Network model (de Cheveigné, 2001).

**Correlation Network model.** This model (hereafter CN model) is an abstract model of auditory processing that allows a range of auditory signal-processing functions to be implemented in a uniform way: pitch, timbre, localization, segregation. It consists of three modules (Fig. 1). The first calculates arrays of running autocorrelation (monaural) and crosscorrelation (binaural) coefficients. The second forms a linear combination of coefficients produced by the first module. The third controls the parameters of the second module while monitoring its output. It is responsible for producing the behavior needed for each function (pitch, etc.). Fast signal processing is limited to the first module, while the second and third handle relatively slowly varying quantities. This might ease mapping of the model to the auditory system (first module to brainstem, second and third to more central levels). The CN model can implement any model that operates on a quadratic statistic (power or correlation) of linear combinations of delayed versions of its inputs. In particular it can implement cancellation models (Durlach, 1963; de Cheveigné, 1993). It has also proved useful as a basis for F0 estimation (de Cheveigné and Kawahara, 2002).

While the CN model addresses auditory processing, here we treat it as a digital signal processing model to demonstrate how the desired functionality can be obtained. The basic ingredients of the CN model are arrays of running autocorrelation (AC) and crosscorrelation coefficients (here only the former are considered). Using a sampled-signal notation, the AC function of the signal at the left ear is:

$$r_{t,W}(\tau) \quad = \quad \sum_{j=t+1}^{t+W} s_j s_{j+\tau} \qquad (1)$$

where $s$ is the signal, $\tau$ the autocorrelation lag parameter and $W$ the size of the integration window (supposed square). In the speech processing literature this quantity is called *autocovariance*. In its original formulation the CN model used a fixed window size but here we suppose that that it can be tuned to an arbitrary value. The first module thus produces an array indexed by time, lag and integration window size ($t$, $\tau$, $W$).

**F0-adaptive spectral estimation.** The first step is to estimate the period $T$, for example using an algorithm also derived from the CN model (de Cheveigné and Kawahara, 2002), or other effective methods (Kawahara et al., 1999b). Next, the

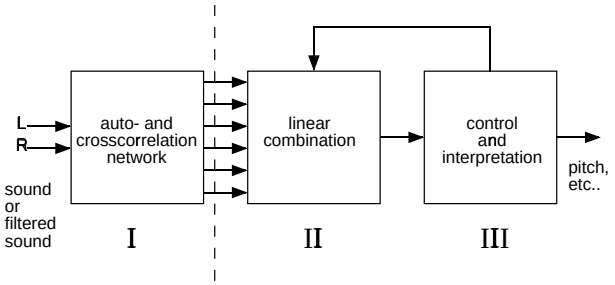


Figure 1: Structure of the Correlation Network model. Fast time-domain processing is limited to the first module (left of the dotted line). Subsequent processing operates on slowly-varying quantities.

AC function is calculated for $[0, T]$ using a window size of $T$. The $T$-sized window ensures that the AC function and estimates derived from it do not fluctuate in time. This function is if necessary *interpolated* to an array of $N+1$ samples equally spaced between 0 and $T$ where $N$ is preferably a power of two. This array is then reversed and added to itself ($r_t(\tau) + r_t(T-\tau)$) to avoid a sharp transition when considered cyclically.

Finally, an $N$-point DFT is taken to estimate the power spectrum at the harmonics of $F_0 = 1/T$. The estimate is *exact* if the signal is perfectly periodic and $T$ is such that interpolation is unnecessary (interpolation may introduce some spectral error). For an imperfectly periodic signal the relation between the coefficients obtained and the actual spectral envelope (whatever that means) is less straightforward. One thing is clear however: the AC calculation technique effectively avoids the spectral splatter that occurs with simpler schemes (such as a DFT applied directly to one period of the waveform). Frequency resolution is $1/T$, the best possible for sampling a spectral envelope. Estimation can be performed at arbitrary time intervals, but useful temporal resolution is limited to $2T$. The spectrotemporal estimate array can then be interpolated for display or resynthesis as in the standard STRAIGHT method (Kawahara et al. 1999a). Alternatively it can be used directly for harmonic-weighted pattern-matching as in the "missing data" model of de Cheveigné and Kawahara (1999a). In any case there are no periodicity-related artifacts.

To the extent that similar processing might be carried out within the auditory system, this scheme can be seen as a model for spectral estimation within the auditory system. There is some evidence that integration windows for pitch are F0-adaptive (Wiegrebe, 2000; Plack and White, 2000). It has also been argued that F0-invariant vowel perception can be

explained only on the basis of F0-dependent processing (de Cheveigné and Kawahara, 1999a).

**Multi-source spectral estimation.** A nice feature of the CN model is that it allows calculating the AC function of various transforms of the signal (or signals in the multichannel case), in particularly those that allow cancellation of interfering sources (e.g. Durlach, 1963; de Cheveigné, 1993). This opens the perspective of accurate F0-adaptive spectral estimation of *multiple* concurrent sources. The cancellation operation introduces errors in the spectral estimate, but they can be either compensated or ignored if F0s of the concurrent signals are known. Estimates of these can also be obtained effectively within the CN model (de Cheveigné and Kawahara, 1999b, 2002).

**Implementation considerations.** A final consideration for simulations and applications is that of efficiency. Major computational costs occur within module 1. Its internal bandwidth is that of the signal multiplied by a factor of $f_s^2 \tau_M W_M$ where $f_s$ is the sampling rate and $\tau_M$ and $W_M$ are the maximum lag and window sizes needed (both are related to the largest expected period). The bandwidth of the interface with module 2 is much smaller as module 2 needs random access to only a few coefficients at a time. Storage requirements of module 1 are on the order of $f_s^3 \tau_M W_M D_M$ where $D_M$ is the largest time span needed, but they can to some extent be reduced by synthesizing each $W$-sized window "on demand" as the sum of (on average) $\log_2(W)/2$ power-of-two-sized subwindows stored in a hierarchical structure.

Bandwidth, storage and computation requirements of module 1 are large even by todays standards. If the CN model turns out to be general and useful, implementation as a hardware device might make sense. This device would input one or more channels of audio and support queries for auto- and crosscorrelation terms with arbitrary $t$, $\tau$, and $W$ (fractional queries being handled by interpolation). As a final refinement, rather than considering this device as a finite-length buffer of running correlation coefficients (deleted at one end as they are formed at the other), "perpetual statistics" can be implemented by accumulating large scale terms of the hierarchical subwindow structure while discarding smaller-scale terms. A proportion of storage would thus be devoted to keeping statistics of ever-larger chunks of signal situated ever-further in the past. This follows the concept of scalable metadata (de Cheveigné and Peeters; de Cheveigné, 2002);

**Summary.** To summarize, the Correlation Network (CN) model offers a uniform basis for a range of useful signal processing operations including source segregation, F0 estimation, and F0-adaptive spectrum estimation. It shows how seemingly diverse themes explored by both authors can be brought together. [This work was supported by the Cognitique programme of the French Ministry of Research and Education.]

## 1. References

[1] de Cheveigné, A. (1993). "Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing," J. Acoust. Soc. Am. 93, 3271-3290.

[2] de Cheveigné, A., and Peeters, G. (1999), "Scale tree," ISO/IEC JTC1/SC29/WG11, MPEG99/m5076 (MPEG-7 proposal document).

[3] de Cheveigné, A., and Kawahara, H. (1999a). "Missing data model of vowel perception," J. Acoust. Soc. Am. 105, 3497-3508.

[4] de Cheveigné, A., and Kawahara, H. (1999b). "Multiple period estimation and pitch perception model," Speech Communication 27, 175-185.

[5] de Cheveigné, A. (2001). "Correlation Network model of auditory processing.", Proc. Workshop on Consistent & Reliable Acoustic Cues for sound analysis (Aalborg, Danmark).

[6] de Cheveigné, A., and Kawahara, H. (2001). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am., submitted.

[7] de Cheveigné, A. (2002). "Scalable metadata for search, sonification and display.", Proc. International Conference on Auditory Display, Kyoto.

[8] Durlach, N. I. (1963). "Equalization and cancellation theory of binaural masking-level differences," J. Acoust. Soc. Am. 35, 1206-1218.

[9] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999a). "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication 27, 187-207.

[10] Kawahara, H., Katayose, H., de Cheveigné, A., and Patterson, R. D. (1999b). "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity.", Proc. Eurospeech, 2781-2784.

[11] Markel , J. D., and Gray, A. H. (1976). "Linear prediction of speech," Berlin, Springer-Verlag.

[12] Plack, C. J., and White, L. J. (2000). "Perceived continuity and pitch perception," J. Acoust. Soc. Am. 108, 1162-1169.

[13] Wiegrebe, L. (2001). "Searching for the time constant of neural pitch integration," J. Acoust. Soc. Am. 109, 1082-1091.