Correlation Network Model of Auditory Processing

Alain de Cheveigné

CNRS - Ircam, 4 place Igor Stravinsky, 75004, Paris, FRANCE

Abstract

The Correlation Network model serves as a framework for models of auditory processing for pitch, timbre, localization and sound segregation. It comprises three modules. The first module calculates arrays of running correlation coefficients (two autocorrelation arrays, one crosscorrelation array). Each array is two dimensional, indexed in time measured relative to a sliding origin (the "present"), and lag. If peripheral frequency analysis is taken into account, the arrays have a third dimension: tonotopy. Integration over a sliding window in the correlation calculation removes most of the fine time structure, so the output of the first module consists of slowly varying values. The second module calculates a weighted sum of its inputs. The third module controls the weights of the second module while monitoring its output, and is responsible for producing the behavior expected from the model. Based upon the Correlation Network model, a wide range of models of pitch, timbre and binaural processing can be implemented, in particular those involving correlation and cancellation. It offers a uniform basis for these operations with a simple mapping to known anatomy (module I to the brainstem, modules II and III to midbrain and beyond). It allows complex models (such as multi-stage cancellation) to be cast in relatively simple and plausible terms. It provides useful inspiration for signal processing tasks such as F0 estimation, spectral estimation and source separation.

1. Introduction

The anatomy of the auditory system comprises the cochlea, that splits the acoustic signal into channels that respond best to narrow bands of frequencies, and several stages of neural processing within the auditory nervous system. Much of this circuitry is suited for the transport and processing of *time-domain* patterns, suggesting that time-domain analysis is carried out within the nervous system. Classic models assume that the cochlea produces only slowly-varying spectral patterns, but recently there has been a development of time-domain neural processing models to explain pitch, timbre and sound segregation. Several of these models are based on correlation involving excitatory-excitatory (EE) neural interaction, or cancellation involving excitatory-inhibitory (EI) interaction.

The binaural localization model of Jeffress [15] is one of the earliest time-domain models. A network of delay lines and neural coincidence counters is fed from both ears. In response to a sound source to one side of the midline plane, peak activity occurs at a position for which an internal delay compensates for the difference between external propagation delays to each ear. This indicates the azimuth of the source. Another early model is the monaural pitch model of Licklider [19], which also postulates delay lines and coincidence counters. In response to a periodic sound, peak activity occurs at a position for which the internal delay matches the period. This indicates the pitch of the sound. The network calculates the equivalent of the autocorrelation function, which is known to carry the same information as the power spectral density (its Fourier transform). Indeed, Meddis and Hewitt [17] suggested that it could be used to identify vowel timbre in a pattern-matching model.

A second class of models involves inhibitory interaction. In the equalization-cancellation (EC) model of Durlach [14], delayed signals from both ears are subtracted, instead of multiplied as in Jeffress's model. This allows a strong interfering source to be canceled so that a weak target can more easily be detected. A similar idea has been applied to monaural processing of mixtures of sounds, such as simultaneous voices or musical sounds [7, 8]. Periodic interference is suppressed so that a weaker target can be perceived. Cancellation can also be used to explain pitch perception [9], and it may be cascaded with other time-domain processing to account for the perception of multiple pitches [11], binaural pitch phenomena ([1], or identification of mixtures of vowels [8].

Together, these time-domain models account for a wide range of processing functions. The present paper attempts to unify them within a common framework.

2. The model

2.1. A basic ingredient: correlation

The basic ingredient is a set of arrays of autocorrelation (AC) and crosscorrelation (CC) coefficients. Using a sampled-signal notation, the AC function of the signal at the left ear is calculated as:

$$r_t^L(\tau) \quad = \quad \sum_{j=t+1}^{t+W} x_j^L x_{j+\tau}^L \tag{1}$$

where x^L is the signal at the left ear, τ the autocorrelation lag parameter and W the size of the integration window. A square window is used for simplicity, but other forms (such as leaky integration) would serve just as well. A similar function $r_t^R(\tau)$ is calculated for the right ear (the superscript may be dropped for monaural models). The interaural CC function is calculated as:

$$c_t(\theta) = \sum_{j=t+1}^{t+W} x_j^L x_{j+\theta}^R \tag{2}$$

where θ is the crosscorrelation lag parameter. These functions are calculated for every time instant t. Thanks to temporal integration they are not expected to fluctuate much with t, but for accuracy they are nevertheless calculated at each instant with a sliding window. Depending on the level of abstraction required, processing is assumed to affect each peripheral filter channel (in a detailed model), or the raw acoustic waveforms (in a more abstract model). To keep things simple the latter is assumed except where noted.

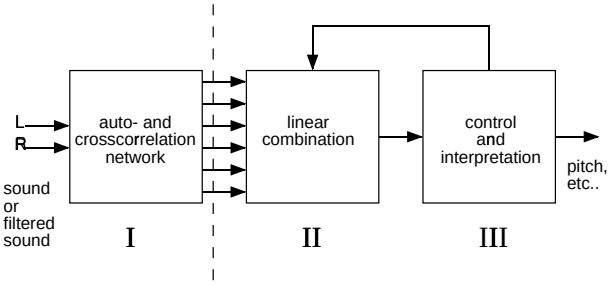

Figure 1: Structure of the Running Correlation Network model. Fast time-domain processing is limited to the first module (left of the dotted line). Subsequent processing operates on slowly-varying quantities.

2.2. Structure

The Correlation Network model involves three modules (Fig. 1). The first produces arrays of AC and CC coefficients. The second forms a linear combination of these coefficients. The third controls the parameters of the second module while monitoring its output, and accounts for the behavior of the model (pitch, timbre, etc.).

Module I delivers all AC and CC coefficients within a certain range of time and lag. As they are temporally smoothed, subsequent modules process slowly-varying quantities, and so fast time-domain processing is restricted to the first module. Module II forms a linear combination with factors that can be positive or negative. Module III controls these factors while monitoring the output of module II.

Modules are distinct for conceptual reasons, but it might be useful to consider implementations in which they are merged, for example as a neural network with Hebbian learning. In a detailed model, modules I to III would operate within each frequency channel produced by peripheral filtering, but to simplify we assume that they operate directly on acoustic signals.

3. Particularizations

The Correlation Network model can be used to implement various known auditory processing models. In some cases implementation is trivial (correlation models), in others it is slightly less obvious.

3.1. Autocorrelation model of pitch

This model was first proposed by Licklider [19] and is currently one of the more popular models of pitch perception. Ignoring peripheral filtering (as in [22]), an AC function is calculated from the acoustic signal sampled at either ear, and the position of the function maximum is used to indicate the period (and thus the pitch) of the source.

$$r_t(\tau) \quad = \quad \sum_{j=t+1}^{t+W} x_j x_{j+\tau} \tag{3}$$

Samples separated by a period are similar and thus tend to produce larger products than samples separated by other intervals, leading to a peak in $r_t(\tau)$ when τ equals the period. The integration smoothes these values over time to produce a stable estimate.

In the present framework, module 2 selects one term of the AC array (left or right) as determined by the lag parameter τ. Module 3 varies this parameter while monitoring the output for a maximum.

3.2. Autocorrelation model of vowel identification

In the model of Meddis and Hewitt [17], a vowel's identity is determined by template matching of a summary pattern obtained by adding AC functions calculated within peripheral channels (see also [11]). The summary AC pattern can be approximated by the AC function of the waveform.

In the present framework, module 2 selects a term of the autocorrelation array determined by the lag parameter τ. Module 3 varies this parameter and matches the pattern of variation of the output of module 2 to a template. The best match indicates the vowel.

3.3. Crosscorrelation model of localization

This model, due to Jeffress [15], is the earliest detailed model of time-domain processing in the auditory nervous system. Ignoring peripheral filtering, the CC function is calculated from signals sampled at both ears, and the position of the function maximum (as a function of lag) is used to indicate the azimuth of the source.

In the present framework, module 2 selects one term of the CC array as determined by parameter θ. Module 3 varies this parameter while monitoring the output for a maximum.

3.4. Cancellation model of pitch

In the previous examples, module II performed the trivial task of selecting one coefficient. This example gives a better idea of its role. The squared difference function is defined as:

$$d_t(\tau) \quad = \quad \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2 \tag{4}$$

In the cancellation model of pitch, the position of the *minimum* of this function is the cue to pitch. The cancellation model can be seen as a sort of "negative" version of Licklider's model. The squared difference can be expanded, and $d_t(\tau)$ expressed in terms of AC coefficients.

$$\begin{aligned} d_t\tau) \quad &= \quad \sum_{j=t+1}^{t+W} [x_j^2 + x_{j+\tau}^2 - 2x_j x_{j+\tau}] \\ &= \quad r_t(0) + r_{t+\tau}(0) - 2r_t(\tau) \end{aligned} \tag{5}$$

In the present framework, module 2 forms this linear combination as determined by parameter τ. Module 3 varies τ while monitoring the output for a minimum.

3.5. EC model of binaural unmasking

The Equalization Cancellation (EC) model of Durlach [14] explains why signals in noise are easier to detect with two ears rather than one. Signals from left and right are equalized (by applying a delay and/or an amplitude factor to either signal) and subtracted to cancel the noise. Supposing that the decision statistic is energy, it is equal to the squared difference function:

$$d_t(\theta, \alpha) = \sum_{j=t+1}^{t+W} (x_j^L - \alpha x_{j+\theta}^R)^2 \tag{6}$$

The squared sum can be developed and the statistic expressed in terms of monaural AC and binaural CC terms:

$$\begin{aligned} d_t(\theta,\alpha) &= \sum_{j=t+1}^{t+W} [(x_j^L)^2 + \alpha^2 (x_{j+\theta}^R)^2 - 2\alpha x_j^L x_{j+\theta}^R] \\ &= r_t^L(0) + \alpha^2 r_{t+\theta}^R(0) - 2\alpha c_t(\theta) \end{aligned} \quad (7)$$

In the present framework, module 2 forms the combination determined by parameters θ and α. Module 3 sets these parameters to obtain the best signal-to-noise ratio, and then monitors the output of module 2 for the presence of a signal.

The original model applied to single channels, but it can be extended to apply uniformly to each channel from the auditory periphery (or to raw acoustic waveforms in a simplified model). Culling and Summerfield [5] proposed a modified EC model (mEC) that departs from the original in two ways: (1) each channel applies the EC operation using its own parameters, and (2) these parameters are based on criteria local to the channel.

The EC or mEC models can be used to detect the presence of a signal (according to the magnitude of the cancellation residual), to produce a binaural pitch (according to the position along the tonotopic axis of a peak in cancellation residual) [6], or to identify the timbre of a vowel (according to an eventual formant pattern in the tonotopic cancellation residue)[5].

3.6. Multiple pitch model

The previous models assumed a single stage of time-domain processing (correlation or cancellation). This model and the next assume several stages in cascade. The multiple pitch perception model of [11] accounts for pitches evoked by mixtures of periodic sounds (voices or instruments playing together). In its iterative version, one sound is first suppressed by cancellation tuned to an initial period estimate T, and the result processed to estimate the period of the remaining sound. Calling $z_t = x_t - x_{t+T}$ the cancellation residue, its AC function is:

$$r_t^z(\tau) = \sum_{j=t+1}^{t+W} z_t z_{t+\tau} = \sum_{j=t+1}^{t+W} (x_t - x_{t+T})(x_{t+\tau} - x_{t+\tau+T})$$

Developing produces the linear combination:

$$r_t^z(\tau) = r_t(\tau) - r_t(T+\tau) - r_{t+\tau}(t+T-\tau) + r_{t+T}(\tau) \quad (8)$$

The AC function $r_t^z(\tau)$ is used to determine the period of a second sound.

In the present framework, module 2 forms the linear combination determined by parameters (T, τ). Module 3 sets T to the latest estimate of the period of source A, and varies τ while monitoring the output for a maximum. This gives an estimate of the period of source B. T is then set to that value, and τ is varied to refine the estimate of the period of source A, etc..

Instead of this iterative algorithm, a "joint estimation" version of the same model can be formulated based on the following difference function [11]:

$$d_t(\tau_1,\tau_2) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau 1} - x_{j+\tau 2} + x_{j+\tau 1+\tau 2})^2$$

which can be expanded into a sum of AC terms

$$\begin{aligned} d_t(\tau_1,\tau_2) = r_t(0) + r_{t+\tau 1}(0) + r_{t+\tau 2}(0) + r_{t+\tau 1+\tau 2}(0) \\ -2r_t(\tau_1) - 2r_t(\tau_2) + 2r_t(\tau_1+\tau_2) \\ +2r_{t+\tau 1}(\tau_2-\tau_1) - 2r_{t+\tau 2}(\tau_1) - 2r_{t+\tau 1}(\tau_2) \end{aligned} \quad (9)$$

In the present framework, module 2 forms the linear combination determined by parameters (τ_1, τ_2). Module 3 varies these parameters while monitoring the output for a minimum. The search is either exhaustive (all pairs are tested) or iterative as before. The original formulation of [11] required a cascade of cancellation stages. The present framework offers the same function with a single stage.

3.7. Concurrent vowel identification model

This model explains why mixtures of vowels are better identified if they have different F_0s rather than the same F_0 [7, 8]. Supposing that the period T of the stronger vowel is known, that vowel is suppressed by forming the difference $z_t = x_t - x_{t+T}$. The autocorrelation function $r_t^z(\tau)$ of z_t is then used to identify the weaker vowel by pattern matching as in the model of Meddis and Hewitt cited above. This function may be calculated according to Eq. 8.

In the present framework, module 2 forms the linear combination determined by parameters (T, τ). Module 3 sets T (determined according to any of the previous period-estimation models) and varies τ while monitoring the output of module 2. The pattern of output as a function of τ is matched to a template.

3.8. Binaural pitch perception model

Binaural pitch phenomena are usually understood by interpreting the binaural stimulus as the sum of an interfering source with high interaural correlation, and a tonal target with an interaural correlation that is low or different from the interference. The interference is suppressed by the EC [14] or mEC [5] operation. According to one popular account, the amount of energy that survives cancellation is used as a tonotopic pattern to derive pitch according to a spectral pitch model [6]. This account is plausible, but paradoxical in the sense that these phenomena have long been cited as evidence *against* spectral models of pitch.

It is possible instead to formulate a fully time-domain account [1]. The output of the EC stage within each channel can be considered as a fast-varying time-domain signal, rather than as one sample of a slowly varying tonotopic pattern, and fed to a pitch model such as those mentioned above, for example Licklider's AC model. Denoting the output of the EC stage as $z_t = x_t^L - x_{t+\theta}^R$, its AC function is:

$$r_t^z(\tau,\theta) = \sum (x_t - y_{t+\theta})(x_{t+\tau}^L - x_{t+\tau+\theta}^R) \quad (10)$$

Developing produces the linear combination:

$$r_t^z(\tau,\theta) = r_t^L(\tau) - c_t(\theta+\tau) - c_{t+\tau}(t+\theta-\tau) + r_{t+\theta}^R(\tau) \quad (11)$$

In the present framework, module 2 forms this combination with parameters (θ, τ). Module 3 first determines θ according to the criteria of the EC or mEC model, and then varies τ while monitoring the output for a maximum to estimate the pitch. It is trivial to introduce an amplitude equalization factor in the EC stage ($z_t = x_t^L - \alpha x_{t+\theta}^R$). Eq. 11 then becomes:

$$\begin{aligned} r_t^z(\tau,\theta) &= r_t^L(\tau) - \alpha c_t(\theta+\tau) - c_{t+\tau}(t+\theta-\tau) \\ &\quad +\alpha^2 r_{t+\theta}^R(\tau) \end{aligned} \quad (12)$$

It is also possible to use a cancellation model to determine the pitch period by minimizing a difference function rather than

maximizing an AC function:

$$\begin{aligned} d_t(\tau,\theta) &= r^L_t(0) + r^R_{t+\theta}(0) + r^L_{t+\tau}(0) + r^R_{t+\theta+\tau}(0) \\ &\quad -2c_t(\theta) - 2r^L_t(\tau) + 2c_t(\theta+\tau) \\ &\quad +2c_{t+\tau}(\theta-\tau) - 2r^R_{t+\theta}(\tau) - 2c_{t+\tau}(\theta) \end{aligned} \quad (13)$$

This allows joint estimation of both parameters by searching for a global minimum. These variants are all implemented by a simple modification of module II.

To summarize, each of these models is easily implemented by adjusting the linear combination (module II) and control algorithm (module III) of the Correlaton Network model. The signal processing module (module I) remains the same.

4. Mapping to physiology

4.1. Structure

Processing of "fast" time-domain patterns is limited to the first module. This is a useful feature since it is known that phase-locking degrades as one proceeds within the auditory system. It is tempting to map the first module to the brainstem (cochlear nucleus and olivary complex) and the next two modules to the midbrain and beyond (inferior colliculus, MGB and cortex), where neural patterns are slower. However given the complexity of the auditory system the mapping cannot be that clean.

The first module requires delay lines and coincidence counters. There is evidence of both, and indeed of correlation-like calculations [21], although the existence of delay lines long enough to address all needs is a subject of debate. The second module requires weighted summation of relatively slow inputs, both excitatory and inhibitory, for which there is also good evidence [3]. The third module implements the algorithmic processing required for each particularization. One can only speculate on how that is implemented, but it is worth noting that the assumption of dynamic modulation of weights is common among neural models.

4.2. Fast inhibition

Cancellation models account for functions not easily explained by other models [7, 8, 10]. However their original formulation calls for fast inhibitory interaction between temporally structured neural patterns. There is some evidence that inhibitory synapses may be slower than excitatory synapses. It is therefore useful to know that the same functions can be implemented with fast *excitatory* interaction only.

This does not mean that fast inhibitory interaction cannot be acommodated in the framework, or indeed be valuable. There is evidence of fast inhibitory interaction in the LSO [16]. At the cost of greater complexity, the Correlation Network model can be extended to allow module I to calculate difference function terms such as:

$$d_t(\tau) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2 \quad (14)$$

The advantage of these terms, if available, is to make certain linear combinations simpler and possibly more accurate. For example, the ten terms of Eq. 9 can be replaced by six:

$$\begin{aligned} d_t(\tau_1,\tau_2) &= d_t(\tau_1) + d_{t+\tau 1}(\tau_2) - 2r_t(\tau_2) + 2r_t(\tau_1+\tau_2) \\ &\quad +2r_{t+\tau 1}(\tau_2-\tau_1) - 2r_{t+\tau 2}(\tau_1) \end{aligned} \quad (15)$$

Nevertheless, all useful functions can be implemented without fast inhibition, so it is not necessary to consider it further here.

4.3. Cascaded processing

Among models mentioned above, several originally required cascaded time-domain processing (for example the binaural pitch model of [1] required binaural cancellation followed by autocorrelation). This is a heavy requirement, as it entails a multistage neural topology capable of maintaining phase locking over several successive synapses. The present model dispenses of this requirement, while offering exactly the same functions.

4.4. Peripheral filtering

Peripheral filtering was ignored in the formulation of the model. It may be incorporated into the model by processing each channel in parallel and summing the results. The outcome is the same as if the model were applied to the waveform, if two conditions are met: (1) peripheral filtering is linear (so it can be swapped with delay or other linear operations), and (2) channels are orthogonal (so the power of the sum is the sum of powers). Both conditions are only approximately fulfilled by the auditory system.

Peripheral filtering may have several functional advantages. One is that its dispersive properties have a "linearizing" effect that compensates for non-linear transduction. For example the half-wave rectification properties of the hair cell entail the loss of half the information (the missing part of the waveform), but this loss is compensated if a second channel responds after a 180^o phase shift.

A second functional advantage results from compression within each channel. The AC function, Fourier transform of the power spectrum, is strongly dominated by high-amplitude parts of the spectrum (for example the first formant of speech). Within-channel compression reduces this dominance and provides a more balanced representation of the spectrum, akin to cubic-root or log transforms that have been found beneficial in speech analysis.

A third functional advantage is that channels dominated by interference can be ignored in the final summation.

There is a similarity between auditory peripheral filtering followed by neural segregation, and recent ICA (Independent Component Analysis) methods that operate in the frequency domain [2, 18], which allows the convolutive mixing problem to be replaced by several scalar mixing problems (see below). If the audio system made use of a similar feature, that would constitute a fourth functional advantage.

4.5. Caveats

A first caveat is that the model assumes *linearity*, particularly to implement cancellation. Non-linearity may affect performance to a degree that is unknown. An advantage of the Correlation Network model (with respect to previous cancellation models) is that cancellation occurs on slowly-varying rather than phase-locked patterns. Linearity may be easier to ensure for slow patterns.

A second caveat is that the mapping of cancellation to autocorrelation (Eqs. 4, 5) works for models that use quadratic statistics such as energy or autocorrelation. This restricts its generality, however it is known that many perception processes seem to involve an energy statistic.

A third caveat, common to other time-domain models of auditory processing, is that we must assume neuronal delay lines on the order of up to 20 ms. The existence of delay lines that long is still controversial.

5. Signal processing

The Correlation Network model may provide useful inspiration for certain signal processing tasks.

5.1. Fundamental frequency estimation

The principles behind the Correlation Network have been applied with success to speech F_0 estimation, leading to a method (YIN) that seems to outperform other methods [12] (see also [13], main conference). They can also be used to implement the multiple-period estimation algorithms of [11]. The advantage is partly computational (efficient use of precalculated AC coefficients), partly a question of flexibility. For example it is easy to incorporate factors to compensate for amplitude variation of a signal:

$$\begin{aligned} d_t(\tau,\alpha) &= \sum_{j=t+1}^{t+W} (x_j - \alpha x_{j+\tau})^2 \\ &= r_t(0) + \alpha^2 r_{t+\tau}(0) - 2\alpha r_t(\tau) \end{aligned} \quad (16)$$

This entails a simple change in the coefficients applied by module II (see [12] for fuller details). Actually, in this example one can go a step further and determine the value of α that minimizes d. Putting the result in Eq. 16 gives:

$$d_t(\tau) = r_{t+\tau}(0)[1 - r_t(\tau)^2 / r_t(0) r_{t+\tau}(0)] \quad (17)$$

The right-hand side is a function of AC coefficients, but not a linear combination. To acommodate it, the model must be extended to combinations other than linear.

5.2. Spectral estimation

We saw that the Correlation Network allows calculation of the AC functions of various linear combinations of signals with arbitrary delays, on the basis of AC (and eventually CC) coefficients of the ingredient signals. The AC function in turn determines the power spectrum from which log spectrum, cepstrum, LPC, PLP, etc. can be derived. The model thus produces an output that is suited for calculating spectral estimates and features.

5.3. Matched FIR filter

Suppose that a task calls for finding a finite impulse response filter such that the power at its output is maximal (or minimal) subject to certain constraints on its parameters. Supposing that the filter has N taps, the filtered signal is a linear combination of N delayed versions of x_t. Generalizing from Eq. 5, the power can be expressed as a sum of $N(N+1)/2$ autocorrelation terms. Thus, to perform the task it is not necessary to apply the actual filters to the signal and measure their output: it is sufficient to test the corresponding combinations of autocorrelation terms. In the present framework, module 2 forms these linear combinations, defined by the delays of the taps and the corresponding factors. Module 3 varies these parameters while monitoring the output for a maximum (or minimum).

5.4. Source segregation and ICA

Section 3 described a number of segregation models that can be applied more widely to source segregation tasks.

Blind source separation (BSS) techniques such as ICA (Independent Component Analysis) address the task of recovering several source signals from observed signals in which they are mixed. Consider N sensors that sample mixtures of M sources, so that the observations $x_n(t)$ are related to the source signals $s_m(t)$ by a mixing matrix X. BSS attempts to find a matrix Y such that $YX = I$, the identity matrix. The process is "blind" in the sense that the mixing matrix X is unknown. Blind separation applies certain criteria to the outputs, and searches for the matrix Y that best fulfills them. In the case of ICA, the criterion is statistical independence between outputs (separated sources) according to various statistical measures. Early BSS attempts addressed only the case of a scalar (instantaneous) mixing matrix, but recently the techniques have been extended to convolutive mixing and unmixing matrices [20].

It is worth pointing out the parallel between cancellation models and ICA. A model such as EC subtracts binaural signals one from another (after delay and an optional amplitude factor), which is a simple case of a convolutive unmixing matrix. The aim is to cancel one source of a pair to better detect the other. Obviously if a source is successfully canceled, the output is then statistically independent from that source. Supposing that the model has a second output that cancels the second source, the two outputs are statistically independent (if inputs are) and thus the EC model tries to fulfill ICA criteria.

The Correlation Network is relevant to ICA in two respects. One is that, once the appropriate unmixing matrix has been found, it provides a convenient way to derive useful features (power, AC function) of the unmixed signals from the AC coefficients of the sensor signals. Second, it may assist the ICA process by cheaply providing the statistics it needs [2, 18].

6. Conclusion

What is new? We showed that the Correlation Network could implement a number of existing auditory processing models, and so one might ask what it offers beyond those models. The contribution of the new model is to unify those models within a common framework that reveals their basic similarities and eases their implementation. Severe requirements such as cascaded phase-locked processing, or fast inhibition, are relaxed. The model structure maps well to the structure of the auditory system, with fast signal processing limited to initial stages and slower processing used subsequently. Perhaps the most interesting feature of the model is its "productiveness", since once the basic ingredients are granted, one can synthesize a wide range of functions by simple tuning of modules II and III.

7. Acknowledgments

This work was supported by the Cognitique programme of the French Ministry of Research and Education. Thanks to Daniel Pressnitzer and others for useful discussions on these ideas.

8. References

[1] Akeroyd, M. A., and Summerfield, A. Q. (2000). "A fully-temporal account of the perception of dichotic pitches," Br. J. Audiol. 33(2), 106-107.

[2] Anemller, J., and Kollmeier, B. (2000). "Amplitude modulation decorrelation for convolutive blind source separation.", Proc. Second international workshop on independent component analysis and blind source separation, 215-220.

[3] Cai, H., Carney, L., and Colburn, S. (1998). "A model for binaural response properties of inferior colliculus neurons. I. A model with interaural time difference-sensitive exci-

tatory and inhibitory inputs.," J. Acoust. Soc. Am. 103, 475-493.

[4] Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust speech recognition.", Proc. ICASSP, 863-866.

[5] Culling, J. F., and Summerfield, Q. (1995). "Perceptual segregation of concurrent speech sounds: absence of across-frequency grouping by common interaural delay," J. Acoust. Soc. Am. 98, 785-797.

[6] Culling, J. F., Summerfield, Q., and Marshall, D. H. (1998). "Dichotic pitches as illusions of binaural unmasking I: Huggin's pitch and the "Binaural Edge Pitch"," J. Acoust. Soc. Am. 103, 3509-3526.

[7] 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.

[8] de Cheveigné, A. (1997). "Concurrent vowel identification III: A neural model of harmonic interference cancellation," J. Acoust. Soc. Am. 101, 2857-2865.

[9] de Cheveigné, A. (1998). "Cancellation model of pitch perception," J. Acoust. Soc. Am. 103, 1261-1271.

[10] de Cheveigné, A. (1999). "Pitch shifts of mistuned partials: a time-domain model," J. Acoust. Soc. Am. 106, 887-897.

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

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

[13] de Cheveigné, A., and Kawahara, H. (2001). "Comparative evaluation of F0 estimation algorithms", Proc. Eurospeech.

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

[15] Jeffress, L. A. (1948). "A place theory of sound localization," J. Comp. Physiol. Psychol. 41, 35-39.

[16] Joris, P. X. (1996). "Envelope coding in the Lateral Superior Olive. II. Characteristic delays and comparison with responses in the Medial Superior Olive.," J. Neurophysiol. 76, 2137-2156.

[17] Meddis, R., and Hewitt, M. J. (1992). "Modeling the identification of concurrent vowels with different fundamental frequencies," J. Acoust. Soc. Am. 91, 233-245.

[18] Murata, N., Ikeda, S., and Ziehe, A. (1998). "An approach to blind source separation based on temporal structure of speech signals.", Proc. ICANN.

[19] Licklider, J. C. R. (1951). "A duplex theory of pitch perception," Experientia 7, 128-134.

[20] Torkkola, K. (1999). "Blind separation of audio signals: Are we there yet?", Proc. Workshop on blind separation and independent component analysis, Aussois, France.

[21] Yin, T. C. T., Chan, J. C. K., and Carney, L. H. (1987). "Effects of interaural time delays of noise stimuli on low-frequency cells in the cat's inferior colliculus. III. Evidence for cross-correlation," J. Neurophysiol. 58, 562-583.

[22] Yost, W. A. (1996). "Pitch of iterated rippled noise," J. Acoust. Soc. Am. 100, 511-518.