Generalized Correlation Network

Alain de Cheveigné, Daniel Pressnitzer

Ircam-CNRS 1 place Igor Stravinsky 75004, Paris, France

last revised: 7 May 2001

Abstract

This draft presents a signal processing network, based on auto- and crosscorrelation, that is intended to serve as a basis for models of auditory processing of pitch, timbre and binaural disparities. The model may also serve as a signal processing model in its own right, for methods of F_0 estimation, spectral estimation, source separation, etc.. The model comprises three modules. The first is a network that calculates arrays of running autocorrelation and running crosscorrelation coefficients. These arrays are two dimensional, indexed in time measured relative to a sliding time origin (the "present") and lag of auto- or crosscorrelation functions. Temporal smoothing with a sliding window removes fine time structure, so the output of this module consists of an array of slowly varying values. The second module is

a linear weighting network that calculates a weighted sum of values calculated by the first module. The third module controls the weights of the second module and interprets whatever there is to interpret. Based upon this network, a wide range of models of pitch, timbre and binaural processing can be implemented, in particular cancellation models. Cancellation models normally require fast inhibitory interaction in the time domain, but the present framework allows it to be replaced by fast excitatory interaction. Furthermore, models that normally require cascaded time-domain processing can be implemented as a single stage. Like other time-domain auditory models, this model assumes the existence of delay lines. It also requires accurately balanced neural processing of excitatory and inhibitory signals. However this balanced processing occurs on slow signals rather than fast.

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 neural circuitry is suited for the transport and processing of accurate time-domain signals, suggesting that analysis of the temporal structure of the acoustic signal is performed for part within the nervous system. Whereas classic models of hearing assumed that the cochlea transforms the fine temporal structure of sound into slowly-varying spectral patterns, there has been a recent development of time-domain neural processing models to explain pitch, timbre and sound segregation.

The binaural crosscorrelation model of Jeffress (1948) is one of the earliest time-domain models. It postulates a network of delay lines and neural coincidence counters fed by both ears. In response to a source situated to the side of the midline plane, a maximum of activity occurs at a position for which internal delays compensate for the external propagation delays to each ear. The monaural autocorrelation model of Licklider (1951) also postulates delay lines and coincidence counters. In response to a periodic sound, a maximum of activity occurs for an internal delay equal to the period of the sound, and the position of this maximum is used as a cue to indicate its pitch. Meddis and Hewitt (1992) have suggested that autocorrelation patterns could also be used to identify vowel timbre in a pattern-matching model. These models are all based on various forms of correlation analysis, implemented by delay lines and excitatory-excitatory (EE)

interaction.

Other models involve inhibition. In the equalization-cancellation (EC) model of Durlach (1963), signals from both ears are subtracted after a delay (and an eventual amplitude compensation) rather than multiplied as in Jeffress's model. Subtraction allows interference to be canceled, so that a weak target can more easily be detected in certain binaural configurations. More recently, the same idea was applied to monaural processing of mixtures of sounds, such as simultaneous voices or musical sounds (de Cheveigné, 1993, 1997). Neural signals are delayed and processed by a neuron with both excitatory and inhibitory synapses, that calculate the neural equivalent of a difference. Periodic interference is thus suppressed, so that a weaker target can be perceived. Cancellation can also be used to explain pitch perception (de Cheveigné, 1998), and cascaded stages of cancellation can account for the perception of multiple pitches (de Cheveigné and Kawahara, 1999).

The present paper attempts to unify these models by casting them all within the same framework based on auto- and crosscorrelation. Certain requirements of cancellation models, such as fast inhibitory interaction, or cascaded stages of processing, can be relaxed, as exactly the same functionality is obtained with a single stage that involves only fast excitatory interaction. Besides unifying previous models, this framework allows new models to be derived.

2 A basic ingredient: correlation

The basic ingredient is set of arrays of autocorrelation and cross-correlation coefficients. Using a sampled-signal notation, the autocorrelation function of the signal at the left ear is calculated as:

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

where x^R is the signal at the left ear, τ the lag parameter and W is the size of the integration window. A similar function is calculated for the right ear:

$$r_t^R(\tau) = \sum_{j=t+1}^{t+W} x_j^R x_{j+\tau}^R \quad (2)$$

where x^R is the signal at the right ear. If the particular ear is not of importance (for models of effects with monaural or diotic presentation) the superscripts can be dropped. The crosscorrelation function is calculated as:

$$c_t(\theta) = \sum_{j=t+1}^{t+W} x_j^L x_{j+\theta}^R \quad (3)$$

where θ is the cross-correlation lag parameter. These functions are calculated for every time instant t. Because of the temporal smoothing produced by integration, the functions are not expected to fluctuate much, but for accuracy the model nevertheless requires them to be calculated at each instant. To keep things simple, peripheral filtering will not be mentioned explicitly in the following. Depending on the level of abstraction required, processing can be assumed to affect the raw acoustic waveforms (in an abstract model), or each individual filter channel (in a more detailed model).

3 Structure

The model involves three modules that are conceptually distinct. The first calculates the arrays of autocorrelation and crosscorrelation coefficients. The second forms a linear combination of selected coefficients. The third module controls the parameters of the second module while monitoring its output, and produces the predictions of the model: pitch, timbre, etc..

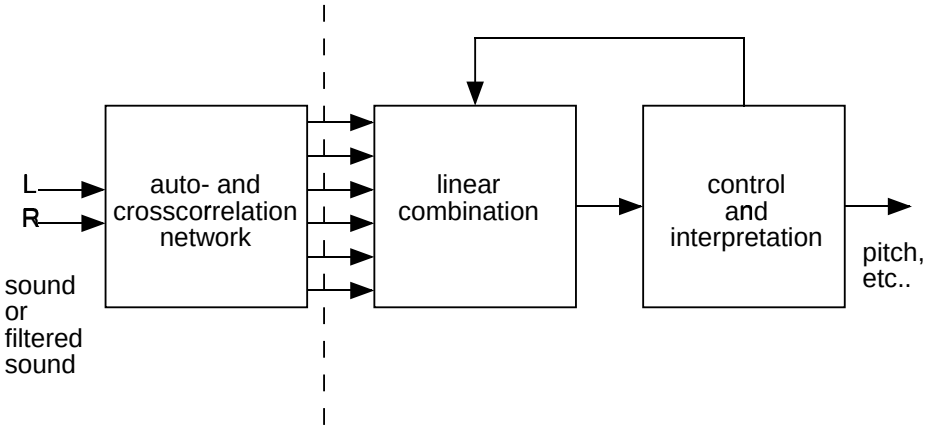

Fig. 1 *Structure of the Running Correlation Network model. Time-domain processing is limited to the first module (left of the dotted line). Subsequent processing operates on slowly-varying quantities.*

The first module performs a massively parallel computation of the correlation coefficients defined in Sect. 2. All coefficients (within a certain range of time and lag) are available simultaneously. The coefficients are temporally smoothed so subsequent modules process slowly-varying quantities. Fast time-domain processing is limited to the first module.

The second module calculates a linear combination of its inputs, with factors that can be positive or negative, integer or fractionnary.

The third module serves to control the parameters of the second module, and to analyze its output to produce whatever result (pitch, etc.) is expected from the model. The second and third modules are separated for conceptual reasons, but one can conceive of an implementation in which they are merged, for example as a neural network in which parameters are determined locally, for example by Hebbian learning.

This processing can be seen as occuring in parallel within all channels produced by peripheral filtering, and in both ears. It is also possible to assume, as an approximation, that the acoustic signals are processed directly. The following assumes this approximation except where noted. Operations are expressed using a digital sampled notation.

4 Particularizations of the model

A Autocorrelation model of pitch (Licklider, 1951)

An autocorrelation function is calculated from the acoustic signal, sampled at either ear, and the position of its maximum is used to indicate the period (and thus the pitch) of the source.

$$r_t(\tau) = \sum_{j=t+1}^{t+W} x_j x_{j+\tau} \tag{4}$$

Signal samples separated by a period are similar and thus tend to produce larger products than samples separated by other intervals. The integration smooths these values

temporally to produce a stable estimate. In terms of the present formalization, module 2 selects a term of the autocorrelation array (left or right or sum of both) as determined by the lag parameter τ. Module 3 varies this parameter while monitoring the output for a maximum.

B Autocorrelation model of vowel identification (Meddis and Hewitt, 1992)

In the concurrent vowel segregation model of Meddis and Hewitt (1992), a vowel's identity was determined by template matching of a summary pattern obtained by adding autocorrelation functions calculated within peripheral channels (see also de Cheveigné and Kawahara, 1999). This summary pattern can be approximated by the autocorrelation function of the waveform. In the present formalization, module 2 selects a term of the autocorrelation array as 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.

C Cancellation model of pitch (de Cheveigné, 1998)

This is a reformulation of Licklider's model in terms of cancellation. A squared difference function is calculated from the acoustic signal, and the position of its minimum is used to indicate the period. The function is defined as:

$$d_t(\tau) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau})^2 \quad (5)$$

The squared difference can be expanded:

$$\begin{aligned} d_t\tau) &= \sum_{j=t+1}^{t+W} x_j^2 + \sum_{j=t+1}^{t+W} x_{j+\tau}^2 - 2\sum_{j=t+1}^{t+W} x_j x_{j+\tau} \\ &= r_t(0) + r_{t+\tau}(0) - 2r_t(\tau) \end{aligned} \quad (6)$$

The difference function $d_t(\tau)$ required by the cancellation model can thus be implemented accurately in terms of autocorrelation. In the present formalization, module 2 forms this linear combination as determined by parameter τ. Module 3 varies τ while monitoring the output for a minimum.

D Cross-correlation model of localization (Jeffress, 1948)

The crosscorrelation function is calculated from signals from both ears, and the position of its maximum used to indicate the azimuth of the source. In terms of the present formalization, module 2 selects one term of the cross-correlation array as determined by interaural delay parameter θ. Module 3 varies this parameter while monitoring the output for a maximum.

E Equalization-cancellation model (Durlach, 1963)

Signals from left and right are equalized (by applying a delay and/or an amplitude factor to either signal) and subtracted to cancel the interference and allow a target to be detected more easily. Let us suppose that the decision statistic is based on energy:

$$d_t(\theta) = \sum_{j=t+1}^{t+W} (x_j^L - \alpha x_{j+\theta}^R)^2 \quad (7)$$

The squared sum can be developped and expressed in terms of left and right autocorrelations, and binaural crosscorrelation:

$$\begin{aligned} d_t(\theta) &= \sum_{j=t+1}^{t+W} (x_j^L)^2 + \alpha^2 \sum_{j=t+1}^{t+W} (x_{j+\theta}^R)^2 - 2\alpha \sum_{j=t+1}^{t+W} x_j^L x_{j+\theta}^R \\ &= r_t^L(0) + \alpha^2 r_{t+\theta}^R(0) - 2\alpha c_t(\theta) \end{aligned} \quad (8)$$

In terms of the present formalization, module 2 calculates this linear combination as 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.

Peripheral filtering was not part of Durlach's original model, which could be seen as applying to acoustic waveforms at both ears, or uniformly to each channel from the auditory periphery. Culling and Summerfield (1995) proposed a modified EC model (mEC) that departs from the original in two ways. The EC operation applies to individual filtered channels, and its parameters are based on a channel-specific criterion of minimum energy after cancellation, rather than a global criterion of maximum signal-to-noise ratio. The model can be used to detect the presence of a signal (according to the magnitude of the cancellation residual), or to produce a binaural pitch (according to the position along the tonotopic axis of a peak in cancellation residual), or to identify the timbre of a vowel (according to the pattern along a tonotopic axis formed by peaks of cancellation residual corresponding to formants).

F Multiple period estimation model (de Cheveigné and Kawahara, 1999)

The aim of the model is to estimate the periods of two or more concurrent periodic sounds. The model has two versions: iterative and joint estimation, that are described in the case of two concurrent sounds.

In the iterative version, the period of one sound is first estimated, and that estimate is used to suppress that sound so that the second period can be estimated. The second sound is then suppressed in turn, and the first period estimate is refined. The process may be repeated several times. Let us suppose that period estimation is based on autocorrelation. Calling $z_t = x_t - x_{t+T}$ the result of suppressing a sound based on an estimate T of its period, the autocorrelation function of z is:

$$\begin{aligned} r_t^z(\tau) &= \sum_{j=t+1}^{t+W} z_t z_{t+\tau} \\ &= \sum_{j=t+1}^{t+W} (x_t - x_{t+T})(x_{t+\tau} - x_{t+\tau+T}) \\ &= r_t(\tau) - r_t(T+\tau) - r_{t+\tau}(t+T-\tau) + r_{t+T}(\tau) \end{aligned} \quad (9)$$

In the present formalization, module 2 forms this linear combination of autocorrelation terms as 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..

In the joint estimation version of the model, the following difference function is calcu-

lated and the position of its minimum as a function of $(\tau 1, \tau 2)$ is used as an estimate of the periods:

$$d_t(\tau 1, \tau 2) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau 1} - x_{j+\tau 2} + x_{j+\tau 1+\tau 2})^2 \quad (10)$$

The squared difference can be expanded into a sum of autocorrelation terms

$$\begin{aligned} d_t(\tau 1, \tau 2) &= r_t(0) + r_{t+\tau 1}(0) + r_{t+\tau 2}(0) + r_{t+\tau 1+\tau 2} \\ &\quad -2r_t(\tau 1) - 2r_t(\tau 2) + 2r_t(\tau 1 + \tau 2) \\ &\quad +2r_{t+\tau 1}(\tau 2 - \tau 1) - 2r_{t+\tau 2}(\tau 1) - 2r_{t+\tau 1}(\tau 2) \end{aligned} \quad (11)$$

In terms of the present formalization, module 2 forms this linear combination as determined by parameters $(\tau 1, \tau 2)$. Module 3 varies these parameters while monitoring the output for a minimum. The search can be exhaustive (all pairs are tested) or iterative as before.

The model can be extended in principle to an arbitrary number of concurrent periodic sounds. The original formulation of the model called for a cascade of two or more cancellation filters, with impulse responses $h_1(t) = \delta(t) - \delta_{\tau 1}(t)$ and $h_2(t) = \delta(t) - \delta_{\tau 2}(t)$. In terms of a neural network model, these filters can be implemented by assuming a cascade of processing stages that each involves fast inhibitory interaction on a coincidence counting neuron. The present formulation avoids the need for cascaded processing. It also allows the fast inhibitory interaction to be replaced, if needed, by fast excitatory interaction only.

G Concurrent vowel identification model (de Cheveigné, 1993, 1997)

The aim of this model is to identify the weaker member of a pair of concurrent vowels. Supposing that the period T of the stronger vowel is known, the signal is passed through a cancellation filter with transfer function $h(t) = \delta(t) - \delta_T(t)$ to suppress that vowel, and an autocorrelation function is calculated from the result. The best match between the short-term part of the autocorrelation function and a set of templates determines the perceived identity of the weaker vowel. Calling $z_t = x_t - x_{t+T}$ the output of the first stage, its autocorrelation function is:

$$\begin{aligned} r_t^z(\tau) &= \sum_{j=t+1}^{t+W} z_t z_{t+\tau} \\ &= (x_t - x_{t+T})(x_{t+\tau} - x_{t+\tau+T}) \\ &= r_t(\tau) - r_t(T+\tau) - r_{t+\tau}(t+T-\tau) + r_{t+T}(\tau) \end{aligned} \quad (12)$$

In terms of the present formalization, module 2 forms this linear combination of autocorrelation terms as determined by parameters (τ, T). Module 3 sets T (as determined by one 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. Once again, the cascade of processing steps postulated by the original cancellation model de Cheveigné (1993, 1997) is avoided, while providing exactly the same function.

H Time-domain model of binaural pitch perception

This is a new model based on the present formalization. The aim is to explain binaural pitch phenomena such as those discussed by Culling et al. (1998a,b, 2000) with a fully time-domain model (as recently formulated by Akeroyd and Summerfield, 2000).

Binaural pitch phenomena that involve noise at both ears can usually be understood by interpreting the binaural stimulus as consisting of the superposition of an interfering source with high interaural correlation, together with a tonal target with low interaural correlation (or at least different from the interference). The interference is suppressed by the EC operation of Durlach (1963) or the mEC operation of Culling and Summerfield (1995). The amount of energy that survives cancellation, as a function of position along a tonotopic axis, constitutes a spectral pattern from which a pitch cue may be derived. This is one of the rare cases for which a spectral model must be invoked to explain pitch, which is paradoxical given that these phenomena have long been cited as evidence *against* spectral models of pitch.

It would be satisfying to apply a time-domain model of pitch instead. For example the output of the EC stage within each channel can be considered as a fast-varying time-domain signal, rather than a slowly varying tonotopic pattern. This output is then fed to a pitch model such as those mentioned above. A problem is that this puts strong constraints on neural processing (interaural cancellation must have a temporally accurate output, and pitch-period estimation must occur beyond that stage). The present framework allows these constraints to be removed.

Suppose that the signal from the right ear is delayed by θ and subtracted from the left to obtain an internal signal z_t that is free from interference and can be processed by the pitch model:

$$z_t = x_t^L - x_{t+\theta}^R \tag{13}$$

The signal z_t has a temporal fine structure that can be exploited to determine pitch, for example with an autocorrelation mechanism:

$$\begin{aligned} r_t^z(\tau,\theta) &= \sum_{j=t+1}^{t+W} z_t z_{t+\tau} \\ &= (x_t - y_{t+\theta})(x_{t+\tau}^L - x_{t+\tau+\theta}^R) \\ &= r_t^L(\tau) - c_t(\theta+\tau) - c_{t+\tau}(t+\theta-\tau) + r_{t+\theta}^R(\tau) \end{aligned} \tag{14}$$

The cascade of binaural cancellation followed by autocorrelation is replaced by a simple sum of autocorrelation and crosscorrelation terms. In the present framework, module 2 forms this linear 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, for example $z_t = x_t^L - \alpha x_{t+\theta}^R$. Eq. 14 then then becomes:

$$r_t^z(\tau,\theta) = r_t^L(\tau) - \alpha c_t(\theta+\tau) - c_{t+\tau}(t+\theta-\tau) + \alpha r_{t+\theta}^R(\tau) \tag{15}$$

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

$$d_t(\tau,\theta) = \sum_{j=t+1}^{t+W} (z_j - z_{j+\tau})^2$$

$$= \sum_{j=t+1}^{t+W} (x_j^L - x_{j+\theta}^R - x_{j+\tau}^L + x_{j+\theta+\tau}^R)^2 \quad (16)$$

The squared sum can be expanded:

$$\begin{aligned} d_t(\tau,\theta) &= r_t^L(0) + r_{t+\theta}^R(0) + r_{t+\tau}^L(0) + r_{t+\theta+\tau}^R(0) \\ &\quad -2c_t(\theta) - 2r_t^L(\tau) + 2c_t(\theta+\tau) + 2c_{t+\tau}(\theta-\tau) \\ &\quad -2r_{t+\theta}^R(\tau) - 2c_{t+\tau}(\theta) \end{aligned} \quad (17)$$

This allows joint estimation of both parameters by searching for a global minimum. In the present framework, module 2 forms this combination with parameters (θ, τ). Module 3 varies these parameters while monitoring the output for a minimum.

5 Transposing to a physiological model

This section tries to work out the details involved in deriving a physiological model from the present framework.

A The role of peripheral analysis

The discussion so far gave no role to peripheral analysis. Peripheral analysis is an important stage of auditory processing, and so one would like to know (a) if the models still work when that stage is included, and (b) whether the system benefits in terms of functionality by doing so.

1 Do the models work if peripheral analysis and transduction are included?

Setting aside the non-linear transduction process and considering only the linear filtering stage, we can remark that this stage can be swapped with any other linear stage. For example a cancellation filter such as:

$$h_\tau(t) = \delta(t) - \delta_\tau(t)) \tag{18}$$

can indifferently be placed before the filterbank (ie applied to the acoustic waveform) or after the filterbank. In terms of power, it is obvious that of the power of $x_t - x_{t+\tau}$, as measured in Eq. 5, is zero, then so is the power of every channel of the filterbank. Thus it makes no difference to the cancellation models (and the models that are derived from them) whether cancellation is performed ideally before peripheral filtering, or after.

However we know that peripheral frequency analysis is followed by a non-linear process within the haircell by which mechanical motion is transduced to instantaneous discharge probability within auditory nerve fibers. It remains to be seen whether the models can still work with this (and other) intervening stage of non-linearity.

There are several reasons to be optimistic. A first is that the dispersive properties of the basilar membrane produce a sharp phase shift (180 degrees ?) for each frequency in the vicinity of its resonance. This in effect allows both polarities to be coded. A second is that simulations have shown that cancellation-type models are effective despite the non-linearities (de Cheveigné, 1993, 1997). A third is that, by splitting power and information over channels with different frequency responses, peripheral filtering effectively linearizes the system [dig up paper by Slaney that more or less says so].

Nevertheless, detailed simulations are necessary to determine how non-linearity of transduction and neural processing affect this sort of processing.

2 What is the functional advantage of peripheral filtering?

Parallel processing of peripheral filter outputs may offer several functional advantages, over direct processing of the acoustic waveform.

The first is the "linearizing" property mentioned above. Non-linear transduction is likely to reduce the effectiveness of neural processing relatively to an ideal linear model, so a scheme that restores the linearity is useful.

A second advantage is related to the amplitude normalization ("automatic gain control") properties of non-linear filtering and haircell transduction. The processing models described above are based on the autocorrelation function, which is the Fourier transform of the power spectrum. The power spectrum puts strong emphasis on high-amplitude parts of the spectrum, and therefore the autocorrelation function is relatively insensitive to details of lower-amplitude parts of the spectrum, which may nevertheless be perceptually important. Peripheral filtering followed by amplitude normalization produces a more balanced representation, closer for example to that provided by the logarithmic spectrum, or the cepstrum that has proven useful for speech analysis.

A third advantage is that individual channels can be selected. If part of the spectrum is dominated by noise or competing sources, analysis restricted to certain channels may be more effective than analysis of the entire spectrum, or of the acoustic waveform. Even in the case that target and interference overlap within the spectrum, the signal-to-noise

ratio may be more favorable within certain channels. This may be crucial if, as one can imagine, neural signal processing has a limited dynamic range.

B Implementation of module 1

Brainstem?

C Implementation of module 2

Midbrain?

D Implementation of module 3

To what extent can parameters be set according to hebbian principles?

6 A role for fast inhibitory interaction?

Relative to previous cancellation models, the present framework dispenses of two requirements: fast inhibitory interaction, and cascaded processing (for complex models such as multiple pitch, or binaural pitch). The fact that they are not required does not mean that they must be ruled out. In particular fast inhibitory interaction, if available, can be used to simplify certain linear combinations. For example, certain terms of Eq. 12 can be grouped by three and replaced by difference terms defined by Eq. 6:

$$\begin{aligned} d_t(\tau 1,\tau 2) &= d_t(\tau 1) + d_{t+\tau 1}(\tau 2) \\ &\quad -2r_t(\tau 2) + 2r_t(\tau 1+\tau 2) + 2r_{t+\tau 1}(\tau 2-\tau 1) - 2r_{t+\tau 2}(\tau 1) \end{aligned} \tag{19}$$

With difference terms, the right hand side has six terms instead of ten, implying a simpler linear combination in module 2. Thus fast inhibitory interaction can be of use, even if it is not required.

7 Exotic models

The following models address more complex tasks. The aim here is to explore the capabilities of the framework more than to model specific auditory mechanisms.

A Variable amplitude period sound

Many musical sounds can be approximated by a periodic signal with a time-varying amplitude. Single-period estimation models are relatively robust with respect to amplitude variations, but a model that more closely matches this kind of signal can be useful. Supposing that the signal x_t is such that $x_{t+T} = ax_t$ for all t in the vicinity of the analysis, then the following function is zero for $(\tau, \alpha) = (T, a)$:

$$\begin{aligned} d_t(\tau,\alpha) &= \sum_{j=t+1}^{t+W} (x_j - \alpha x_{j+\tau})^2 \\ &= r_t(0) + \alpha^2 r_{t+\tau}(0) - 2\alpha r_t(\tau) \end{aligned} \quad (20)$$

If α is known (as may be the case for instruments with stereotyped temporal envelopes), estimation with this model may be more accurate and reliable than with the standard, fixed amplitude model. If α is unknown, then it can be estimated by minimizing $d(\tau, \alpha)$. In the current formalization, module 2 forms the linear combination of Eq. 20 as determined

by parameters (τ, α). Module 3 varies these parameters while monitoring the output for a minimum.

However, a variational argument shows that, for a given τ, d_t is minimum as a function of α if:

$$\alpha = r_t(\tau)/r_{t+\tau}(0) \tag{21}$$

which can be substituted in Eq. 20:

$$d_t(\tau) \;=\; r_{t+\tau}(0)(1 - r_t(\tau)^2/r_t(0)r_{t+\tau}(0)) \tag{22}$$

Thus it is necessary only to explore the search space of τ. As before, the right hand side is a function of autocorrelation terms, but it is no longer a linear combination. To fit the present formalization, module 2 must be extended to implement a wider class of functions.

B Concurrent, variable amplitude periodic sounds

Whereas single-period estimation is tolerant of amplitude variations, multiple-period estimation is much more fragile. The period of an amplitude-varying sound is hard to estimate in the presence of a concurrent sound, and once its period has been estmated the sound itself is hard to cancel, so the period of the other sound is also hard to estimate. The iterative model of Sect. F can be adapted to the case where one or the other sound, or both, vary in amplitude.

Supposing that the signal x_t is the sum of a periodic sound x'_t with period $T1$, and a variable-amplitude periodic sound x''_t such that $x''_{t+T2} = ax''_t$ for all t, then the following

function is zero for $\tau 1 = T1$, $\tau 2 = T2$, $\alpha = a1$:

$$d_t(\tau 1, \tau 2, \alpha) = \sum_{j=t+1}^{t+W} (x_j - x_{j+\tau 1} - \alpha x_{j+\tau 2} + \alpha x_{j+\tau 1+\tau 2})^2 \quad (23)$$

The squared sum can be developped:

$$\begin{aligned} d_t(\tau 1, \tau 2, \alpha) &= r_t(0) + r_{t+\tau 1}(0) + \alpha^2 r_{t+\tau 2}(0) + \alpha^2 r_{t+\tau 1+\tau 2}(0) \\ &\quad -2r_t(\tau 1) - 2\alpha r_t(\tau 2) + 2\alpha r_t(\tau 1 + \tau 2) \\ &\quad +2\alpha r_{t+\tau 1}(\tau 2 - \tau 1) - 2\alpha r_{t+\tau 2}(\tau 1) - 2\alpha^2 r_{t+\tau 1}(\tau 2) \end{aligned} \quad (24)$$

If α is known, then search for $T1$ and $T2$ may proceed as in the constant amplitude case (Sect. F). If α is unknown, it can be derived as in the previous section by setting the derivative of $d_t(\tau 1, \tau 2, \alpha)$ with respect to α to zero:

$$\begin{aligned} \alpha = &[r_t(\tau 2) - r_t(\tau 1 + \tau 2) - r_{t+\tau 1}(\tau 2 - \tau 1) + r_{t+\tau 2}(\tau 1)] \\ &/[r_{t+\tau 2}(0) + r_{t+\tau 1+\tau 2}(0) - 2r_{t+\tau 1}(\tau 2)] \end{aligned} \quad (25)$$

This value can then be inserted in Eq. 24 to obtain an equation (not shown) that depends only on $\tau 1$ and $\tau 2$. Thus it is necessary only to explore the search space of $(\tau 1, \tau 2)$. As in the previous example, the expression is a function of autocorrelation terms but not a linear combination, so it does not fit the present formalization unless module 2 is allowed to include functions other than linear.

This example can be extended to the case where the signal is the sum of more than two periodic components, one of which varies in amplitude. It can also be extended to the case where the signal is the sum of two periodic signals that both vary in amplitude, with parameters $\alpha 1$ and $\alpha 2$. In that case, however, the expression obtained for $\alpha 1$ involves $\alpha 2$

and vice-versa, so minimization is more complex. In any case, these examples show that the basic model is quite flexible.

C 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. 6, 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 formalization, 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).

8 Signal processing applications

A Fundamental frequency estimation

This framework has been successfully applied to the task of speech fundamental frequency estimation, yielding error rates much lower than competing methods (de Cheveigné and Kawahara, 2001).

B Speech processing

The examples given above showed that the autocorrelation function of a source can be isolated after removal of contributions of interfering sources. From the autocorrelation function one can derive the power spectrum, from which quantities such as the cepstrum, MFCC coefficients, etc. can be derived. The autocorrelation function can also be used to derive LPC coefficients. Thus, the framework could be applied to speech recognition within acoustically cluttered scenes, by simultaneous suppression of interference (based on harmonicity or spatial cues) and analysis of target cues.

One word of caution: the filtering that suppresses the interference also suppresses certain components of the target. The target autocorrelation function that is retrieved is not equal to that of the target in isolation. However we note that (a) all interference contributions are removed, and (b) the distortion is known (or can be modeled) and taken into account when the target is processed, for example using missing-data techniques (Cooke et al., 1997).

C Relations to blind separation and ICA

TBD

9 Discussion

The framework unifies a series of time-domain models of auditory perception, and provides a basis from which to derive more. All fast signal processing is done at an early stage.

This is an attractive feature, as it is known that within the auditory nervous system, accurate time-domain is not transmitted beyond the level of the brainstem. This stage is simple and involves only delay lines and coincidence counters. Ample evidence for these ingredients has been found (in the MSO of mammals or nucleus laminaris of birds), although there is yet no firm evidence for delay lines as long as required for periodicity analysis.

Sophisticated processing is deferred to a later stage, where it can be performed on slowly-varying quantities, rather than on signals with fine time structures. The price to pay is complexity (of the linear combinations) but this is perhaps not such a drawback if one considers the complexity of neural circuitry. It is hard to match complex neural circuitry with simple models.

The framework offers the option to avoid fast inhibitory interaction, although it does not prohibit using it if available. More importantly, it avoids the cascaded processing postulated by some cancellation models. Cascaded time-domain processing puts strong constraints on neural anatomy and physiology, and it is welcome that exactly the same functionality can be provided, if necessary, without those constraints.

References

- 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.

- Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust

speech recognition.", Proc. ICASSP, 863-866.

- 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.

- Culling, J. F., Marshall, D., and Summerfield, Q. (1998). "Dichotic pitches as illusions of binaural unmasking II: the Fourcin pitch and the Dichotic Repetition Pitch," J. Acoust. Soc. Am. 103, 3525-3539.

- 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.

- Culling, J. F. (2000). "Dichotic pitches as illusions of binaural unmasking. III. The existence region of the Fourcin pitch," J. Acoust. Soc. Am. 107, 2201-2208.

- 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.

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

- de Cheveigné, A. (2001). "Running autocorrelation model of F0 estimation.", Proc. ASA.

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

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

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

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

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

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