Cocktail-party processing based on the Correlation Network

Alain de Cheveigné, CNRS and Ircam.

Hideki Kawahara, Media Design Informatics Group, Design Information Science Dept. Faculty of Systems Engineering, Wakayama University.

Alexis Baskind, Ircam.

Send all correspondence to:

Alain de Cheveigné, Ircam - CNRS, 1 place Igor Stravinsky, 75004, Paris.

cheveign@ircam.fr

Last revised: 27 Aug 2002

Number of pages (including figures):

Number of Figures:

Number of Tables:

Abstract (English)

This paper attempts to apply ideas from a recent model of auditory processing, the Correlation Network (CN) model, to the task of estimating spectral parameters of concurrent voices or instruments. The CN model is an abstract model for monaural and binaural auditory processing (pitch, timbre, localization and segregation) comprising three modules. The first calculates arrays of running correlation coefficients (monaural autocorrelation and binaural crosscorrelation). The second module forms a weighted sum of the correlation terms produced by the first module. The third module controls the parameters of the second module while monitoring its output, and is responsible for producing the model's behavior (i.e. predict pitch, timbre, localization, etc.). By adjusting the second and third modules, the model may be configured to account for a wide range of tasks performed by the auditory system. The signal processing techniques described in this paper are the transposition of the ideas behind that model. An initial processing module calculates arrays of running autocorrelation and/or crosscorrelation coefficients indexed by lag, time and window size. On the basis of these coefficients, it is possible to obtain accurate estimates of the fundamental frequencies (F_0s) and spectral envelopes of each voice or instrument within a mixture. Estimates are obtained with good accuracy and spectral and temporal resolution, although in some cases they may be incomplete due to collisions between harmonics. Incomplete estimates may be used directly for weighted pattern-matching, or else missing values may be reconstructed using a model.

Abstract (Français)

Cet article essaye d'appliquer les principes d'un modèle récent de traitement auditif, le modèle Réseau de Corrélation (RC), à la tâche d'estimer les caractéristiques spectrales de plusieurs voix ou instruments concurrents. Le modèle Réseau de Corrélation (RC) est un modèle abstrait de traitement monaural et binaural (hauteur, timbre, localisation et ségrégation) comprenant trois modules. Le premier calcule des séries de coefficients de corrélation (autocorrélation monaurale et corrélation croisée binaurale). Le second forme une somme pondérée des coefficients calculés par le premier. Le troisième contrôle les paramètres du second tout en observant sa sortie, et assure le comportement attendu pour chaque tâche (c.a.d. prédire la hauteur, timbre, localisation, etc.). En modifiant les deuxième et troisième modules, le modèle RC peut être configuré pour rendre compte d'une large gamme de tâches effectuées par le système auditif. Les techniques de traitement de signal décrits dans cet article sont la transposition des idées du modèle RC. Un premier module calcule des séries de coefficients d'autocorrélation et/ou corrélation croisée, indexées par le délai, le temps et la taille de fenêtre. Sur la base de ces ingrédients, on peut obtenir des estimations des fréquences fondamentales (F_0) et enveloppes spectrales de chaque voix ou instrument d'un mélange. Ces estimations sont précises et ont une bonne résolution temporelle et fréquentielle, mais dans certains cas elles peuvent être incomplètes du fait de collisions entre harmoniques. Des estimations incomplètes peuvent être utilisées directement pour la reconnaissance de formes pondérée, ou alors des estimations complètes peuvent être reconstruites à l'aide d'un modèle.

1 Introduction

This paper addresses the task of analyzing the speech of several overlapping speakers, a task for which human listeners still behave better than machines (Lippmann, 1997). The expression “cocktail party effect” was coined by Cherry (1953) to designate the ability of listeners to follow a voice in the presence of background noise or competing speakers (Brokx and Nooteboom, 1982; Scheffers, 1983; Darwin and Culling, 1990). It is an example of the more general process of Auditory Scene Analysis (ASA) (Bregman (1990). There have been efforts to replicate ASA mechanisms computationally in what are known as Computational Auditory Scene Analysis (CASA) systems. Pioneering work was done by Parsons (1976), Lyon (1984) or Weintraub (1985), and recent efforts are reviewed by and Cooke and Ellis (2001).

Among the strongest cues for perceptual segregation is *harmonicity* such as occurs in voiced portions of speech. Pairs of concurrent voices are more intelligible if they have different F_0s (Brokx and Nooteboom, 1982; Scheffers, 1983), in which case there are two distinct time-varying harmonic structures, one for each voice. F_0 cues appear to help in at least two ways: at each instant they reduce the masking of one voice by the other, and over time they allow various parts belonging to the same voice to be grouped together (Darwin and Hukin, 2000). For the first role (instantaneous unmasking), it seems that the important factor is not the harmonic structure of the target voice, but rather that of the masker (Lea, 1992; Summerfield and Culling, 1992, de Cheveigné et al., 1995). Many systems have tried to use harmonic structure to separate mixed speech, or to improve the performance of automatic speech recognition (ASR) systems on mixed speech (Parsons,

1976; Weintraub, 1985; Stubbs and Summerfield, 1988, 1990, 1991; Nakatani and Okuno, 1995; etc.).

Harmonic structure is important also for *spectral estimation*, even in the case of a single voice. Supposing that speech is produced according to a simplified source-filter model (a slowly time-varying filter excited by a source with a flat spectral envelope), the goal of spectral estimation is to recover the spectrotemporal envelope of the filter, independently from the source. If the source is periodic, the filter's transfer function is sampled at harmonics of F_0, and this creates a number of difficulties. Estimates usually have a residual harmonic structure and temporal fluctuations. These may be reduced by smoothing, but depending on the degree of smoothing the spectra may still show ripples along the time and/or frequency axes, or else important spectrotemporal features may be blurred. The tradeoff is F_0-dependent, and indeed the best estimates are obtained with F_0-adaptive smoothing (Kawahara, 1997; Kawahara et al., 1999). However even such optimal smoothing is subject to aliasing effects due to sparse sampling of the original spectral envelope. If the task is pattern-matching, aliasing can be avoided by removing the smoothing stage and using the F_0-based samples directly (de Cheveigné and Kawahara, 1999b).

Spectral estimation is obviously more difficult when several sources are present because statistics derived from the waveform (autocorrelation function, spectrum, etc.) reflect more than one source. This paper applies ideas embodied in a recently-proposed auditory processing model, the Correlation Network model (de Cheveigné, 2001). The CN model is first outlined, then its basic principles are transposed to the practical task of estimating spectra of two mixed voices.

2 Correlation Network model

The Correlation Network (CN) model is described in detail in de Cheveigné (2001) and de Cheveigné and Pressnitzer (2002), and is only outlined here. It attempts to unify a number of models of auditory signal processing for pitch (Licklider, 1959; de Cheveigné, 1998; de Cheveigné and Kawahara, 1999a; Akeroyd and Summerfield, 2000), timbre (Meddis and Hewitt, 1992), localization (Jeffress, 1948), and segregation, either binaural (Durlach, 1963) or monaural based on harmonicity (de Cheveigné, 1993, 1997). These models have in common that they use second-order statistics of various combinations of signals from one or both ears.

2.1 Basic ingredient: correlation

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

$$r_{L}(\tau,t) \quad = \quad (1/W)\sum_{i=t+1}^{t+W} x_{L}(i)x_{L}(i-\tau) \tag{1}$$

where $x_{L}(t)$ is the signal at the left ear, τ the 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 $r_{R}(\tau,t)$ is calculated for the right ear (the subscript may be dropped for monaural models). The interaural CC function is calculated as:

$$c(\theta,t) = (1/W)\sum_{i=t+1}^{t+W} x_{L}(i)x_{R}(i-\theta) \tag{2}$$

where θ is the crosscorrelation lag parameter. These functions are calculated for every time instant t, using a sliding window. Depending on the level of abstraction required, processing is assumed to affect the raw acoustic waveforms (in a simplified model) or each peripheral filter channel (in a more detailed model). To keep things simple, the former is assumed except where noted.

2.2 Structure of the CN model

The CN 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. The third module is also responsible for the behavior of the model (pitch, timbre, etc.).

[Fig. 1 about here]

Module I delivers all AC and CC coefficients within a certain range of time and lag. As the coefficients are temporally smoothed (integration in Eqs. 1 and 2), modules II and III process only slowly-varying quantities. Module II forms a linear combination with factors that may be positive or negative. Module III controls these factors while monitoring the output of module II. The three modules are distinct for conceptual reasons, but they might be merged within a neuronal implementation.

2.3 Particularizations

The CN model may be used as a common processing substrate to implement several known auditory processing models. A feature common to these models is that they involve *second-order statistics* (power or autocorrelation) of time-domain patterns of auditory nerve fiber activity (or in more abstract models: acoustic waveforms).

As a first example, the autocorrelation model of Licklider (1951) is currently a popular explanation for pitch perception (Meddis and Hewitt, 1991, Cariani and Delgutte, 1996). The model assumes that AC functions are calculated within each channel of a peripheral filter bank, but in a simplified account it may be seen as operating on the AC function of the acoustic waveform at either ear (Yost, 1996). The maximum of the AC function as a function of lag τ serves as a cue to pitch. This model may be implemented trivially in terms of the CN model: module I calculates the AC function, module II selects one value with parameter τ, and modules III varies τ while monitoring the output of module II for a maximum.

As a second example, the binaural detection model of Durlach (1963) explains why detection is sometimes better with two ears than with one. The model (known as the Equalization-Cancellation, or EC model) assumes that the signal from one ear is internally delayed by a lag θ and scaled by a factor α (equalization) and subtracted from the signal from the other ear (cancellation). The decision statistic is the *power* of the remainder:

$$d(\theta, \alpha, t) = \sum_{i=t+1}^{t+W} \left(x_{_L}(i) - \alpha x_{_R}(i-\theta)\right)^2 \tag{3}$$

Expanding the squared sum on the right hand side, and distributing the summation over terms, this statistic may be expressed as a combination of monaural AC and binaural CC

terms:

$$d(\theta, \alpha, t) \quad = \quad r_{_L}(0,t) + \alpha^2 r_{_R}(0, t-\theta) - 2\alpha c(\theta, t) \tag{4}$$

In the framework of the CN model, module I calculates the necessary terms and module II combines them as in Eq. 4 with parameters θ and α. Module III adjusts these parameters to obtain the best signal-to-noise ratio, and then monitors the output of module II for the presence of a signal. Durlach's original model processed the *acoustic* waveforms at both ears, but more recent models assume that similar processing occurs within narrow-band channels produced by cochlear filtering (Culling and Summerfield, 1995; Breebart, 2001).

Other models of auditory processing may be implemented within the framework of the CN model, as summarized in Table 1. There are several advantages in doing so. First, it is not necessary to postulate radically different processes for different functions: the same structure serves for all. This makes sense if we view the auditory system as a general-purpose processor that evolved to address a wide range of sound-processing tasks. Second, "fast" signal processing is confined to the first module, while the second and third deal with more slowly varying quantities. This may facilitate the mapping between these functional models and auditory anatomy and physiology. It is known that accurate temporal synchrony is mainly found at the level of the brainstem, but not at higher levels. Third, the CN model avoids the need to assume a cascade of accurate time domain processing, that certain models require (de Cheveigné, 1997, de Cheveigné et al. 1999, Akeroyd and Summerfield, 2000). That assumption is expensive in terms of physiology, and it is useful to know that the same functionality can be attained without that assumption.

The reader is referred to de Cheveigné (2001) and de Cheveigné and Pressnitzer

(2002) for a more detailed discussion. The rest of this paper attempts to transpose the essential ideas behind the CN model to practical tasks of signal processing for speech and music.

Table 1 *Eight time-domain models of auditory processing. Each uses a quadratic statistic (column 2) of a linear combination of delayed inputs from one or both ears. These may be reformulated as combinations of correlation terms calculated from the raw waveforms (column 3). Notation:* $x(t)$ *is either the acoustic waveform (in simple models), or the temporal pattern within one channel of the auditory periphery (in more detailed models). Subscripts* L *and* R *indicate the left and right ear respectively, while the lack of subscript indicates that the ear is indifferent. The term* $r(\tau, t)$ *represents the autocorrelation function calculated for lag* τ *at time* t*, while* $c(\theta, t)$ *is the crosscorrelation function at interaural delay* θ *and time* t*. The factor* α *is applied internally to one ear to compensate for external amplitude differences. The first module of the CN model calculates the necessary AC and CC coefficients, while the second module forms their linear combination (column 3).*

Model	*Statistic*	*Linear combination in Module 2*
Licklider (1958), pitch	autocorrelation of $x(t)$	$r(\tau,t)$
Meddis and Hewitt (1992), vowel timbre	autocorrelation of $x(t)$	$r(\tau,t)$
Jeffress (1948), localization	crosscorrelation of $x_{L}(t)$ and $x_{R}(t)$	$c(\theta,t)$
Durlach (1963) binaural detection	power of $x_{L}(i)-\alpha x_{R}(i-\theta)$	$r_{L}(0,t)+\alpha^2 r_{R}(0,t-\theta)-2\alpha c(\theta,t)$
de Cheveigné (1998), pitch	power of $x(t)-x(t-\tau)$	$r(0,t)+r(0,t-\tau)-2r(\tau,t)$
de Cheveigné and Kawahara (1999) multiple pitches	power of $[x(t)-x(t-\tau)]-[x(t-\upsilon)-x(t-\tau-\upsilon)]$	$r(0,t)+r(0,t-\tau_1)+r(0,t-\tau_2)+r(0,t-\tau_1-\tau_2)-2r(\tau_1,t)-2r(\tau_2,t)+2r(\tau_1+\tau_2,t)+2r(\tau_2-\tau_1,t-\tau_1)-2r(\tau_1,t-\tau_2)-2r(\tau_2,t-\tau_1)$
de Cheveigné (1993), vowel segregation	autocorrelation of $x(t)-x(t-T)$	$r(\tau,t)-r(\tau-T,t-T)-r(\tau+T,t)+r(\tau,t-T)$
Akeroyd and Summerfield (2000), binaural pitch	autocorrelation of $x(t)-x(t-\tau)$	$r_{L}(\tau,t)-\alpha c(\tau-\theta,t-\theta)-\alpha c(\tau+\theta,t)+\alpha^2 r_{R}(\tau,t-\theta)$

3 "Cocktail-party" signal processing

The aim is to characterize one or more sources on the basis of one or more signals in which they are superposed. Tasks are accurate F_0 estimation and spectral envelope estimation for each source.

As a starting point, we assume the existence of a general purpose module that calculates arrays of correlation coefficients from the input signals. This module, dubbed "Correlation Machine", is the equivalent of module I of the CN model, with two additional refinements that were not in the original CN model. The first is that coefficients are calculated with *variable window size* (W) and indexed by that value in addition to time and lag. The second is that more than two inputs are allowed (although we mostly consider the single-channel case). This module is described in the next section. Subsequent sections describe processing schemes that use the coefficients that it produces.

3.1 The "Correlation Machine"

Suppose that N input signals $(x_n(t))$ are transduced by N microphones. From them, all possible combinations of running CC and AC coefficients are calculated:

$$r_{jk}(\tau, t, W) \quad = \quad (1/W) \sum_{i=t+1}^{t+W} x_j(i) x_k(i-\tau) \tag{5}$$

The calculation is performed for all values of window size W within a given range, so the coefficients are indexed by W in addition to τ and t. For N input signals there are N^2 such arrays of correlation coefficients (Fig. 2). To simplify notation, channel indices are merged when they are the same (e.g. r_j rather than r_{jj}) or dropped when not necessary. W and t may also be omitted.

[Fig. 2 about here]

The set formed by these coefficients (indexed by channel pair, time, lag, and window size) is the basic ingredient used by algorithms described in rest of this paper. The module is expected to support queries for non-integer indices using interpolation. Coefficients may be calculated directly "on the fly", or precalculated and buffered for efficiency: scheduling is under the responsability of the module.

Implementation issues are not our main concern, but it is worth considering the processing costs. Coefficients are typically needed for values of lag τ and window size W within a certain range, and for arbitrary values of t. Integration over the W-sized window (Eq. 5) represents the major computational cost as it must be repeated for each index. For fixed W, the cost may be reduced by calculating coefficients for consecutive values of t as a moving average. For variable W, the following scheme may be used. For each (j, k, τ), denote as $p_{jk}(\tau, t)$ the product $x_j(t)x_k(t-\tau)$ and consider it as a time series. Replace each even-order coefficient by the sum of two consecutive coefficients ($p_{jk}(2t) \leftarrow p_{jk}(2t) + p_{jk}(2t+1)$). Repeat the process for pairs of even-order coefficients ($p_{jk}(4t) \leftarrow p_{jk}(4t) + p_{jk}(4t+2)$), and so on with increasing powers of two, up to some order that depends on the maximum window size required. Once this has been done, for any W the coefficient $r_{jk}(\tau, t, W)$ may be obtained as a sum of (on average) $log2(W)/2$ terms. Using such techniques, it is possible to keep computation and storage costs within reason. This much having been said, we consider implementation issues no further.

Much of this paper deals with the single-channel case ($N = 1$) for which the module

produces a single array $r(\tau, t, W)$ of running autocorrelation coefficients:

$$r(\tau, t, W) \quad = \quad (1/W) \sum_{i=t+1}^{t+W} x(i)x(i-\tau) \tag{6}$$

Note that this definition differs from that of the short-term autocorrelation function:

$$r'(\tau, t, W) \quad = \quad (1/W) \sum_{i=t+1}^{t+W-\tau} x(i)x(i-\tau) \tag{7}$$

The latter definition is more common, in particular because it is efficiently calculated by FFT. However the former definition has the property that the AC function of a linear combination of signals is the linear combination of their AC functions (Papoulis, 1984, p. 300). That property, valid for true AC functions (defined over all time), holds also for running autocorrelation as defined by Eq. 6 but not for short-term autocorrelation as defined by Eq. 7. That is why we use Eq. 6.

3.2 Single and two-voice F_0 estimation

Algorithms for single-voice and multiple-voice F_0 estimation are described by de Cheveigné and Kawahara (1999a, 2002). They work by searching within the parameter space of a linear combination of delayed versions of the input signal for a minimum of output power (Fig. 3). The necessary statistics can be expressed in terms of input AC coefficients, and thus they fit well within the present framework. The methods may be extended to handle variable amplitude periodic sources, and to estimate the amplitude variation factors together with the periods (de Cheveigné and Kawahara, 2002). The reader is referred to those papers for details. Using these algorithms or others (e.g. Kawahara, 1999), accurate period estimates may be obtained with subsample resolution. We therefore assume available the F_0 estimates required by the following algorithms.

[Fig. 3 about here]

3.3 Spectrum estimation and the AC function

Algorithms in this paper produce AC functions that approximate the AC functions of target sources. Short-term and running AC functions are both used in linear predictive coding (LPC) analysis (the latter is usually called "autocovariance" in this context; Markel and Gray, 1976). The true AC function (calculated over all time) and power spectrum form a Fourier transform pair, so if the former is given the latter can be derived, and from it other spectral or cepstral representations. The same is true of the short term AC function calculated according to Eq. 7 with respect to the short term power spectrum. The running AC function (Eq. 6) does not allow such a simple relation in general, but if the signal is periodic *and* if calculations are performed with a window of size W multiple of the period, then the running AC function is equal to the true AC function. In that case the power spectrum can be derived exactly.

3.4 The spectral envelope of a periodic sound

Suppose that the observed signal $x(t)$ was produced by a slowly-varying filter excited by a periodic source with period T and a flat spectral envelope. The goal is to estimate the amplitude transfer function of the time-varying filter, independently from the source, with the best possible time and frequency resolution. There are several obstacles to this goal. One is the well-known tradeoff between spectral and temporal resolution, another has to do with the particular problems created by periodic excitation.

The time-frequency tradeoff is a less serious problem than commonly thought. Articulatory and auditory constraints, respectively, limit the spectro-temporal detail of what can be produced and perceived. Certain features (e.g. plosive bursts) may require good temporal resolution and others (e.g. low-frequency formant patterns) good spectral resolution, but we have the option to perform several analyses or to use non-uniform transforms such as wavelets. For speech it is common to use representations with *less* resolution than allowed by the time-frequency tradeoff (e.g. mel-scale filter bank coefficients). The time-frequency resolution limit is not a major concern.

Periodic excitation, on the other hand, is both an advantage and a drawback. The deterministic nature of periodic excitation avoids the random error inherent in noise excitation, and additive noise may be reduced by averaging over periods. On the other hand, the spectral envelope of the filter is *sampled* at harmonics of F_0, and information encoded in the waveform is limited to those sample points (Fig. 4). Other aspects of the filter shape are missing and cannot be retrieved. Nevertheless, spectral analysis methods generally produce "complete" envelopes as a result of interpolation or smoothing. This implies "guessing" unknown values, and the result may differ from the true envelope. For applications that don't require a complete envelope, such as pattern matching, it may be better to skip smoothing and use directly the estimated harmonic sample points. For applications that do require a complete envelope, it is arguably best to perform interpolation explicitly on their basis (for example using a model of the production process), rather than rely on the smoothing implicit in the estimation process. In both cases, accurate estimation of the sample points should be the goal.

Envelopes estimated by standard methods commonly have a superimposed fine struc-

ture that consists of spectral or temporal "ripples" related to the periodic excitation. These are particularly troublesome for resynthesis, as their interaction with the harmonic structure of the new source ("moiré effect") is a source of noise. Fine structure is usually attenuated by smoothing (based on the assumption that the original envelope was smooth, which may or may not be true). Heavy smoothing may eliminate useful spectrotemporal features, while light smoothing may leave F_0-related structure. The optimal amount of smoothing depends upon F_0. Indeed, when high-quality F_0-adaptive smoothing is applied (as in the STRAIGHT system of Kawahara et al., 1999), remarkably good quality synthetic speech results.

3.5 F_0-adaptive spectral estimation based on the AC function

Accurate F_0-adaptive spectrum estimation can be performed using the AC function. Stages of the estimation algorithm are outlined in Fig. 5 (a). First, the period T is estimated. Second, AC coefficients are calculated with an integration window size of *one period*, for a range of lags from 0 to T. Window size and lag range are illustrated respectively as full and dotted lines below the waveform in Fig. 5 (b). Third, the AC coefficient arrays are resampled (by cubic or spline interpolation) so as to cover the period exactly with an integer number of samples (c). For computational convenience this number may be chosen equal to a power of two. Fourth, the power spectrum is calculated by DFT (eventually FFT or fast DCT). The result is a set of (frequency, amplitude) pairs that constitute the estimate (d).

If the signal is periodic, integration over a T-size window produces an unbiased estimate of the true AC function. The DFT then produces exact estimates of the power of

each partial. The amount of signal necessary to obtain this result, $2T$, can be reduced to $3T/4$ by noting that the AC function of a periodic function is even, and thus determined by its values over $[0, T/2]$. This determines the shortest duration of a locally periodic segment for which exact estimates can be obtained using this method.

If the signal is imperfectly periodic, the notions of partial or "envelope sample" are approximations. The previous method may give poor results as a result of the spectral splatter effects that arise when using square windows. These may be reduced by applying a raised-cosine window (hanning) to a $2T$ portion of the AC function before the DFT, keeping one sample out of every two of the result. Granted the symmetry of the AC function, this requires a $2T$ segment of signal. We describe below a measure of periodicity that can be used to indicate when the assumption of perfect period holds, and to help choose which strategy (square or hanning window) is appropriate for a given segment of signal.

To summarize, the AC function may be used to accurately estimate the spectral envelope of a periodic source at harmonics of F_0. A useful feature is that the method can be extended to several concurrent periodic signals, as illustrated in Sects. 3.8 and 3.10. In preparation, the next section describes the notion of segregation by harmonic cancellation.

3.6 Segregation by harmonic cancellation

Suppose that the observed signal z is the sum of two signals x and y, the second of which is periodic with period T_y. The contribution of that second source may be removed from z by applying a filter with impulse response $(\delta(t) - \delta(t - T_y))/2$. The result is equal to a *comb-filtered* version x' of the input signal x, free from any influence of y, but that differs

from x as a result of the comb filtering. The spectral effect of applying a comb-filter with delay T is illustrated in Fig. 6 (a). The power transfer function has zeros at all multiples of $1/T$, and these suppress all partials of y if $T = T_y$. This filtering also causes the spectrum of x' to differ from x.

Of interest for the present discussion is that it is not necessary to actually comb-filter the input signal. Supposing its AC coefficients are available, the AC function of x' is given by:

$$r_{x'}(\tau, t) = r(\tau, t) - r(\tau + T_y, t) - r(\tau - T_y, t - T_y) + r(\tau, t - T_y) \tag{8}$$

From this the power spectrum of x' may be derived. It differs from that of x to a degree that can be estimated if x is also periodic, a case treated in Sect. 3.8. The next section considers any quasi-periodic sound to be the sum of two "sources", one periodic and the other aperiodic, and uses harmonic cancellation to split them.

[Fig. 6 about here]

3.7 Periodic / aperiodic decomposition

Consider a quasiperiodic signal x_t, and T an estimate of its period. It is possible to express x_t as the sum of two signals p_t and a_t:

$$p_t = (x_t + x_{t-T})/2, \qquad a_t = (x_t - x_{t-T})/2 \tag{9}$$

Clearly, $x(t) = p(t) + a(t)$. We have $p_t = x_t$ if x is purely periodic, whereas a is non-zero only if x is *not* perfectly periodic. It thus makes sense to call p and a the "periodic" and

"aperiodic" parts of x_t (note however that p itself is not necessarily periodic). Denoting:

$$||x_t||^2 = (1/W) \sum_{j=t+1}^{t+W} x_j^2 \tag{10}$$

it is easy to verify that:

$$(||x_t||^2 + ||x_{t-T}||^2)/2 = ||p_t||^2 + ||a_t||^2 \tag{11}$$

The left hand can be understood as an estimate of the signal power (averaged over two windows), while terms on the right measure the power of the aperiodic and periodic parts respectively. Signals p and a thus also define a partition of signal *power*. A measure of waveform aperiodicity may then be defined as

$$\epsilon = ||a||^2/(||a||^2 + ||p||^2) \tag{12}$$

This measure equals 0 for a periodic signal, 0.5 for noise, and 1 for an "anti-periodic" signal (such that $x_{t+T} = -x_t$).

Of interest for the present discussion is that these statistics may be derived from the AC coefficients of the input signal:

$$\begin{aligned} ||a|| &= [r(0,t) + r(0,t+T) - 2r(t,T)]/2 \\ ||p|| &= [r(0,t) + r(0,t+T) + 2r(t,T)]/2 \end{aligned} \tag{13}$$

For best accuracy, input AC functions should be calculated with a T-sized window. Parseval's relation implies that the power spectra of p and a also form a partition of the power spectrum of x (Fig. 6). "Periodic" and "aperiodic" spectra may be derived from AC functions r_p and r_a, themselves derived from input AC coefficients:

$$\begin{aligned} r_p(\tau,t) &= r(\tau,t) + r(\tau+T,t) + r(\tau-T,t-T) + r(\tau,t-T) \\ r_a(\tau,t) &= r(\tau,t) - r(\tau+T,t) - r(\tau-T,t-T) + r(\tau,t-T) \end{aligned} \tag{14}$$

Using the method of Sect. 3.5, estimates of both spectra are obtained at multiples of $1/T$ along the frequency axis. By interpolation one can derive periodic and aperiodic "spectrograms". Each estimate frame requires a signal segment of size $3T$.

Note that other definitions of "periodic part" and "aperiodic part" are possible. For example if the signal is quasi-periodic over N cycles, then $\sum_{k=1}^{N} x(t+kT)$ defines one cycle of the best periodic approximation. The definition used here (which corresponds to $N = 2$) offers the best possible temporal resolution.

3.8 Two-voice pitch-adaptive spectral estimation

Suppose that the observed signal z is the sum of two periodic signals x and y with periods T_x and T_y respectively. The following algorithm allows accurate estimation of the spectral envelopes of *comb-filtered* versions of the input signals: x' filtered by a filter with impulse response $\delta(t) - \delta(t - T_y)$, and y' filtered by a filter with impulse response $\delta(t) - \delta(t - T_x)$.

Supposing that all necessary AC coefficients $r(\tau, t, W)$ have been calculated from z, the algorithm proceeds as follows:

(1) Estimate both periods, T_x and T_y

(2) Based on AC coefficients calculated with $W = T_x$:

  (2a) calculate the AC function of x'

  (2b) derive the spectral envelope of x' at multiples of $1/T_x$

(3) Based on AC coefficients calculated with $W = T_y$:

  (3a) calculate the AC function of y'

  (3b) derive the spectral envelope of y' at multiples of $1/T_y$

Of interest in the present context is that the AC function of x' may be calculated using Eq. 8 based on AC coefficients $r(\tau, t) = r(\tau, t, T_x)$ calculated with a window of size T_x. The spectrum estimate of x' is exact if x and y are perfectly periodic. However it differs from the spectrum of x by the effect of the comb-filter, as illustrated in Fig. 7 (c). The ratio between the two may be represented by the power transfer function:

$$e(f) = (1 - cos(2\pi f/|T_x - T_y|)/2) \tag{15}$$

illustrated in Fig. 7(d).

The AC function of y' may be likewise calculated using Eq. 8 based on AC coefficients calculated with a window of size T_x. The estimate of y' differs from that of y by the same factor $e(f)$. In principle, estimates of x and y may be obtained by applying the corrective factor $1/e(f)$, but this obviously fails where $e(f) = 0$, and the result is prone to noise for $e(f)$ small. Section 3.11 examines ways to make effective use of spectrally distorted estimates. Figure 7 (e) illustrates the effect of spectral distortion on the power spectrum envelope of a vowel.

3.9 Variable-amplitude periodic signals

The methods described so far assume periodicity and may behave poorly otherwise. In particular, the cancellation remainder of an imperfectly periodic source cannot be distinguished from interference of other sources. One important source of aperiodicity does nevertheless allow accurate estimation: amplitude variation.

First, we note that F_0-estimation methods may be extended to estimate the period(s) of one or several variable-amplitude sources (de Cheveigné and Kawahara, 2002). Harmonic

cancellation may also be extended. Suppose for example that a signal y may be expressed locally as:

$$y(t) = y_0(t)e^{bt} \tag{16}$$

where y_0 is periodic. y may be removed from the mixture $z = x + y$ by applying the filter $\delta(t) - e^{bT}\delta(t - T_y)$. Of interest for the present discussion is that Eq. 8 can be extended to obtain $r_{x'}$ from the input AC coefficients:

$$\begin{aligned} r_{x'}(\tau, t) &= r(\tau, t) - e^{bT_y} r(\tau + T_y, t) - e^{bT_y} r(\tau - T_y, t - T_y) \\ &\quad + e^{2bT_y} r(\tau, t - T_y) \end{aligned} \tag{17}$$

From $r_{x'}$ the spectrum of x' can be derived. It differs of course from that of x, but the distortion is less severe than in the strictly periodic case.

Period-adaptive spectral estimation may also be extended to estimate the spectral envelope of the underlying periodic sound (before application of the variable amplitude factor). Suppose that x may be expressed locally as:

$$x(t) = x_0(t)e^{at} \tag{18}$$

where x_0 is periodic with period T. If a is known, the AC function r_0 of x_0 may be obtained from r:

$$r_0(\tau, t) = r(\tau, t)e^{-a\tau}e^{-at} \tag{19}$$

If a is unknown it may be estimated by comparing amplitudes over two windows separated by T. Alternatively, Eq. 6 may be replaced by:

$$r(\tau, t) = (1/W) \sum_{i=t+1}^{t+W} x(i + \tau/2)x(i - \tau/2) \tag{20}$$

The amplitude factors balance each other out, so that this formula gives directly the AC function of $x_p(t)$.

The two-voice estimation algorithm of Sect. 3.8 may likewise be extended. Supposing that the variation rates a and b of x and y are known, Eq. 8 may be modified to accurately suppress the variable-amplitude y:

$$\begin{aligned} r_{x'}(\tau,t) &= r(\tau,t) - e^{bT_y} r(\tau+T_y,t) - e^{bT_y} r(\tau-T_y,t-T_y) \\ &\quad + e^{2bT_y} r(\tau,t-T_y) \end{aligned} \tag{21}$$

A similar equation can suppress x to obtain $r_{y'}$. The values thus obtained may then be corrected according to Eq. 19 to obtain accurate estimates of the AC functions of x'_0 and y'_0, comb-filtered versions of x_0 and y_0. The amplitude variation rates a and b may be knowN before hand (for example from instrument models), or they may be estimated as a by-product of the F_0-estimation process (de Cheveigné and Kawahara, 2002). Amplitude compensation of the *interfering* voice is crucial to reduce talk-through, while that of the target is less crucial.

3.10 Multiple channels

Obvious parallels may be drawn between within-channel self-similarity (due to periodicity), and across-channel similarity (due to source-microphone propagation). If both are available, both may be used and combined. The full range of possibilities is well beyond the scope of this paper. Here we give just one example.

Suppose two spatially distinct sources $x(t)$ and $y(t)$, and two spatially distinct sensors

picking up signals $s_1(t)$ and $s_2(t)$. Neglecting room effects, we have:

$$\begin{aligned} s_1(t) &= x(t) + a_y y(t - \theta_y) \\ s_2(t) &= a_x x(t - \theta_x) + y(t) \end{aligned} \tag{22}$$

Where a_x and θ_x are the attenuation and delay of source x in sensor 2 relative to 1, and a_y and θy are the attenuation and delay of source y in sensor 1 relative to sensor 2. Time and amplitude scales are adjusted so that x and y have a delay of 0 and a scale of 1 at sensors s_1 and s_2 respectively.

If delays and attenuation are known, it is possible to suppress either source and obtain *comb-filtered* versions x' filtered by a filter with impulse response: $\delta(t) - a_y\delta(t - \theta_y)$, and y' filtered by a filter with impulse-response $\delta(t) - a_x\delta(t - \theta_x)$. These are spectrally distorted versions of x and y. Taking example on Sect. 3.8, the distortion of either source can be accurately estimated if that source is periodic.

Of interest for the present discussion is that it is not necessary to actually filter the input signals: AC functions of x' and y' may be derived from the AC and CC coefficients of sensor signals s_1 and s_2. For example:

$$\begin{aligned} r_{x'}(\tau, t) &= r_{11}(\tau, t) - a_y r_{12}(\tau + \theta_y, t) - a_y r_{21}(\tau - \theta_y, t - \theta_y) + a_y^2 \\ &\quad r_{22}(\tau, t - \theta_y) \end{aligned} \tag{23}$$

Depending on the number, position and nature (periodicity) of sources, a wide range of schemes may be devised to take advantage of periodicity and spatial cues while compensating for the limitations of each. A full taxonomy is beyond the scope of this paper. Our purpose here is mainly to point out that they can be implemented based on the sensor signal correlation coefficients produced by the “correlation machine”.

3.11 Pattern-matching, smoothing, reconstruction, sharing

The previous algorithms may produce spectral estimates that are incomplete and/or distorted and/or uncertain. This section briefly discusses approaches to make the best of these estimates.

Pattern-matching. An incomplete pattern may be matched to a template by applying a non-uniform weight that ignores the missing parts in the distance calculation. Appropriate matches correctly yield a distance of zero, but inappropriate matches may also incorrectly do so if they differ from the template by a dimension that is missing. This form of error is unavoidable.

Distortion, if it is known, may be accommodated by applying the same distortion to the template. Such is the case for example where the distortion is the result of comb-filtering with known parameters, as in Sect. 3.8. For example, the distorted envelope of a vowel (dotted line Fig. 3.8 (e)) can be used as a template to recognize that vowel accurately despite the distortion. As an alternative, it is also possible to apply an *inverse filter* to the distorted pattern before matching, but that approach is more complex and prone to noise.

Uncertainty may be accommodating within the more general framework of statistical pattern matching. Uncertainty is in part due to the presence of aperiodicity, and can be estimated from measures such as described in Sect. 3.7.

Models. Models may be constrained by fitting their parameters based on incomplete or uncertain evidence. The most useful case is where the model is sufficiently rigid to be

fully specified in this way. Examples are acoustic models of voice production, or dynamic models of articulation.

Smoothing. Smoothing may be seen as the application of a simple model of a "smooth" spectrum, constrained by the available samples. The model may be local (polynomial fitting) or global (low-pass filtering, cepstral smoothing, etc.). While a sophisticated model may be more accurate, simple smoothing may be adequate to attenuate F_0-related spectrotemporal fluctuations. An effective example is the F_0-adaptive smoothing of Kawahara (1997; Kawahara et al., 1999).

Splitting shared data. In the two-voice situation of Sect. 3.8, harmonics of either voice that share the same frequency cannot be estimated directly. The only knowledge available is the amplitude $|z|$ of their sum, which puts a weak upper bound on the absolute difference of their amplitudes $|(|x| - |y|)|$. However if the amplitude of either voice is known, for example from a model, that of the other may be given narrower bounds. For example if $|y|$ is known then $|x|$ is bounded by $|z| \pm |y|$.

Finally, if the amplitude of both partials is known, an estimate of their relative *phases* may be derived by comparing $|z|$ to $|x| + |y|$. For this purpose it is useful to have an estimate of $|z|$ free of fluctuations related to either period, a goal that may be attained by calculating $r(0, t, T_x)$, and then averaging this quantity over a T_y-sized window. These cascaded smoothing steps are sufficient to remove all residual fluctuation related to either T_x or T_y.

4 Discussion

This paper described the transposition of a model of auditory perception, the CN model, to tasks of signal processing. Other algorithms exist for the same tasks, so it is worth outlining what is specific to this approach, in terms of processing principles or performance.

Key ideas are: (a) Calculation at each time step of AC and CC coefficients using "running" rather than standard short-term necessary) to the period of a source. (c) Cancellation of interfering sources by linear combination of delayed input signals. (d) Indirect calculation of useful statistics (power and/or correlation) of these combinations as a function of the precalculated AC and CC coefficients. (e) Period estimation, and period-adaptive spectral estimation based on these statistics. (f) Handling of the resulting incomplete, distorted or uncertain spectral estimates according to missing-data techniques.

Each of these has roots in previous work. The AC function is a standard basis for spectral estimation (Oppenheim and Schafer, 1989), and in particular linear prediction (Markel and Gray, 1976). It is the basis of auditory models of pitch and timbre (Licklider, 1951; Meddis and Hewitt, 1991, 1992). Cancellation has also been proposed for auditory models (e.g. Durlach, 1963; de Cheveigné, 1993) as well as for signal processing. The relation between the covariance matrix of the output of a linear function and that of its input is well known. Period-adaptive spectral estimation has been proposed by e.g. Beauchamp and Madden (2000) and Kawahara et al. (1999). Missing-data techniques have been developed for automatic speech recognition by Cooke et al. (1997).

The present work goes beyond previous work in several ways. The use of "running" formulae allows the AC function of linear combinations of delayed inputs to be calculated

from the AC and CC functions of those inputs (something not easily done with standard blockwise formulae based on FFT). This in turn allows accurate and efficient implementation of cancellation techniques to remove interference. Accurate period estimation (de Cheveigné and Kawahara, 2002) is a key ingredient in making these techniques effective.

4.1 Caveats

Periodicity is exploited either to allow cancellation of an interfering source, or to improve spectral estimation accuracy of the target. Departures from periodicity of the interference reduce the effectiveness of cancellation, and for the target they make period-adaptive estimation less effective. Slow amplitude variations may be however be accommodated as explained in Sect. 3.9.

The DFT applied to the AC function produces a power spectrum, whereas applied to the waveform (as in standard methods) it produces an amplitude spectrum. Neglecting phase, one can be transformed to the other so the procedures should yield equivalent results. However spectra obtained from the AC function tend to have a less wide dynamic range and to be more sensitive to windowing effects than those obtained from the waveform.

Among the techniques described in this paper, only single-voice F_0 estimation has been formally evaluated (de Cheveigné and Kawahara, 2002). The others still await formal evaluation, a task beyond the scope of this exploratory paper.

4.2 Relations with auditory processing

The basic CN model did not involve the variable integration window size postulated in Sect. 3.1 and found useful in Sects. 3.5 and 3.8. However it is interesting to note that evidence has been found for variable window size integration in pitch perception (Plack and White, 2000; Wiegrebe, 2001).

Cochlear filtering splits the waveform at each ear into an array of narrow-band channels, each transduced to neural firing patterns. Most time-domain models (such as can be implemented by the CN model) make little direct use of such filtering. One can imagine at least four functional advantages that the auditory system might derive from having a frequency-selective cochlea. (1) Immediate access to a limited-resolution spectral representation. (2) Compensation for the information loss due to the non-linearity of mechanical-to-nervous transduction. Intuitively, it is easy to believe that a π phase shift due to dispersion along the basilar membrane would compensate for any loss due to half-wave rectification (the half-wave missing in one channel being present in another). More generally, it is known that narrow-band signals can be reconstructed after severe non-linearities such as infinite clipping (e.g. Mallat, 1991). (3) Compensation for the fact that the AC function (Fourier transform of the *power* spectrum) is dominated by high-amplitude parts of the spectrum. Adaptive compression within each channel, as occurs in the auditory system, leads to a more balanced representation of weak and strong parts. (4) Segregation on the basis of spectral content, by attenuating or suppressing certain channels that are dominated by interference. Points (1) and (2) are of little use in the context of this paper, but points (3) and (4) might worth considering.

To summarize, this paper extrapolated ideas from a model of auditory processing (CN model) to the task of estimating the characteristics of several sources from their mixture. Formal evaluation is incomplete, but encouraging results are available for F_0-estimation (de Cheveigné and Kawahara, 2002). It is hoped that development of these signal-processing schemes will in return provide insight into the processes of auditory perception.

5 Acknowledgments

This paper is based on work presented at the Workshop on Consistent and Reliable Acoustic Cues for Sound Analysis (CRAC) (de Cheveigné, 2001), and a meeting (May 2001) of the Hearnet network funded by the British ESPRC. This research was partially funded by the Cognitique programme of the french Ministry of Research. Author Hideki Kawahara is supported by the CREST programme of the Japan Science and Technology Corporation.

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] Beauchamp, J. W., and Madden, T. J. (2000). "A real-time/non-real-time spectrum analyzer for musical sounds," J. Acoust. Soc. Am. 107, 2843.

[3] Bregman, A. S. (1990). "Auditory scene analysis," Cambridge, Mass., MIT Press.

[4] Brokx, J. P. L., and Nooteboom, S. G. (1982). "Intonation and the perceptual separation of simultaneous voices," Journal of Phonetics 10, 23-36.

[5] Cherry, E. C. (1953). "Some experiments on the recognition of speech with one, and with two ears," J. Acoust. Soc. Am. 25, 975-979.

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

[7] Cooke, M., and Ellis, D. (2001). "The auditory organization of speech and other sources in listeners and computational models," Speech Comm 35, 141-177.

[8] Darwin, C. J., and Culling, J. F. (1990). "Speech perception seen through the ear," Speech Communication 9, 469-475.

[9] Darwin, C. J., and Hukin, R. W. (2000). "Effectiveness of spatial cues, prosody and talker characteristics in selective attention," J. Acoust. Soc. Am. 107, 970-977.

[10] 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.

[11] de Cheveigné, A., McAdams, S., Laroche, J., and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. 97, 3736-3748.

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

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

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

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

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

[17] de Cheveigné, A., and Kawahara, H. (2002). "YIN, a fundamental frequency estimator for speech and music," J. Acoust. Soc. Am. 111, 1917-1930.

[18] de Cheveigné, A., and Pressnitzer, D. (2002). "A correlation network model of auditory signal processing," J. Acoust. Soc. Am. in preparation

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

[20] Hermansky, H., and Morgan, N. (1994). "RASTA processing of speech," IEEE trans Speech and Audio Process. 2, 578-589.

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

[22] Kawahara, H. (1997). "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," Proc. ICASSP, 1303-1306.

[23] Kawahara, H., Masuda-Katsuse, I., and de Cheveigné, A. (1999). "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.

[24] Lea, A. (1992), "Auditory models of vowel perception," Nottingham unpublished doctoral dissertation.

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

[26] Licklider, J. C. R. (1959). "Three auditory theories," in "Psychology, a study of a science," Edited by S. Koch, New York, McGraw-Hill, I, 41-144.

[27] Lippmann, R. P. (1997). "Speech recognition by machines and humans," Speech Comm. 22, 1-16.

[28] Lyon, R. (1984). "Computational models of neural auditory processing," Proc. IEEE ICASSP, 36.1.(1-4).

[29] Mallat, S. (1991). "Zero-Crossings of a Wavelet Transform," IEEE Trans. Information Theory 37, 1019-1033.

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

[31] Meddis, R., and Hewitt, M. J. (1991). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. I: Pitch identification," J. Acoust. Soc. Am. 89, 2866-2882.

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

[33] Nakatani, T., Okuno, H. G., and Kawabata, T. (1995). "Residue-driven architecture for computational auditory scene analysis," Proc. IJCAI, 165-172.

[34] Oppenheim, A. V., and Schafer, R. W. (1989). "Discrete-time signal processing," Englewood Cliffs, New Jersey, Prentice-Hall.

[35] Parsons, T. W. (1976). "Separation of speech from interfering speech by means of harmonic selection," J. Acoust. Soc. Am. 60, 911-918.

[36] Papoulis, A. (1984). "Signal analysis," New York, McGraw-Hill.

[37] Plack, C. J., and White, L. J. (2000). "Pitch matches between unresolved complex tones differing by a single interpulse interval," J. Acoust. Soc. Am. 108, 696-705.

[38] Rosenthal, D. F., and Okuno, H. G. (1997). "Computational auditory scene analysis," Lawrence Erlbaum.

[39] Scheffers, M. T. M. (1983a), "Sifting vowels," Groningen unpublished doctoral dissertation.

[40] Stubbs, R. J., and Summerfield, Q. (1988). "Evaluation of two voice-separation algorithms using normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 84, 1236-1249.

[41] Stubbs, R. J., and Summerfield, Q. (1990). "Algorithms for separating the speech of interfering talkers: Evaluations with voiced sentences, and normal-hearing and hearing-impaired listeners," J. Acoust. Soc. Am. 87, 359-372.

[42] Stubbs, R. J., and Summerfield, Q. (1991). "Effects of signal-to-noise ratio, signal periodicity, and degree of hearing impairment on the performance of voice-separation algorithms," J. Acoust. Soc. Am. 89, 1383-1393.

[43] Summerfield, Q., and Culling, J. F. (1992). "Periodicity of maskers not targets determines ease of perceptual segregation using differences in fundamental frequency," Proc. 124th meeting of the ASA, 2317(A).

[44] Weintraub, M. (1985), "A theory and computational model of auditory monaural sound separation," Stanford unpublished doctoral dissertation.

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

[46] Yost, W. A., Patterson, R., and Scheft, S. (1996). "A time domain description for the pitch strength of iterated rippled noise," J. Acoust. Soc. Am. 99, 1066-1078.

Figure captions:

Fig. 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.*

Fig. 2 *"Correlation machine". Autocorrelation coefficients are on the diagonal and cross-correlation coefficients are in the upper and lower triangular parts. Each coefficient is calculated for a range of lags (τ), times (t) and window sizes (W). Here, 4 inputs are illustrated, but most methods described in the paper assume a single input (i.e. only $r_{11}(\tau, t, W)$ is calculated).*

Fig. 3 *Illustration of the principle of F_0 estimation of a single voice (a) or two voices (b) using the algorithms of de Cheveigné and Kawahara (1999, 2002). (a): power of the combination $x(t) - x(t-\tau)$ as a function of τ. The minimum indicates the period T. (b): power of the combination $x(t) - x(t-\tau_1) - x(t-\tau_2) + x(t-\tau_1-\tau_2)$ as a function of τ_1 and τ_2 (gray scale, darker means closer to zero). The first minimum in the lower half quadrant indicates the periods T_1 and T_2.*

Fig. 4 *Simplified model of the production of a periodic vowel sound. The amplitude power transfer function of the vocal tract (a) modifies the amplitudes of the partials of the source (b), supposed of constant amplitude, to produce the line spectrum of the vowel (c). Information encoded in the waveform is limited to the sample points: no other aspect of the envelope can be retrieved.*

Fig. 5 *(a) Flow-chart of single-voice spectral estimation. (b) Waveform. The line below the waveform indicates the portion used to calculate one period of the AC function (the full portion represents the integration window, the dotted portion the range of samples used. (c) Autocorrelation function. (e) Series of partial power estimates obtained by taking the DFT of (c).*

Fig. 6 *Power transfer functions of comb filters* $(\delta(t) - \delta(t-T)/2$ *(a) and* $(\delta(t) + \delta(t-T))/2$ *(b). The former has zeros at multiples of* $1/T$*. The latter has maxima at those values. The two sum to one.*

Fig. 7 *(a) Power spectra of sounds* x *(period* T_x*) and* y *(period* T_y*, thick bars) with flat spectral envelopes. The power spectrum of* $z = x + y$ *is the sum of these power spectra. (b) The dotted line represents the power transfer function of the filter* $\delta(t) - \delta(t - T_x)$ *tuned to remove* x*. The bars represent the power spectrum of* y'*, comb-filtered version of* y*. (c) The dotted line represents the power transfer function of filter* $\delta(t) - \delta(t - T_y)$*, tuned to remove* y*. The bars represent the power spectrum of* x'*, comb-filtered version of* x*. (d) Amplitude transfer function representing the spectral distortion undergone by both* x' *and* y'*. (e) Effect of this spectral distortion on the spectrum of a vowel. The full line represents the power spectral envelope of vowel /a/. Supposing the vowel is produced with fundamental* $1/T_x$*, the full symbols represent the amplitude of its partials. After filtering with a filter tuned to* T_y*, the amplitude of partials are represented by the open symbols. The dotted line represents the overall effect of spectral distortion on the vowel's envelope.*

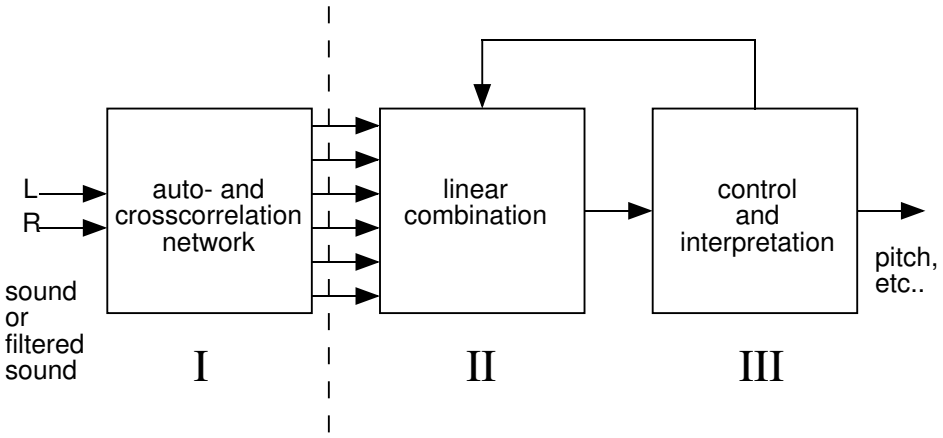
L
R
sound
or
filtered
sound
auto- and
crosscorrelation
network
linear
combination
control
and
interpretation
pitch,
etc..
I
II
III

Fig. 1

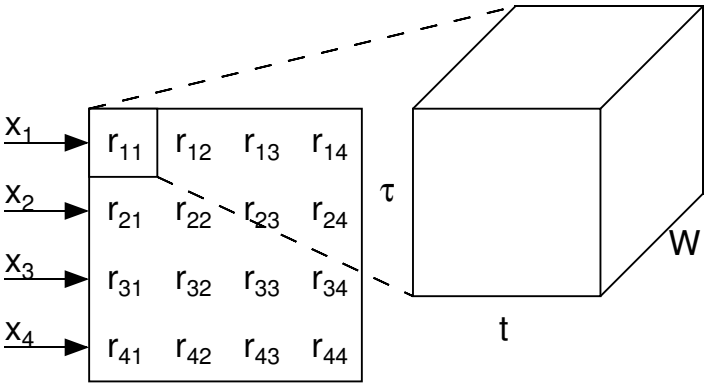
x1
x2
x3
x4
r11 r12 r13 r14
r21 r22 r23 r24
r31 r32 r33 r34
r41 r42 r43 r44
τ
W
t

Fig. 2

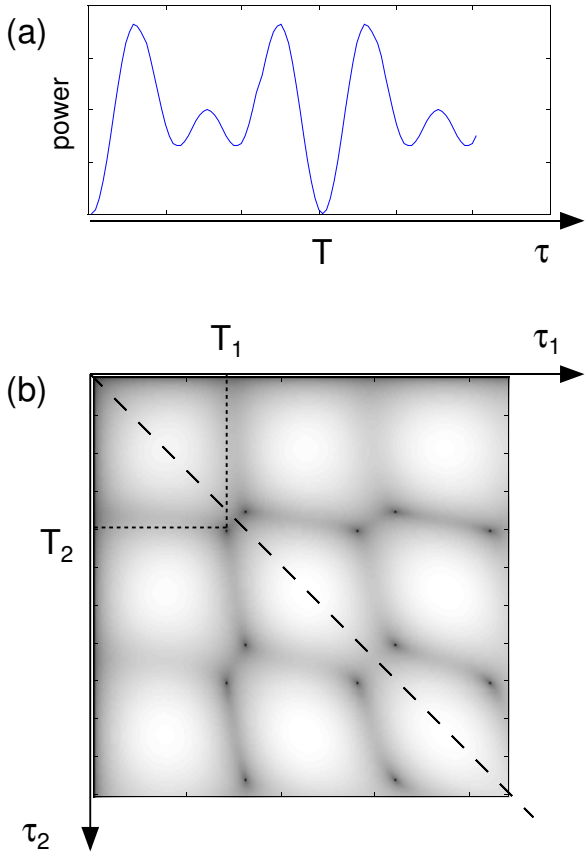

Fig. 3

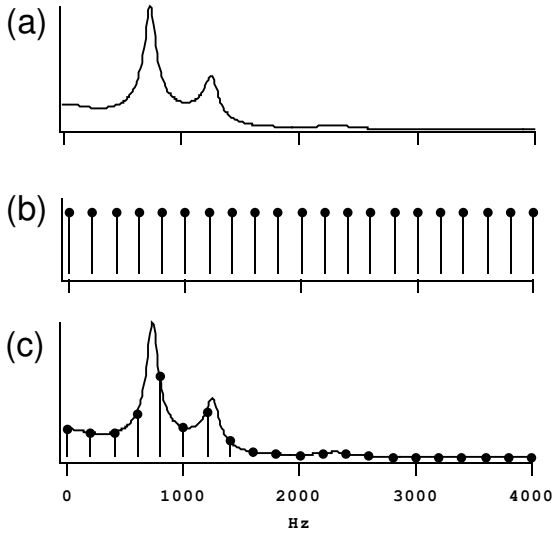

Fig. 4

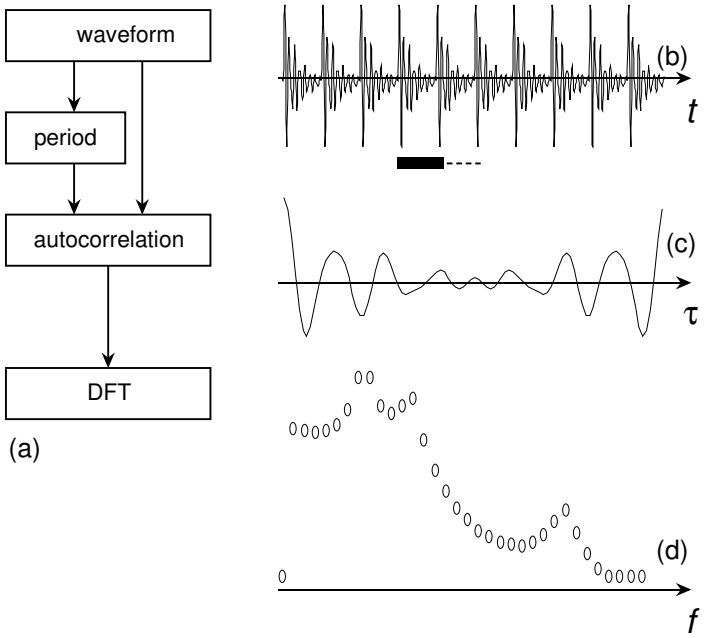

Fig. 5

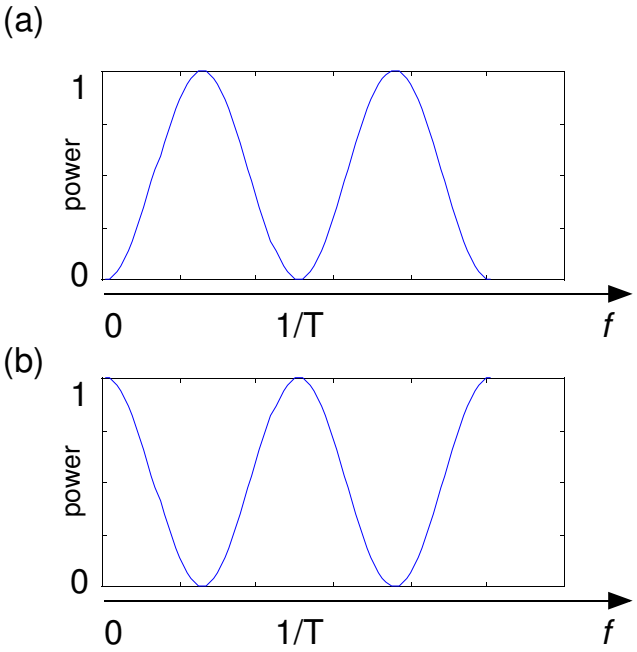

Fig. 6

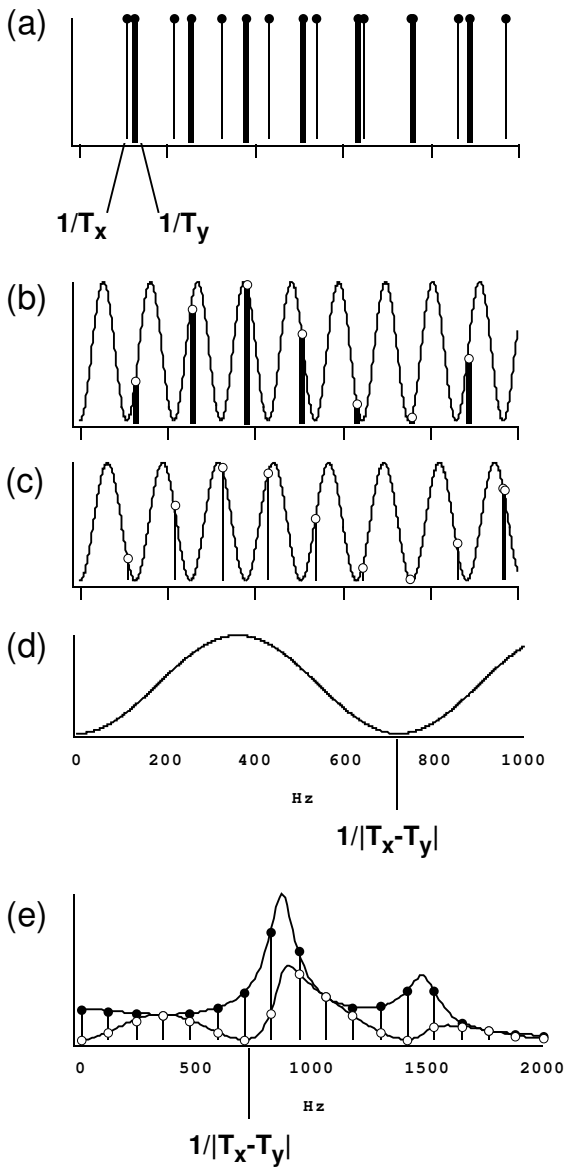
(a)
1/T_x
1/T_y
(b)
(c)
(d)
0
200
400
600
800
1000
Hz
1/|T_x-T_y|
(e)
0
500
1000
1500
2000
Hz
1/|T_x-T_y|

Fig. 7