Multiple period estimation and pitch perception model

Alain de Cheveigné,

CNRS/Université Paris 7, and ATR Human Information Processing Research Laboratories

Hideki Kawahara,

Media Design Informatics Group, Design Information Science Dept. Faculty of Systems Engineering, Wakayama University, and ATR Human Information Processing Research Laboratories

Send all correspondence to:

Alain de Cheveigné, Laboratoire de Linguistique Formelle, case 7003, 2 place Jussieu, 75251, Paris, FRANCE.

alain@linguist.jussieu.fr

Last revised: 8 July 1998

Number of pages (including figures):

Number of Figures:

Number of Tables:

Abstract (English)

The pitch of a periodic sound is strongly correlated with its period. To perceive the multiple pitches evoked by several simultaneous sounds, the auditory system must estimate their periods. This paper proposes a process in which the periodic sounds are canceled in turn (multistep cancellation model) or simultaneously (joint cancellation model). As an example of multistep cancellation, the pitch perception model of Meddis and Hewitt (1991a,b) can be associated with the concurrent vowel identification model of Meddis and Hewitt (1992). A first period estimate is used to suppress correlates of the dominant sound, and a second period is then estimated from the remainder. The process may be repeated to estimate further pitches, or else to recursively refine the initial estimates. Meddis and Hewitt's models are spectrotemporal (filter channel selection based on temporal cues) but multistep cancellation can also be performed in the spectral or time domain. In the *joint cancellation model*, estimation and cancellation are performed together in the time domain: the parameter space of several cascaded cancellation filters is searched exhaustively for a minimum output. The parameters that yield this minimum are the period estimates. Joint cancellation is *guaranteed* to find all periods, except in certain situations for which the stimulus is inherently ambiguous.

Abstract (Français)

La hauteur d'un son périodique est étroitement liée à sa période. Pour percevoir les hauteurs multiples de plusieurs sons simultanés, le système auditif doit estimer leurs périodes. Cet article propose un processus d'estimation par lequel les sons périodiques sont annulés les uns après les autres (modèle d'annulation successive), ou simultanément (modèle d'annulation simultanée). Comme exemple de modèle d'annulation successive, le modèle de perception de la hauteur de Meddis et Hewitt (1991, a,b) peut être associé au modèle d'identification de voyelles concurrentes de Meddis et Hewitt (1992). Une première estimation de la période sert à supprimer les corrélats du premier son, et la deuxième période est estimée à partir du reste de cette opération. L'opération peut être répétée pour estimer d'autres périodes, ou pour affiner les estimations initiales de façon récursive. Les modèles de Meddis et Hewitt sont de type spectro-temporel (sélection de canaux de filtre selon des critères temporels), mais l'annulation successive peut s'opérer aussi bien dans le domaine temps ou fréquence. Dans le modèle d'*annulation simultanée*, estimation et annulation se font ensemble dans le domaine temps: l'espace des paramètres d'une cascade de filtres d'annulation est parcouru exhaustivement, à la recherche d'un minimum du signal de sortie. Les paramètres de ce minimum sont les estimations des périodes. Le succès du modèle d'annulation simultanée est *guaranti*, sauf dans certaines situations où le stimulus est ambigu.

1 Introduction

Pitch perception has played a central role in hearing theory, and inspired much effort and controversy. A wide variety of models have been put forward to explain how a single pitch is evoked by a quasi-periodic stimulus, such as the note of a musical instrument (Schouten 1970; de Boer, 1976; Evans, 1978; Moore 1982; Houtsma, 1995). However Darwin and Ciocca (1992) remark that instruments are usually played together, and often evoke several pitches at the same time. Nordmark

(1978) presented subjects with a mixture of two complex tones with fundamental frequencies (F_0) separated by a semitone. Spectral envelopes were the same and there were no onset disparities. The pitch of both could be heard separately, and each could be matched to an isolated tone with high accuracy. On the other hand Kubovy (1979) reported that a mixture of six pure tones sounded like "noise", the pitch of each tone emerging only when its binaural properties were switched between presentations. Some musicians claim to "hear" 10 or more individual notes within a chord, although we are not aware of a formal test of such claims. They may do so indirectly, based on prior knowledge of the timbre of chords of different composition, without explicit perception of the pitches of all the notes in presence. Dynamic cues such as onset asynchrony may also ease the task. Nevertheless, it is clear that certain subjects can hear several pitch-like entities within a single steady state sound with no cues other than fundamental frequency. Classic pitch perception models don't explain how this is possible.

Several authors have investigated the conditions in which several pitches may be heard (Rasch, 1978; Assmann and Paschall, 1998; McAdams, 1984; Moore, Peters and Glasberg 1986; Hartmann, McAdams and Smith, 1990; Kubovy, 1979) or the way that the pitch of a sound may be affected by the presence of other sounds (Lamoré, 1978; Hartmann and Doty 1996; Lin and Hartmann 1997; Darwin, Buffa, Dierdre and Ciocca 1992; Houtsma and Beerends 1992; Beerends and Houtsma, 1988, 1989; Darwin, Ciocca and Sandell, 1994; Darwin and Ciocca 1992; Peters, Moore and Glasberg 1983; Moore, Peters and Glasberg 1986).

Various schemes to estimate multiple periods have been proposed (see de Cheveigné, 1993 for a review). Single-period estimation models can be extended to estimate two periods by using secondary cues such as the second-largest peak in an autocorrelation pattern (Weintraub 1985; Assmann and Summerfield, 1990). However this approach is not too effective. For one thing, the "secondary cue" is often absent, or else not unique, or its position may not quite correspond to the pe-

riod. For another, the "primary cue" itself may be degraded when two or more periodic sounds are present. A better idea is to use an initial period estimate to guide a model of harmonic sound segregation to suppress one voice, and then estimate the period of the second voice from the remainder. The purpose of this paper is to elaborate that idea. The multistep "estimate-cancel-estimate" principle is illustrated by combining two recent models of pitch perception (Meddis and Hewitt, 1991a,b) and harmonic sound segregation (Meddis and Hewitt, 1992) to perform the steps of estimation and cancellation, respectively. We then suggest that both steps can be performed together, and this leads to the *joint cancellation* model.

The models are of a "mechanistic" flavor, designed to give insight into the mechanism that makes perception of multiple pitches possible, rather than "blackbox" models designed for quantitative predictions of human performance. Although our ultimate focus is on understanding the perception of multiple pitches in humans, the related engineering problem of multiple period estimation is also of interest, and in places the discussion may wander to considerations such as computational cost, etc., that are not strictly pertinent to a pitch perception model.

2 Multistep cancellation model

Consider a stimulus composed of the superposition of N periodic sounds of periods T_i:

$$S(t) = \sum_{i=1}^{N} s_i(t), \tag{1}$$

$$\forall t, s_i(t) = s_i(t + T_i) \tag{2}$$

Suppose that we know how to reliably estimate *one* of the periods from the combined stimulus, for example T_N. Suppose further that, given that knowledge, we know how to design a linear filter Φ_N that can cancel that sound:

$$\forall t, \Phi_N(s_i(t)) = 0 \tag{3}$$

We can swap the filter and summation stages:

$$\Phi_N(\sum_{i=1}^{N} s_i(t)) = \sum_{i=1}^{N} \Phi_N(s_i(t)) \quad (4)$$

$$= \sum_{i=1}^{N-1} \Phi_N(s_i(t)) + 0 \quad (5)$$

The filter gets rid of the last term, of period T_N. The periodicity of the other terms is not modified, except in certain cases considered below in Section 4.1. In the general case we are left with the same problem but with $N-1$ terms rather than N. By recursion we can find all N periods.

This scheme depends critically on the assumption that at least *one* period can be accurately estimated from any mixture, which may not be true in practice. If the first period is not correctly estimated it won't be properly canceled, and the other period estimates may also be wrong. However, by repeating the estimation and cancellation steps it is possible to refine the estimates recursively. Each estimate T_j is refined by first filtering out periods $T_k, k \neq j$, based on their (possibly incorrect) estimates, before re-estimating T_j. The process is repeated for each period until the pattern of estimates starts repeating.

In the following we suggest three possible formulations of the multistep model: spectro-temporal, spectral, and temporal. We then move on to the joint-cancellation model.

2.1 Spectrotemporal formulation

Meddis and Hewitt (1991a,b) proposed a pitch perception model based on autocorrelation. After peripheral filtering and hair-cell transduction, the autocorrelation function (ACF) of instantaneous firing probability is calculated within each channel. The ACFs are summed to obtain a summary autocorrelation function (SACF), and a period estimate is derived from the position of the maximum of that function. The model is derived from that of Licklider (1951). Meddis and Hewitt's (1991) model was designed to estimate one pitch, not several. Assmann and Summerfield (1990) extended it to derive a second estimate based on the second largest peak in the SACF. However the

second estimate was often incorrect. A better scheme is to use a first period estimate to drive a harmonic segregation stage to suppress one voice, and then estimate the period of the other voice from the remainder. An appropriate harmonic segregation stage is that of Meddis and Hewitt's (1992) model of concurrent vowel identification.

In Meddis and Hewitt's (1992) model, an estimate of the period of the dominant vowel was derived from the position of the largest peak in the SACF (as in the same author's pitch perception model). Channels that showed a peak in their ACF at the same delay were attributed to the dominant vowel and removed. The ACFs of the remaining channels were summed to form a "residual" SACF, which was matched to a template to determine the identity of the second vowel. Although not mentioned by the authors, it is evident that the residual SACF could also be used to to derive a second period estimate.

Meddis and Hewitt's two models can thus be combined into a multistage pitch perception model. The algorithm involves two steps: a) estimate a period from the peak in the SACF, and b) remove all channels that have peaks at that period. The steps can be repeated a number of times, either to estimate other periods, or else to refine the previous estimates. Each period estimate determines a pitch.

2.2 Spectral formulation

The same principle was applied in the frequency domain by Parsons (1976). Starting from a complex spectrum based on a 51.2 ms Hanning-weighted FFT, Parsons constructed a table of peaks. The table was refined by "splitting" any peak that could be interpreted as the superposition of peaks belonging to closely spaced components. Once the table was constructed, a Schroeder histogram was formed: for each peak, all its potential fundamentals were noted and their positions accumulated within the histogram (Schroeder 1968). The largest peak in the histogram served to estimate

the first talker's fundamental frequency (F_0). Then all its harmonics were removed from the table, and a second histogram was constructed to estimate the second F_0. A similar method was proposed by Nagabuchi (1979). Parsons' method can be recast as a perception model by replacing the relatively high resolution FFT by a cochlear excitation pattern, such as exploited by place models of pitch perception (Terhardt 1974; Goldstein 1973; Wightman 1973). To handle multiple sounds, Duifhuis, Willems and Sluyter (1982) proposed a "harmonic sieve" to select from a table of excitation pattern peaks only those peaks that belonged to a harmonic series. This method was adapted to the task of estimating the pitches of concurrent voices by Scheffers (1973a,b). However, when tested on pairs of static synthetic vowels, Scheffers' model did not perform very well. Depending on the F_0 difference (ΔF_0), pitch of one voice was correct on up to 98 % of all trials, but the correct rate for the second never went beyond 42%. Assmann and Summerfield (1990) also tested an implementation of Scheffer's model, and concluded that the resolution of the cochlear excitation pattern was too poor to support reliable F_0 estimation (or F_0-guided segregation of the vowel spectra).

2.3 Temporal formulation

The channel selection scheme of Meddis and Hewitt (1992) may be replaced by a within-channel neural cancellation scheme proposed by de Cheveigné (1993, 1997a). The latter works even when all channels are dominated by a single period, whereas Meddis and Hewitt's (1992) model requires a partition between channels, and fails when that partition cannot be made. Apart from that, the two schemes are quite similar. The neural filter is quite effective in extracting periodicity features of a weaker voice, judging from examples reported by de Cheveigné (1997a, Fig 5(a), or 1993, Figs. 7(e) and 8(e)). A temporal formulation has the advantage that it does not depend critically on frequency analysis. The *spectral* formulation required frequency analysis with sufficient resolution to resolve individual harmonics of each series. This is difficult to achieve if there are many com-

ponents and the F_0s are close. The *spectrotemporal* formulation had less severe requirements, but frequency analysis had to be sharp enough so that each source dominated at least some channels. The *temporal* formulation has no such requirement, as it does not depend on frequency analysis in any fundamental way. Frequency analysis (such as occurs in the cochlea) may nevertheless be of use in a physiological implementation of a time-domain model, to compensate for the limited accuracy and dynamic range due to the non-linearity and stochastic nature of transduction and neural processing.

The weakness of the multistep algorithm is that it is not guaranteed to succeed, even when estimates are refined. The joint cancellation model does not have this weakness.

3 Joint cancellation model

Consider as before a stimulus S composed of the superposition of periodic sounds of periods T_i, $i = 1, ..., N$. Suppose that, given the period of any of these sounds, we know how to design a linear filter that can cancel it:

$$\forall t, \Phi_i(s_i(t)) = 0 \tag{6}$$

If we cascade all these filters

$$\Phi = \Phi_1 \circ \Phi_2 \circ ... \circ \Phi_N \tag{7}$$

and apply the result to the compound stimulus, the output is zero:

$$\forall t, \Phi(S(t)) = 0 \tag{8}$$

Suppose finally that each filter is specific to one period, that is, it won't cancel a signal with another period (in practice this may sometimes not be true). The principle of estimation is simple: search the N-dimensional parameter space of the filters until a zero is found. The periods T_i specific to the filters Φ_i are the period estimates.

The filter Φ_i can be implemented as a time-domain comb filter of lag parameter T_i defined by the impulse response:

$$h_i(t) = \delta(t) - \delta(t - T_i) \tag{9}$$

This filter has zeros at the frequency $f_i = 1/T_i$ and all its multiples, and its response to a periodic signal of period T_i is everywhere zero. It can be applied directly to the sound signal or equivalently (with uniform parameters) to all channels of a linear cochlear filter bank.

The algorithm is computationally expensive because of the exhaustive search. This is not a obstacle if it is implemented in a massively parallel architecture, as might be the case of a physiological implementation. Computation is saved at the expense of accuracy if periods are estimated stepwise rather than jointly, which is precisely the principle of the step-wise cancellation model we saw previously.

3.1 Neural formulation

In the neural formulation, each linear filter is replaced by a "neural cancellation filter" based on delay lines and inhibitory synapses (Fig. 1). A gating neuron is fed via two pathways, one direct and excitatory, the other delayed and inhibitory. Spikes arriving along the direct pathway are transmitted, except when a spike arrives simultaneously along the delayed pathway (de Cheveigné 1993). The statistics of spike transmission of a real neuron would depend on many factors that are conveniently ignored in a simplified semilinear model that relates input and output spike probability densities:

$$o_\tau(t) = MAX[0, i(t) - i(t - \tau)] \tag{10}$$

Where $i(t)$ and $o_\tau(t)$ are input and output densities, respectively (de Cheveigné, 1997a). The $MAX[0, .]$ operation (half-wave rectification) reflects the fact that probability cannot be negative. Neither the

neural filter nor its simplified model are linear, so the analysis of the previous section does not apply rigorously. Nevertheless, simulation of the filtering process shows that the neural formulation of the joint cancellation model is effective.

3.2 Single period case

It is interesting to consider the limit of single period estimation. Minimum square output of the cancellation filter corresponds to minimization of a "squared difference function" of the form:

$$SDF(\tau) = \sum_{0}^{N} (s_i - s_{i-\tau})^2 \tag{11}$$

By expanding the square one can show (Ney 1982) that this is approximately equivalent to maximization of the autocorrelation function:

$$ACF(\tau) = \sum_{0}^{N} s_i s_{i-\tau} \tag{12}$$

Multiplication is replaced by subtraction, and the search for a maximum is replaced by a search for the minimum. In this way, classic autocorrelation models of pitch perception may be reformulated in cancellation terms (de Cheveigné, 1998).

4 Illustration

4.1 Two-period estimation

In this example, processing is performed directly on the waveform. Period estimation is performed by searching for a minimum in the output of a comb filter, squared and summed over a square window (Eq. 11). The sampling rate is 10 kHz. The search range is from 40 to 100 samples (100 to 250 Hz), and the window size is 100 samples. Cancellation is applied by applying the comb filter (Eq. 9) tuned to the period to cancel.

The waveform consists of the sum of two periodic signals, each made up of the first ten harmonics of its fundamental with equal amplitude and sine phase. The waveform is displayed in Fig. 2 for F_0s of 111 and 119 Hz (periods are 90 and 84 samples respectively). In this case, the initial period estimate based on the SDF is 91 samples (Fig. 3 (a)). Subsequent estimates are obtained after application, in each case, of a comb filter tuned to the previous estimate (Figs. 3 (b,c)). In this example, the correct estimates (84, 90 samples) were obtained in a few steps. In other cases the estimate may take more steps, and in some cases the algorithm may get caught in a local minimum and give incorrect estimates.

Multistep estimation was tested on a set of 2352 waveforms, each consisting of a sum of two equal-amplitude 10-component complexes with periods ranging from 51 to 99 samples (periods different). This range covers slightly less than an octave between 100 and 200 Hz. The algorithm was designed to iterate until the sequence of estimates produced started repeating itself. Error rates are given in Table 1. Without recursive refinement, the error rate was 38 % (implying that a single-period algorithm applied to a two-period stimulus often fails to estimate either period). With recursive refinement, the error rate fell to 8.42 %. In most cases, the estimates were off by one sample and would have been corrected if the algorithm had included tests for "one-off" estimates in its search rules. In other cases, however, both estimates were clearly wrong. It is interesting to note that the error rate falls dramatically if the amplitude of one sound is scaled relative to the other: 1.36 % for 3 dB level difference, and 0 % for 6 dB. When one sound is stronger, its period estimate is more likely to be correct, in which case cancellation is perfect and the second estimate is also correct. In a physiological implementation of the multistep model, the auditory system might make use of peripheral filtering to isolate spectral regions where one sound dominates the other (supposing they differ in spectral envelope), thus enhancing estimation reliability. It might also take advantage of onset and envelope disparities in a similar fashion.

The *joint cancellation* algorithm was also implemented with, as expected, an error rate of 0%. In all cases estimation was performed on the 300-sample segment shown in Fig. 2, a relatively small window of data (see Sect. 5.2). It must be stressed that several factors contributed to ease the estimation task. The component sounds were perfectly periodic and there was no noise. Furthermore all periods were exact multiples of the sampling period, and the algorithm searched the parameter space at sampling-period multiples. Finally, the search range was limited to less than an octave, thus avoiding the "failure scenarios" described in Sect. 5.1. See de Cheveigné (1993) for results on the more difficult task of estimating F_0 from concurrent natural voices.

4.2 Three-period estimation

An experiment was performed using stimuli consisting of the sum of *three* periodic sounds with distinct periods in the range 51 to 99. There were 18424 wveforms, each corresponding to a different period triplet. As before, each periodic sound consisted of ten harmonics with equal amplitudes in sine phase. Each waveform was 400 samples in duration.

For three period estimation, the multistep estimation algorithm was implemented in the following way. A variable-size list of period candidates was built and initialized arbitrarily with two values. Pairs of estimates from this list were used to set the parameters of a cascade of two comb filters which were applied to the waveform. A new period estimate was derived from the residual. If the new estimate was different from all previous estimates, it was added to the list. The algorithm terminated when all pairs within the list had been tested, and no new estimates had appeared. For each triplet of candidate periods, the ratio of RMS output to input of three cascaded comb filters tuned to these periods was calculated. The triplet with the smallest measure was chosen as the final estimate. The algorithm usually terminated in 5-10 steps, but in some cases it tested several hundred pairs before terminating. In our simulation the number of steps was limited to 50 to save

time.

The error rate of the multistep algorithm was 6.25 %. In many cases only one estimate was incorrect, implying that, had the number of steps not been limited, the algorithm would have eventually found the third estimate also, had the number of steps not been limited. In other cases all three estimates were incorrect. The error rate of the joint cancellation algorithm was again 0%.

5 Discussion

5.1 Failure scenarios

Given perfectly periodic waveforms, the joint cancellation algorithm is *guaranteed* to succeed in estimating all periods, except in a number of cases where the stimulus is ambiguous. The simplest case is when one period (T_1) is multiple of another period (T_2). Their sum is periodic with period T_1, and a filter tuned to T_1 is sufficient to cancel both periods, so the presence of T_2 cannot be determined. A similar situation arises when two periods have a common multiple T within the search range. In that case the algorithm will report T in place of the original periods. Another failure scenario is when every partial of a constituent sound belongs to the harmonic series of some other sound. In this case the algorithm will skip that constituent. In a variant of this scenario, a subset of the partials of a sound is captured by competing harmonic series, leaving a remainder with a period that is only a fraction of its original period. This smaller period will be reported instead of the original period. In all of these cases the stimulus is inherently ambiguous, and *any* period estimation algorithm must fail.

5.2 Implementation considerations

If Eq. 8 is satisfied by a set of period estimates T_i, it is satisfied by all of their multiples. The basic algorithm does not reject these multiples, so a practical implementation needs to include a mech-

anism to do so. This usually takes the form of a bias that favors short periods over long, and this may introduce new errors (for example if the algorithm incorrectly locks on to a strong harmonic).

Sound waveforms are never perfectly periodic, and noise may be present which may impair the success of the algorithm. The limited sampling resolution of a digital implementation may impair the quality of the cancellation, and a "physiological" implementation would most certainly also have limited linearity and temporal resolution. The practical success of the algorithm depends on these factors. A two-dimensional version of this algorithm was nevertheless effective in the difficult task of estimating both pitches of concurrent natural voiced speech (de Cheveigné 1993).

In practice the criterion of Eq. 8 is tested by integration over a limited time window. Reliable estimation requires that the integration window be at least as large as the largest expected period T_{max}. Supposing that the search ranges for all N periods are the same (this need not be the case), the total duration of signal necessary for the estimation is $(N+1)T_{max}$.

The joint-estimation algorithm relies on exhaustive search within an N-dimensional parameter space, which requires on the order of $(f_s T_{max})^{N+1}$ operations for each frame (where f_s is the sampling rate). As pointed out previously, this cost is not an obstacle for a massively parallel architecture such as might be hypothesized to exist in the auditory nervous system. However it is a formidable obstacle if the algorithm is to be used in practice to estimate many periods. The multi-step cancellation algorithm is one example of a "smarter" search strategy that requires on the order of $N(f_s T_{max})^2$ operations per frame. This algorithm is less costly but also less reliable: one cannot guarantee that it won't fall into a local minimum.

5.3 Finding N

Both algorithms suppose that the number N of periodic sounds is known. N may be found *in principle* by trying different values and searching in each case for a zero output of the filter cascade.

N is taken as the smallest number of cascaded cancellation filters that produces a zero output. In practice, of course, this scheme is extremely time-consuming, and unreliable because of imperfect periodicity or processing inaccuracy (limited sampling resolution). One can speculate that this task (and period estimation itself) might be solved more effectively by exploiting continuity or timbre constraints.

5.4 Period estimation and pitch perception

Our discussion has been mainly about multiple *period* estimation, rather than pitch. It is important to keep a clear conceptual distinction between the two notions, especially in the less well understood situation of concurrent sounds. Nevertheless, pitches perceived are generally in a one-to-one correspondence with the periods of the component sounds (at least such is the goal of the musician trying to "hear out" the notes). Multiple period estimation is the hard part of any model of multiple pitch perception, and that is what this paper is about.

There is more to multiple pitch perception than our discussion suggests. For one thing, the ability to hear multiple pitches is very subject-dependent. At one extreme, musically untrained subjects may be unable to hear even a single pitch, or unable to perform a meaningful task related to whatever they do hear. At the other, certain musically trained subjects can hear out a large number of notes, in some cases more than were included in the stimulus. For example McAdams (1989) reported that musically trained subjects sometimes heard four to six pitches within a stimulus formed by the sum of three steady state harmonic vowels spaced at 5 semitone intervals. It must be noted that certain isolated periodic stimuli also have an ambiguous pitch.

Musicians can take advantage of onset asynchrony between notes of a chord (Rasch 1978). The isolated portion of a note that starts first might support perception of that note's pitch, if the portion were long enough [several periods are required for accurate pitch perception (Robinson and Patter-

son, 1995)]. However that does not explain how the other pitches are perceived. It is also possible that musicians may identify the pitches of notes within a chord indirectly, from prior knowledge of the timbre of different chords.

The present paper showed that, given ideal stimuli, the multiple period estimation task can be solved perfectly, except in certain special cases where no other method could solve it. Cues such as onset disparitities, musical expectations, regularities, etc. may make the task easier, and even be indispensable in practice (for imperfectly periodic stimuli, in the case of limited computation resources, etc.). They are not indispensable in principle.

5.5 Pitch shifts

Within the context of a cancellation model of multiple pitch perception, it is possible to account for certain effects observed experimentally, such as the pitch shifts of mistuned partials (Hartmann and Doty 1996). This question is explored in de Cheveigné (1997b).

5.6 Relation to Auditory Scene Analysis and CASA models

The models reported here used only periodicity cues. Listeners probably benefit also from other well known auditory scene analysis cues, such as onset asynchrony or other envelope disparities, and higher-level knowledge of the timbre and envelope of individual instrumental notes. Such cues might also be of use in a note tracking system. However our models tell us that a lot can be done without them.

Most CASA models start with a spectrotemporal representation, usually based on autocorrelation. Segregation occurs as spectrotemporal regions are mapped out and assigned (on the basis of common periodicity, continuity, common fate, etc.) to one or another source. The emphasis is on *bringing together* elements that belong to the same source. In contrast, the pitch perception models

discussed in this paper support a vision of hearing as a process of *suppression* of elements within the auditory scene. This fits well with the view that perceptual processes adapt in such a way as to be sensitive to *novelty* or departure from a habitual state (Barlow 1990), and with certain computational auditory scene analysis models such as the residue-driven architecture of Nakatani et al. (1995).

We presented one version of the multistep model that relied on a spectrotemporal representation (based on Meddis and Hewitt's models). However other versions of both models did not: a spectrotemporal map is not indispensable for this task. Segregation was performed without any manipulation of spectral "elements" such as partials, etc.. We should perhaps think twice about their necessity in auditory scene analysis models.

Conclusion

Two models were presented to explain the perception of multiple pitches evoked by concurrent periodic sounds. In the first (multistep cancellation) each period is estimated and the corresponding sound is then canceled, until no more sounds remain. In the second (joint cancellation), estimation and cancellation of all sounds are performed together. The former algorithm may be implemented based on two recent models of pitch perception (Meddis and Hewitt 1991a,b) and harmonic sound segregation (Meddis and Hewitt 1992). The latter may be implemented as a cascade of neural cancellation filters (de Cheveigné 1993; 1997a). In its linear formulation, the latter algorithm is guaranteed to find all periods in presence, except for in certain cases when the stimulus is inherently ambiguous. Both models illustrate a principle of auditory scene analysis according to which elements of an auditory scene are suppressed in turn.

6 Acknowledgments

Steve McAdams, Richard Parncutt, John Culling, Malcolm Slaney, and four anonymous reviewers made useful comments on preliminary versions of this paper.

References

[1] Assmann, P. F., and Paschall, D. D. (1997). "Pitches of concurrent vowels," J. Acoust. Soc. Am. (in preparation).

[2] Assmann, P. F., and Summerfield, Q. (1990). "Modeling the perception of concurrent vowels: Vowels with different fundamental frequencies," J. Acoust. Soc. Am. 88, 680-697.

[3] Barlow, H. B. (1990). "A theory about the functional role and synaptic mechanism of visual after-effects," in "Vision: coding and efficiency," Edited by C. Blakemore, Cambridge, England, Cambridge University Press, 363-375.

[4] Beerends, J. G., and Houtsma, A. M. J. (1988). "The influence of duration on the perception of single and simultaneous two-tone complexes," in "Basic issues in hearing," Edited by H. Duifhuis, J. W. Horst and H. P. Wit, London, Academic, 380-385.

[5] Beerends, J. G., and Houtsma, A. J. M. (1989). "Pitch identification of simultaneous diotic and dichotic two-tone complexes," J. Acoust. Soc. Am. 85, 813-819.

[6] Darwin, C. J., and Ciocca, V. (1992). "Grouping in pitch perception: effects of onset asynchrony and ear of presentation of a mistuned component," J. Acoust. Soc. Am. 91, 3381-3390.

[7] Darwin, C. J., Ciocca, V., and Sandell, G. J. (1994). "Effects of frequency and amplitude modulation on the pitch of a complex tone with a mistuned harmonic," J. Acoust. Soc. Am. 95, 2631-2636.

[8] Darwin, C. J., Buffa, A., Williams, D., and Ciocca, W. (1992). "Pitch of dichotic complex tones with a mistuned frequency component," in "Auditory physiology and perception," Edited by L. D. Y. Cazals, K. Horner, Oxford, Pergamon press, 223-229.

[9] 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.

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

[11] de Cheveigné, A. (1997b). "Harmonic fusion and pitch shifts of inharmonic partials," J. Acoust. Soc. Am. 102, 1083-1087.

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

[13] de Boer, E. (1976). "On the "residue" and auditory pitch perception," in "Handbook of sensory physiology," Edited by W. D. Keidel and W. D. Neff, Berlin, Springer-Verlag, 479-583.

[14] Duifhuis, H., Willems, L. F., and Sluyter, R. J. (1982). "Measurement of pitch in speech: an implementation of Goldstein's theory of pitch perception," J. Acoust. Soc. Am. 71, 1568-1580.

[15] Evans, E. F. (1978). "Place and time coding of frequency in the peripheral auditory system: Some physiological pros and cons," Audiology 17, 369-420.

[16] Goldstein, J. L. (1973). "An optimum processor theory for the central formation of the pitch of complex tones," J. Acoust. Soc. Am. 54, 1496-1516.

[17] Hartmann, W. M., McAdams, S., and Smith, B. K. (1990). "Hearing a mistuned harmonic in an otherwise periodic complex tone," J. Acoust. Soc. Am. 88, 1712-1724.

[18] Hartmann, W. M., and Doty, S. L. (1996). "On the pitches of the components of a complex tone," J. Acoust. Soc. Am. 99, 567-578.

[19] Houtsma, A. J. M. (1995). "Pitch perception," in "Hearing," Edited by B. C. J. Moore, London, Academic Press, 267-295.

[20] Houtsma, A. J. M., and Beerends, J. G. (1992). "The role of aural frequency analysis in pitch perception with simultaneous complex tones," in "Auditory frequency selectivity," Edited by B. C. J. Moore and R. D. Patterson, New York, Plenum Press, 237-444.

[21] Kubovy, M. (1979). "Concurrent pitch-segregation and the theory of indispensable attributes," in "Perceptual organization," Edited by M. Kubovy and J. Pomeranz, Hillsdale, N.J., Lawrence Erlbaum, xxxx-xxxx.

[22] Lamoré, P. J. J. (1977). "Pitch and masked thresholds in octave complexes in relation to interaction phenomena in two-tone stimuli in general," Acustica 37, 250- 257.

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

[24] Lin, J. L., and Hartmann, W. M. (1997). "The pitch of mistuned harmonics: evidence for a template model," J. Acoust. Soc. Am. in preparation,

[25] McAdams, S. (1984), "Spectral fusion, spectral parsing, and the formation of auditory images," Stanford unpublished doctoral dissertation.

[26] McAdams, S. (1989). "Segregation of concurrent sounds. I: Effects of frequency modulation coherence," J. Acoust. Soc. Am. 86, 2148-2159.

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

[28] Meddis, R., and Hewitt, M. J. (1991b). "Virtual pitch and phase sensitivity of a computer model of the auditory periphery. II: phase sensitivity," J. Acoust. Soc. Am. 89, 2883-2894.

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

[30] Moore, B. C. J. (1982). "An introduction to the psychology of hearing," London, Academic Press.

[31] Moore, B. C. J., Peters, R. W., and Glasberg, B. R. (1986). "Thresholds for hearing mistuned partials as separate tones in harmonic complexes," J. Acoust. Soc. Am. 80, 479-483.

[32] Nagabuchi, H., Kobayashi, T., and Yamamoto, H. (1979). "Speech enhancement and suppression in mixed speech," Transactions of the IECE (Japan) 62(10), 627- 634 (in Japanese).

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

[34] Nakatani, T., Goto, M., Ito, T., and Okuno, H. G. (1995). "Multi-agent based binaural sound stream segregation.", Proc. IJCAI Workshop on Computational Auditory Scene Analysis, 84-91.

[35] Ney, H. (1982). "A time warping approach to fundamental period estimation," IEEE Trans. SMC 12, 383-388.

[36] Nordmark, J. O. (1978). "Frequency and periodicity analysis," in "Handbook of perception, vol. IV - Hearing," Edited by E. C. Carterette and P. P. Friedman, New York, Academic Press, 243-282.

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

[38] Patterson, R. D. (1987). "A pulse-ribbon model of monaural phase perception," J. Acoust. Soc. Am. 82, 1560-1586.

[39] Peters, R. W., Moore, B. C. J., and Glasberg, B. R. (1983). "Pitch of components of complex tones," J. Acoust. Soc. Am. 73, 924-924.

[40] Rasch, R. A. (1978). "The perception of simultaneous notes such as in polyphonic music," Acustica 40, 21-33.

[41] Robinson, K., and Patterson, R. D. (1995). "The stimulus duration required to identify vowels, their octave, and their pitch chroma," J. Acoust. Soc. Am. 98, 1858-1865.

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

[43] Scheffers, M. T. M. (1983b). "Simulation of auditory analysis of pitch: an elaboration on the DWS pitch meter," J. Acoust. Soc. Am. 74, 1716-1725.

[44] Schouten, J. F. (1970). "The residue revisited," in "Frequency analysis and periodicity detection in hearing," Edited by R. Plomp and G. F. Smoorenburg, Sijthoff, 41-58.

[45] Schroeder, M. R. (1968). "Period histogram and product spectrum: new methods for fundamental-frequency measurement," J. Acoust. Soc. Am. 43, 829-834.

[46] Terhardt, E. (1974). "Pitch, consonance and harmony," J. Acoust. Soc. Am. 55, 1061-1069.

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

[48] Wightman, F. L. (1973). "The pattern-transformation model of pitch," J. Acoust. Soc. Am. 54, 407-416.

single step	*multistep*	*joint*
38 %	8.42 %	0 %

Table 1 *Error rates for three versions of the two-period estimation algorithm.*

Figure captions:

Fig. 1 *Schematic "neural cancellation filter" consisting of a gating neuron with excitatory (direct) and inhibitory (delayed) synapses.*

Fig. 2 *waveform consisting of the sum of two periodic waveforms of periods 90 and 84 samples (111 and 119 Hz). Each is the sum of ten harmonics in sine phase.*

Fig. 3 *Three steps in the multistep period estimation algorithm. (a) The SDF (average squared difference function) calculated from the raw waveform. The global minimum within the search range is at 91 samples (line). (b) The SDF calculated after filtering the waveform to cancel the 91-sample period. The global minimum is at 84 samples. (c) The SDF calculated after filtering the waveform to cancel the 84-sample period. The global minimum is at 90 samples.*

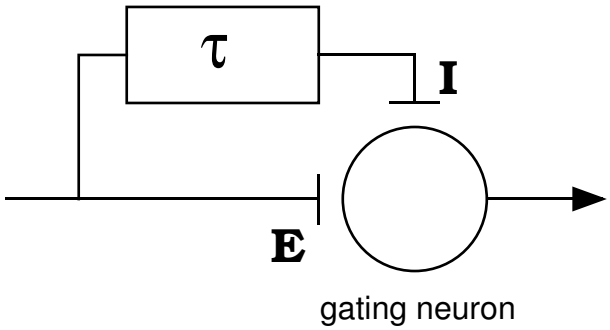

Fig. 1

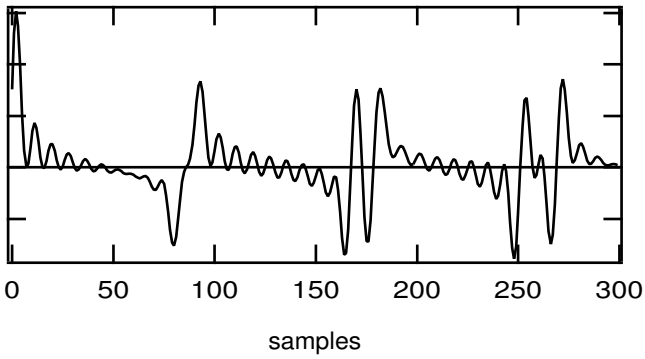

Fig. 2

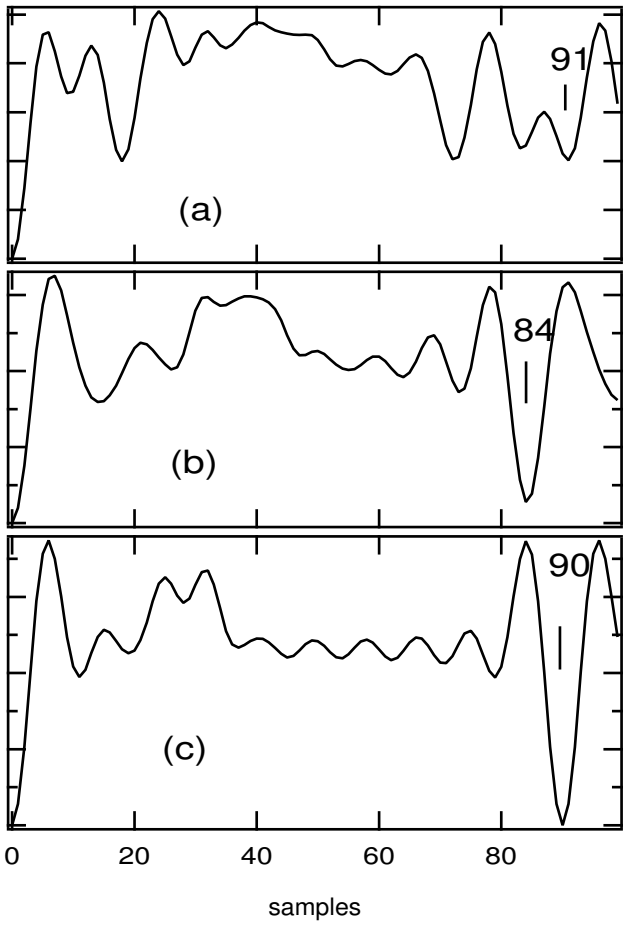

Fig. 3