Separation of concurrent harmonic sounds: Fundamental frequency estimation and a time-domain cancellation model of auditory processing

Alain de Cheveigné

Laboratoire de Linguistique Formelle, CNRS/Université Paris 7, tc 806, 2 place Jussieu, 75251, Paris, France, and ATR Auditory and Visual Perception Research Laboratories, Sanpeidani, Inuidani, Seika-cho, Soraku-gun, Kyoto, 619-02 Japan

(Received 4 December 1991; revised 20 November 1992; accepted 5 February 1993)

Signal-processing methods and auditory models for separation of concurrent harmonic sounds are reviewed, and a processing principle is proposed that cancels harmonic interference in the time domain. The principle is first formulated in signal processing terms as a time-domain comb filter. The critical issue of fundamental frequency estimation is investigated and an algorithm is proposed. Tested on a restricted database of natural voiced speech, the algorithm successfully found estimates correct within 3% of an octave for 90% of all frames. Next, the principle is formulated in physiological terms. A hypothetical "neural comb filter" is described, based on neural delay lines and inhibitory synapses, and tested using auditory-nerve fiber discharge data obtained in response to concurrent vowels [A. R. Palmer, J. Acoust. Soc. Am. **88**, 1412–1426 (1990)]. Processing successfully suppresses the correlates of either vowel in the response of fibers that respond to both, allowing the other vowel to be better represented. The filter belongs to the class of "cancellation models" for which predictions can be made concerning the outcome of certain psychoacoustic experiments. These predictions are discussed in relation to recent experimental results obtained elsewhere.

PACS numbers: 43.64.Bt, 43.66.Hg, 43.71.Pc, 43.71.Cq

INTRODUCTION

Listeners can follow and understand the speech of one speaker among many, even when binaural information is not available (Cherry, 1953). This aspect of our ability to organize the sound environment has received renewed interest of late (Assmann and Summerfield, 1988, 1989, 1990; Darwin and Culling, 1990; McAdams, 1989; Meddis and Hewitt, 1992; Palmer, 1988, 1990, 1992; Stubbs and Summerfield, 1988, 1990, 1991; Summerfield and Assmann, 1991).

Among other effects of speech on speech, one might expect vowels to be affected by the presence of a concurrent vowel, because identification depends on the spectrum envelope that is strongly affected by the presence of the other vowel. The frequency components of both are intimately mixed, so simple filtering strategies that partition the spectrum into discrete regions are not effective. However, listeners can identify concurrent synthetic vowels at levels significantly above chance, particularly if there is a difference in fundamental frequency (Assmann and Summerfield, 1990; Summerfield and Assmann, 1991; Scheffers, 1983). The benefits to speech understanding of a difference in fundamental frequency were confirmed for synthetic and natural speech by Brokx and Nooteboom (1982). A wide variety of schemes have been proposed to explain or reproduce this capability, with varied success. This paper attempts to identify some of the factors involved in the task. Given a mixture of two voices, one can *enhance* one voice using its harmonicity, or else *cancel* the other voice using that voice's harmonicity. Which strategy is better? Which strategy (perhaps both) does the auditory system use? Another question is that of frequency analysis. Most models and methods involve frequency analysis and depend critically on its resolution. Is that step indispensable, or is time-domain processing possible, in particular in the auditory system? A third issue is that of estimation of the fundamental period information that most models require. This is a difficult task for a single voice, how well can it be done when two voices are mixed?

The paper begins with a review of the literature in which we discuss schemes and classify them according to how they work. In particular, we distinguish *harmonic enhancement* models from *harmonic cancellation* models. We then propose a simple processing principle for canceling harmonic interference in the time domain. We formulate it first in speech signal-processing terms as a time-domain comb filter, and investigate the important problem of $F0$ estimation from mixed speech. We propose an algorithm for that purpose that we evaluate on natural speech, but we do not attempt to design a complete system for speech separation. Instead, we formulate the same processing principle in physiological terms, and propose a simple neural "filter" based on delay lines and inhibitory synapses. The filter cancels the correlates of one vowel within auditory-nerve discharge patterns, and thus allows those of the other vowel to be better represented. We test this filter using data recorded in the guinea pig in response to mixed vowel stimuli (Palmer, 1990). We then briefly discuss a variant that enhances rather than cancels. Finally, we discuss the two basic strategies, cancellation and enhance-

 0001-4966/93/063271-20$06.00

ment, in relation to recent results obtained in psychoacoustics (Lea, 1992; Lea and Summerfield, 1992). Throughout the paper, we attempt to keep a clear distinction between signal processing methods and hearing models. While both are of interest to harmonic sound and speech separation, they do not address quite the same tasks, and their evaluation follows different conventions.

I. REVIEW OF METHODS AND MODELS OF SPEECH SEPARATION

A. Speech signal processing methods

Parsons (1976) described a method by which the Fourier transforms of 51.2-ms windows of mixed speech were "dissected" into spectral peaks. The peaks were accumulated in a table that was used to construct a histogram of potential fundamentals of all spectral components (Schroeder, 1968). The *F*0 of a first speaker was determined from this histogram, and then that of the second speaker was obtained by removing from the table the harmonics belonging to the first speaker, and repeating the histogram calculation on the remaining peaks. The voice of each speaker was then resynthesized by reverse Fourier transformation of its share of the spectrum. The method was tested with natural speech and evaluated informally. Intelligibility and naturalness were reported as being good.

Stubbs and Summerfield (1988) proposed a vowel separation method based on cepstral filtering, and compared it with an implementation of Parson's harmonic selection algorithm. Formal evaluation was done with normal-hearing and hearing-impaired subjects, using synthetic vowels and natural CV words with synthetic-vowel backgrounds. Both methods improved recognition rates, but harmonic selection performed better. For the first method, *F*0 estimation was done by searching for two peaks in the cepstrum. In further papers Stubbs and Summerfield (1990, 1991) emphasize the difficulties of *F*0 tracking, but show that if these difficulties are overcome (using *F*0 values obtained from the speech before mixing) natural speech sentences can be separated quite successfully using either cepstral filtering or harmonic selection.

Nagabuchi *et al.* (1979) proposed a method similar to that of Parsons, in which *F*0 estimates were obtained from peaks in the autocorrelation function of the LPC residual. *F*0 estimation was done in two steps: in a first step the highest peak of the autocorrelation function indicated the *F*0 of the dominant voice. That estimate served to determine a frequency-domain comb filter that eliminated the first voice, allowing the second *F*0 to be measured in turn. Applied to natural speech (male speech mixed with female speech), processing improved preference scores.

Hanson and Wong (1984) estimated the spectrum of the interfering voice using *F*0 information derived from the interfering signal before mixing, and then subtracted it from the spectrum of the mixed speech. Formal evaluation with naturally spoken sentences showed that processing improved intelligibility. Childers and Lee (1987) improved on this scheme by adding a minimum cross-entropy "spectral tailoring" stage that reduced spectral distortion. Informal listening tests indicated an improvement in intelligibility, but "quality was not retained." Naylor and Boll (1987) added an *F*0 estimation stage to Hanson and Wong's system, for which they tested four different estimation methods (cepstrum, maximum likelihood, harmonic matching, and an auditory model-based method). They retained the second. Informal tests with natural speech showed improved intelligibility, particularly when the signal-to-noise ratio was poor.

Min *et al.* (1988) used an analysis window size tailored (on the basis of *F*0) to a multiple of the period of either voice to achieve better spectral sampling. *F*0 was estimated using the ACF and AMDF methods (autocorrelation function, and average magnitude difference function, defined later on) supplemented by a "look-forward and look-backward double check" scheme. Min *et al.* consider applying their method to separate three voices. Silva and Almeida (1990) proposed a sinusoidal decomposition model in which the overlap between harmonics caused by limited spectral resolution was resolved using a stationary least-squares estimation technique. These last two methods were not formally evaluated.

The previous methods all operate in the frequency domain. In contrast, Frazier *et al.* (1976) used an adaptive time-domain comb filter "tuned" to the fundamental period of the target voice, to enhance it. This parameter was obtained from a glottal accelerometer attached to the speaker's throat, and the authors did not suggest how it might be estimated directly from the acoustic signal.

Weintraub (1985, 1986) applied Lyon's (1984) model of cochlear filtering and nerve firing coincidence analysis to the task of speech separation. Nerve fiber discharge was simulated at the output of a filter bank, and discrete autocorrelation (AC) functions were calculated for each channel as in Licklider's (1956, 1959, 1962) model of pitch perception. According to Licklider, the correlate of pitch is a ridge that spans the frequency dimension of the two-dimensional AC pattern at a delay equal to the period. In Weintraub's system, the AC pattern was summarized by summing the AC functions across channels, and the *F*0s were derived from the largest and second-largest peaks in this sum. Weintraub proposed two versions of his system that differed in their segregation principles. In the first, filter channels were *selected* according to whether they were dominated by one periodicity or the other. Lyon (1983/1989) had used a similar approach for segregating spatially distinct sources in a binaural processing model. In the second version, individual channels were *shared* between voice streams by modeling the relative contributions of the component voices. This second version incorporated a dynamic programming *F*0 estimation algorithm that required the component voices to be in different registers. Tested over a database of natural speech (digits), it gave a first *F*0 estimate correct within five samples (3% at 100 Hz) for 88.8% of all frames. The second estimate was correct for 74.3% of all frames.

The previous methods relied on periodicity information. Kopec and Bush (1989), on the other hand, attempted to recognize constituent voices without using *F*0

information, by calculating LPC poles of the compound spectrum, and matching *subsets* of these poles to LPC poles of templates. They tested their method using naturally spoken digits masked by whole sentences. The error rate was reduced by a factor of 2.6 compared to a conventional Euclidean cepstral distance, when the target-interference ratio was 0 dB and dynamic time warping alignment was used to match targets and templates.

B. Auditory processing models

Weintraub and Lyon worked from the standpoint of speech processing, but they were also interested in understanding the auditory system. One can argue that a successful processing scheme is a good basis for an auditory model, because it has proved itself capable of actually performing a task in realistic conditions. One can also argue the contrary: An improvement in intelligibility with processing is proof that the auditory system cannot perform the same processing, or at least not to the same degree of precision. Be that as it may, most of the auditory processing models to be reviewed in this section are related to one or the other of the previous signal processing methods.

Scheffers (1983) studied the psychophysics of voiced speech separation by measuring recognition rates for pairs of synthetic vowels, chosen from a set of eight Dutch vowels and synthesized at various fundamental frequencies. Rates for both vowels correct improved from 45% for equal *F*0s to 62% for a semitone difference. Scheffers then developed a processing model similar to Parson's method, in which all spectral components within 3% of multiples of the *F*0 of a constituent vowel were selected by a "harmonic sieve." The *F*0s were estimated using an extension of the DWS method (Duifhuis *et al.*, 1982). Tested on mixtures of synthetic steady-state vowels, the method was successful in identifying at least one *F*0 but had difficulties with the second (96% and 24% of all frames correct within 3%, respectively).

Assmann and Summerfield (1990) repeated Scheffers's vowel recognition experiments for pairs of synthetic English vowels (from a set of five). When stimuli were 200 ms in duration, the recognition rate for both vowels correct improved from about 60%, for equal *F*0s, to about 75% for a difference of two semitones. With shorter stimuli (51.2 ms) they failed to find a similar improvement. Summerfield and Assmann (1991) repeated the same experiment, but in one condition the 200-ms pairs of vowels were presented dichotically. This appears to enhance the benefit of a difference in *F*0: for one subject, the rate improved from 60% for equal *F*0s to close to 100% for a difference of 0.5 semitones. In monaural conditions, the same subject's rates plateaued at about 80%. These experiments all indicate that rates tend to improve with differences in *F*0, but Assmann and Summerfield (1988, 1989; Summerfield and Assmann, 1991) also found, as did Scheffers, that with equal *F*0s both members of a vowel pair are identified with an accuracy significantly greater than chance. To explain this fact, Assmann and Summerfield (1989) proposed several template matching models, similar in spirit to the method of Kopec and Bush (1989) mentioned earlier.

The performance of place models such as Scheffers's depends on the resolution of spectral analysis. Given current estimates of the selectivity of the peripheral auditory system, as determined from psychoacoustics (Moore *et al.* 1983) or physiology (Carney and Yin, 1988), one can wonder if resolution is sufficient for a place model to work. Assmann and Summerfield (1990) explored the question by incorporating a computer model of peripheral auditory filtering into a place model of vowel segregation. The *F*0s were derived, using a harmonic sieve, from the peaks in the "excitation pattern" (average filter output as a function of center frequency), and the vowels were segregated by sampling this pattern at harmonics of one or the other *F*0. They found that the model performed poorly, mainly because the spectral representation lacked resolution. Among other things, the model had difficulty finding the *F*0s of both the component vowels. Assmann and Summerfield also tested a "place-time" model akin to the second version of Weintraub's auditory signal-processing model. Output channels of the filter bank were processed to obtain a set of autocorrelation (AC) functions, one for each channel. For pitch estimation, these functions were summed across channels and two pitch estimates were derived from the largest and second-largest peaks in the sum. A vowel's spectrum was then isolated by sampling the AC functions at a lag corresponding to that vowel's period. This place-time model was more successful than the place model.

Meddis and Hewitt (1992) proposed a place-time model similar to Weintraub's first method. The AC functions of channel outputs of a cochlea model were summed across channels, as in Assmann and Summerfield's place-time model, but this time a *single* period estimate was derived from the largest peak. Segregation was achieved by selecting all channels for which the AC function had its largest peak at that period. These channels were assigned to one voice, their AC functions summed, and the identity of the corresponding vowel determined by template-matching of the short-delay portion of the sum. Remaining channels were assigned to the other voice.

Palmer (1990) investigated vowel separation from the point of view of physiology. He recorded the activity of a population of cochlear nerve fibers in the guinea pig in response to concurrent vowels /a/ and /i/, with respective fundamentals of 100 and 125 Hz, and tested several models of vowel segregation directly on the data. All models involved *F*0 estimation as a first step. One class of models was based on the average localized synchrony rate (ALSR) (Young and Sachs, 1979). The ALSR combines place and synchrony information, and has an aspect similar to that of the average across fibers of the Fourier transform of the period histogram. The ALSR pattern was submitted to (a) a harmonic selection method similar to Parson's, (b) a harmonic sieve method, (c) "cepstral" analysis. All three methods yielded correct estimates of both *F*0s. Palmer also tried two "place-time" methods based on autocorrelation of fiber discharge patterns: (a) a histogram of the delays at which peaks occurred in AC patterns within channels, and (b) a pooled AC function similar to that used by Weintraub and Assmann and Sum-

merfield. These place-time methods were disappointing, mainly because they failed to simultaneously estimate both *F*0s. However, using the same data, Palmer (1992) tested Meddis and Hewitt's model that only requires *one F*0, and found it quite successful.

Lea (1992) proposed a model of vowel separation based on an array of AC functions calculated from filter bank channels. A first period estimate was obtained from the sum across channels of the AC functions. Then, a "synthetic autocorrelation" array was calculated that approximated the contribution of the first vowel. This synthetic array was subtracted from the AC array, a second pitch estimated from the residue, and a second array synthesized and subtracted in turn from the AC array, leaving a second residue. The two residues were used to estimate the constituent vowels, by template matching of the short-lag portion of the sum across channels of the AC functions as in Meddis and Hewitt's model. Subtraction of a synthetic pattern is reminiscent of a technique used by Weintraub (1985) to eliminate nonsignal-specific portions of the AC histogram. It is also similar in principle to spectral subtraction as used by Hanson and Wong (1984).

C. Discussion

In this section, we shall outline some basic similarities and differences between the methods and models we have reviewed.

1. Fundamental frequency estimation

In all schemes, two steps can be logically distinguished: *fundamental frequency estimation* and *separation.* In some cases, *F*0s are estimated directly from the mixed speech, before and independently from separation. For example, the cepstral filtering method of Stubbs and Summerfield (1988) determines both *F*0s from peaks in the cepstrum of the mixed speech signal. Weintraub (1985), Min *et al.* (1988), and the place-time model of Assmann and Summerfield (1989) also estimate *F*0s directly, as do methods that only require one *F*0: Childers and Lee (1987), Naylor and Boll (1987) and Meddis and Hewitt (1992).

Direct estimation of both fundamental frequencies before speech separation allows a clean, modular design. However estimation of a single *F*0 from a single voice is known to be troublesome (Hess, 1983), and estimation of two *F*0s (or even a single one) from mixed speech is likely to be even more difficult. Extracting the *F*0 of one voice is easier if the other voice is first attenuated, and for this reason separation and *F*0 estimation often work hand in hand according to an iterative scheme: (1) estimate period of voice A, (2) use period of voice A to remove voice A, (3) estimate period of voice B (from residue of step 2), and (4) use period of voice B to remove voice B, then go to step 1.

Some methods repeat these steps several times to refine the estimates, others perform steps (1)–(3) only once (we nevertheless classify them as "iterative" also, because generalizing them to repeat several times is trivial). In this loose sense, iterative *F*0 estimation is proposed by Parsons (1976), Nagabuchi *et al.* (1979), Scheffers (1983), and Lea (1992).

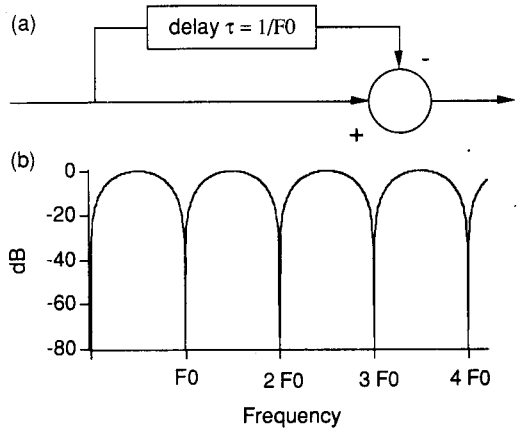

FIG. 1. (a) Time domain comb filter. (b) Magnitude transfer function of time-domain comb filter.

2. Separation strategies

The review uncovered a wide variety of schemes for separating harmonic sounds. They all boil down to two primitive strategies: enhancement and cancellation, together with various heuristics to estimate or reconstruct the spectra.

a. Enhancement versus cancellation. Given a mixture of two periodic sounds, *harmonic enhancement* uses the periodicity of the sound we wish to hear (the target) to enhance it relatively to the background. In the frequency domain, this corresponds to selecting harmonics of the target and enhancing them relatively to others. In the time domain, the processing can be done with a filter defined by the following impulse response:

$$h(t)=\left(\frac{1}{K}\right)\sum_{n=0}^{K-1}\delta(t-nT), \tag{1}$$

where t is time, T is the period of the target voice, δ is the Dirac delta function, and K is the number of "prongs" in the comb-shaped impulse response. This filter adds up copies of the signal delayed by multiples of the period of the target voice. Components of that voice add up in phase, while others add up less favorably. Enhancement works with any kind of interference, but offers at best a modest improvement in signal-to-noise ratio unless the impulse response is long (see Appendix A). Speech may not be sufficiently stationary to allow that.

With *harmonic cancellation* on the other hand, the harmonic nature of the *interfering* voice is used to cancel it. In the frequency domain, this corresponds to setting the harmonics of that voice to zero [Fig. 1(b)]. In the time domain, the same result can be obtained by a comb filter that offers infinite rejection with a relatively short impulse response [Fig. 1(a)]:

$$h(t)=\tfrac{1}{2}\left[\delta(t)-\delta(t-T)\right]. \tag{2}$$

Cancellation is of no help if the interference is not harmonic. For example it will not improve speech intelligibility in a background of wind or rain, and it is not clear how it could handle several voices simultaneously at a cocktail party. However, Lea (1992) remarks that cancellation is appropriate to eliminate a major source of interference: one's own voice. In general, separation is most needed when the signal-to-noise ratio is poor. In that case, the period of the interference is easier to estimate than the period of the target, and cancellation therefore easier to perform than enhancement.[1]

The two strategies are not mutually exclusive, and a number of methods and models employ both. Often enhancement is used to extract the stronger voice, and cancellation the weaker. This is the course taken by the model of Meddis and Hewitt (1992): Channels dominated by the periodicity of the stronger voice are selected to estimate that voice (enhancement), and the weaker voice is taken from the remainder (cancellation). As the amplitude ratio varies, the "dominant" voice may change, so that a segregated voice is alternatively the product of one strategy or the other. The result of segregation is not the same for the two strategies (see next paragraph) and therefore switching between the two may complicate the task of a matching stage.

b. Shared harmonics. Shared harmonics are components that cannot be attributed uniquely to one voice because they belong to the harmonic series of both. A shared harmonic (or filter channel) may be handled in one of four ways: (1) give it to both voices, (2) give it to neither, (3) give it to either, or (4) share it between the two voices.

The first two strategies are "primitive" in the sense that they only require knowledge of the fundamental frequency and of the frequency of the harmonic. They correspond to enhancement and cancellation, respectively. Strategy (3) supposes in addition an exclusion rule ("if the component belongs to voice A then it does not belong to voice B"). Strategy (4) requires additional "well-formedness" constraints, such as spectral or temporal continuity, to determine the share appropriate for each voice. For example, Parsons (1976) adjusts the amplitude of shared peaks by interpolating between adjacent harmonics, whereas Weintraub (1985) goes a step further and minimizes a spectro-temporal cost function in an iterative refinement process.

Once the share of one voice is determined, the other can be estimated by subtraction, *if the phases are known.* Weintraub remarks that this is usually not the case, so one can at best estimate lower and upper bounds of the vector difference.[2] This form of subtraction and strategy (3) both obey Bregman's (1990) "exclusive allocation" principle, whereas strategy 1 corresponds to "conjoint allocation."

c. Interference subtraction. Rather than estimating the target, some methods estimate the interference and subtract it from the mixture (Hanson and Wong, 1984; Naylor and Boll, 1987; Lea, 1992). It is easy to see that "subtraction of interference estimated by harmonic cancellation" is equivalent to "harmonic enhancement of the target." Likewise, "subtraction of interference estimated by enhancement" is equivalent to "cancellation of the interference." Thus, such subtraction strategies do not differ basically from the "primitive" strategies described above, unless of course constraints other than harmonicity contribute to determine the interference that is subtracted. For example, the spectrum of an interfering sound that starts before the target can be estimated during the interval in which it is alone, and later subtracted from the mixture. This corresponds to Bregman's (1990) "old plus new" heuristic.

3. Spectral, spectro-temporal, and time-domain processing

Spectral and spectro-temporal models use frequency analysis, but differ in the way they "label" channels, and in the degree of frequency resolution they require. Spectro-temporal models label channels according to their dominant time-domain periodicity. They work if some channels are dominated by one voice and others by the other, and therefore require that channel bandwidths be narrower than the regions in which one or the other voice's spectrum is dominant. Spectral models on the other hand label channels according to their position on a frequency scale relative to a "harmonic sieve," and require frequency resolution fine enough to isolate individual partials in the compound spectrum.

There is an alternative to frequency analysis. We saw that harmonic enhancement or cancellation can be performed in the time domain by filters with very simple impulse responses. For speech separation methods, operating in the time domain avoids having to decide the length and shape of a spectral analysis window. For auditory processing models, time-domain processing opens a new perspective where performance is no longer completely determined by the resolution of peripheral frequency analysis. The rest of this paper explores these ideas.

II. TIME-DOMAIN CANCELLATION MODEL OF MIXED SPEECH SEPARATION

The time-domain harmonic sound cancellation principle is first formulated in this section as a *speech processing* method, and later reformulated as an auditory processing model in Sec. III. Speech is probably the most important harmonic stimulus that we encounter, and efforts toward implementation of speech separation can reveal problems that are not apparent with simple stimuli. This is particularly true for $F0$ estimation: The difficulties due to the imperfect periodicity of natural speech are not readily apparent when steady-state synthetic stimuli are utilized.

A. Voice separation model

Mixed voiced speech is modeled as the sum of two periodic signals. A periodic signal is by definition invariant by translation along the time axis by a period or its multiple. If such a signal is fed to a time domain comb-filter (Fig. 1), the output must be zero everywhere. This can

also be understood in the frequency domain by remarking that the transfer function of the filter has zeros at the fundamental $1/T$ and all harmonics.

Thus, if mixed voiced speech is fed through a comb filter tuned to the period of one voice, that voice is canceled. The other voice undergoes a certain distortion due to the multiplication of its spectrum by the transfer function of the comb filter. However, informal listening tests (using speech filtered by a comb filter tuned to the $F0$ track of another speaker) show that this form of distortion has little effect on intelligibility. More serious is the fact that the interference may not be perfectly periodic, and therefore not perfectly canceled. The same is true if the period is not precisely estimated: in both cases, the filter output contains residual components superimposed on the target voice.

This method is similar in effect to cancellation methods that work in the frequency domain (Parsons, 1976; Stubbs and Summerfield, 1988). A possible advantage of time-domain processing, besides simplicity, is that cancellation requires a filter with a short impulse response, whereas spectral methods may need a large window to ensure frequency resolution. Also, a time-domain filter can be modified to handle amplitude variations of the interfering voice [by giving different weights to the delta functions in Eq. (2)]. As in other voice-separation algorithms, the knowledge of the *fundamental period* of the interfering component is critical for the algorithm to work. This is the subject of the next section.

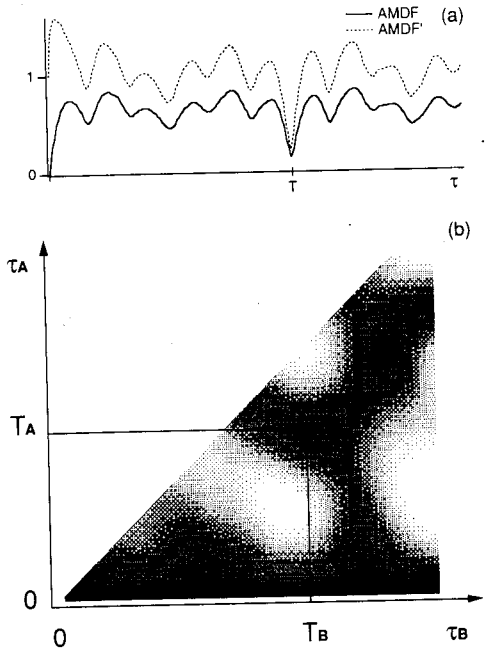

FIG. 2. (a) Continuous line: AMDF (average magnitude output of a comb filter as a function of the lag parameter) for a frame of natural voiced speech. Dotted line: mean-normalized AMDF (see Appendix B 2). The dip indicates the period. (b) DDF (average magnitude output of two cascaded comb filters as a function of lag parameters τ_A and τ_B) for an input consisting of a sum of two periodic signals of period T_A and T_B. Black signifies low output. Search is restricted to the lower half quadrant ($\tau_A < \tau_B$).

B. Mixed voiced speech F0 estimation

1. Algorithm

First, let us consider a single voice. If voiced speech is modeled as a periodic signal and fed through a comb filter with the following impulse response:

$$h(t) = \delta(t) - \delta(t-\tau),$$

where t is time, τ is the lag, and δ is the Dirac delta function, the output should be zero everywhere if the lag equals the period or one of its multiples. Other nonzero values of the lag should produce a non-null output. This is the basis of a classic speech $F0$ estimation algorithm known as the average magnitude difference function (AMDF) method (Ross *et al.*, 1974; Hess, 1983). The lag parameter space of a time-domain comb filter is searched for a minimum in the filter output (full-wave rectified and averaged over a window), and the position of this minimum gives the period estimate [Fig. 2(a)].

Mixed voiced speech can be similarly modeled as the sum of *two* periodic functions. If the signal is fed through two cascaded comb filters, with lag parameters equal to the periods (or their multiples), the output is everywhere zero. It is nonzero for all other values of the lag parameters. This is the basis of the present $F0$ estimation algorithm (referred to hereafter as the DDF—double difference function—algorithm). The two-dimensional lag parameter space of two cascaded comb filters is searched for a minimum in output, and the coordinates of this minimum give the estimates of the periods. Figure 2(b) shows the DDF as a function of the lag parameters τ_A and τ_B, a darker shade meaning a smaller output. The coordinates (T_A, T_B) of the minimum furnish two period estimates (de Cheveigné, 1990, 1991). For reasons of symmetry the search space can be limited to a half quadrant.

If the speech components are truly periodic (supposing their periods are multiples of the sampling period) the algorithm is *guaranteed* to find both periods, unless they happen to be equal or multiples of a common subharmonic within the search range. In that case, a single filter is sufficient to cancel both signals, so the parameter of the second filter is unconstrained. In actual practice voiced speech is often not very periodic so one might wonder whether the algorithm works at all.

2. Evaluation

The algorithm was evaluated by comparing its $F0$ estimates to those obtained by a reference algorithm from the individual voices before mixing. The reference algorithm produced, as a by-product, a "strength-of-periodicity" measure (PM) that was used, in an early stage, to select a subset of "clean" voiced data. Evaluation proceeded according to the following steps: (1) process isolated speech data with the reference $F0$ estimation algorithm; (2) extract 225-ms portions of speech with good periodicity (PM > threshold); (3) pair and add to obtain mixed speech

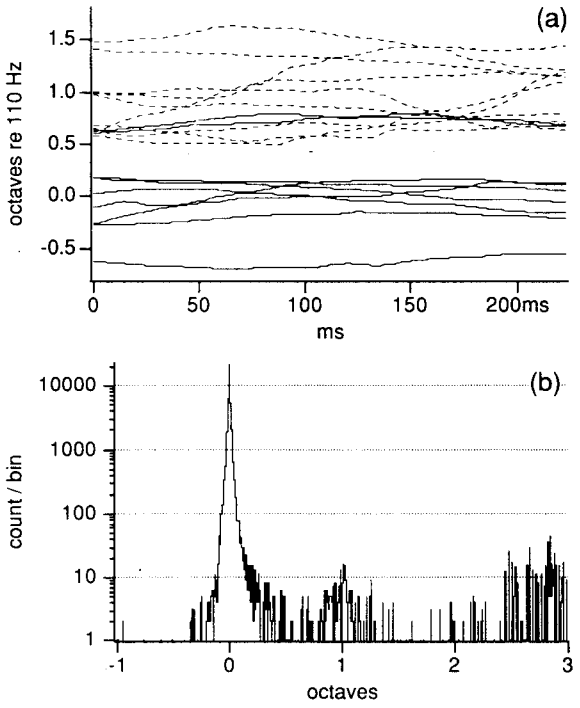

FIG. 3. (a) Fundamental frequency tracks for all tokens in database. Dotted lines: female, continuous lines: male. (b) Histogram of errors (deviation in octaves between DDF $F0$ estimate and closest reference estimate or subharmonic) over the database. Note that the vertical scale is logarithmic.

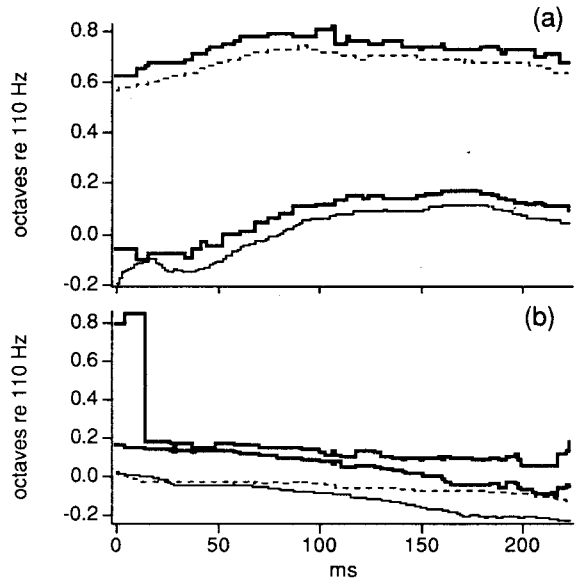

FIG. 4. Fundamental frequencies of component voices of two tokens of mixed speech. Thin lines are "reference" values obtained directly from components before mixing, offset vertically for clarity by −0.05 oct (top) or −0.2 oct (bottom). Thick lines are estimates produced by the DDF algorithm. Both examples are for the male speaker. In (b) when the $F0$ tracks of the component voices coincide one DDF estimate correctly takes their common value, the other takes a spurious value.

tokens; (4) estimate both periods from the mixed speech using the DDF algorithm; and (5) compare these estimates with those from step (1).

The test data, described in Appendix B 1, consisted of "clean" voiced tokens taken from natural speech pronounced by one male and one female speaker. The reference algorithm is described in Appendix B 2. All $F0$ estimates were expressed on an octave scale relative to 110 Hz, and comparisons done along this scale. The $F0$s of the test tokens covered a range of over two octaves [Fig. 3(a)].

The detailed implementation of the mixed voice DDF algorithm is described in Appendix B 3. The search range for both $F0$s (55–440 Hz) started 10% below the lowest $F0$ in the database (61 Hz) and covered 3 oct. The algorithm does not distinguish periods and superperiods, and can therefore lock on to the subharmonic of an $F0$ if it falls within the search range. To prevent such an occurrence from counting as an error, each DDF estimate was compared to all subharmonics of the corresponding reference estimates. Figure 3(b) displays the histogram of deviations (in octaves) between DDF estimates and their closest reference estimate (or subharmonic) over the database. It is plotted on a log scale: On a linear scale the histogram is too sharp for interpretation: 90% of all estimates fell within 3% of a reference value. In Fig. 4, $F0$ plots for two individual pairs can be seen. The thin lines are reference $F0$ estimates (offset vertically for clarity) and the thick lines are estimates produced by the DDF algorithm from the mixed speech signal. They appear to follow very closely the reference $F0$ tracks, except at one point in Fig. 4(b) where the tracks coincide and one DDF estimate breaks down.

3. Comparison with other methods

Here, we compare the exhaustive search strategy of the DDF algorithm with three alternative strategies.

a. Two F0 estimates produced by a conventional F0 estimation algorithm. Conventional $F0$ estimation algorithms can be adapted to produce two estimates from mixed speech (Weintraub, 1985; Stubbs and Summerfield, 1988; Min *et al.*, 1988; and the place-time model of Assmann and Summerfield, 1989). Following this principle, we adapted a classic $F0$ estimation algorithm based on the AC function to derive two period estimates: a first one corresponding to the largest peak within the search range, and a second one corresponding to the second-largest peak. Details are given in Appendix B 4. With this algorithm, 62% of all estimates missed the 3% criterion, about six times more than for the DDF algorithm.

b. Iterative version of the DDF algorithm. The DDF algorithm is computationally expensive because it does an exhaustive search, but it is also readily parallelizable so it should run fast on parallel hardware. Alternatively, a "smarter" algorithm can be devised. For example, in the iterative scheme of Sec. I C 1, single-voice $F0$ estimation alternates with voice cancellation to produce two $F0$ estimates that are refined in turn. A version of the DDF algorithm was implemented following this principle (see Appendix B 5). With this iterative version, 1.4 times more estimates missed the 3% criterion than for the exhaustive-search DDF algorithm. Investigation using simple signals revealed that for some inputs, the iterative algorithm gets caught in a cycle of incorrect estimates that misses the

TABLE I. Percentage of estimates that fell further than 10%, 3%, or 1% of an octave from a reference estimate when *both* $F0$s were estimated. Three algorithms were tried: The conventional ACF algorithm adapted to produce two estimates, an iterative version of the DDF algorithm, and the exhaustive-search DDF algorithm. Within each column a smaller percentage indicates that the algorithm was more accurate.

	>10%	>3%	>1%
ACF	44%	62%	76%
DDF (iterative)	3.1%	14%	31%
DDF (exhaustive)	1.7%	10%	27%

global minimum. A hybrid algorithm (exhaustive search near values obtained by the iterative algorithm) gave results closer to those of the full-search algorithm. Results for various two-$F0$ estimation algorithms are summarized in Table I.

c. Single F0 estimate produced by a single-voice F0 algorithm. Some voice separation schemes only need one $F0$ estimate (Childers and Lee, 1987; Naylor and Boll, 1987; Meddis and Hewitt, 1992). Is a classic algorithm designed for single voices sufficient for this task? Or would one estimate from the pair produced by the DDF algorithm be more accurate? To answer this question, we implemented three single-voice algorithms, based on ACF, AMDF, and an approximation of the $F0$ estimation stage used by Meddis and Hewitt (1992). Results were similar for all, so we report only the latter method, described in detail in Appendix B 6. In Table II, the results of this algorithm are compared with those obtained by taking the lower estimate of the iterative version of the DDF algorithm, or either the lower or the higher of the estimates of the standard DDF algorithm. Almost four times more estimates failed the 3% criterion for the single-voice algorithm than for the lower $F0$ estimate of the standard DDF algorithm.

In summary, the exhaustive search strategy of the standard DDF algorithm was more effective than the alternative strategies we tested, even when the task was that of obtaining a *single* $F0$ estimate from mixed speech.

C. Discussion

The DDF algorithm was successful in finding the fundamental frequencies of both voices for a large proportion of frames. The restricted conditions of the evaluation must be stressed: only "clean" voiced speech was used (according to a criterion that eliminated 25% of voiced speech, see Appendix B 1), and the search range was restricted to 3 oct. On the other hand, the algorithm performed its task on a frame-by-frame basis: continuity or voice-register constraints could possibly enhance reliability in a more complex implementation. The algorithm produces an unordered pair of estimates, whereas a practical system would need to assign each estimate to a voice and track it when the $F0$s cross. That task cannot be accomplished using frame-wise information only, but it could perhaps be done using $F0$ continuity constraints (Parsons, 1976), intonation rules (Brokx and Nooteboom, 1982), or spectral continuity of the separated speech on either side of the $F0$ crossing point (Stubbs and Summerfield, 1991). One might also choose, among alternative parses, whichever one makes better "sense" after further processing. A related problem is recognizing how many voices are present, and detecting when voicing starts and stops. These and other issues must be addressed in the design of a complete speech separation system. As part of such a system, the time-domain cancellation scheme presented in Sec. II A is likely to perform similarly to the frequency-domain schemes described in the review: Insofar as both assume harmonicity, both should be equally affected by the inharmonicity of actual speech. The performance of a real system depends on the skill with which missing information is reconstructed "intelligently" when signal-driven separation fails, and that is beyond the scope of the present paper.

We now consider whether a similar form of processing takes place in the auditory system.

TABLE II. Percentage of estimates that fell further than 10%, 3%, or 1% of an octave from a reference estimate when a *single* $F0$ was estimated. Estimates were made using a single-voice $F0$ estimation algorithm (based on gammatone filtering and autocorrelation), or by taking one estimate (higher or lower) of the DDF algorithm. Within each column, a smaller percentage indicates that the algorithm was more accurate.

	>10%	>3%	>1%
Gammatone/ACF	3.0%	24%	46%
DDF (iterative, lower estimate)	1.5%	10%	23%
DDF (exhaustive, lower estimate)	0.69%	6.5%	18%
DDF (exhaustive, higher estimate)	2.7%	14%	35%

III. CANCELLATION MODEL OF AUDITORY PROCESSING

A. The "transparency of sound"

People can often attend to one sound within an auditory scene as if that sound were presented alone. Sounds within the scene, though perceptible if attended to, can also be ignored, and this is what Bregman (1990) calls "transparency of sound" by contrast to what happens in visual scenes where interfering objects are generally opaque. The mechanism that organizes the sound scene apparently removes the correlates of the interfering sounds from those of the total scene, leaving the correlates of the target sound. Whether subtraction actually occurs at a peripheral level (on a low-level physiological representation of sound) or is simulated at a cognitive level (reconstruction of the percept of the target "as if" it were alone) is not clear, but many experiments can be explained according to Bregman's "old plus new" heuristic (Bregman, 1990): When two sounds start and stop at the same time they tend to fuse, and both are difficult to hear separately. When one starts earlier however, not only is it easier to hear, but the other sound is also easier to hear. An interpretation is that the spectrum of the first sound was estimated during the interval when it was alone, and then *subtracted* from the spectrum of the composite sound.

Of course, "subtraction" operates on a representation internal to the listener and not on the spectrum itself. Implicit is therefore the idea that the representation (periph-

eral or central) used by the separation mechanism is isomorphic to a spectrum. A variety of physiological results (Sachs and Young, 1979; Carney and Yin, 1988; Møller, 1977a,b; Horst *et al.*, 1986; Palmer, 1988, 1990) suggest that the spectral resolution of the auditory periphery may not be sufficient for a spectral model along the lines of that of Parsons (1976) to work. This was confirmed by Assmann and Summerfield (1990) using parameters based on psychoacoustic measures of selectivity. Spectro-temporal schemes require less resolution and are therefore more successful (Weintraub, 1985; Assmann and Summerfield, 1990; Meddis and Hewitt, 1992; Palmer, 1990, 1992). This section examines a further possibility, that of a time-domain method of separation.

The mechanism to be described is not incompatible with peripheral frequency analysis. A comprehensive model might employ filtering followed by time-domain processing, to make up for limitations of the former (limited frequency selectivity) and the latter (limited linearity). For reasons of clarity, we shall discuss the time-domain mechanism as if it operated alone, but by this *we do not mean to imply that peripheral frequency analysis plays no role, or only a minor role, in harmonic sound separation.*

B. Time-domain cancellation model

One could imagine a version of the "old-plus-new" heuristic in which a time-domain "template" is subtracted, but it is difficult to imagine how such a template could be generated with the right phases for subtraction to be effective. An exception is the case of a periodic stimulus, where a template exists in the form of the pattern left by the stimulus a period earlier in time. Assuming an appropriate form of short-term time-domain sensory memory, cancellation could be achieved by subtracting the stored pattern from the current pattern, at a delay of one period. This running "subtraction," similar in principle to the comb-filter described in Sec. II A, might operate on a time-domain physiological representation of the sound such as that carried by the auditory nerve. In that form, the model is reminiscent of the equalization and cancellation (EC) model of binaural interaction (Durlach, 1963) that has been successful in accounting for many aspects of binaural phenomena (Colburn and Durlach, 1978). According to the EC model, patterns coming from both ears are "equalized" (mainly by a time delay) and "subtracted" to attenuate masker components.

The cancellation process requires knowledge of the fundamental period of the interference. In the following, we shall assume that this knowledge is derived according to an "iterative" scheme (Sec. I C 1) involving Licklider's pitch model (Licklider, 1956, 1959, 1962) and a neural cancellation filter to be described in the next section. In Licklider's model, each channel of the auditory nerve is processed by a coincidence network that calculates the equivalent of an autocorrelation function. This results in a pattern of activity over the two dimensions of frequency (inherited from cochlear analysis), and lag (of the autocorrelation process). A periodic stimulus gives rise to a ridge spanning the frequency dimension at a lag equal to the period, and this is the cue to pitch (Moore, 1982; van Noorden, 1982; Lyon, 1984; de Cheveigné, 1986; Meddis and Hewitt, 1988, 1991a,b, 1992). The implementation of the "iterative" scheme does not have to be sequential: physiology would more likely have it perform in parallel.

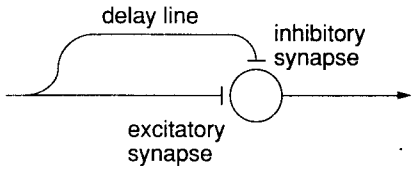

FIG. 5. "Neural" comb filter. Every spike arriving at the excitatory synapse is transmitted, unless a spike arrives simultaneously at the inhibitory synapse.

C. Neural cancellation model

1. Neural "comb filter"

Time domain cancellation requires subtraction. A mechanism that comes to mind immediately is inhibition, and a "neural comb filter" that operates according to this principle is shown in Fig. 5. The neuron is fed via two paths: a direct path with an excitatory synapse, and a delayed path with an inhibitory synapse. The characteristics of the neuron and synapses are such that the neuron is certain to fire on every spike arriving on the direct pathway, unless a spike arrives simultaneously (with a certain margin) along the delayed pathway.

2. Data and analysis methods

The model was tested using data recorded in the guinea pig in response to double vowel stimuli (Palmer, 1990). Palmer measured responses of guinea pig auditory-nerve fibers to stimuli consisting of two synthetic vowels /a/ and /i/, at *F*0s of 100 and 125 Hz, respectively. The formant frequencies and bandwidths are given in Table III. Vowels were presented either alone or simultaneously in 500-ms bursts at 1-s intervals, until either 100 presentations were completed or 5000 spikes had been collected. Nerve fiber discharge times relative to stimulus onset were recorded with 10-μs accuracy.

Of the 315 fibers for which Palmer obtained data in response to double vowels, we selected 170 for which data

TABLE III. Frequency and bandwidth in Hertz of the formants *F*1–*F*5 of synthetic vowels /i/ and /a/.

	Freq.	BW	Freq.	BW	Freq.	BW	Freq.	BW	Freq.	BW
/i/	270	90	2290	110	3010	170	3300	250	3850	200
/a/	730	90	1090	110	2440	170	3300	250	3850	200

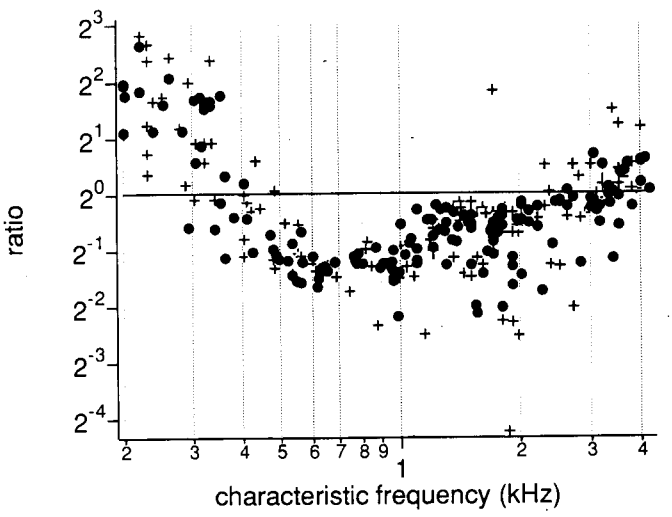

FIG. 6. Ratio of synchrony measures SM(8 ms)/SM(10 ms), for all fibers as a function of characteristic frequency. Response is dominated by /i/ (8 ms) at low and high frequencies, and by /a/ (10 ms) at intermediate frequencies. Dots correspond to the 170 selected fibers that responded to both double and single vowels (see text), crosses represent the remaining 145 fibers for which responses were available for double vowels only.

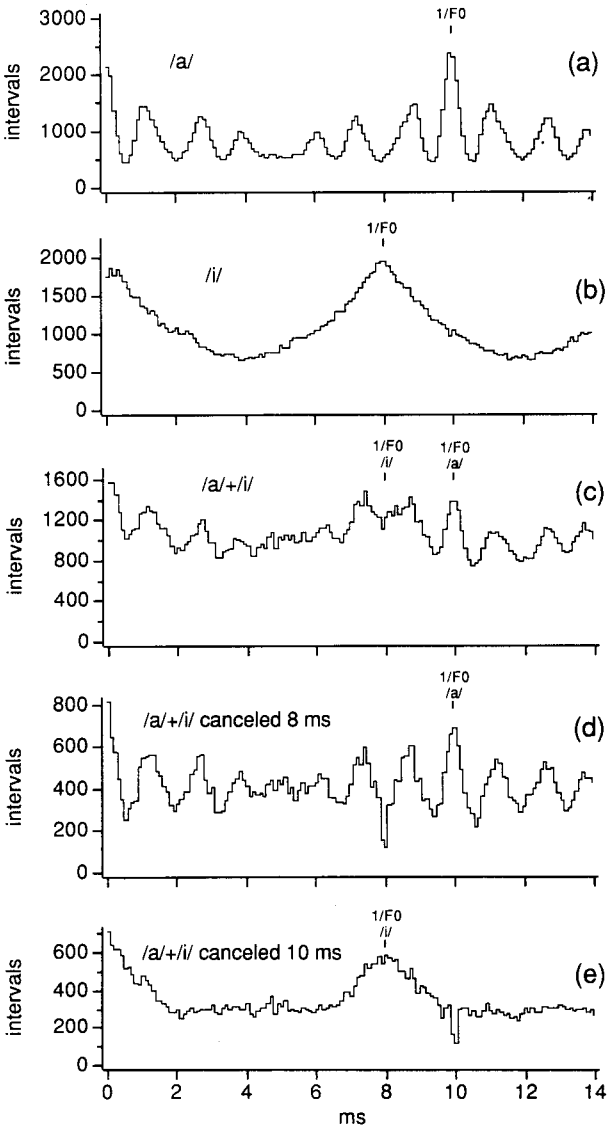

FIG. 7. AC histograms for a single fiber (055017, cf=3.34 kHz, spontaneous rate=92.7 s/s) for vowels /a/ and /i/ in isolation and mixed. Markers above histograms indicate their periods. (a) Response to vowel /a/. (b) Response to vowel /i/. (c) Response to mixed vowel /a/+/i/. Note the split "peak" at 8 ms, and the smaller but sharper peak at 10 ms. This histogram is also plotted at a different scale in Fig. C1 (c). (d) Same as (c), but spike data were filtered by a "neural" comb filter of lag 8 ms. The peaks around 8 ms are reduced. (e) Same as (c), but spike data were filtered by a "neural" comb filter of lag 10 ms. Most evidence of synchrony to 10 ms has disappeared. Bin width is 0.1 ms.

were also available in response to both isolated vowels (this is not always the case because a fiber can get lost before a complete stimulus set is presented). The data were processed in three steps (Appendix C). In the first step, the data for each fiber were sorted according to spike occurrence time relative to stimulus onset, to combine the data from successive presentations of the stimulus into a single run. This operation simulates the result of recording simultaneously from several identical fibers (as many as there were presentations) in response to a single presentation. In the second step, the sorted data for each fiber were processed by the "neural comb filter" of Fig. 5. In the third step, the filtered data were processed to obtain four different representations: *autocoincidence (AC) histograms* for single fibers (also called autocorrelograms, autocorrelation histograms, or discrete autocorrelation functions) (Ruggero, 1973; Evans, 1983; Palmer, 1990), *pooled AC histograms, pooled period histogram Fourier transforms (PHFT), and average localized synchrony rate (ALSR) patterns* (Young and Sachs, 1979; Palmer, 1990). This range of representations allows comparisons to be made with other studies and models, but it displays the information somewhat redundantly, so the reader may wish to concentrate on a single representation (for example the pooled AC histogram).

Synchrony to the fundamental of one or the other vowel, before and after processing, was estimated according to a "synchrony measure" (SM) derived from the AC histogram. Its definition, and the motivation for using this measure rather than the more familiar *synchronization index* (ratio of Fourier components at $F0$ and 0 Hz), are explained in Appendix C. The ratio of synchrony measures at the fundamental periods of /i/ and /a/ [SM(8 ms)/SM(10 ms)] for the double vowel stimulus is plotted in Fig. 6, as a function of CF, for all 315 fibers of the original data set. The 170 selected fibers are represented by dots, the others by crosses. Synchrony at the fundamental of /i/ (8 ms, or 125 Hz) dominates most fibers below 400 Hz and many above 2.5 kHz, but the bulk of the fibers in between these frequencies synchronize to /a/ (10 ms, or 100 Hz). This can be understood by examining the spectrum of the stimulus: /a/ exceeds /i/ between 500 Hz and 2 kHz by as much as 36 dB (Palmer, 1990, Fig. 2).

3. Response to single vowels

Figure 7(a) shows an AC histogram in response to vowel /a/ for a single fiber. The response at the fundamen-

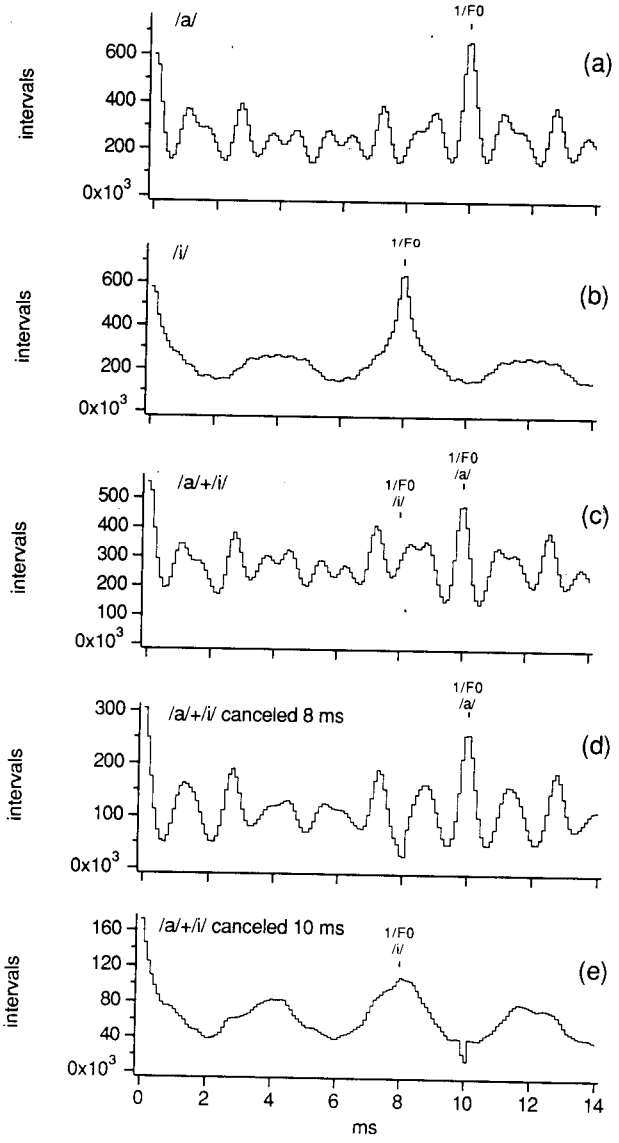

FIG. 8. AC histograms pooled by summing across the selected population of 170 fibers (see text). Markers above histograms indicate the periods of the constituent vowels. (a) Response to vowel /a/. (b) Response to vowel /i/. (c) Response to mixed vowel /a/+/i/. Note that the pooled response appears mostly dominated by /a/. (d) Same as (c), but spike data were filtered by a "neural" comb filter of lag 8 ms. (e) Same as (c), but spike data were filtered by a "neural" comb filter of lag 10 ms. Most of the evidence of the vowel /a/ that dominated the response in (c) has disappeared, and the shape is more similar to the response of /i/ alone. Bin width is 0.1 ms.

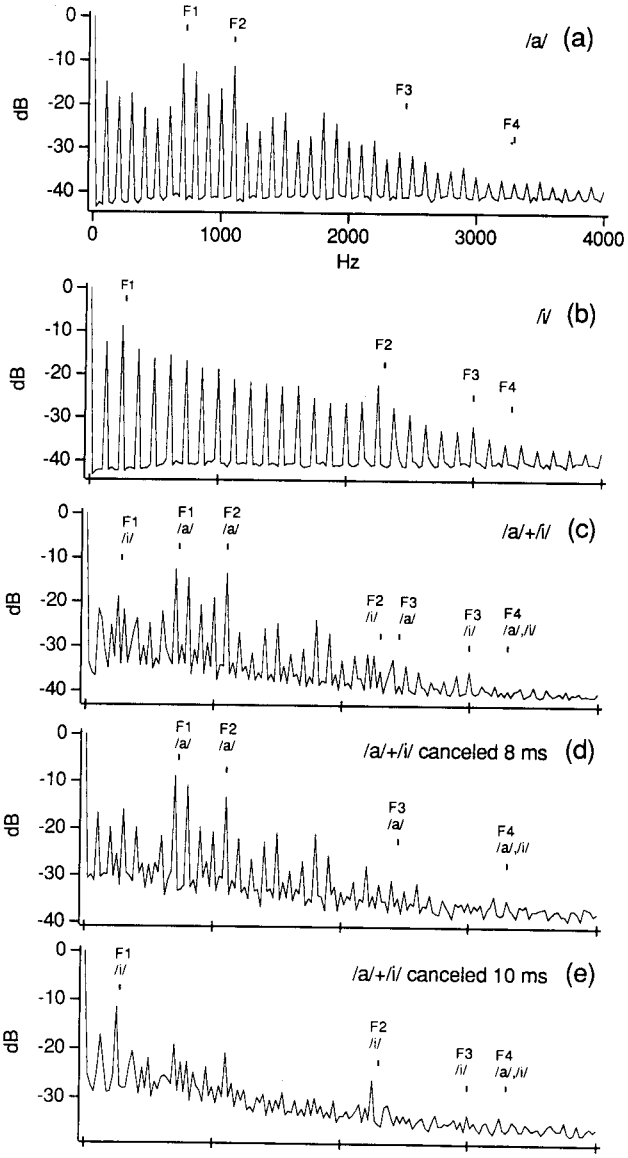

FIG. 9. Pooled period histogram Fourier transforms (PHFT) obtained by summing Fourier transform magnitudes of period histograms across the selected population of 170 fibers (see text). The pooled PHFTs are normalized relative to their zero frequency component and plotted on a decibel scale. Markers above graphs indicate the formants of the constituent vowels. (a) Response to vowel /a/ alone. (b) Response to vowel /i/ alone. (c) Response to mixed vowels /a/+/i/. (d) Same as (c), but spike data were filtered by a "neural" comb filter of lag 8 ms. (e) Same as (c), but spike data were filtered by a "neural" comb filter of lag 10 ms. Note the prominent peaks near the first two formants of /i/ (270 and 2290 Hz), and the lack of evidence of the dominant vowel /a/.

tal (10 ms) is strong, as confirmed by a synchrony measure of 2.8. Across the population of fibers the measure ranges from 1.7 to 7. Figure 7(b) shows the same fiber responding to /i/. The response at 8 ms is strong. The synchrony measure is 1.78 for this fiber and ranges from 1.1 to 12 across the population. The pooled AC histograms shown in Fig. 8(a) and (b) display the same strong synchrony to the fundamental periods.

Figure 9(a) and (b) shows pooled PHFTs (period histogram Fourier transforms) for the same data. Intense harmonics near the first two formants are clearly visible for both vowels. The third formant of /i/ (3010 Hz) can also be made out, but the third formant of /a/ is no more prominent than the spurious peaks at 1500, 1800, or 2200 Hz. The ALSR patterns [Fig. 10(a) and (b)] are similar to the pooled Fourier transforms, with the difference that the first three formants stand out slightly more clearly, especially in the case of /i/.

4. Response to mixed vowels /a/+/i/

Figure 7(c) shows an autocoincidence histogram for a fiber that responded more or less equally to both vowels. [The same data are plotted in Fig. C1 (c) at a different

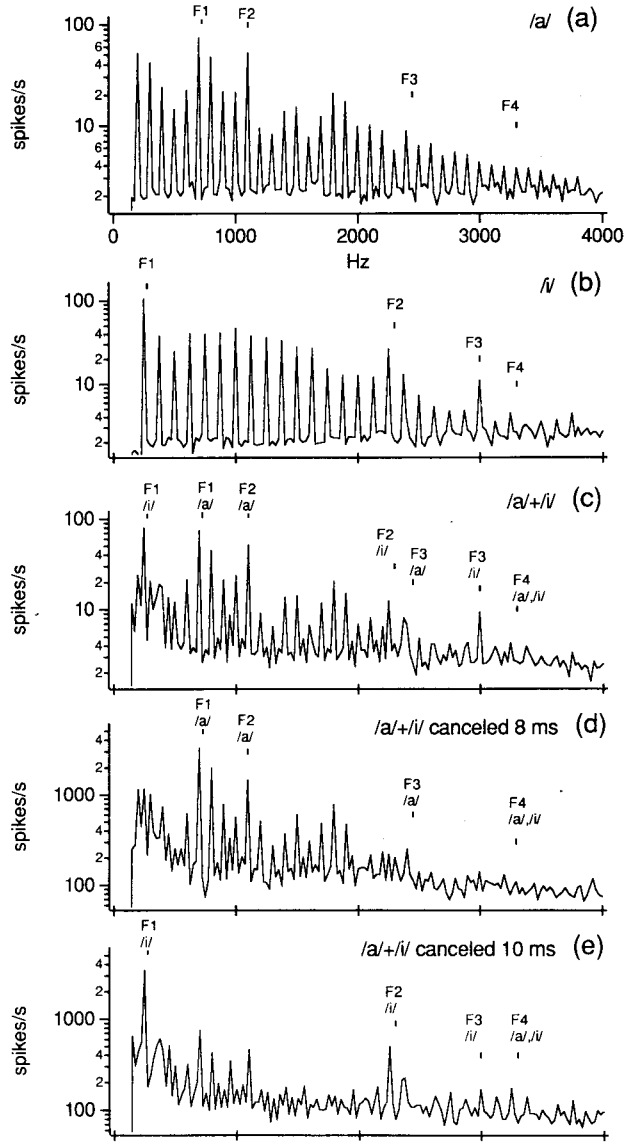

FIG. 10. Average localized synchronized rate (ALSR). Markers above the graph indicate formants of constituent vowels. (a) Response to /a/ alone. (b) Response to /i/ alone. (c) Response to /a/+/i/. (d) Same as (c), but spike data were filtered by a "neural" comb filter tuned to 8 ms. The peaks near *F*1, *F*2, and *F*3 of /i/, prominent in the double vowel response, are attenuated here. (e) Same as (c), but spike data were filtered by a "neural" comb filter tuned to 10 ms. The peaks near *F*1 and *F*2 of /a/, prominent in the double vowel response, are attenuated here, but the peak at the *F*3 of /i/ is also attenuated. Plots (a), (b), and (c) were obtained from unsorted data, (d) and (e) were obtained from sorted (merged) data and have therefore larger ordinate values due to the larger combined rate.

scale. In that figure, small sharp peaks are visible at 10 ms and multiples, and one can make out a pattern of "humps" at multiples of 8 ms. A clear peak occurs at 40 ms, which is the true period of the composite stimulus.]

The pooled AC histogram in Fig. 8(c) is quite dominated by the response to /a/. The pooled PHFT of Fig. 9(c) shows the same dominance: The peaks occur clearly at 100-Hz intervals (except in low- and possibly high-frequency ranges), but the peak-to-trough ratio is much smaller than for the single vowel /a/. Intense harmonics occur near the first two formants of /a/, but the first formant of /i/ is barely visible. The second formant of /i/ is practically invisible, except as a disruption in the periodicity of the spectrum near 2200 Hz. The ALSR [Fig. 10(c)] is similar in shape, but the formants of /i/ are more prominent, including the third formant at 3010 Hz.

5. Responses to mixed vowels filtered at 8 ms

Figure 7(d) shows the response of the single fiber to the double vowel, after "comb filtering" with a lag parameter of 8 ms. The histogram is now clearly dominated by the periodicity of /a/. The pooled AC histogram response in Fig. 8(d) is also dominated by /a/, but that was already the case of the unfiltered response in Fig. 8(c). Both resemble the shape of the pooled histogram for /a/ alone. The pooled PHFT [Fig. 9(d)] is also not very different from that of the unfiltered response. The main difference is a more regular spacing of peaks near the first and second formants of the interfering vowel /i/ (270 and 2290 Hz). The ALSR Fig. 10(d) is similar in shape to the pooled PHFT, except that the ripple indicating periodicity to 100 Hz is slightly less clear.

6. Responses to mixed vowels filtered at 10 ms

Figure 7(e) shows the responses of the single fiber to the double vowel, after "comb filtering" with a lag parameter of 10 ms. The histogram is now clearly dominated by the periodicity of /i/. The same is true of the pooled AC histogram [Fig. 8(e)], whose shape differs radically from the unfiltered response: It is now closer to that of the response to /i/ alone, and no longer reflects the presence of the concurrent vowel /a/. The shape of the low time (below 4.5 ms) part of the pooled AC histogram is the cue to the identity of a vowel according to the model of Meddis and Hewitt (1992). The pooled PHFT in Fig. 9(e) is also very much changed: Harmonics near the first formant of /i/ are prominent, and the second formant shows as a peak at 2250 Hz. The regular 125 Hz spacing of components is visible at low frequencies, but remains unclear in most of the rest of the spectrum. Attenuated peaks remain visible near the formants of /a/ (730 and 1090 Hz). The ALSR in Fig. 10(e) is similar to the pooled PHFT, but the *F*2 peak of /i/ is more prominent, and the ripple indicating a 125-Hz *F*0 is evident over more of the frequency range.

7. Comb filtering for harmonic enhancement

The previous neural comb filter performs harmonic cancellation. It is worth examining briefly a possible physiological implementation of the alternative strategy of *harmonic enhancement.* Consider a neural circuit similar to Fig. 5, but in which the neuron has two excitatory synapses and fires only if discharges arrive simultaneously at both. Licklider's model (1956) employs an array of such circuits, and Delgutte (1984) described a similar filter for neural processing. Figure 11(a) shows the effect of this enhancement filter on the pooled AC histogram of the double vowel response. The pattern resembles that of /i/ alone, except that there is a strong ripple at a period cor-

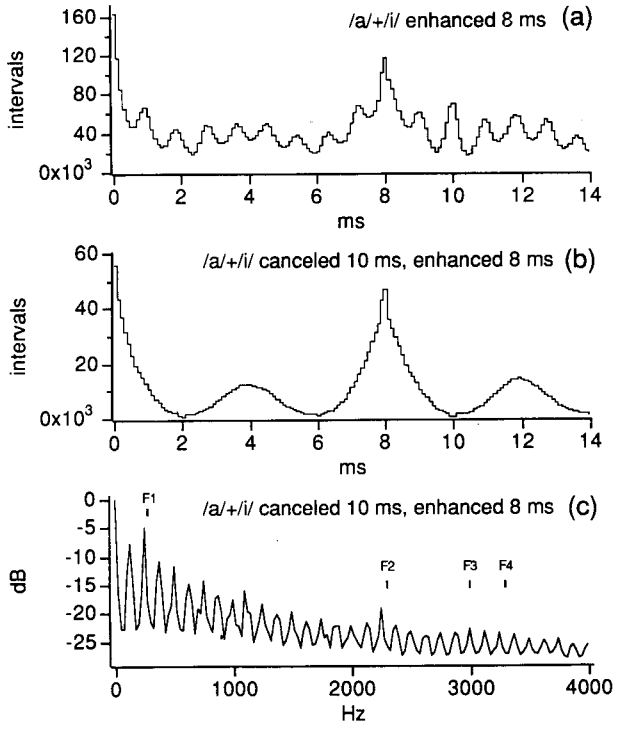

FIG. 11. Enhancement filtering: (a) pooled AC histogram of the response to /a/+/i/ enhancement-filtered at 8 ms. The pattern approaches that of /i/ alone, except for a strong ripple due to the 1090 Hz $F2$ of /a/ [compare with Fig. 8(b)]. (b) Pooled AC histogram of the response to /a/+/i/, cancellation-filtered at 10 ms then enhancement-filtered at 8 ms. The pattern is quite similar to that of /i/ alone [Fig. 8(b)], and no evidence of /a/ is visible. (c) Pooled PHFT of the response to /a/+/i/ cancellation-filtered at 10 ms, then enhancement-filtered at 8 ms [compare with the response to /i/ alone in Fig. 9(b)].

responding to the second harmonic of /a/, 1090 Hz. Figure 11(b) shows the result of applying this enhancement filter after a cancellation filter at 10 ms. The ripple has disappeared, and the result is now strikingly close to the response to /i/ alone [Fig. 8(b)]. The same filtered data are displayed as a pooled PHFT in Fig. 11(c). Peaks near $F1$ and $F2$ of /i/ are strong and there is little evidence of /a/. Enhancement filtering results in a strong peak at 8 ms in the AC histogram [Fig. 11(a) and (b)], and a clear 125-Hz ripple in Fig. 11(c), but we would be wrong to conclude that it could help us estimate the period of /i/, as that information must be available before the filter can be applied.

8. Discussion

The cancellation variety of "neural comb filter" removes evidence of either vowel from the response to their sum, and thus allows the periodicity of the other vowel to stand out. For this particular stimulus, the "iterative" $F0$ estimation mechanism invoked earlier would find both fundamentals in just one pass: a first period estimate from the peak in the pooled AC histogram [Fig. 8(c)] would allow a comb-filter to remove that periodicity, resulting in a peak at 8 ms that indicates the second period [Fig. 8(e)]. The enhancement filter, on the other hand, is useless for $F0$ estimation because the $F0$ must be known beforehand.

The cancellation filter was helpful in restoring the spectral shape of /i/. The vowel /a/ already dominated the response to the mixed vowel and had little to gain. The enhancement filter suffers from crosstalk when used alone [Fig. 11(a)], but in combination with a cancellation filter it is remarkably effective [Fig. 11(b) and (c)].

The neural filters work by partitioning the spike population into subpopulations. This simple strategy is also at work in the channel-selection process of Meddis and Hewitt (1992), but following a different criterion. In that model, spikes are selected according to whether they belong to a channel dominated by a given period, whereas in the present model the partition is according to discharge history. For these particular data, the two partitions are quite similar, because most fibers are completely dominated by one periodicity or the other (Fig. 6). Indeed, Palmer's (1992) implementation of Meddis and Hewitt's model using these data gave results similar to those we report here. However, with other stimuli, the partitions might be less similar.

A "neural comb filter" requires delay lines as long as the longest periods that need canceling. Meddis and Hewitt (1991b) and others argue that there is little physiological evidence for delay lines that long. It is difficult for a nonphysiologist to judge the weight of such a lack of evidence: What are the chances that delay lines might exist but go unnoticed? De Ribaupierre *et al.* (1980) found transmission latencies up to 12 ms in the medial geniculate body of the cat. Langner (1981), Langner and Schreiner (1988), and Schreiner and Langner (1988a,b) found evidence that could be interpreted as supporting Licklider's model which also uses delay lines for periodicity analysis. Strong support has been found for the related cross-correlation model of binaural interaction, within the medial superior olive and inferior colliculus of the cat and guinea pig and related centers in other animals (Jeffress, 1948; Kuwada *et al.*, 1980; Chan *et al.*, 1987; Yin *et al.*, 1987; Konishi *et al.*, 1988; Carney and Yin, 1989; Carr and Konishi, 1990; Palmer *et al.*, 1990; Yin and Chan, 1990). However, delays required for binaural analysis are generally much shorter than those needed for periodicity analysis. Palmer *et al.* (1990) found that best delays in the inferior colliculus were generally longer (in a range of 100–800 μs) than the animal's maximum interaural delay (90 μs). They also found neurons with dips in their response as a function of delay, suggesting a form of inhibitory interaction. It might be that such responses to binaural stimuli reflect mechanisms not intended for binaural processing (McFadden, 1973). The second ingredient required by the neural filter, gating by inhibition, seems less controversial (Wicksberg and Oertel, 1990).

Peripheral frequency analysis does not participate explicitly in the filter process. We found the filter to be effective when applied to a single fiber that responded to both vowels (peripheral filtering having in this case failed to separate them) as well as when applied systematically to all fibers before derivation of a synthetic pattern that ignores tonotopy. However, peripheral frequency analysis affects the time-domain representation that the neural filter

works on, and certainly plays a major role in the system of which this neural filter is a hypothetical part.

D. Predictions: Whispered/voiced mixed vowel recognition

Recognition of concurrent vowels depends on many levels of processing in addition to separation (peripheral filtering, transduction, template matching, etc.), so quantitative agreement between predictions and measured recognition rates may tell us rather little about the separation process itself. Perhaps more interesting are some strong qualitative predictions that can allow us to choose between competing strategies. As pointed out earlier, a fundamental distinction can be made between harmonic cancellation and harmonic enhancement. If we mix whispered and voiced vowels, an enhancement model predicts that *a voiced vowel (V) is easier to hear than a whispered vowel (W)*, whatever the nature of the other vowel that happens to be present. The harmonicity of the target gives it an advantage. A cancellation model predicts instead that *a target is easier to hear if the background is voiced (V) than if it is whispered (W)*, whatever the nature of the voice we want to hear. The harmonicity of the interference gives the target an advantage.

Specifically, if $R(A|B)$ stands for the recognition rate of vowel A mixed with vowel B, the enhancement model predicts

$$R(\mathrm{V}|\mathrm{W}) > R(\mathrm{W}|\mathrm{W}),$$

while the cancellation model predicts

$$R(\mathrm{W}|\mathrm{V}) > R(\mathrm{W}|\mathrm{W}).$$

Furthermore, if voiced vowels V1 and V2 of differing fundamentals are mixed together or with a whispered vowel, the enhancement model predicts:

$$R(\mathrm{V1}|\mathrm{V2}) > R(\mathrm{W}|\mathrm{V2}),$$

and the cancellation model:

$$R(\mathrm{V1}|\mathrm{V2}) > R(\mathrm{V1}|\mathrm{W}).$$

This experiment was carried out by Lea (1992). Synthetic voiced and whispered vowels were produced by the Klatt cascade synthesizer for five English vowels. Spectra were matched by comparing the "excitation pattern" calculated according to the procedure described by Moore and Glasberg (1983) (other procedures to adjust the continuous spectra of the whispered vowels to the line spectra of the voiced vowels gave similar results). Voiced vowels were produced at $F0$s of 100 and 112 Hz. Subjects heard a single presentation of a 200-ms vowel pair, and were required to report both constituents. The results were the following:

(1) $R(100\ \mathrm{Hz}|100\ \mathrm{Hz}) = 79.4\%$,
(2) $R(112\ \mathrm{Hz}|100\ \mathrm{Hz}) = 83.1\%$,
(3) $R(100\ \mathrm{Hz}|112\ \mathrm{Hz}) = 84.3\%$,
(4) $R(\mathrm{whispered}|\mathrm{whispered}) = 79.0\%$,
(5) $R(100\ \mathrm{Hz}|\mathrm{whispered}) = 80.1\%$,
(6) $R(\mathrm{whispered}|100\ \mathrm{Hz}) = 86.2\%$.

The first three lines renew the classic result that recognition of voiced vowels is better when $F0$s differ. The fourth and sixth lines together demonstrate a significant advantage when the background is voiced (the similar pattern between the third and fifth is not statistically significant), and confirm *the cancellation hypothesis*. The fourth and fifth lines show no significant advantage when the target is voiced (the second and sixth line even suggest—nonsignificantly—the opposite). These results *fail to support the enhancement hypothesis*. This is rather surprising, as it suggests that voicing is of no particular use in protecting speech from noise interference. One would have expected the auditory system to use both strategies to some extent.

E. Discussion

The suggestion that the auditory system uses the harmonic nature of the interference, but not that of the target, to separate harmonic sounds comes as a surprise. However, we should realize that interference rejection is most necessary when the signal-to-noise ratio is poor. Under such conditions, the harmonic structure of the interference is easier to extract than that of the target, so cancellation is appropriate. Cancellation is also at work in binaural interaction mechanisms according to the equalization and cancellation model of Durlach (1963). A version of the "neural comb filter" of Fig. 5 could in fact implement the cancellation stage of that model. A comb filter could also be applied to the responses of fibers dominated by a strong vowel formant (for example the $F1$ of /a/), to attenuate that formant and allow the periodicity of other formants to appear. Just how far one can go with this kind of analysis remains to be investigated, but it might be productive to consider, as a working hypothesis, that cancellation is a general form of auditory analysis. As each sound is "heard," it is canceled and removed from the auditory scene, with the ecological advantage that the auditory scene is kept clear, so new sounds stand out as soon as they appear. To explore this hypothesis physiologically might require shifting the emphasis from the search of maxima to that of *minima* of response (as in responses of cells in the IC to binaural stimuli, Palmer *et al.*, 1990).

In binaural experiments, the release from masking obtained in favorable configurations is usually less than 12–15 dB. This may indicate the limits imposed on a neural cancellation process by the imprecision of the neural representation of sound. Peripheral frequency analysis may compensate for this by linearly separating the correlates of various sources according to their frequency content, before transduction. In this sense, the role of the cochlea in sound separation may be that of a "last linear stage" before transduction to a less linear neural code, as suggested by Møller (1977b). Between 1 and 5 kHz, evidence of synchrony in neural discharges disappears, therefore, neural filtering is only imaginable below a certain frequency limit. Binaural masking level differences typically fall from about 12 dB below 500 Hz to about 3 dB above 2 kHz (Colburn and Durlach, 1978). This may indicate an upper frequency limit of neural cancellation mechanisms, although it could

also reflect a specifically binaural limit (related for example to the maximum acoustic delay between the ears).

Auditory scene analysis phenomena are usually interpreted in terms of grouping of frequency components resolved in the auditory periphery. Our results suggest that peripheral channels are not "atomic," but may be further analyzed within the auditory system by time-domain processing. Thus scene analysis mechanisms may exist for which prior resolution into discrete components is not essential.

IV. SUMMARY

(1) A review of the literature showed that most methods and models of voiced speech separation rely on frequency analysis, and that many have difficulty estimating a critical ingredient: the *F*0 of constituent voices.

(2) A time-domain processing principle for harmonic sound separation was presented and formulated in signal processing terms as a time-domain comb filter tuned to the *F*0 of the interfering voice.

(3) An algorithm for the estimation of the *F*0s of mixed voice speech was presented, and evaluated over a restricted data base. The method proved successful for a large proportion of frames. Further work is required to verify whether this algorithm, together with the preceding separation scheme, can be of benefit in speech enhancement or recognition systems.

(4) The processing principle was reformulated in physiological terms as a "neural comb filter" involving a delay line and a gating neuron with inhibitory synapses, and operating on a representation similar to that carried by the auditory nerve.

(5) The filter was tested using guinea pig auditory-nerve fiber discharge data obtained in response to mixed vowel stimuli. It proved successful in isolating the periodicity of either vowel by canceling correlates of the other. It was also partially successful in isolating their spectra.

(6) Predictions were made for a psychoacoustic experiment (Lea, 1992) that can support or contradict the class of cancellation models to which this model belongs. The results support the harmonic cancellation hypothesis.

(7) Time-domain neural processing can complement the spectral or spectro-temporal mechanisms assumed by other models.

ACKNOWLEDGMENTS

Much of this work was carried out while the author was staying at the ATR Auditory and Visual Perception Research Laboratories. The author wishes to thank ATR for their kind hospitality and support, and the Centre National de la Recherche Scientifique (CNRS) for leave of absence. Thanks are also due to the Institut de Recherche et de Coordination Acoustique/Musique (IRCAM) for access to their facilities. Eric Bateson, Nick Campbell, Bertrand Delgutte, Laurent Demany, Tatsuya Hirahara, Andrew Lea, Dominic Massaro, Stephen McAdams, Ray Meddis, Harald Singer, Quentin Summerfield, Yoh'Ichi Tohkura, Minoru Tsuzaki, and an anonymous reviewer all made useful comments on earlier versions of the manuscript. Alan Palmer of the MRC Institute of Hearing Research in Nottingham generously made available the guinea pig data and offered useful suggestions for their interpretation. Andrew Lea, also of Nottingham, shared his unpublished experimental results and contributed useful discussions. John Holdsworth and Roy Patterson of the MRC Applied Psychology Unit in Cambridge made available the gammatone filter software. Jean Laroche of the Ecole Nationale Supérieure des Télécommunications (ENST) contributed the derivation in Appendix A.

APPENDIX A: TRADEOFF BETWEEN ENHANCEMENT RATIO OF A FILTER AND "COMPACTNESS" OF ITS IMPULSE RESPONSE

We wish to show that a filter implementing the enhancement strategy requires a long impulse response to be effective. The harmonic enhancement ratio of a continuous filter can be defined as

$$\alpha = \sum_{k=-\infty}^{+\infty} |H(kf_0)|^2 \Big/ T_0 \int_{-\infty}^{\infty} |H(f)|^2 \, df, \tag{A1}$$

where $H(f)$ is the transfer function and $f_0 = 1/T_0$ is the fundamental frequency of the harmonic series to be enhanced. Let $g(\tau)$ be the autocorrelation function of the impulse response $h(t)$ of the filter

$$g(\tau) = \int_{-\infty}^{\infty} h(t)h(t+\tau)dt.$$

Its Fourier transform $G(f)$ is equal to $|H(f)|^2$, so we have

$$\sum_{k=-\infty}^{\infty} |H(kf_0)|^2 = \sum_{k=-\infty}^{\infty} G(kf_0).$$

But (Poisson formula)

$$\sum_{k=-\infty}^{\infty} G(kf_0) = \left(\frac{1}{f_0}\right) \sum_{k=-\infty}^{\infty} g\left(\frac{k}{f_0}\right)$$

$$= T_0 \sum_{k=-\infty}^{\infty} g(kT_0).$$

And since the integral of the squared amplitude of $H(f)$ is $g(0)$, the enhancement ratio can be expressed as a function of $g(\tau)$ sampled at multiples of the period T_0

$$\alpha = \sum_{k=-\infty}^{\infty} g(kT_0) \Big/ g(0). \tag{A2}$$

For α to be large, there must be many terms in the numerator, which is possible only if there are many terms in the sampled impulse response, which must therefore be long. If we take for example the filter defined by Eq. (1), we find $\alpha = K$, that is, the enhancement ratio is proportional to the number of "prongs" in the impulse response.

APPENDIX B: IMPLEMENTATION DETAILS OF *F0* ESTIMATION ALGORITHMS

1. Test data

Test data were taken from three Japanese sentences pronounced according to five different intonation patterns, by one male speaker (known as MYI) and one female speaker (known as FST). They were taken from the ATR database (Kuwabara *et al.*, 1989). Speech sampled at 20 kHz, 12-bit resolution, was processed by the reference algorithm. The periodicity measure PM was then scanned for portions that exceeded an arbitrary threshold (PM >1.4) for a duration of 225 ms or more. The corresponding portions of speech were trimmed to 225 ms. For each speaker, nine such portions were selected, paired, and summed with all other portions to obtain tokens of "mixed speech" (36 male–male tokens, 36 female–female tokens, and 81 male–female tokens, representing a total of 22 950 analysis frames). No attempt was made to equalize the levels of the voices, or to avoid mixtures with crossing *F0* paths. The primary motivation for selecting portions with a good periodicity was to ensure that the reference *F0* tracks used for evaluation were reliable. Within the speech data, 75% of all voiced frames satisfied the criterion ("voiced" being defined permissively as any run with PM >0.5 for more than 30 ms).

2. Reference single-voice *F0* estimation algorithm

This algorithm was used to process the database to obtain the reference *F0* estimate and a "degree of periodicity" measure that was used for selecting portions of speech for evaluation. Speech was first smoothed by convolution with a 1-ms rectangular window, then the AMDF was calculated using overlapping 20-ms rectangular windows at 1.5-ms intervals. The fixed window was left-aligned on the analysis point, and the moving window moved to the right (negative lag):

$$\mathrm{AMDF}(i,k)=\sum_{m=0}^{N-1}\left|S_{i+m}-S_{i+m+k}\right|,$$

where i is the index of the analysis point, k is the lag, and N is the window size. This function was calculated for a range of lags corresponding (arbitrarily) to a minimum *F0* of 100 Hz for the female speaker and 60 Hz for the male speaker. The AMDF value at each lag was divided by the mean of values for shorter lags, to amplitude normalize the function, eliminate the zero at zero lag, and attenuate spurious dips at short lags:

$$\mathrm{AMDF}'(i,k)=\mathrm{AMDF}(i,k)\Big/\frac{1}{k}\sum_{m=1}^{k}\mathrm{AMDF}(i,m).$$

The lag at the minimum of this new function was taken to be the period [Fig. 2(a)]. The algorithm can easily but inappropriately choose a multiple of the period (subharmonic). This was avoided by further requiring a period minimum to be less than 0.9 times the value at 1/2 or 1/3 of its lag. No other smoothing or error correction was used, but the results were checked by visual inspection. Search was performed over a range of lags corresponding to *F0*s of up to 300 Hz for the male speaker and up to 600 Hz for the female speaker. Period values were transformed to a base 2 logarithmic scale expressing octaves relative to 110 Hz (A2 on the musical scale).

This algorithm produces a by-product that can be interpreted as a measure of degree of periodicity. This is defined as

$$\mathrm{PM}=-\log_2[\mathrm{AMDF}'(T)],$$

where T is the period estimate. The periodicity measure is large (2 to 6) during "steady" portions of voiced speech, and small (close to 0) at transitions and during unvoiced portions. It gives an indication of the reliability of the *F0* estimate produced by the algorithm.

3. DDF mixed-voice *F0* estimation algorithm

This algorithm provides two period estimates from mixed speech. Speech was first smoothed by convolution with a 1-ms window, and then the double difference function was calculated as follows:

$$\mathrm{DDF}(i,k,l)=\sum_{m=0}^{N-1}\left|s_{i+m}-s_{i+m+k}-s_{i+m+l}+s_{i+m+k+l}\right|,$$

where i is the index of the analysis point, k and l are the lags, and N is the window size. Window size, alignment, and analysis increment were the same as for the reference algorithm. The search range for each lag parameter was set to a 3-oct range excluding lags longer by 10% or more than the longest period in the database (based on the reference *F0* data obtained before mixing). This corresponded to a 55- to 440-Hz frequency range. As in the case of the reference algorithm, period estimates were transformed to a base 2 logarithmic scale expressing octaves relative to 110 Hz.

4. Double voice *F0* estimation algorithm based on ACF

This algorithm derives two period estimates from mixed speech from two major peaks in the ACF pattern. The ACF was calculated as the scalar product of a fixed window and a moving window. Window size, alignment, and analysis increment were as in the reference algorithm. Two *F0* estimates were derived as follows.

(1) The lag at the largest peak was taken as a first tentative estimate.

(2) If the value of the ACF at a lag within three samples of the first estimate divided by n ($n=2,3,4,5$) was greater than 0.95 times the ACF at the first estimate, the first estimate was replaced by that lag. This gave period estimate A.

(3) All peaks corresponding to this estimate and its multiples were set to zero by setting to zero all values of the ACF that decreased monotonically with distance from the top of the peak.

(4) The lag at the largest peak of the remaining pattern gave a second tentative estimate.

(5) This estimate was refined as in (2), to give period estimate B.

The search range for this algorithm was set to 3 oct, excluding lags longer by 10% or more than the longest period of either voice.

5. Iterative version of the DDF algorithm

This algorithm provides two period estimates from mixed speech, by performing alternately period estimation and comb filtering. Comb filtering was performed by subtracting from each value a value T samples to the right, where T is an estimate of the period of the voice to cancel. This convention is coherent with the definition of lag used in the $F0$ estimation. For each voice, $F0$ estimates were derived from the global minimum of the AMDF within the search range. Window size and increment were as in the other algorithms. The algorithm was bootstrapped by taking an initial estimate without filtering. The search range was set to a 3-oct range excluding lags larger than 10% of the largest period.

6. Single-voice *F0* estimation algorithm using a gammatone filter bank and ACF

This algorithm estimates a "dominant" $F0$ from the double voice speech signal. The speech signal was fed through a 52 channel gammatone filter bank (Holdsworth *et al.*, 1988; Patterson *et al.*, 1992), after smoothing by convolution with a 1-ms window. Channels were 1 ERB (equivalent rectangular bandwidth) wide, and spaced at 0.25-ERB intervals from 80 Hz to 1 kHz (smoothing removes most energy above that frequency). The filter software was configured so as to align the peaks in the envelopes of the impulse responses of all channels. Each channel output was amplitude-normalized (by dividing each sample by the mean of the absolute magnitude calculated over a 75-ms centered window), half-wave rectified, and smoothed by convolution with a 1-ms window. For each channel the ACF was calculated as the scalar product of a fixed window and a sliding window. Window size and alignment, and analysis increment were as in the other algorithms. The ACFs were summed, and the sum was searched for a maximum. The search range covered 3 oct and excluded lags longer by 10% or more than the longest period of either component (based on the reference $F0$ data). No provision was made to prevent the algorithm from choosing a super-period instead of the period (subharmonic error, sometimes improperly called "octave error"), but as for other algorithms such mistakes were not counted as errors in the evaluation process. Filter delay was compensated for when $F0$ estimates were aligned with reference estimates.

APPENDIX C: PROCESSING OF SPIKE DATA

The steps of processing are detailed in Fig. C1 (a).

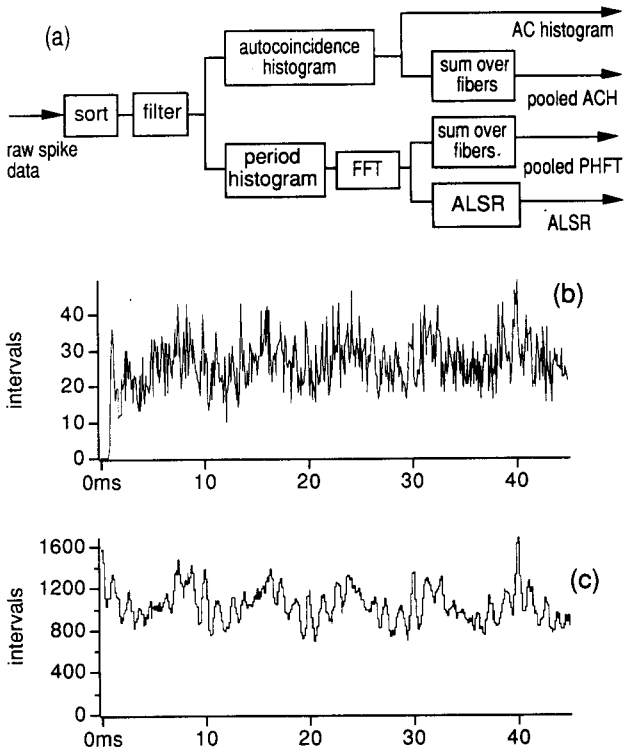

FIG. C1. (a) Flow chart of spike data processing. (b) AC histogram of unsorted data obtained in response to the mixed vowels /a/+/i/, for a single fiber. (c) AC histogram calculated after sorting the same data (the same histogram is plotted at a different scale in Fig. 7). Note the improved regularity and the disappearance of the gap at intervals shorter than the refractory period. Bin width is 0.1 ms.

1. Spike sorting

Data for each fiber, representing spike occurrence times relative to stimulus presentation onset for repeated 500-ms presentations, were trimmed to a 40- to 480-ms post-onset-time range, and data for successive presentations were sorted according to increasing spike occurrence time. The result of sorting simulates the recorded concurrent activity of as many identical fibers as there were stimulus presentations (assuming statistical independence of firing between these hypothetical fibers and, for the actual fiber from which data were collected, between presentations). This operation greatly improves the resolution of AC histograms because the density of *intervals* in a spike train varies with the square of the density of spikes [Fig. C1(c)]. It has no effect on the shape of period histograms or ALSR patterns, but multiplies their ordinate by a factor equal to the number of presentations. The number of presentations was typically about 40 and the apparent discharge rate after sorting was usually in the range of 5000–9000 spikes/s.

2. Neural filter

To simulate the neural cancellation filter, all spikes preceded by a spike T seconds in the past (within a window of 0.1 ms) were deleted. In the enhancement version of the filter (Sec. III C 7), spikes that did *not* fulfill this condition were deleted instead. Applying this processing to

sorted data corresponds actually to a modified version of the filter in Fig. 5, in which gating neurons receive gating synapses from *several* delayed members of a homogeneous population of fibers.

3. Single fiber and pooled AC histograms

Intervals between spikes, consecutive or not, were accumulated in bins according to their length to form the autocoincidence histogram. Bin width was 0.1 ms. Histograms were calculated for all individual fibers, then added across the 170 selected fibers to obtain pooled AC histograms.

4. Pooled Fourier transforms of period histograms

Data for each fiber were processed to obtain 512-bin period histograms locked to the 40-ms period of the two-vowel stimulus, and Fourier transforms were calculated. The Fourier transform magnitudes for all 170 selected fibers were summed to obtain a pooled period histogram Fourier transform (pooled PHFT).

5. ALSR (average localized synchrony rate)

The ALSR represents the average discharge synchronized to frequency components (at 25-Hz intervals) for neurons whose characteristic frequency (CF) is within 0.25 oct of that frequency (Young and Sachs, 1979; Palmer, 1990). The ALSR was derived from the magnitude Fourier transforms for the 170 selected fibers, taking into account their CFs. In some cases (as noted), the ALSR was calculated from spike data before rather than after sorting. This difference has no effect on the shape of the pattern.

6. Synchrony measure and synchronization index

The synchrony measure reflects the height of the AC histogram at a given lag τ, relative to the rest of the histogram:

$$\mathrm{SM}(\tau) = \mathrm{AC}(\tau)/\mathrm{mean}_{0\text{–}40\ \mathrm{ms}}(\mathrm{AC}).$$

We chose to use this measure in place of the *synchronization index* (ratio of Fourier components at $F0$ and 0 Hz) for two reasons. The first is that several fibers can each respond exclusively to a single harmonic but adequately represent the fundamental if taken together, as already noted by Fletcher (1929, quoted by Schubert, 1978). Such fibers each have a small synchronization index, but the $F0$ can nevertheless easily be extracted (Goldstein and Srulovicz, 1977; Moore, 1982; van Nooden, 1982; de Cheveigné, 1986; Meddis and Hewitt, 1991a). The second is that the probability of discharge within a single fiber can be periodic without there being a strong component at $F0$ (although this component is usually present to a certain degree, due to rectification). These points can be verified by pooling AC histograms from fibers with a low synchronization index in response to either vowel: The result shows a clear synchrony to $F0$.

[1]The terminology would sound more symmetrical if the strategies were called "selection"/"cancellation," or else "enhancement"/"attenuation." However, whereas perfect harmonic cancellation is possible with a short impulse response, Appendix A shows that perfect harmonic selection requires an infinite-length impulse response. "Enhancement" and "cancellation" therefore better describe the result of processing using filters with short impulse responses. The term "subtraction" used by Lea (1992) subsumes both our "cancellation" and spectral subtraction techniques that are not specific to harmonic interference.

[2]However, if components of the voices are slightly mistuned their phase relationship varies, so that the amplitude of the vector sum slowly alternates between the sum and the difference of their amplitudes. This is in principle sufficient to determine both amplitudes.

Assmann, P. F., and Summerfield, Q. (**1988**). "Pitch-pulse asynchrony and the perceptual segregation of competing voices," Speech 88 Conf. (7th FASE), Edinburgh, pp. 531–538.

Assmann, P. F., and Summerfield, Q. (**1989**). "Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency," J. Acoust. Soc. Am. **85**, 327–338.

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.

Bregman, A. S. (**1990**). *Auditory Scene Analysis* (MIT, Cambridge, MA).

Brokx, J. P. L., and Nooteboom, S. G. (**1982**). "Intonation and the perceptual separation of simultaneous voices," J. Phon. **10**, 23–36.

Carney, H., and Yin, T. C. T. (**1988**). "Temporal coding of resonances by low-frequency auditory nerve fibers: single fiber responses and a population model," J. Neurophysiol. **60**, 1653–1677.

Carney, L. H., and Yin, T. C. T. (**1989**). "Responses of low-frequency cells in the inferior colliculus to interaural time differences of clicks: excitatory and inhibitory components," J. Neurophysiol. **62**, 144–161.

Carr, C. E., and Konishi, M. (**1990**). "A circuit for detection of interaural time differences in the brain stem of the barn owl," J. Neurosci. **10**, 3227–3246.

Chan, J. C. K., Yin, T. C. T., and Musicant, A. D. (**1987**). "Effects of interaural time delays of noise stimuli on low-frequency cells in the cat's inferior colliculus. II. Responses to band-pass filtered noises," J. Neurophysiol. **58**, 543–561.

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

Childers, D. G., and Lee, C. K. (**1987**). "Co-channel speech separation," Proc. IEEE ICASSP, 181–184.

Colburn, H. S., and Durlach, N. I. (**1978**). "Models of binaural interaction," in *Handbook of Perception*, edited by E. C .Carterette and M. P. Friedman (Academic, New York), pp. 467–518.

Darwin, C. J., and Culling, J. F. (**1990**). "Speech perception seen through the ear," Speech Commun. **9**, 469–475.

de Cheveigné, A. (**1986**). "A pitch perception model," Proc. IEEE ICASSP, 897–900.

de Cheveigné, A. (**1990**). "*F*0 estimation from mixed speech," ATR Auditory and Visual Perception Research Labs. Tech. Rep. TR-A-0097.

de Cheveigné, A. (**1991**). "A mixed speech *F*0 estimation algorithm," Proc. ESCA (Eurospeech), Genova, pp. 445–448.

Delgutte, B. (**1984**). "Speech coding in the auditory nerve: II. Processing schemes for vowel-like sounds," J. Acoust. Soc. Am. **75**, 879–886.

de Ribaupierre, F., Rouiller, E., Toros A., and de Ribaupierre, Y. (**1980**). "Transmission delay of phase-locked cells in the medial geniculate body," Hear. Res. **3**, 65–77.

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. 1568–1580.

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

Evans, E. F. (**1983**). "Pitch and cochlear nerve fibre temporal discharge patterns," in *Hearing-Physiological Bases and Psychophysics*, edited by R. Klinke and R. Hartmann (Springer-Verlag, Berlin), pp. 140–146.

Fletcher, H. (**1929**). *Speech and Hearing* (van Nostrand, New York).

Frazier, R. H., Samsam, S. Braida, L. D., and Oppenheim, A. V. (**1976**). "Enhancement of speech by adaptive filtering," Proc. IEEE ICASSP, 251–253.

Goldstein, J. L., and Srulovicz, P. (**1977**). "Auditory-nerve spike intervals as an adequate basis for aural frequency measurement," in *Psycho-*

physics and Physiology of Hearing, edited by E. F. Evans and J. P. Wilson (Academic, London), pp. 337–347.

Hanson, B. A., and Wong, D. Y. (**1984**). "The harmonic magnitude suppression (HMS) technique for intelligibility enhancement in the presence of interfering noise," IEEE ICASSP **2**, 18A.5.1-4.

Hess, W. (**1983**). *Pitch Determination of Speech Signals* (Springer-Verlag, Berlin).

Holdsworth, J., Nimmo-Smith, I., Patterson, R. D., and Rice, P. (**1988**). "Implementing a Gamma Tone filter bank," MRC Applied Psychology Unit Tech. Rep.

Horst, J. W., Javel, E., and Farley, G. R. (**1986**). "Coding of spectral fine structure in the auditory nerve. I. Fourier analysis of period and interspike interval histograms," J. Acoust. Soc. Am. **79**, 398–416.

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

Konishi, M., Takahashi, T. T., Wagner, H., Sullivan, W. E., and Carr, C. E. (**1988**). "Neurophysiological and anatomical substrates of sound localization in the owl," in *Auditory Function—Neurobiological Bases of Hearing*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York), pp. 721–745.

Kopec, G. E., and Bush, M. A. (**1989**). "An LPC-based spectral similarity measure for speech recognition in the presence of co-channel speech interference," Proc. IEEE ICASSP, 270–273.

Kuwabara, H., Sagisaka, Y., Takeda, K., and Abe, M. (**1989**). "Construction of ATR Japanese speech database as a research tool," ATR Interpreting Telephony Research Laboratories Technical Report TR-I-0086.

Kuwada, S., Yin, T. C. T., Haberly, L. B., and Wickesberg, R. E. (**1980**). "Binaural interaction in the cat inferior colliculus: physiology and anatomy," in *Psychophysical, Physiological and Behavioral Studies in Hearing*, edited by G. v. d. Brink and F. A. Bilsen (Delft U.P., Delft, The Netherlands), pp. 401–411.

Langner, G. (**1981**). "Neuronal mechanisms for pitch analysis in the time domain," Exp. Brain Res. **44**, 450–454.

Langner, G., and Schreiner, C. E. (**1988**). "Periodicity coding in the inferior colliculus of the cat. I. Neuronal mechanisms," J. Neurophysiol. **60**, 1799–1822.

Lea, A. (**1992**). "Auditory models of vowel perception," unpublished doctoral dissertation, University of Nottingham.

Lea, A. P., and Summerfield, Q. (**1992**). "Monaural segregation of competing voices," Proc. ASJ Committee on Hearing H-92-31, pp. 1–7.

Licklider, J. C. R. (**1956**). "Auditory frequency analysis," in *Information Theory*, edited by C. Cherry (Butterworth, London), pp. 253–268.

Licklider, J. C. R. (**1959**). "Three auditory theories," in *Psychology, a Study of a Science*, edited by S. Koch (McGraw-Hill, New York), pp. 41–144.

Licklider, J. C. R. (**1962**). "Periodicity pitch and related auditory process models," Int. Audiol. **1**, 11–36.

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

Lyon, R. F. (**1983/1988**). "A computational model of binaural localization and separation," Proc. IEEE ICASSP, reproduced in *Natural computation*, edited by W. Richards (MIT, Cambridge, MA), pp. 319–327.

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

McFadden, D. (**1973**). "Precedence effects and auditory cells with long characteristic delays," J. Acoust. Soc. Am. **54**, 528–530.

Meddis, R., and Hewitt, M. (**1988**). "A computational model of low pitch judgement," in *Basic Issues in Hearing*, edited by H. Duifuis, J. W. Horst, and H. P. Witt (Academic, London), pp. 148–153.

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.

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.

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

Min, K., Chien, D., Li, S., and Jones, C. (**1988**). "Automated two speaker separation system," Proc. IEEE ICASSP, 537–540.

Moore, B. C. J. (**1982**). *An Introduction to the Psychology of Hearing* (Academic, London).

Moore, B. C. J., and Glasberg, B. R. (**1983**). "Suggested formulae for calculating auditory-filter bandwidths and excitation patterns," J. Acoust. Soc. Am. **74**, 750–753.

Møller, A. R. (**1977a**). "Frequency selectivity of single auditory-nerve fibers in response to broadband noise stimuli," J. Acoust. Soc. Am. **62**, 135–142.

Møller, A. R. (**1977b**). "Frequency selectivity of the basilar membrane revealed from discharges in auditory nerve fibers," in *Psychophysics and Physiology of Hearing*, edited by E. F. Evans and J. P. Wilson (Academic, London), pp. 197–207.

Nagabuchi, H., Kobayashi, T., and Yamamoto, H. (**1979**). "Speech enhancement and suppression in mixed speech," Trans. IECE (Japan) **62**, 627–634 (in Japanese).

Naylor, J. A., and Boll, S. F. (**1987**). "Techniques for suppression of an interfering talker in co-channel speech," Proc. ICASSP, 205–208.

Palmer, A. R. (**1988**). "The representation of concurrent vowels in the temporal discharge patterns of auditory nerve fibers," in *Basic Issues in Hearing*, edited by H. Duifhuis, J. W. Horst, and H. P. Wit (Academic, London), pp. 244–251.

Palmer, A. R. (**1990**). "The representation of the spectra and fundamental frequencies of steady-state single- and double-vowel sounds in the temporal discharge patterns of guinea pig cochlear-nerve fibers," J. Acoust. Soc. Am. **88**, 1412–1426.

Palmer, A. R. (**1992**). "Segregation of the responses to paired vowels in the auditory nerve of the guinea-pig using autocorrelation," in *Audition Speech and Language*, edited by B. Schouten (Mouton-DeGruyter, Berlin) (in press).

Palmer, A. R., Rees, A., and Caird, D. (**1990**). "Interaural delay sensitivity to tones and broad band signals in the guinea-pig inferior colliculus," Hear. Res. **50**, 71–86.

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

Patterson, R. D., Robinson, K., Holdsworth, J., McKeown, D. Zhang, C., and Allerhand, M. (**1992**). "Complex sounds and auditory images," in *Auditory Physiology and Perception*, edited by Y. Cazals, L. Demany, and K. Horner (Pergamon, Oxford), pp. 429–446.

Ross, M. J., Shaffer, H. L. Cohen, A., Freudberg, R., and Manley, H. J. (**1974**). "Average magnitude difference function pitch extractor," IEEE Trans. Acoust. Speech Signal Process. **ASSP-22**, 353–362.

Ruggero, M. A. (**1973**). "Response to noise of auditory nerve fibers in the squirrel monkey," J. Neurophysiol. **36**, 569–587.

Sachs, M. B., and Young, E. D. (**1979**). "Encoding of steady-state vowels in the auditory nerve: representation in terms of discharge rate," J. Acoust. Soc. Am. **66**, 470–479.

Scheffers, M. T. M. (**1983**). "Sifting vowels," Ph.D. thesis, University of Gröningen.

Schreiner, C. E., and Langner, G. (**1988a**). "Coding of temporal patterns in the central auditory nervous system," in *Auditory Function—Neurobiological Bases of Hearing*, edited by G. M. Edelman, W. E. Gall, and W. M. Cowan (Wiley, New York), pp. 337–361.

Schreiner, C. E., and Langner, G. (**1988b**). "Periodicity coding in the inferior colliculus of the cat. II. Topographical organization," J. Neurophysiol. **60**, 1823–1840.

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

Schubert, E. (**1978**). "History of research on hearing," in *Handbook of Perception*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), pp. 41–80.

Silva, F. M., and Almeida, L. B. (**1990**). "Speech separation by means of stationary least-squares harmonic estimation," Proc. IEEE ICASSP, 809–812.

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.

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.

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.

Summerfield, Q., and Assmann, P. F. (**1991**). "Perception of concurrent vowels: effects of harmonic misalignment and pitch-period asynchrony," J. Acoust. Soc. Am. **89**, 1364–1377.

van Noorden, L. (**1982**). "Two channel pitch perception," in *Music, Mind, and Brain*, edited by M. Clynes (Plenum, London), pp. 251–269.

Weintraub, M. (**1985**). "A theory and computational model of auditory monaural sound separation," Ph.D. thesis, Stanford University.

Weintraub, M. (**1986**). "A computational model for separating two simultaneous sounds," Proc. IEEE ICASSP, Tokyo **1**, 3.1.1-4.

Wickesberg, R. E., and Oertel, D. (**1990**). "Delayed, frequency-specific inhibition in the cochlear nuclei of mice: a mechanism for monaural echo suppression," J. Neurosci. **10**, 1762–1768.

Yin, T. C. T., and Chan, J. C. K. (**1990**). "Interaural time sensitivity in medial superior olive of cat," J. Neurophysiol. **64**, 465–488.

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

Young, E. D., and Sachs, M. B. (**1979**). "Representation of steady-state vowels in the temporal aspects of the discharge patterns of populations of auditory-nerve fibers," J. Acoust. Soc. Am. **66**, 1381–1403.