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, 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 Brox 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 fiber 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-
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 F0 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 Parsons'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, F0 estimation was done by searching for two peaks in the cepstrum. In further papers Stubbs and Summerfield (1990, 1991) emphasize the difficulties of F0 tracking, but show that if these difficulties are overcome (using F0 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 F0 estimates were obtained from peaks in the autocorrelation function of the LPC residual. F0 estimation was done in two steps: in a first step the either voice to achieve better filter spectral sampling. F0 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 F0s 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
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 F0s to 62% for a semitone difference. Scheffers then developed a moving model similar to Beranek's method.

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 F0s 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 F0. They found that the model performed poorly, mainly because the spectral representation lacked resolution. Among other things, the model had difficulty finding the F0s 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 func-
merfield. These place-time methods were disappointing, mainly because they failed to simultaneously estimate both F0s. However, using the same data, Palmer (1992) tested Meddis and Hewitt’s model that only requires one F0, 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, F0s 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 F0s 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 F0s directly, as do methods that only require one F0: 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 F0 from a single voice is known to be troublesome (Hess, 1983), and estimation of two F0s (or even a single one) from mixed speech is likely to be even more difficult. Extracting the F0 of one voice is easier if the other voice is first attenuated, and for this reason, separation and F0 estimation often work hand in hand.

![Figure 1](image)

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

loose sense, iterative F0 estimation is proposed by Parsons (1976), Nagabuchi et al. (1979), Scheffers (1983), and Lea (1992).

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:

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

where \(t\) is time, \(T\) is the period of the target voice, \(\delta\) 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 suf-
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 $F_0$ 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.

**B. Mixed voiced speech $F_0$ 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, $\tau$ is the lag, and $\delta$ 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 $F_0$ 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 $F_0$ 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 $\tau_A$ and $\tau_B$, 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 $F_0$ 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 $F_0$ estimation algorithm; (2) extract 225-ms portions of speech with good periodicity (PM > threshold); (3) pair and add to obtain mixed speech.
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 F0s 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 F0s (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. Convention 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. IC 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 1: Parameters of emission data (for ethanol. 100, 20, or 10% Annovia, B 1% and the search range was restricted to 3
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; Möller, 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 su-

FIG. 5. "Neural" comb filter. Every spike arriving at the excitatory synapse is transmitted, unless a spike arrives simultaneously at the inhibitory synapse.
Ratio of occurrence, $\frac{EM(\text{Car} \text{a})}{EM(10 \text{ car})}$ for all
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 "windmill" type filter of 8 ms. (e) Same as (d), but spike data were filtered by a "windmill" type filter of 10 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 prongs indicate the formants of the vowels.
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 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 F2 peak

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 F1, F2, and F3 of /i/, prominent in the double vowel response, are attenuated here. (e) Same as (c), but spike data were filter.
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 non-physiologist 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 psycho-
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

The first three lines renew the classic result that recognition of voiced vowels is better when FOs 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
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 MY1) 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 > 0.3) for a duration of 905 ms.

Search was performed over a range of lags corresponding to F0s 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

$$PM = -\log_2[AMDF'(T)]$$

where $T$ is the period estimate. The periodicity measure is large (2 to 6) during “steady” portions of voiced speech,
(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

![](image)

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
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

1The 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.

2However, 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.


