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, 7, pl. 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
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 (Schröeder, 1968). The FO 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, FO estimation was done by searching for two peaks in the cepstrum. In further papers Stubbs and Summerfield (1990, 1991) emphasize the difficulties of FO tracking, but show that if these difficulties are overcome (using FO values obtained from the speech before mixing) natural speech sentences can be separated quite successfully using either cepstral filtering or harmonic selection.

Nagahuchi et al. (1979) proposed a method similar to that of Parsons, in which FO estimates were obtained from peaks in the autocorrelation function of the LPC residual. FO estimation was done in two steps: in a first step the highest peak of the autocorrelation function indicated the FO of the dominant voice. That estimate served to determine a frequency-domain comb filter that eliminated the first voice, allowing the second FO 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 FO 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 FO 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 FO) to a multiple of the period of either voice to achieve better spectral sampling. FO 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 auto-correlation (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 FOs 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 FO estimation algorithm that required the component voices to be in different registers. Tested over a database of natural speech (digits), it gave a first FO 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 FO
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 processing model similar to Parson’s method, in which all spectral components within 3% of multiples of the F0 of a constituent vowel were selected by a “harmonic sieve.” The F0s 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 F0 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 F0s, 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 F0: for one subject, the rate improved from 60% for equal F0s 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 F0, but Assmann and Summerfield (1988, 1989; Summerfield and Assmann, 1991) also found, as did Scheffers, that with equal F0s 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 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 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 F0 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 F0s. 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 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 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

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

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:

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

(1)

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 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)]:

\[
h(t) = \frac{1}{T} [\delta(t)-\delta(t-T)].
\]

(2)

A. de Cheveigné: Time-domain harmonic sound separation
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 enhacement" 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
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 meth-
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.

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 (error in octaves) of the DDF estimates with their closest F0 reference.

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.

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 somewhat simplified DDF algorithm, which...
TABLE I. Percentage of estimates that fell further than 10%, 3%, or 1% of an octave from a reference estimate when both F0s 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.

<table>
<thead>
<tr>
<th></th>
<th>&gt; 10%</th>
<th>&gt; 3%</th>
<th>&gt; 1%</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACF</td>
<td>44%</td>
<td>62%</td>
<td>76%</td>
</tr>
<tr>
<td>DDF (iterative)</td>
<td>3.1%</td>
<td>14%</td>
<td>31%</td>
</tr>
<tr>
<td>DDF (exhaustive)</td>
<td>1.7%</td>
<td>10%</td>
<td>27%</td>
</tr>
</tbody>
</table>

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

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 F0s 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. JLA is likely
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; Moller, 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 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).
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.

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 occur-
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 /l/. 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 8(a) and (b) shows pooled PHFTs (periods...
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 cancellation filter was helpful in restoring the spectral shape of /a/. 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 Ribaucour 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

\[ R(V|W) > R(W|W) , \]

The first three lines renew the classic result that recognition of voiced vowels is better when F0s 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 fragment (for example the F1 of /a/) to attenuate
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 script. 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
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 > 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 22950 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):

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

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:

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

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 (sub-harmonic). 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 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, 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:

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

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 0.23-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 (sub-harmonic 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).

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 ACF 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 synaptic inputs 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 autocorrelation 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 \( \tau \), relative to the rest of the histogram:

\[
SM(\tau) = AC(\tau)/\text{mean}_{0.40 \text{ ms}}(AC).
\]

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 Srluvicz, 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.

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 "rectification" used by Lea (1992) subsumes both our "cancellation" and spectral subtraction techniques that are not specific to harmonic interference.

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.


Goldstein, J. L., and Srluvicz, P. (1977). "Auditory-nerve spike intervals as an adequate basis for aural frequency measurement," in Psycho-
A. de Cheveigne: Time-domain harmonic sound separation


