Missing-data model of vowel identification[a)]

Alain de Cheveigné[b)]
Laboratoire de Linguistique Formelle, CNRS/Université Paris 7, 2 place Jussieu, case 7003, 75251, Paris, France and ATR Human Information Processing Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

Hideki Kawahara
Media Design Informatics Group, Design Information Science Department, Faculty of Systems Engineering, CREST/Wakayama University, Sakaedani, Wakayama 640-8510, Japan and ATR Human Information Processing Research Laboratories, 2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan

(Received 9 October 1998; revised 15 January 1999; accepted 5 February 1999)

Vowel identity correlates well with the shape of the transfer function of the vocal tract, in particular the position of the first two or three formant peaks. However, in voiced speech the transfer function is *sampled* at multiples of the fundamental frequency (F_0), and the short-term spectrum contains peaks at those frequencies, rather than at formants. It is not clear how the auditory system estimates the original spectral envelope from the vowel waveform. Cochlear excitation patterns, for example, resolve harmonics in the low-frequency region and their shape varies strongly with F_0. The problem cannot be cured by smoothing: lag-domain components of the spectral envelope are aliased and cause F_0-dependent distortion. The problem is severe at high F_0's where the spectral envelope is severely undersampled. This paper treats vowel identification as a process of pattern recognition with *missing data*. Matching is restricted to available data, and missing data are ignored using an F_0-dependent weighting function that emphasizes regions near harmonics. The model is presented in two versions: a frequency-domain version based on short-term spectra, or tonotopic excitation patterns, and a time-domain version based on autocorrelation functions. It accounts for the relative F_0-independency observed in vowel identification. © *1999 Acoustical Society of America.*
[S0001-4966(99)00906-6]

PACS numbers: 43.71.An, 43.72.Ar, 43.66.Ba [JMH]

INTRODUCTION

In voiced speech, the vocal tract resonator is excited with a regular train of glottal pulses, due to opening and closing of the glottis at a rate equal to the fundamental frequency (F_0). According to the acoustic theory of speech production (Fant, 1970), speech is the result of filtering this train by the resonator. The shape of glottal pulses depends on the mode of phonation and characteristics of the speaker. This shape can be included mathematically within the vocal tract impulse response, and the vocal tract is then seen as excited by a train of pulses, infinitely narrow in time. In the following, the term ''vocal tract'' implies this resonator-*cum*-glottal-pulse-shape equivalent. In the frequency domain, if F_0 is constant, the spectrum of the excitation is a series of equal-amplitude peaks at multiples of F_0. The speech spectrum therefore also consists of peaks with amplitudes determined by the amplitude of the transfer function. In other words, the transfer function of the vocal tract is *sampled* at multiples of F_0 (Fig. 1).

The timbre and identity of a sustained vowel are determined by the shape of the vocal tract transfer function, particularly the positions of the first two or three formants. However, the listener has no access to this shape, but only to the waveform or auditory representations derived from it. Figure 2 shows the rms output of a bank of gammatone filters in response to the Japanese vowel /a/. This pattern can be taken as approximating the activity evoked by the vowel over a tonotopic dimension within the auditory system (excitation pattern). At $F_0 = 50$ Hz (top), the pattern is smooth with two clear peaks corresponding to the formants. At $F_0 = 200$ Hz (middle), these peaks are still present, but slightly shifted and there are many other smaller peaks. At $F_0 = 216$ Hz (bottom), the peaks at $F1$ and $F2$ of /a/ are no more prominent than other peaks, and it is not clear what aspect of the excitation pattern might be used to characterize the vowel. Upon listening, the vowel's timbre or identity do not change strikingly between 200 and 216 Hz.

Despite certain interactions (see the Discussion), vowel quality is on the whole remarkably independent of F_0. One could make the hypothesis that the auditory system, by some process that is yet to be understood, forms an internal representation that is invariant over variations of F_0. For example, summation of activity of converging nerve fibers might smooth out the ripples visible in Fig. 2. Indeed, the figure of 3.5 bark has been proposed as an appropriate integration range for vowel spectrum matching. In this paper, we argue against smoothing for several reasons: (a) Undersampling implies a genuine *loss of data*: information about the transfer function at frequencies other than multiples of the F_0 is lost, and smoothing cannot retrieve it. (b) Smoothing and interpolation attempt to guess missing samples based on an *a priori* model of what they should look like, and this may

[a)]Part of this work was presented in an ATR technical report (de Cheveigné and Kawahara, 1998).
[b)]Electronic mail: cheveign@ircam.fr

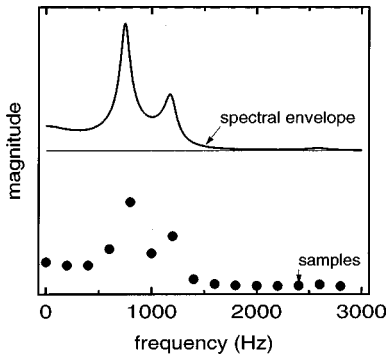

FIG. 1. Line: spectral envelope of vowel /a/. Dots: spectral envelope sampled at intervals of $F_0 = 200$ Hz.

be misleading. (c) In particular, aliasing due to undersampling may produce F_0-dependent distortion that interferes with the recognition process. (d) Aliasing is avoided if the auditory system puts a nonuniform weight on the unsmoothed representations that it derives from the waveform. This weighting function requires an estimate of the F_0.

A. Sampling of the spectral envelope

The shape of a spectral envelope can be described in the Fourier domain along a dimension of *lag* (time interval or inverse frequency). This dimension is also known as *spatial frequency*, or *quefrency* in the context of cepstrum analysis (Rabiner and Schafer, 1978). Components at short lags represent gross features of the spectral envelope that vary gradually along the frequency axis, whereas components with larger lag values represent finer details.[1] We use the term *sampling lag* to designate the inverse T_0 of the spacing F_0 between samples of the spectral envelope. From the sampling theorem, we know that the envelope is adequately represented by the sample points if its shape contains no components beyond half the sampling lag. This limit, $T_0/2$, will be referred to as *Nyquist lag*.

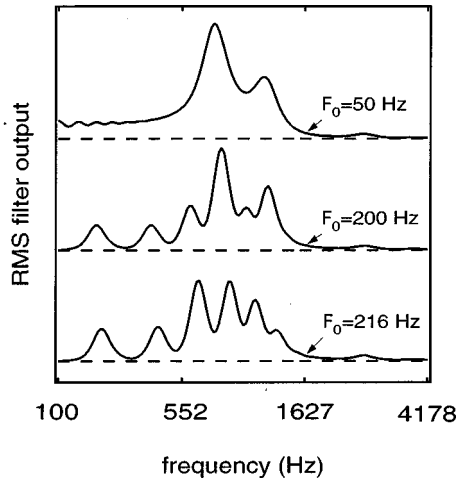

FIG. 2. Magnitude of output of gammatone filter bank as a function of channel frequency (excitation pattern). The filterbank had 150 channels uniformly spaced on a scale of equivalent rectangular bandwidth (ERB) from 100 to 4178 Hz. Each curve is for a different F_0. Note the peaks at harmonics for the higher two F_0's, and the lack of unambiguous evidence for $F1$ and $F2$ at $F_0 = 216$ Hz.

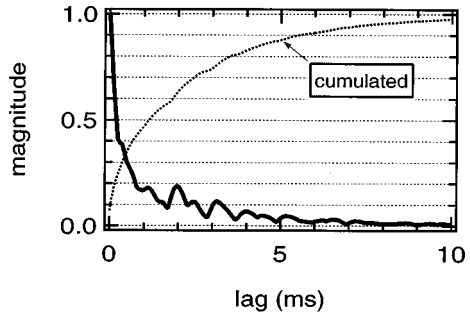

FIG. 3. Full line: magnitude of the lag-domain components of the vowel spectral envelope as a function of lag, averaged over the five Japanese vowels (/a/, /e/, /i/, /o/, /u/). Dotted line: same, cumulated over lags (proportion of lag-domain components with lags smaller than a given lag). The zero-lag value (dc component) is not included in the cumulated-distribution calculation, nor in the plots.

Figure 3 shows the distribution of spectral envelope lag-domain components, averaged over synthetic envelopes of the five Japanese vowels /a/, /e/, /i/, /o/, /u/.[2] The figure was obtained by calculating the magnitude of the Fourier transform of the envelope of each vowel, and averaging over vowels. The low-lag region dominates, but components are also present at larger lags. The cumulated distribution is plotted as a dotted line (the zero-lag point was excluded from both distributions because it represents a dc offset rather than the shape). On average, about 10% of lag-domain components lie beyond 5 ms which is the Nyquist lag for $F_0 = 100$ Hz. These are not adequately represented in the sampled spectrum when $F_0 = 100$ Hz. At 200 Hz (2.5 ms), the proportion is about 25%, and at 300 Hz (1.67 ms), about 40%.

Consider the short-term spectrum of a synthetic vowel /a/ at $F_0 = 100$ Hz [Fig. 4(a), full line at top]. It was calculated by taking the Fourier transform of a 100-ms portion of the waveform. The sampling rate was 40 960 Hz. The short-term spectrum resembles the spectral envelope (dotted line) sampled along the frequency axis at intervals of 100 Hz. If the spectral envelope contained no lag-domain components beyond the Nyquist lag, its shape could be accurately reconstructed by smoothing the short-term spectrum (filtering in the lag domain) to remove components beyond that lag. The result of such smoothing is shown as the lower line in Fig. 4(a). Smoothing was performed by taking the Fourier transform of the short-term magnitude spectrum, setting values beyond the Nyquist lag to zero, then applying the inverse transform. The slight difference between this and the original spectral envelope (dotted line at top) implies that the original did in fact contain components beyond the Nyquist lag. The difference is small, suggesting that little was lost by sampling the envelope at 100-Hz intervals.

At $F_0 = 200$ Hz [Fig. 4(b)], the peaks in the smoothed spectrum (bottom) are wider and there is a strong ripple with a period inverse of the Nyquist lag (2.5 ms). At $F_0 = 300$ Hz [Fig. 4(c)], the reconstructed envelope (full line at bottom) is severely distorted, indicating that spectral envelope components beyond the Nyquist lag (1.67 ms) were necessary to describe the original shape. The significance of the dotted line is discussed in the next paragraph.

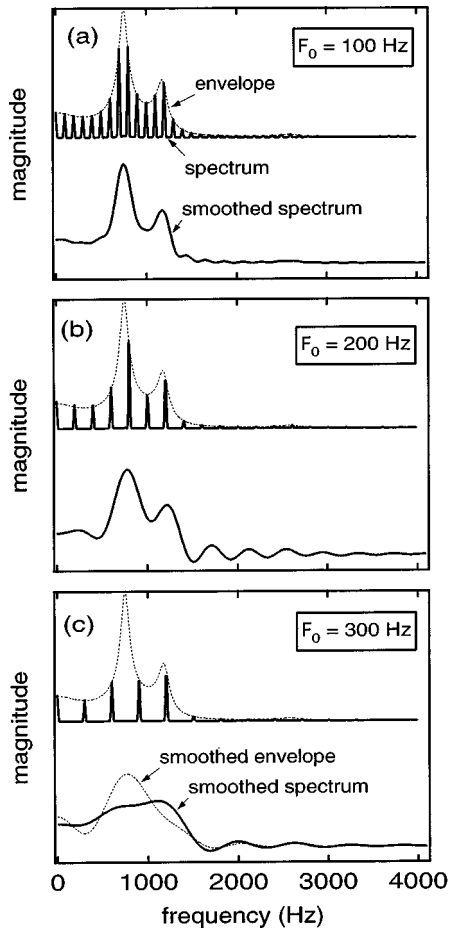

FIG. 4. (a) Top: envelope (dotted line) and short-term magnitude spectrum of /a/ at $F_0 = 100$ Hz (full line). Bottom: smoothed short-term spectrum. Smoothing was performed by taking the Fourier transform of the magnitude spectrum, setting it to zero for lags larger than the Nyquist lag $T_0 = 1/2F_0$ (5 ms), and taking the inverse Fourier transform. The smoothed spectrum consists entirely of components below the Nyquist lag. (b) Same, at $F_0 = 200$ Hz. Note the ripples with a period corresponding to the inverse of the Nyquist lag (2.5 ms), that indicate that aliasing is taking place. (c) Same, at $F_0 = 300$ Hz, with the addition of the smoothed spectral envelope (dotted curve at bottom). The spectral envelope was smoothed by removal of lag components beyond the Nyquist lag. The difference between smoothed envelope and smoothed spectrum is the result of aliasing.

B. Smoothing considered harmful

Undersampling produces data that are incomplete, but correct in the sense that each sample corresponds to a value of the spectral envelope. This is no longer the case if the sampled spectrum is smoothed. Consider the smoothed spectra of Fig. 4. They differ from the original spectral envelope partly because of the absence of high-lag components, and partly because components beyond the Nyquist lag are *aliased*; that is, reflected with respect to the Nyquist lag and reintroduced into the smoothed envelope. The respective contribution of both factors is illustrated in Fig. 4(c). The lower dotted line represents the Nyquist-smoothed spectral envelope, that differs from the spectral envelope (upper dotted line) merely because it lacks components beyond the Nyquist lag. The lower full line (Nyquist-smoothed short-term spectrum) differs from it by the additional factor of aliasing. The effects of aliasing are highly F_0-dependent and may interfere with the identification process, particularly at high F_0's. In other words, smoothing replaces data that are incomplete, but correct, with data that are complete but incorrect.

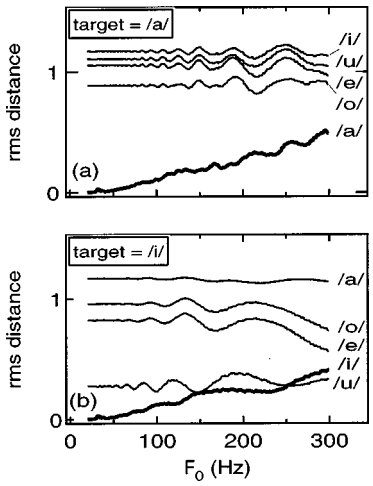

FIG. 5. Simple vowel identification model. (a) Distance between reference templates for vowels /a/, /e/, /i/, /o/, and /u/, and the smoothed short-term spectrum calculated from a target /a/ waveform, as a function of F_0. Smoothing was performed by removing all components beyond the Nyquist lag ($1/2F_0$). (b) Same as (a), for a target /i/ waveform.

To illustrate the difficulties that this causes for identification, a simple vowel-identification model was implemented. The reference template for each vowel was the spectral envelope of that vowel. Target vowels were synthesized at F_0's ranging from 20 to 300 Hz in 1-Hz steps. Their short-term spectra were smoothed by removing components above the Nyquist lag (the value of which depended on the target's F_0), and compared to each template by scaling target and template to the same rms value (1.0) and calculating their rms difference. Target-template distances for target = /a/ are plotted in Fig. 5(a). The distance from competing templates /e/, /i/, /o/, and /u/ remains relatively large, despite some fluctuations. The distance from the ''correct'' template /a/ is smaller, but it increases steadily with F_0, indicating that the estimated envelope is less and less faithful to the original.

A similar plot for the target /i/ is shown in Fig. 5(b). Here, at high F_0's the estimated envelope is actually closer to the incorrect /u/ template than to the correct /i/ template. The model thus fails. It should be stressed that the task is comparatively easy: the vowel set is small, its members well separated in $F1-F2$ space, and there is no variability. In more realistic conditions, the vowel space might be more densely populated and there would be many sources of variability, making the task even more difficult. This model is simple and allows ample room for improvement, but it serves to deliver a message that should be clear: F_0-dependent effects of spectral sampling can be severe, particularly at high F_0, and they cannot be eliminated by smoothing.

I. MISSING-DATA VOWEL IDENTIFICATION MODEL

The model acknowledges that important spectral information was lost due to undersampling. Instead of interpolating or otherwise trying to estimate the missing data, pattern

matching proceeds using the available data only, giving zero weight to missing data in the pattern-matching process. A similar idea underlies ''missing data'' or ''missing feature'' techniques that have been proposed recently in speech recognition to cope with deleted spectro-temporal features (Cooke *et al.*, 1994, 1997; Lippmann, 1997; Morris *et al.*, 1998). Two versions of the model are proposed: one works in the frequency domain, and the other in the time domain. Both require an estimate of the F_0 of the vowel.

A. Frequency-domain version

The frequency-domain model is straightforward. Spectral templates are assumed to be available for all vowel classes. The following steps occur when a vowel is recognized: (a) its short-term spectrum is estimated, (b) its F_0 is estimated, (c) based on F_0, a spectral weighting function is calculated that emphasizes regions near multiples of F_0, and (d) the short-term spectrum is compared to all templates using the weighting function. The template that yields the smallest distance determines the vowel that is ''recognized.'' Templates are defined over the whole spectrum, but comparison is restricted to certain frequencies depending on the F_0.

The weighting function $W(f)$ and spectral distance $D(T,T_i)$ from target T to template T_i might be defined as

$$W(f)=\sum_{n=0}^{\infty}\delta(f-n\hat{F}_0), \tag{1}$$

$$D(T,T_i)=\int (T(f)-T_i(f))^2 W(f)\,df, \tag{2}$$

where $\delta(\,)$ is the Dirac delta function, $\hat{F}_0$ is the fundamental frequency estimate, $T(f)$ is the short-term spectrum measured from the waveform, and $T_i(f)$ is the spectral envelope of the ith vowel. The infinitely narrow peaks of $W(f)$ in Eq. (1) are satisfactory in theory. In practice, the peaks might be made to widen gradually with f to accommodate inevitable inaccuracy in F_0 estimation. With a square shape and relative width of 3%, such a weighting function is equivalent to the harmonic sieve of Duifhuis *et al.* (1982), that has been proposed as a mechanism to select information in the context of pitch perception (Moore *et al.*, 1984, 1985; Darwin *et al.*, 1992) and concurrent vowel identification (Scheffers, 1983).

The model requires an F_0 estimate. This is impossible to obtain for stimuli that are too short, whispered, or otherwise nonstationary. However, in those cases no special processing is called for: the short-term spectrum is unbiased, and the sampling operation is not necessary. A complete formulation of the model must explain how it switches from one mode to the other. For example, the weighting function might be uniform by default, and replaced by a gradually sharpening spectral comb as reliable F_0 information is obtained. Sharpness of the spectral comb could be under the control of a ''periodicity measure'' (many F_0 estimation methods produce such a measure as a by-product). When periodicity is good, F_0-estimation errors are unlikely; the model is therefore robust.

As an illustration, a simple missing-data vowel identification model was built, similar to that of the Introduction, subsection B but with smoothing replaced by nonuniform weighting. The weighting function was 1.0 at multiples of F_0, and 0.0 elsewhere. Figure 6 shows the weighted distance of a synthetic /i/ target vowel to each of the templates, as a function of F_0. The distance to the correct template /i/ is not quite zero, because of limited sampling resolution in the implementation, but it remains smaller than the distances to incorrect templates. The vowel /i/ is identified correctly at all F_0's [compare with Fig. 5(b)].

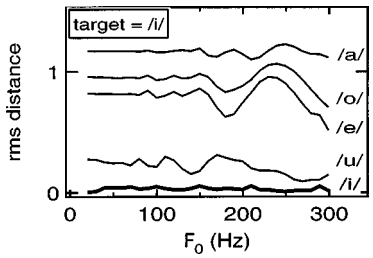

FIG. 6. Simple missing-data vowel identification model. Weighted distance between reference templates for vowels /a/, /e/, /i/, /o/, and /u/, and the unsmoothed short-term spectrum estimated from a target /i/ waveform, as a function of F_0. The weighting function was 1.0 at multiples of F_0, and 0 elsewhere. The nonzero values of the distance to the /i/ template are due to sampling error in the spectrum calculations.

Physiologically, one can imagine that short-term spectral estimation occurs in the cochlea, and that the harmonic sieve is applied along a tonotopic dimension at some point in the auditory system, based on an F_0 estimate, itself possibly derived from tonotopic information. The main difficulty, apart from the issue of frequency resolution, is to imagine how the variable-pitch harmonic sieve is constructed based on the F_0 estimate, and how it is deployed across frequency channels. In the following section, we consider an alternative model based on autocorrelation that might be implemented physiologically using time-domain processing in the auditory system.

B. Autocorrelation version

The Fourier-domain reasoning of the Introduction, subsection A that was applied to the *magnitude* of the vocal-tract transfer function can be applied equally well to its *squared magnitude*. The Fourier transform of the squared magnitude is the autocorrelation of the vocal-tract impulse response, $\mathrm{ACF}_{\mathrm{tract}}$, plotted as a thin line in Fig. 7(a). When a vowel is produced with a constant F_0, the squared magnitude vocal-tract transfer function is sampled at multiples of F_0. The sampling theorem tells us that the samples describe uniquely a function bandlimited to lags smaller than $T_0/2$. In other words, the information about the vocal tract available in the samples is also represented in the $\tau<T_0/2$ portion of the autocorrelation function $\mathrm{ACF}_{\mathrm{tract}}$ [Fig. 7(a), thick line].

In the vowel identification model of Sec. I A, magnitude spectra could be replaced by squared-magnitude spectra. Parseval's theorem tells us that the Euclidean distance between such spectra is the same as that between the corresponding autocorrelation functions. One can thus in turn replace squared-magnitude spectra by autocorrelation functions in the vowel identification model. The spectral

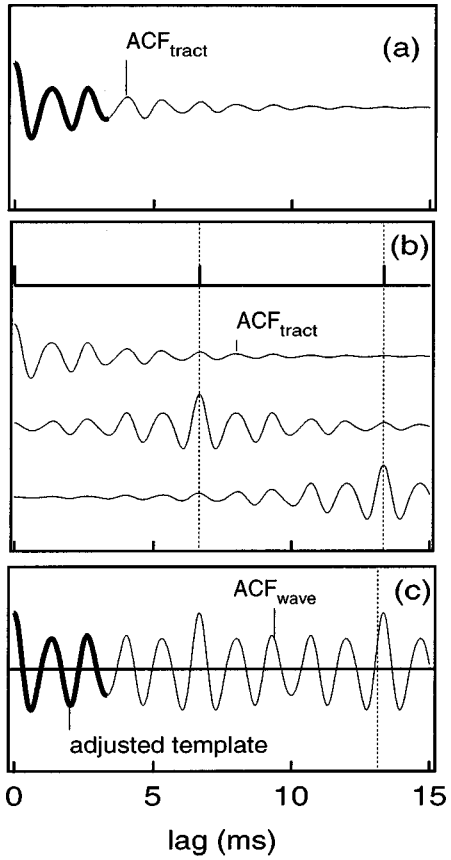

FIG. 7. (a) Autocorrelation of vocal tract transfer function (ACF_{tract}). (b) Illustration of the convolution process by which ACF_{wave} is derived from ACF_{tract} in the case where $F_0 = 150$ Hz. Copies of ACF_{tract} are shifted and added to obtain ACF_{wave} [thin line in (c)]. The vertical dotted lines indicate multiples of the period, 6.67 ms. The thick line in (c) is an adjusted template (see the text).

weighting function is replaced by a lag-domain weighting function restricted to lags $\tau < T_0/2$.

A difficulty remains. The autocorrelation function of the impulse response of the vocal tract (ACF_{tract}) is not directly observable. Observable is the autocorrelation of the wave-form (ACF_{wave}), that differs from ACF_{tract} but is related to it by the following relation:

$$\mathrm{ACF}_{\mathrm{wave}}(\tau) = \mathrm{ACF}_{\mathrm{tract}}(\tau) \circ \sum_{k=-\infty}^{\infty} \delta(\tau - kT), \tag{3}$$

where $\circ$ represents convolution. As illustrated in [Fig. 7(b)], copies of (ACF_{tract}) are shifted to multiples of T_0, and added up to obtain ACF_{wave}. Because of overlap between the shifted functions, ACF_{wave} differs from ACF_{tract} in the important region $\tau < T_0/2$. The difference depends on F_0.

For this reason, ACF_{wave} cannot make a perfect match to the templates, even if it is restricted to $\tau < T_0/2$. However, if F_0 is known, it is possible to *adjust* the templates to obtain a perfect match. This is done by adding up appropriately shifted versions of the templates, exactly as in the convolution illustrated in Fig. 7(b). In this way, an accurate match is obtained between the correct vowel template and the observed ACF_{wave} [Fig. 7(c)]. We can thus formulate an autocorrelation version of the missing-data vowel perception model. The following steps occur when a vowel is recognized: (a) ACF_{wave} is estimated from the wave-form, (b) F_0 is estimated, (c) based on F_0, the set of ACF_{tract} templates are adjusted, and (d) each one is compared to ACF_{wave} over the $\tau < T_0/2$ range of lags. The closest match indicates the ''recognized'' vowel.

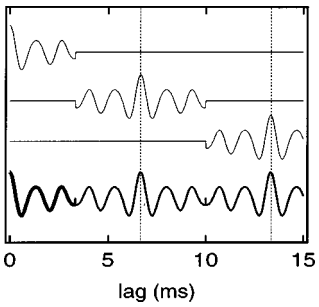

FIG. 8. Illustration of the hypothetical case of a vocal tract with a transfer function band-limited to lags smaller than 3.33 ms. ACF_{wave} is derived from ACF_{tract} by convolution, but the $\tau < 3.33$ ms portion is unaffected by the convolution and remains equal to ACF_{tract}. Template adjustment would be unnecessary in this hypothetical case.

The F_0 estimate is involved in two places: template adjustment and determination of the range of lags to be matched. The adjustment step would be unnecessary in the hypothesis that ACF_{tract} is limited to $T_0/2$, as illustrated in the top of Fig. 8. If ACF_{wave} is zero beyond $T_0/2$, ACF_{wave} and ACF_{tract} are equal for $\tau < T_0/2$ (thick line at the bottom of Fig. 8), and adjustment is unnecessary. Omitting the adjustment stage is thus equivalent to putting faith in the assumption that the squared-magnitude spectrum is bandlimited to $\tau < T_0/2$ in the lag domain. This assumption is all the more incorrect as F_0 is high.

In the squared-magnitude spectrum, high-amplitude parts of the spectrum are emphasized at the expense of others [Fig. 9(a)]. The first formant is well represented, and this accounts for the ripple that dominates ACF_{tract}. The second formant is less well represented, and higher formants hardly at all. The magnitude spectrum of the Introduction, subsection B is a slightly more balanced representation [Fig. 9(b)]. The *log magnitude spectrum* represents both peaks and valleys in equal detail, whatever their amplitude [Fig. 9(c)], and its inverse Fourier transform, the *cepstrum*, is widely used in speech processing. The success of the cepstrum in speech-

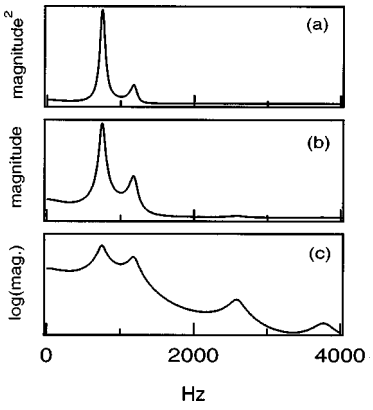

FIG. 9. (a) Squared-magnitude transfer function of /a/. (b) Magnitude transfer function of /a/. (c) Log-magnitude transfer function of /a/.

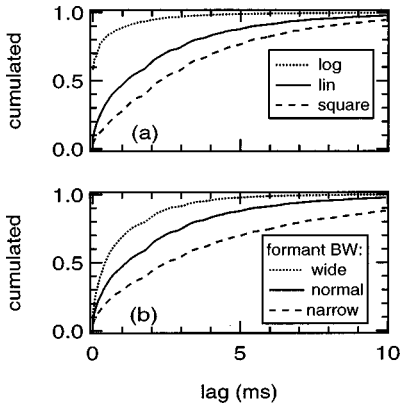

FIG. 10. (a) Cumulated magnitude of lag-domain components of the spectral envelope, for a linear (full line), log (dotted line), or squared (dashed line) spectral envelope. The smoother log spectrum has fewer high-lag components, whereas the sharper squared spectrum has more. (b) Cumulated magnitude of lag-domain components of the spectral envelope for vowels with formant bandwidths that are standard (full line), twice standard (dotted line), or half standard (dashed line). Narrow formants imply more high-lag components; wide formants imply less.

processing applications suggests that the log magnitude spectrum (and cepstrum) are more effective substrates for pattern matching than the squared-magnitude spectrum (and autocorrelation). The excessive emphasis on $F1$ of the autocorrelation function is alleviated in the model of Sec. I C, where autocorrelation functions are calculated within channels of a basilar-membrane/hair-cell model.

As illustrated in Fig. 10(a), the lag-domain composition of the spectral envelope differs from that of its squared and log transforms. The squared version is richer in high-lag components, whereas the log version is poorer. As illustrated in Fig. 10(b), the lag-domain composition also depends on formant bandwidth. Narrow formants require more high-lag components, wide formants less. The severity of the aliasing effects described in this paper thus depend on other factors in addition to F_0. The relative lack of high-lag components of the log spectrum is another possible reason for its success (and that of the cepstrum) in speech processing. The perceptual severity of aliasing would depend on the representation used by the auditory system. Autocorrelation functions are *a priori* highly sensitive, but this sensitivity might be alleviated by compressive mechanisms in a physiological representation (next section).

Both the autocorrelation model and the spectral model of Sec. I A are perfectly accurate. In a task lacking variability (other than due to F_0), both would perform perfectly at any F_0. However, it is clear from Fig. 7 that identification at higher F_0 has less information to go on, and is likely to degrade as soon as variability is introduced into the task.

C. Multichannel autocorrelation version

In Sec. I A we suggested that the spectral version of the model might be implemented by frequency-domain processing within the auditory system, based on a tonotopic representation. Here, we describe how the autocorrelation version might be implemented by *time-domain* processing within the auditory system, based on the temporal structure of nerve-fiber discharge patterns.

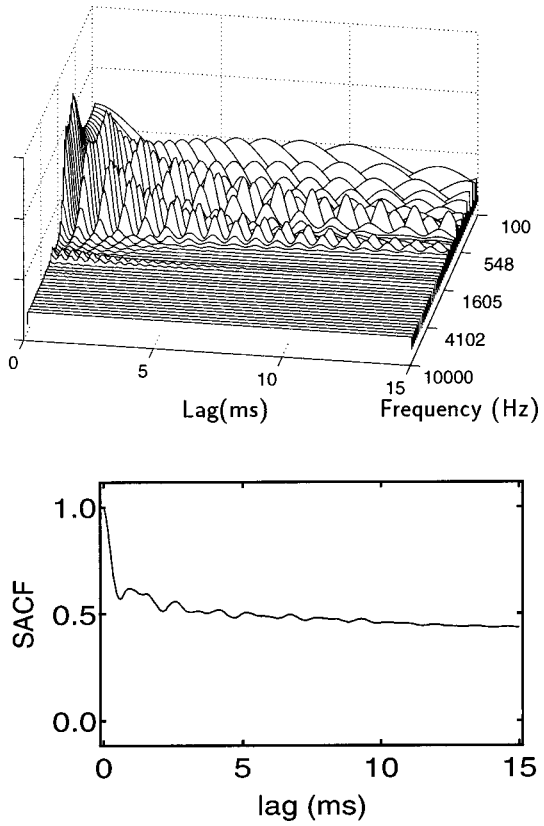

FIG. 11. Top: Array of autocorrelation functions of auditory-nerve fiber discharge probability indexed by channel frequency, in response to a single impulse of the vowel /a/. Probabilities were produced by a model of peripheral filtering and hair-cell transduction, with 50 channels uniformly spaced between 100 and 10 000 Hz on a scale of equivalent rectangular bandwidth (ERB). Bottom: summary autocorrelation function (SACF).

Autocorrelation of nerve-fiber discharge patterns has been suggested as a basis for pitch perception (Licklider, 1951; Meddis and Hewitt, 1991a, b; Cariani and Delgutte, 1997a, b), and cross correlation is an accepted mechanism for localization (Jeffress, 1948; Yin *et al.*, 1987). In the pitch model of Meddis and Hewitt (1991a, b), autocorrelation functions (ACF) of auditory-nerve discharge probability were calculated within each channel of a model of basilar-membrane filtering and hair-cell transduction, and ACFs for all channels were added up to form a summary autocorrelation function (SACF). The pitch was derived from the position of the highest peak in the SACF. Many aspects of pitch phenomena are well accounted for by that model or others that are closely related (de Cheveigné, 1998a). The SACF was also proposed as a substrate for *vowel identification* in the concurrent vowel-identification model of Meddis and Hewitt (1992). In their model, vowels were identified by matching the ''low-lag'' portion of the SACF ($\tau<4$ ms), a scheme that was also used with success by de Cheveigné (1997).

Figure 11 (top) shows an array of autocorrelation functions calculated from the instantaneous discharge probability functions produced by a model of peripheral filtering and hair-cell transduction (Slaney, 1993). The model had 40 channels uniformly distributed on a scale of equivalent rectangular bandwidth (ERB) between 100 and 10 000 Hz. The stimulus was a single-impulse response of the vocal tract

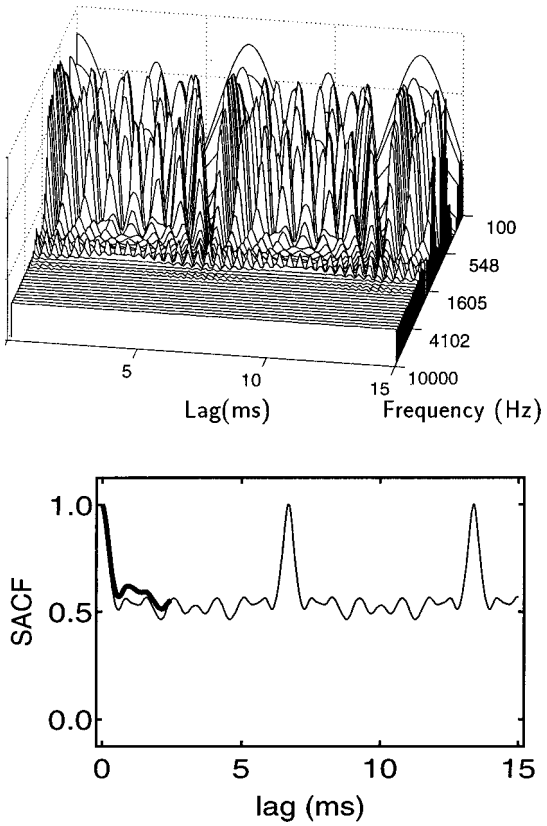

FIG. 12. Same as Fig. 11, in response to vowel /a/ at $F_0 = 150$ Hz.

corresponding to the vowel /a/. The combination of vocal tract and basilar-membrane filter leads to a much longer impulse response than the vocal tract alone, which explains the slow decay of the ACFs in Fig. 11 compared to Fig. 7(b), particularly in channels tuned to low frequencies and/or near a formant. Because of the rectifying property of the hair-cell model, ACFs are never negative.

Figure 11 (bottom) shows the corresponding summary autocorrelation function (normalized by dividing by the value at zero lag). The SACF decays to a relatively high baseline due to the summation of non-negative ACFs for different channels. Compared to the autocorrelation of the waveform [Fig. 7(b)], the SACF lacks the strong ripple at the period of $F1$. This is probably a consequence of the saturating properties of the basilar-membrane model, which equalize contributions of different channels. Compared to waveform autocorrelation, the SACF is affected relatively less by $F1$, and more by other components.

Figure 12 shows the response of the model to the vowel /a/ at $F_0 = 150$ Hz. Each channel shows a peak at multiples of T_0, as does the SACF. Over the interval $[0 - T_0/2]$, the SACF resembles the SACF obtained in response to the impulse response (thick line), justifying the choice of Meddis and Hewitt (1992) of the low-lag portion of the SACF as a substrate for vowel identification. There are, nevertheless, differences between the two SACFs, as we observed previously for the waveform-based autocorrelation functions. Analysis of these differences is complicated by the presence of the nonlinear hair-cell transduction stage. Because of the nonlinearity, within-channel ACFs in response to the periodic stimulus cannot accurately be calculated as a convolution, as in Sec. I B. However, this does not prevent postulating a model similar to that of Sec. I B, in which templates are tabulated rather than calculated.

It is customary to use the SACF instead of the full ACF array as a basis for pitch (Meddis and Hewitt, 1991a, b) or vowel identification (Meddis and Hewitt, 1992), but this is not mandatory. Matching could just as well be performed on the full ACF array, or better still, on an array of ''sub-SACFs'' calculated over sub-bands wide enough to avoid gaps between harmonics at high F_0. One advantage might be better discrimination, as the two-dimensional array is a richer pattern than the one-dimensional SACF. Another is the possibility to factor out spectral tilt and other transmission channel characteristics (thanks to within-channel compression or automatic gain control, as performed by the hair-cell model). A third is that parts of the array may be weighted differently in the spectral domain to handle ''missing features,'' due to bandlimited noise or filtering (Cooke *et al.*, 1994, 1997; Lippmann, 1997; Morris *et al.*, 1998). The full ACF or sub-SACF array is a representation in which correlates of sources and interference can be weighted differentially in both the lag and the frequency domain. It is thus a flexible starting point for sophisticated models of identification and segregation.

II. DISCUSSION

The identity of a vowel depends on the shape of its spectral envelope, essentially the position of the first two or three formants. However, when a vowel is produced at a constant pitch, its spectral envelop is sparsely sampled, implying that details of this shape may not be well represented. Spectral representations derived from the waveform (short-term spectra, auditory excitation patterns, etc.) have a harmonic structure that interferes with the determination of the spectral envelope. Harmonics masquerade as formant peaks, and at high F_0's the short-term spectrum bears little resemblance to the vowel's spectral envelope. Nevertheless, to a first approximation, the vowel's identity does not depend on the fundamental frequency. F_0 typically varies from 90 to 200 Hz for male speakers, or 150 to 310 Hz for women (Howard, 1991), although much wide ranges have been reported, in particular for children (Fairbanks, 1940; Keating and Buhr, 1978). Sundberg (1982) states that steady-state sung vowels retain a degree of intelligibility up to 520 Hz (or even 1000 Hz in consonant–vowel–consonant (CVC) context). The problem of vowel perception at high F_0 has been recognized by many authors (Carlson *et al.*, 1975; Carré and Lancia, 1975; Traunmüller, 1981, 1990; Bladon, 1982; Ryalls and Lieberman, 1982; Klatt, 1982; Sundberg, 1982; Darwin and Gardner, 1985; Gottfried and Chew, 1986; Assmann and Neary, 1987; Beddor and Hawkins, 1990; Perec, 1991; Assmann, 1991; Hirahara and Kato, 1992; Hirahara, 1993; Rosner and Pickering, 1994; Hirahara *et al.*, 1996).

Two qualifications must be made. The first is that in natural speech vowels are rarely sustained. Harmonics in the spectrum of a nonstationary sound are less sharp, and the bias due to undersampling is less severe. In the limit, a single vowel period has no harmonic structure, yet it is sufficient

for vowel identification (McKeown and Patterson, 1995; Robinson and Patterson, 1995). Summerfield *et al.* (1984) remarked that a long-duration synthetic vowel sounds vowel-like at onset, but may then lose its identity. The identity is partially regained at offset. It has been suggested that modulation of the F_0 (e.g., *vibrato*) improves identification by giving samples of the *derivative* of the spectral envelope at F_0 multiples, in addition to the magnitude. However, Sundberg (1982) reported that *vibrato* did not improve identification of high-pitched sung vowels. Hillenbrand and Gayvert (1993) found that vowels were more intelligible when synthesized with a falling rather than static F_0 contour, but the advantage was small. McAdams and Rodet (1988) suggested that FM-induced envelope tracing might enhance identification of concurrent vowels, but Marin and McAdams (1991) failed to find support for the idea. In natural speech, dynamic cues such as consonant–vowel transitions play an important role in vowel identification (Strange *et al.*, 1976, 1998; Hillenbrand and Neary, 1999). Intelligibility of high-F_0 sung vowels is better when they are in CVC context (Sundberg, 1982; Gottfried and Chew, 1986), and in the extreme, there is evidence that a vowel may be identified from the unvoiced portion of a consonant with which it is articulated (Bonneau, 1996). In natural speech, the spectral undersampling problem is thus less severe than it might seem from consideration of monotone sustained vowels.

The second qualification is that, to a second approximation, vowel quality *is* dependent on F_0. From a production point of view, there are several sources of correlation between spectral envelope and F_0. Within a population of speakers, women and children tend to have higher-pitched voices and shorter vocal tracts than men (Peterson and Barney, 1952). For a given speaker, a change in F_0 may imply a change in glottal pulse shape. If the glottal shape is mathematically included in the ''vocal tract'' transfer function (Introduction), the latter would vary with F_0 even if the shape of the vocal tract did not. Variation in larynx height concurrent with F_0 variations may also affect the spectral envelope, and in the case of singing, professional singers are known to intentionally raise the $F1$ of a high-pitched vowel to ensure that it does not fall below the F_0, in which case it would not be adequately excited (Sundberg, 1982). It is also possible that different vowels are systematically given different F_0's (''intrinsic F_0'').

From a perceptual point of view, numerous experiments have found interactions between F_0 and vowel quality. In an experiment that investigated both musical timbre and vowel quality, Slawson (1967) found that, for an F_0 increase of one octave, $F1$ and $F2$ should be increased by about 10% to minimize the change in quality. Carlson *et al.* (1975) measured the perceptual $F1$ boundary between Swedish /i/ and /e/, and found that it increased from about 300 to 350 Hz for $F0$'s varying from 100 to 160 Hz. Ainsworth (1975) determined that a one-octave increase in F_0 had to be accompanied by a 3.4% increase of $F1$ and a 1.2% increase of $F2$ (figures that he argued might be underestimated by a factor of 2). Neary (1989) likewise found a 7%–9% shift for $F1$ and a 1.5% shift for $F2$. Assmann *et al.* (1982) found that inclusion of an F_0-based parameter improved discriminant analysis of sets of acoustic parameters of vowels, and Hirahara and Kato (1992) obtained data points that were better clustered (both for production and for perception) when expressed in a plane of $(F1-F_0)\times(F2-F_0)$ bark differences, rather than an $F1\times F2$ plane. Miller (1989) proposed a vowel perception model in which formant frequencies were normalized by the cubic root of F_0, but several authors have argued that this implies an exagerated F_0 dependency (Neary, 1989; Rosner and Pickering, 1994). Other effects are reported by Traunmüller (1981), Fahey *et al.* (1996), Hoemke and Diehl (1994), Kewley-Port (1996), Kewley-Port *et al.* (1996). Overall, effects of F_0 are rather small. However, to the extent that an orderly relation exists between F_0 and vowel quality, and that listeners exploit it, complete insensitivity to F_0 is not necessarily desirable in a vowel perception model.

Intelligibility usually gets worse at high F_0. Ryalls and Liebermann (1982) found that vowels synthesized with formants appropriate for male and female voices were less intelligible at 250 Hz than at 185 Hz (female), 135 Hz (male), or 100 Hz (both). With synthetic vowels, Sundberg *et al.* (1982) found a decrease in intelligibility between 260 and 700 Hz. With natural vowels sung by a soprano, Benolken and Swanson (1990) also found a steady decrease from 262 to 1047 Hz. Sundberg (1982) notes that intelligibility is better when vowels are pronounced in a CVC context. Gottfried and Chew (1986) asked a counter tenor to pronounce ten vowels in ''$h\wedge d$'' context (''$\wedge$'' signifying the vowel), in both a full voice (130 to 330 Hz) and a head voice (220 to 520 Hz). For both voices the error rate increased with F_0, but in the region of overlap (220 to 330 Hz) the head voice was more intelligible. Whatever the mechanism of vowel perception, a decrease in intelligibility with increases in F_0 is understandable from the progressively sparser sampling of the spectral envelope. Apart from a few exceptions (see the histograms of Benolken and Swanson, 1990), the decrease is gradual and one does not see the sort of irregularities that might be expected as the peaks of the spectral envelope are swept by an expanding comb of voice harmonics.

As a counterpoint, it is interesting to note that Hillenbrand and Nearey (1999) found only minor differences in identification rate between vowels pronounced by male (M), female (F), or child (C) speakers, despite F_0 differences on the order of an octave. Moreover, the ranking was inconsistent according to whether the vowels were natural [rates ranked as (M,F)>C] or resynthesized with formant trajectories that were natural [M>(F,C)] or static [(M,C)>F]. Resynthesized vowels used the original vowels' F_0 contour. If identifiability decreased with F_0, one should expect instead the ranking M>F>C to prevail uniformly. Identification was better for natural than for static formant contours, which confirms the importance of dynamic cues. However, it was better still for natural vowels, suggesting that cues available in natural vowels are not completely exhausted by dynamic formant and F_0 values.

Attempts have been made to relate the observable harmonic peaks of a short-term spectrum with the underlying (but not directly observable) formant peaks. The problem is different for front vowels, for which $F1$ is relatively low and

well separated from other formants, and back vowels, for which $F1$ and $F2$ are close. For front vowels, the difficulty is that of estimating $F1$, given that harmonic spacing is relatively wide compared to formant bandwidth and peripheral resolution. The second formant is less of an issue because that formant is wider, peripheral filtering is less sharp, and $F2$ is less in need of resolution from other formants, given that the assumption that higher formants are grouped to form an ''$F2'$'' works quite well (Carlson *et al.*, 1975). It has been proposed that $F1$ is determined from the position of the single most prominent harmonic (Mushnikov and Chistovich, 1972), or a combination of the two or three most prominent harmonics in the $F1$ region (Carlson *et al.*, 1975). Assmann and Neary (1987) found that a weighted sum of two harmonics was a better predictor than either one harmonic or three. However, Darwin and Gardner (1985) found evidence that more remote harmonics also influenced the perceived $F1$ position.

The question of whether one, two, or three harmonics affect $F1$ is a good empirical question, but an awkward basis for building models of vowel perception. A model exploiting the weighted sum of two harmonics, for example, must locate them. Criteria might be spectral peaks, or a harmonic series, but the model must solve the limit case where F_0 is low and individual harmonics are not well resolved, as well as the opposite case when only one harmonic (possibly none) belongs to the formant. ''Return on investment'' is limited because the model cannot be applied to vowels for which $F1$ and $F2$ are close (see below). Finally, such a model is likely to predict irregular variations of $F1$ boundaries when F_0 varies. A one-harmonic model predicts a saw-toothed variation of the $F1$ estimate: as the most prominent harmonic shifts down the slope of the $F1$ peak, it is suddenly replaced by a different harmonic. A two-harmonic model has a similar problem (its severity depends on the weight of the weaker harmonic relative to the stronger). Hirahara (1993) did in fact find an irregularity in the variation of the $F1$ boundary of a Japanese /i/–/e/ continuum. The boundary shifted from 320 to 400 Hz as F_0 increased from 100 to 150 Hz, stayed at 400 Hz as F_0 increased further to 250 Hz, then increased to about 520 Hz as F_0 increased to 450 Hz. This behavior is indeed irregular, but not quite what one expects from a harmonic tracking mechanism. One expects instead a general *decrease* in $F1$ boundary to compensate for the upward shift of the harmonics, with a local increase at each switch between harmonics. To summarize, schemes based on individual harmonics or their weighted sums are an incomplete answer to the problems posed by harmonic structure.

For back vowels the problem is yet more complex, as pointed out by Assmann (1991). $F1$ and $F2$ are close and harder to relate individually to the set of harmonics than in the case of an isolated $F1$. It has been proposed that when $F1$ and $F2$ are closely spaced (less than 3 to 3.5 bark), they are ''merged'' into a single spectral prominence, characterized by its center of gravity (Chistovich and Lublinskaya, 1979). The position of the center of gravity (COG) should be affected by both the frequencies and the amplitudes of both formants. Assmann (1991) tested this hypothesis with negative results. In one of his experiments, spectral integration effects compatible with the COG hypothesis were found at $F_0 = 250$ Hz, but not at 125 Hz. However, contrary to the criteria of Chistovich and Lublinskaya, they were *larger* for formants spaced more than 3–3.5 bark rather than smaller. Correct or not, the COG hypothesis is an awkward basis for a vowel perception model. If the COG is derived from formant estimates (with the convention that formants are grouped if closer than 3.5 bark), the problem of formant estimation remains entire. If the COG is derived from a smoothed spectrum, then at least three questions arise. First, how does one distinguish a peak reflecting an isolated formant ($F1$) from that reflecting a closely spaced pair ($F1-F2$)? Second, what is the appropriate form of smoothing? Third, is such smoothing not prone to the aliasing effects described earlier?

Klatt (1982) points out that resolution of the auditory periphery is a poor guide: it is much too sharp in the low-frequency region, except for the lowest F_0's. The *ad hoc* value of 3.5 bark handles F_0's up to 350 Hz, but no further, yet for lower F_0's this smoothing may be unnecessarily severe. Logically, the best amount of smoothing depends on F_0, and indeed there is some evidence that the auditory system applies wider spectral integration at higher F_0's (Assmann, 1991). The ''simple vowel classifier'' of Scheffers (1983) and the PEAK procedure of Assmann and Summerfield (1989) both imply F_0-dependent smoothing. Both reconstruct an envelope by linear interpolation between samples of the spectrum at harmonics, an operation equivalent to smoothing by convolution with a triangular window of width $2F_0$ [similar smoothing is employed in the STRAIGHT analysis system of Kawahara (1997)]. Such precisely tuned F_0-dependent smoothing is certainly more effective than the fixed smoothing assumed by the COG hypothesis (or the ''second integration'' of Rosner and Pickering). However, two things are worth noting. First, the ambiguity between a single formant and two closely spaced formants is not resolved [consider for example interpolating between peaks of Fig. 2(b) or (c)]. Second, as other smoothing schemes, this one is susceptible to the aliasing problems analyzed in the Introduction.

The ''missing-data'' model is based on spectral samples that do not necessarily coincide with formant peaks. In this sense, it resembles the ''whole spectrum'' model of Bladon (1982). In that model, the spectrum was smoothed by convolution with an ''auditory filter'' (Bladon and Lindblom, 1981) which makes it susceptible to the aliasing problems pointed out in the Introduction. The missing-data model avoids them, and can thus be seen as an F_0-insensitive implementation of Bladon's ideas. The model belongs to the ''top-down'' or ''analysis-by-synthesis'' variety (Bell *et al.*, 1961; Rabiner and Schafer, 1978), in the sense that it does not try to extract an F_0-invariant representation from the waveform. Instead, it synthesizes a pattern (F_0-sampled spectrum, or F_0-adjusted autocorrelation function) to be compared with incoming patterns.

Bladon's model does not account for the insensitivity to formant bandwidth, formant amplitude, or spectral tilt observed by Klatt (1982), and the present model in its spectral version (Sec. I A) is prone to the same criticism. The

autocorrelation-based version offers some flexibility to deal with this problem. Differences in formant bandwidth affect mainly the high-lag part of the autocorrelation function, and sensitivity to this parameter can be reduced by ignoring features beyond a certain fixed lag. It is tempting to treat spectral tilt and formant amplitude in the same way, by ignoring lag-domain features *below* a certain lag, a technique that works well with the Fourier transform of the log magnitude spectrum, or cepstrum, and is commonly used in speech processing (e.g., Tohkura, 1987). Unfortunately this idea makes less sense applied to a transform of the square-magnitude spectrum (autocorrelation). Nevertheless, the filter-bank implementation of Sec. I C does offer a degree of insensitivity to gross spectral features, thanks to amplitude compression in the hair-cell model. A whole-spectrum model does not account for the fact that spectral peaks (formants) are known to carry a stronger weight than spectral valleys. However, the autocorrelation function puts a strong weight on spectral peaks (essentially $F1$ and $F2$), that are emphasized in the square-magnitude spectrum. In the multichannel version of Sec. I C, this rather extreme emphasis is softened by the automatic gain control (AGC) properties of the hair-cell model, that allow weaker features to be better represented.

A key idea is that the auditory system applies *variable* weights to different parts of incoming evidence according to their reliability. For example, relatively fine features of the spectral envelope may be used at low F_0 but ignored at high F_0. Assmann (1991) found stronger evidence for the center-of-gravity hypothesis (that emphasizes gross spectral features) at 250 Hz than at 125 Hz. He also suggested that the reduced amplitude of the higher formant region (>1 kHz) might serve as a cue for the presence of two closely spaced low formants (<1 kHz), but only *when* F_0 *is high*. When F_0 is low, evidence of the presence of two formants would be derived instead from the detailed spectral shape in the low-frequency region. Different sources of evidence would thus be weighted differently according to their reliability. The same principle can account for the fact that in general (for low F_0) the amplitude balance between widely spaced parts of the spectrum has little effect on vowel identity (Chistovich, 1985; Klatt, 1982). It can also explain the finding of Beddor and Hawkins (1990) that overall spectral shape was important for vowel matching when spectral peaks were broad (nasal or wide formants), whereas detailed spectral shape (formant position) was important when formants were narrow.

Versnel and Shamma (1998) presented evidence for a Fourier-transform representation of spectral shape in the primary auditory cortex of the ferret. Responses to vowels could be predicted from responses to sine-wave ripple stimuli. The principle is related to that of our autocorrelation scheme, the greatest difference being that their representation used a *logarithmic* axis for frequency. In other words, ripples had a constant period in octaves rather than Hz. Vowel stimuli were either voiced, extracted from the TIMIT database with F_0's in the range 100–130 Hz, or unvoiced (synthesized with a 20 component/octave carrier). The authors noted little difference between voiced and unvoiced responses.

The spectral-domain version of the model is similar to the ''harmonic sieve'' of Duifhuis *et al.* (1982). The harmonic sieve was proposed by Duifhuis as a means to select the components of a sound to be included in the calculation of its pitch. Scheffers (1983) used it to assign components of a mixed-speech spectrum to each voice, and Moore *et al.* (1984, 1985), and Darwin and Ciocca (1992) showed that a harmonic sieve with a width of about 3% determined which components of a sound contribute to its pitch. Darwin and Gardner (1986) found that mistuning a component of a vowel by 3% reduced its contribution to the vowel's quality. In those cases, the harmonic sieve played an important role in *segregating* the harmonic sound from competing components. Here, we suggest that it also plays a role in handling the bias due to F_0 in the identification of isolated vowels. Note that this proposition seems to contradict results that show that the identification of a member of a concurrent vowel pair is no better when that vowel is harmonic rather than inharmonic (de Cheveigné *et al.*, 1997).

III. CONCLUSIONS

(1) Because of spectral undersampling, the information available to describe the spectral envelope of a steady-state voiced vowel is incomplete. It is limited to a series of F_0-spaced samples in the frequency domain, or in the lag domain to lag components below the Nyquist lag $(1/2F_0)$.

(2) If the sampled spectrum is smoothed, lag-domain components beyond the Nyquist lag are aliased and contribute spurious F_0-dependent components to the smoothed spectrum. Aliasing is more severe if F_0 is high, and if the spectral envelope is rich in high-lag components (narrow formants).

(3) *Sampling* is evident in the harmonic structure of representations such as the short-term spectrum or auditory excitation pattern. *Aliasing* is evident in the fact that smoothing or interpolation of these patterns does not produce an F_0-invariant representation.

(4) In a vowel identification model, effects of aliasing can be eliminated by (a) avoiding spectral smoothing, and (b) restricting pattern matching to the available samples. The model can be implemented in the spectral domain using a harmonic sieve based on an estimate of F_0.

(5) The model can also be implemented in the lag domain based on the autocorrelation function of the waveform. Pattern matching is restricted to lags smaller than the Nyquist lag. Templates must be adjusted based on the F_0 estimate to compensate for lag-domain aliasing.

(6) Both versions of the model ensure F_0-independent pattern matching. They do not address other known sources of F_0 dependency of vowel production or perception. They also do not deal with the loss of information due to sampling. Sparse sampling at high F_0 should lead to progressively degraded indentification, as is indeed observed experimentally.

(7) The autocorrelation model might be implemented physiologically by neural spike coincidence counting circuits within channels of the auditory nerve, in a manner similar to Licklider's pitch-perception model. Nonlinearity of hair-cell transduction makes template adjustment more

difficult, but the extra tonotopic dimension provides a rich substrate for pattern matching, information weighting, and source segregation.

ACKNOWLEDGMENTS

Part of this work was carried out at ATR Human Information Processing Research Laboratories, within a research agreement with the Center National de la Recherche Scientifique and the University of Paris 7. The author thanks ATR for its kind hospitality, and the CNRS for leave of absence. Thanks to Malcolm Slaney for the AuditoryToolbox software, and to Minoru Tsuzaki, Hiroaki Kato, Tatsuya Hirahara, Erik McDermott, Peter Assmann, Chris Darwin, and Kuldip Paliwal for comments on previous versions of this manuscript. Thanks also to Hartmut Traunmüller, two anonymous reviewers, and the editor for their useful criticism.

[1]Here, we consider the Fourier transform of the magnitude spectrum. Later on in the paper, we consider the Fourier transform of the *squared* magnitude spectrum, or autocorrelation function. It is also common to consider the Fourier transform of the *log* magnitude spectrum, or cepstrum. The principle is the same in each case. The representations differ, however, in their lag-domain composition, and the result of filtering or smoothing varies between representations.

[2]Frequencies of formants $F1$, $F2$, $F3$, $F4$ were taken from Hirahara and Kato (1992). They were, respectively (750, 1187, 1595, 3781 Hz) for /a/, (469, 2031, 2687, 3375 Hz) for /e/, (281, 2281, 3187, 3781 Hz) for /i/, (468, 781, 2656, 3281 Hz) for /o/, and (312, 1219, 2469, 3375 Hz) for /u/. Formants $F5$ of all vowels were set to 4200 Hz. Bandwidths of formants had identical values for all vowels: 90, 110, 170, 250, and 300 Hz, respectively, for formants $F1$ to $F5$.

Ainsworth, W. A. (**1975**). ''Intrinsic and extrinsic factors in vowel judgments,'' in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 103–113.

Assmann, P. F., Neary, T. M., and Hogan, J. T. (**1982**). ''Vowel identification: orthographic, perceptual, and acoustic effects,'' J. Acoust. Soc. Am. **71**, 975–989.

Assmann, P. F., and Neary, T. M. (**1987**). ''Perception of front vowels: the role of harmonics in the first formant region,'' J. Acoust. Soc. Am. **81**, 520–534.

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. (**1991**). ''The perception of back vowels: centre of gravity hypothesis,'' Q. J. Exp. Psychol. **43A**, 423–448.

Beddor, P. S., and Hawkins, S. (**1990**). ''The influence of spectral prominence on perceived vowel quality,'' J. Acoust. Soc. Am. **87**, 2684–2704.

Bell, C. G., Fujisaki, H., Heinz, J. M., Stevens, K. N., and House, A. S. (**1961**). ''Reduction of speech spectra by analysis-by-synthesis techniques,'' J. Acoust. Soc. Am. **33**, 1725–1736.

Benolken, M. S., and Swanson, C. E. (**1990**). ''The effect of pitch-related changes on the perception of sung vowels,'' J. Acoust. Soc. Am. **87**, 1781–1785.

Bladon, R. A. W., and Lindblom, B. (**1981**). ''Modeling the judgment of vowel quality differences,'' J. Acoust. Soc. Am. **69**, 1414–1422.

Bladon, A. (**1982**). ''Arguments against formants in the auditory representation of speech,'' in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström (Amsterdam, Elsevier), pp. 95–102.

Bonneau, A. (**1996**). ''Identification of vowels from french stop bursts,'' in Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, pp. 133–136.

Cariani, P. A., and Delgutte, B. (**1996a**). ''Neural correlates of the pitch of complex tones. I Pitch and pitch salience,'' J. Neurophysiol. **76**, 1698–1716.

Cariani, P. A., and Delgutte, B. (**1996b**). ''Neural correlates of the pitch of complex tones. II Pitch shift, pitch ambiguity, phase-invariance, pitch circularity, rate-pitch and the dominance region for pitch,'' J. Neurophysiol. **76**, 1717–1734.

Carlson, R., Fant, G., and Granström, B. (**1975**). ''Two-formant models, pitch and vowel perception,'' in *Auditory Analysis and Perception in Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 55–82.

Carré, R., and Lancia, R. (**1975**). ''Perception of vowel amplitude transients,''in *Auditory Analysis and Perception of Speech*, edited by G. Fant and M. A. A. Tatham (Academic, London), pp. 83–90.

Chistovich, L. A., and Lublinskaja, V. V. (**1979**). ''The 'center of gravity' effect in vowel spectra and critical distance between the formants: psychoacoustical study of the perception of vowel-like stimuli,'' Hearing Res. **1**, 185–195.

Chistovich, L. A. (**1985**). ''Central auditory processing of peripheral vowel spectra,'' J. Acoust. Soc. Am. **77**, 789–805.

Cooke, M., Green, P., Anderson, C., and Abberley, D. (**1994**). ''Recognition of occluded speech by hidden markov models,'' University of Sheffield Department of Computer Science Technical Report, TR-94-05-01.

Cooke, M., Morris, A., and Green, P. (**1997**). ''Missing data techniques for robust speech recognition,'' Proceedings of the ICASSP, pp. 863–866.

Darwin, C. J., and Gardner, R. B. (**1985**). ''Which harmonics contribute to the estimation of first formant frequency?,'' Speech Commun. **4**, 231–235.

Darwin, C. J., and Gardner, R. B. (**1986**). ''Mistuning of a harmonic of a vowel: Grouping and phase effects on vowel quality,'' J. Acoust. Soc. Am. **79**, 838–845.

Darwin, C. J., and Ciocca, V. (**1992**). ''Grouping in pitch perception: effects of onset asynchrony and ear of presentation of a mistuned component,'' J. Acoust. Soc. Am. **91**, 3381–3390.

de Cheveigné, A., McAdams, S., and Marin, C. (**1997**). ''Concurrent vowel identification II. Effects of phase, harmonicity and task,'' J. Acoust. Soc. Am. **101**, 2848–2856.

de Cheveigné, A. (**1997**). ''Concurrent vowel identification III. A neural model of harmonic interference cancellation,'' J. Acoust. Soc. Am. **101**, 2857–2865.

de Cheveigné, A. (**1998**). ''Cancellation model of pitch perception,'' J. Acoust. Soc. Am. **103**, 1261–1271.

de Cheveigné, A., and Kawahara, H. (**1998**). ''A model of vowel perception based on missing feature theory,'' ATR-HIP Technical Report, TR-H-252.

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

Fahey, R. P., Diehl, R. L., and Traunmüller, H. (**1996**). ''Perception of back vowels: effects of varying $F1-F0$ bark distance,'' J. Acoust. Soc. Am. **99**, 2350–2357.

Fairbanks, G. (**1940**). ''Recent experimental investigations of vocal pitch in speech,'' J. Acoust. Soc. Am. **11**, 457–466.

Fant, G. (**1970**). *Acoustic Theory of Speech Production* (Mouton, The Hague).

Gottfried, T. L., and Chew, S. L. (**1986**). ''Intelligibility of vowels sung by a countertenor,'' J. Acoust. Soc. Am. **79**, 124–130.

Hillenbrand, J., and Gayvert, R. T. (**1993**). ''Identification of steady-state vowels synthesized from the Peterson and Barney measurements,'' J. Acoust. Soc. Am. **94**, 668–674.

Hillenbrand, J. M., and Nearey, T. M. (**1999**). ''Identification of resynthesized /hVd/ utterances: Effects of formant contour,'' J. Acoust. Soc. Am. (in press).

Hirahara, T., and Kato, H. (**1992**). ''The effect of $F0$ on vowel identification,'' in *Speech Perception, Production and Linguistic Structure*, edited by Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka (Ohmsha, Tokyo), pp. 89–112.

Hirahara, T. (**1993**). ''On the role of relative harmonics level around the $F1$ of high vowel identification,'' Proceedings of the ARO abstracts (ISSN 0742-3152), abstract #258, p. 65.

Hirahara, T., Cariani, P., and Delgutte, B. (**1996**). ''Representation of low-frequency vowel formants in the auditory nerve,'' Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, pp. 83–86.

Hoemeke, K. A., and Diehl, R. L. (**1994**). ''Perception of vowel height: the role of $F1-F0$ distance,'' J. Acoust. Soc. Am. **96**, 661–674.

Howard, I. (**1991**). ''Speech fundamental period estimation using pattern classification,'' London, unpublished doctoral dissertation.

Jeffress, L. A. (**1948**). ''A place theory of sound localization,'' J. Comp. Physiol. Psychol. **41**, 35–39.
Kawahara, H. (**1997**). ''Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited,'' Proceedings of the ICASSP, pp. 1303–1306.
Keating, P., and Buhr, R. (**1978**). ''Fundamental frequency in the speech of infants and children,'' J. Acoust. Soc. Am. **63**, 567–571.
Kewley-Port, D. (**1996**). ''Psychophysical studies of vowel formants,'' Proceedings of the ESCA Workshop on the Auditory Basis of Speech Perception, pp. 148–153.
Kewley-Port, D., Li, X., Zheng, Y., and Neel, A. T. (**1996**). ''Fundamental frequency effects on thresholds for vowel formant discrimination,'' J. Acoust. Soc. Am. **100**, 2462–2470.
Klatt, D. H. (**1982**). ''Speech processing strategies based on auditory models,'' in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström (Elsevier, Amsterdam), pp. 181–196.
Licklider, J. C. R. (**1951**). ''A duplex theory of pitch perception,'' Experientia **7**, 128–134.
Lippmann, R. P., and Carlson, B. A. (**1997**). ''Using missing feature theory to actively select features for robust speech recognition with interruptions, filtering, and noise,'' Proceedings of ESCA Eurospeech, KN-37–40.
Marin, C., and McAdams, S. (**1991**). ''Segregation of concurrent sounds. II. Effects of spectral envelope tracing, frequency modulation coherence, and frequency modulation width,'' J. Acoust. Soc. Am. **89**, 341–351.
McAdams, S., and Rodet, X. (**1988**). ''The role of FM-induced AM in dynamic spectral profile analysis,'' in *Basic Issues in Hearing*, edited by H. Duifhuis, J. Horst, and H. Wit (Academic, London), pp. 359–369.
McKeown, J. D., and Patterson, R. D. (**1996**). ''The time course of auditory segregation: Concurrent vowels that vary in duration,'' J. Acoust. Soc. Am. **98**, 1866–1877.
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.
Miller, J. D. (**1989**). ''Auditory-perceptual interpretation of the vowel,'' J. Acoust. Soc. Am. **85**, 2114–2134.
Moore, B. C. J., Glasberg, B. R., and Shailer, M. J. (**1984**). ''Frequency and intensity difference limens for harmonics within complex tones,'' J. Acoust. Soc. Am. **75**, 550–561.
Moore, B. C. J., Peters, R. W., and Glasberg, B. R. (**1985**). ''Thresholds for the detection of inharmonicity in complex tones,'' J. Acoust. Soc. Am. **77**, 1861–1867.
Morris, A. C., Cooke, M. P., and Green, P. D. (**1998**). ''Some solutions to the missing feature problem in data classification, with application to noise robust ASR,'' Proceedings of the ICASSP, pp. 737–740.
Mushnikov, V. N., and Chistovich, L. A. (**1972**). ''Method for the experimental investigation of the role of component loudnesses in the recognition of a vowel,'' Sov. Phys.-Acoust. **17**, 339–344.
Neary, T. M. (**1989**). ''Static, dynamic, and relational properties in vowel perception,'' J. Acoust. Soc. Am. **85**, 2088–2113.
Perec, G. (**1991**). ''Experimental demonstration of the tomatopic organization in the soprano (*Cantatrix sopranica L.*),'' in *Cantatrix Sopranica L.*, edited by G. Perec (Paris), Seuil, pp. 11–33.
Peterson, G. E., and Barney, H. L. (**1952**). ''Control methods in a study of the vowels,'' J. Acoust. Soc. Am. **24**, 175–184.
Rabiner, L. R., and Schafer, R. W. (**1978**). *Digital Processing of Speech Signals* (Prentice-Hall, Englewood Cliffs, NJ).
Robinson, K., and Patterson, R. D. (**1995**). ''The stimulus duration required to identify vowels, their octave, and their pitch chroma,'' J. Acoust. Soc. Am. **98**, 1858–1865.
Ryalls, J. H., and Lieberman, P. (**1982**). ''Fundamental frequency and vowel perception,'' J. Acoust. Soc. Am. **72**, 1631–1634.
Rosner, B. S., and Pickering, J. B. (**1994**). *Vowel Perception and Production* (Oxford University Press, Oxford).
Scheffers, M. T. M. (**1983**). ''Sifting vowels,'' Gröningen unpublished doctoral dissertation.
Slaney, M. (**1993**). ''An efficient implementation of the Patterson–Holdsworth auditory filter bank,'' Apple Computer Technical Report, pp. 35.
Slawson, A. W. (**1967**). ''Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency,'' J. Acoust. Soc. Am. **43**, 87–101.
Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (**1976**). ''Consonant environment specifies vowel identity,'' J. Acoust. Soc. Am. **60**, 213–224.
Strange, W., and Bohn, O.-S. (**1998**). ''Dynamic specification of coarticulated German vowels: Perceptual and acoustical studies,'' J. Acoust. Soc. Am. **104**, 488–504.
Summerfield, Q., Haggard, M., Foster, J., and Gray, S. (**1984**). ''Perceiving vowels from uniform spectra: phonetic exploration of an auditory aftereffect,'' Percept. Psychophys. **35**, 203–213.
Sundberg, J. (**1982**). ''Perception of singing,'' in *The Psychology of Music*, edited by D. Deutsch (Academic, Orlando, FL), pp. 59–98.
Sundberg, J., and Gauffin, J. (**1982**). ''Amplitude of the voice fundamental and the intelligibility of super pitch vowels,'' in *The Representation of Speech in the Peripheral Auditory System*, edited by R. Carlson and B. Granström, pp. 223–228.
Tohkura, Y. (**1987**). ''A weighted cepstral distance measure for speech recognition,'' IEEE Trans. Acoust., Speech, Signal Process. **35**, 1414–1422.
Traunmüller, H. (**1981**). ''Perceptual dimension of openness in vowels,'' J. Acoust. Soc. Am. **69**, 1465–1475.
Traunmüller, H. (**1990**). ''A note on hidden factors in vowel perception experiments,'' J. Acoust. Soc. Am. **88**, 2015–2019.
Versnel, H., and Shamma, S. (**1998**). ''Spectral-ripple representation of steady-state vowels,'' J. Acoust. Soc. Am. **103**, 5502–2514.
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.