english

A model of vowel perception based on missing feature theory

Alain de Cheveigné, Hideki Kawahara

Abstract

Vowel identity correlates well with the shape of the transfer function of the vocal tract, or spectral envelope, rather than with the short-term spectrum, which contains peaks at multiples of the fundamental frequency (F_0) that sample the spectral envelope. It is not clear how the auditory system estimates the original spectral envelope from representations that it derives from the vowel waveform. Cochlear excitation patterns, for example, have high resolution in the low frequency region, and their shape varies strongly with F_0. The problem is acute at high F_0s where the spectral envelope is highly undersampled. This paper treats vowel indentification as a form of pattern recognition with missing data. Rather than trying to interpolate the spectral envelope from an incomplete set of samples, we perform pattern matching restricted to available samples. Missing data points are ignored. In other words, a non-uniform weighting function, dependent on F_0, is used to emphasize spectral regions near harmonics of the fundamental, at the expense of other regions. 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 well for the fact that vowel identification is relatively insensitive to F_0-related features of the short-term spectrum.

1 Introduction

In voiced speech, the vocal tract 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, 1960), speech is the result of filtering this train by the vocal tract. Glottal pulses have a shape that depends on the mode of phonation and characteristics of the speaker. This shape can be included mathematically within the vocal tract impulse response, in which case the vocal tract may be seen as excited by a train of infinitely narrow pulses in the time domain. In the frequency domain, if F_0 is constant, the spectrum of the excitation consists of a series of equal-amplitude peaks at multiples of F_0. The spectrum of speech therefore also consists of peaks, but with amplitudes determined by the amplitude of the transfer function at multiples of F_0. In other words, the transfer function is sampled at multiples of F_0 (Fig. 1).

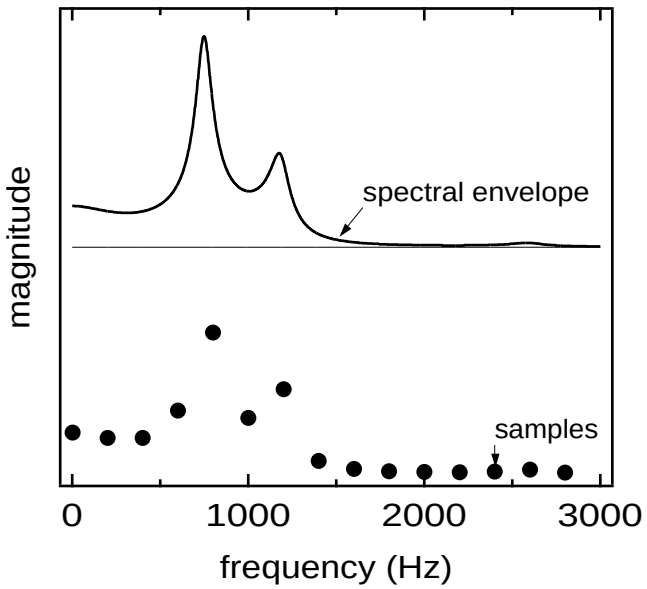

Fig. 1 Line: spectral envelope of vowel /a/. Dots: spectral envelope sampled at F_0=200 Hz.

The timbre and identity of a sustained vowel are determined by the shape of the vocal tract transfer function, and particularly by the positions of the first two or three formants. However the listener hasn't 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 vowel

/a/. The filterbank has 150 channels, equally spaced from 100 Hz to 4178 Hz. The pattern can be taken as approximating the activity evoked by the vowel over a tonotopic dimension (excitation pattern) (see Hirahara et al., 1996 for examples of responses recorded in the auditory system of the cat). At F_0 = 50 Hz (top curve), the pattern is smooth with two clear peaks corresponding to the formants. At F_0 = 200 Hz (middle curve) these peaks are still present, but slightly shifted, and there are many other smaller peaks at low channel frequencies. At F_0 = 216 Hz, however, the peaks at F1 and F2 of /a/ are no more prominent than any 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 does not change strikingly between 200 and 216 Hz.

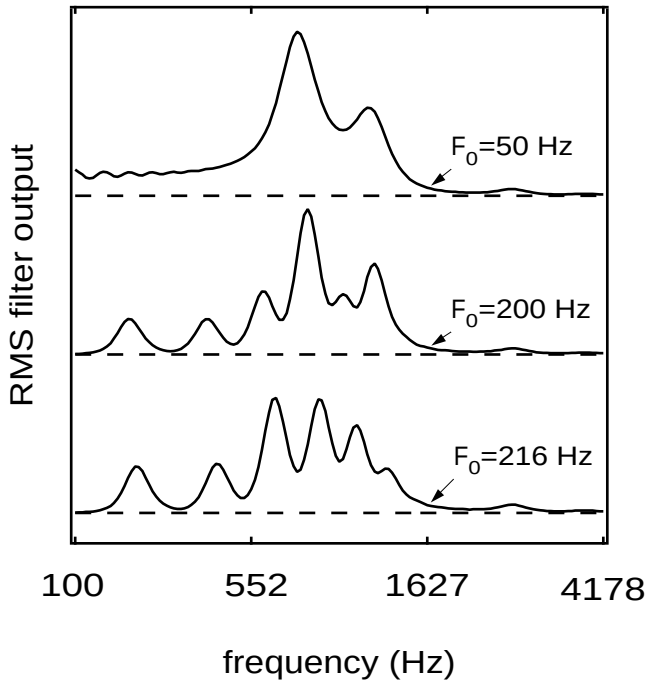

Fig. 2 Magnitude of output of gammatone filter bank as a function of channel frequency (excitation pattern). The filterbank had 75 channels uniformly spaced on a scale of equivalent rectangular bandwidth (ERB) from 100 Hz to 4178 Hz. Each curve is for a different F_0. Note the peaks at harmonics for the higher two F_0s, and the lack of unambiguous evidence for F1 and F2 at F_0 = 216 Hz.

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 this hypothesis for the following two reasons. (a) The sampling of the vocal tract implies a genuine loss of data: the waveform contains no information about the transfer function at frequencies other than multiples of the F_0. Paradoxically, this loss is a consequence of repeated excitation of the vocal tract, and it is all the more severe as the repetition rate is high. Interpolation or smoothing cannot recover this information. (b) Interpolation, etc. are attempts to guess missing data based on an a priori model of what the missing data should look like. To the extent that the guess is incorrect, interpolated data will be misleading.

1.1 The sampled spectral envelope

The shape of a spectral envelope can be described in the Fourier domain along a dimension of inverse frequency (time interval, often referred to as lag, or quefrency). A smooth spectral envelope has mainly components at short lags. We shall use the term sampling lag to designate the sampling rate along the frequency axis, inverse of the spacing (F_0) between sampling points. From the sampling theorem, we know that a spectral envelope sampled at a sampling lag $T_0 = 1/F_0$ is adequately represented by its sample points if it contains no components beyond half the sampling lag, or Nyquist lag, $T_0/2$.

Consider for example the short-term spectrum of the vowel /a/ at F_0 = 100 Hz [Fig. 3(a), top]. It resembles the spectral envelope sampled along the frequency axis at intervals of 100 Hz, with a sampling lag of 10 ms and a Nyquist lag of 5 ms. Supposing that the spectral envelope contains no components (ripples) with lags larger than the Nyquist lag, it's shape can be accurately reconstructed from the short-term spectrum by filtering in the lag domain to remove components beyond the Nyquist lag. The lower curve in Fig. 3(a) shows the result of such filtering. Filtering was performed by taking the Fourier transform of the short-term spectrum, setting values beyond the Nyquist lag to zero, then applying an inverse transform. The result is not quite the same as the original (Fig. 1), implying that the original spectral envelope did in fact contain components beyond the Nyquist lag. However the differences are small, suggesting that little is lost by sampling at 100 Hz intervals.

At 200 Hz [Fig. 3(b)] the peaks 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. 3(c)] the reconstructed envelope is severely distorted, indicating that the spectral envelope contained a large proportion of components beyond the Nyquist lag (1.67 ms), that were necessary to describe the original shape. Those components

were lost in the voice production stage.

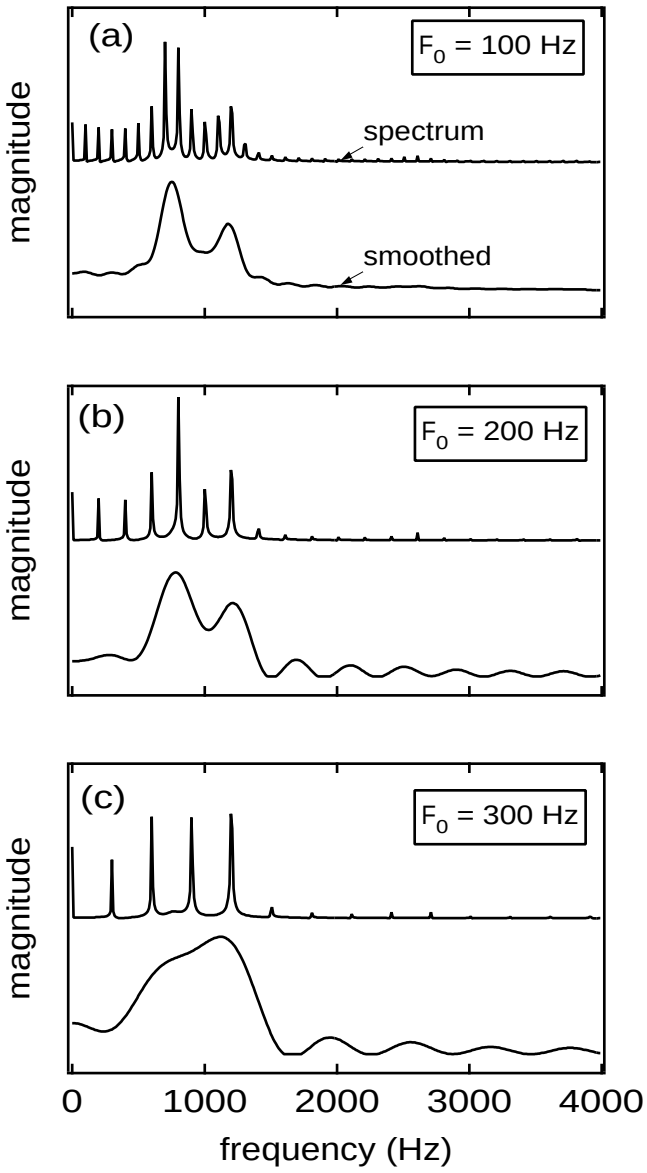

Fig. 3 (a) Short-term magnitude spectrum of /a/ at $F_0 = 100$ Hz, and smoothed short-term spectrum. Smoothing was performed by taking the Fourier transform of the magnitude spectrum, setting it to zero for lags larger than the Nyquist lag $T_0 = 1/2F_0$ (5 ms), and taking the inverse Fourier transform. The smoothed spectrum consists entirely of components below the Nyquist lag. (b) Magnitude spectrum and smoothed magnitude spectrum 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) Magnitude spectrum and smoothed magnitude spectrum at $F_0 = 300$ Hz. Note the distorted shape and lack of evidence of F1 and F2.

1.2 Smoothing considered as harmful

Smoothing is often proposed as a means to remove F_0-related structure from the short-term spectrum, in order to obtain an F_0-invariant representation that can be matched to templates. Low-pass filtering in the lag domain (previous paragraph) is one example of smoothing The same idea underlies cepstral smoothing, performed by manipulating the cepstrum (Fourier transform of the log magnitude spectrum), and in Sect. 2.2 we suggest similar manipulations of the autocorrelation function, that is the Fourier transform of the squared magnitude spectrum.

Smoothing or interpolation may be misleading. To see why, consider the smoothed envelopes produced by low-pass filtering in the previous paragraph. They embody the assumption that components of the envelope beyond the Nyquist lag are zero. That assumpton is incorrect, as evident from the mismatch between original and estimated envelopes. In other words, smoothing has replaced data that were incomplete but correct by data that are complete but incorrect.

To further illustrate this problem, a simple vowel identification model was implemented for the set of Japanese vowels /a/, /e/, /i/, /o/, /u/. The reference template for each vowel was the spectral envelope of that vowel (magnitude of transfer function). Target vowels were synthesized at F_0s ranging from 20 to 300 Hz in 1 Hz steps. The short-term spectrum was calculated and smoothed by removing components below the Nyquist lag (which depends on the target's F_0). The smoothed spectrum was compared to each template, using a spectral distance calculated by scaling both target and template to an RMS value of 1, and then calculating their RMS difference. Target-template distances for /a/ are plotted in Fig. 4(a). The distance between the target /a/ and 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.

Similar plots for the target /i/ are shown in Fig. 4(b). In this case, at high F_0s the estimated envelope is actually closer to the incorrect /u/ template than to the correct /i/ template. The model thus fails. One might conjecture that the problem lies in the dissimilarity between smoothed and unsmoothed spectral envelopes, and that it might be cured if smoothed spectral envelopes were used as templates. Figure 4(c) shows similar data for templates smoothed at a cutoff lag of 10 ms (Nyquist lag corresponding to a $F_0 = 200$ Hz). The estimated envelope is now close to the correct template /i/ at $F_0 = 200$ Hz, however at $F_0 = 300$ Hz the incorrect template /u/ is a better match. Not only that, but distances from rival templates are now similar at low F_0s, suggesting

that smoothing erased features useful for discrimination.

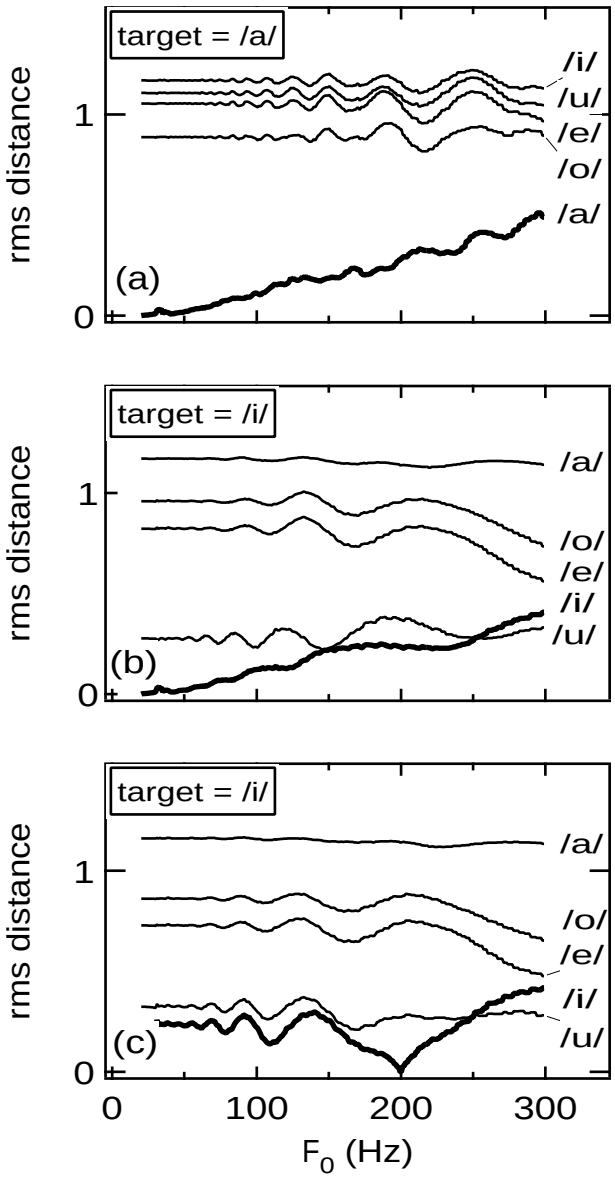

Fig. 4 (a) Distance between reference templates for vowels /a/, /e/, /i/, /o/ and /u/, and the smoothed short-term spectrum estimated from a synthetic /a/ waveform, as a function of F_0. Smoothing was performed by removing all components beyond the Nyquist lag ($1/F_0$). (b) Same as (a), for a synthetic /i/ waveform. (c) Same as (b), but reference templates were low-pass filtered at a lag of 1/400 Hz.

The astute reader may have guessed a solution. As evident from the dip at 200 Hz for /i/ in Fig. 4(c), a perfect match would be obtained at all frequencies

if the templates were smoothed at the same cutoff lag as the target. Reliable identification could thus be achieved, provided that an estimate of the F_0 of the vowel to be matched was available. The model proposed in this paper is equivalent to such a procedure.

2 Missing-data vowel identification model

The model acknowledges that important spectral information was lost due to sampling at F_0 multiples, and that the loss is irretrievable. Instead of trying to interpolate or otherwise estimate the missing parts of the transfer function, pattern matching proceeds using available features only, by giving zero weight to missing features in the pattern matching process. A similar idea underlies "missing feature" techniques that have been proposed recently in speech recognition to cope with deleted spectro-temporal features (Cooke et al., 1996, 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.

2.1 Frequency domain version

The frequency domain version of the 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) a spectral weighting function is calculated that emphasizes regions near multiples of F_0, (d) the short-term spectrum is compared to all templates, using the weighting function. Templates are defined for all frequencies, but comparison is restricted to certain frequencies.

The weighting function can be defined as

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

where $\delta()$ is the Dirac delta function and $\hat{F}_0$ is the estimate of F_0. The spectral distance from target S to template T_i can be calculated as:

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

where $S(f)$ is the target's short-term spectrum and $T_i(f)$ is the ith template.

The infinitely narrow peaks of $W(f)$ in Eq 1 are satisfactory in theory. In practice, to accomodate inevitable inaccuracy in F_0 estimation, the peaks should widen gradually with f. With a square shape and relative widths of 3%,

the weighting function would be equivalent to the harmonic sieve of Duifhuis et al. (1982), that has 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).

Reliable F_0 estimation is impossible for stimuli that are too short, whispered, or highly non-stationary, and in general at stimulus onset. However, in those cases the short-term spectrum is closer to the spectral envelope, and less affected by F_0. Non-uniform weighting is unnecessary. Weighting should thus be uniform by default, and gradually sharpen to emphasize F_0 multiplies if, and when, reliable F_0 information is obtained. Many F_0 estimation schemes produce, as a by-product, reliability estimates that could be used in such a scheme.

Physiologically, one could imagine short-term spectral estimation as being performed by the cochlea, sampled by non-uniform weighting of channels along a tonotopic dimension, based on an F_0 estimate (itself possibly be derived from tonotopic information). The main difficulty with this scheme is to imagine how the set of variable-pitch harmonic sieves is implemented across frequency channels, and how the appropriate sieve is selected based on the F_0 estimate. In the following section we consider an alternative model based on autocorrelation, that might be implemented physiologically using time domain processing.

2.2 Autocorrelation version

The Fourier-domain reasoning that was applied in Sect. 1.1 to the magnitude of the vocal-tract transfer function can be applied equally well to its square. The Fourier transform of the squared magnitude transfer function is the autocorrelation of the vocal-tract impulse response (ACF_{tract}). The squared transfer function of the vowel /a/ is plotted in Fig. 7(a), and the ACF_{tract} is plotted as a thin line in Fig. 5(a). When a vowel is produced with a constant F_0, the squared vocal tract transfer function is sampled at multiples of F_0. The sampling theorem tells us that the samples describe uniquely a spectral function that is band-limited to lags smaller than $T_0/2$. In other words, the information available about the vocal tract in the samples is entirely contained in the $\tau < T_0/2$ portion of the ACF_{tract} [Fig. 5(a), thick line].

According to Parseval's theorem, the Euclidean distance between square-magnitude spectra is the same as that between the corresponding autocorrelation functions. One can thus use autocorrelation templates instead of spectral templates to build a vowel identification model.

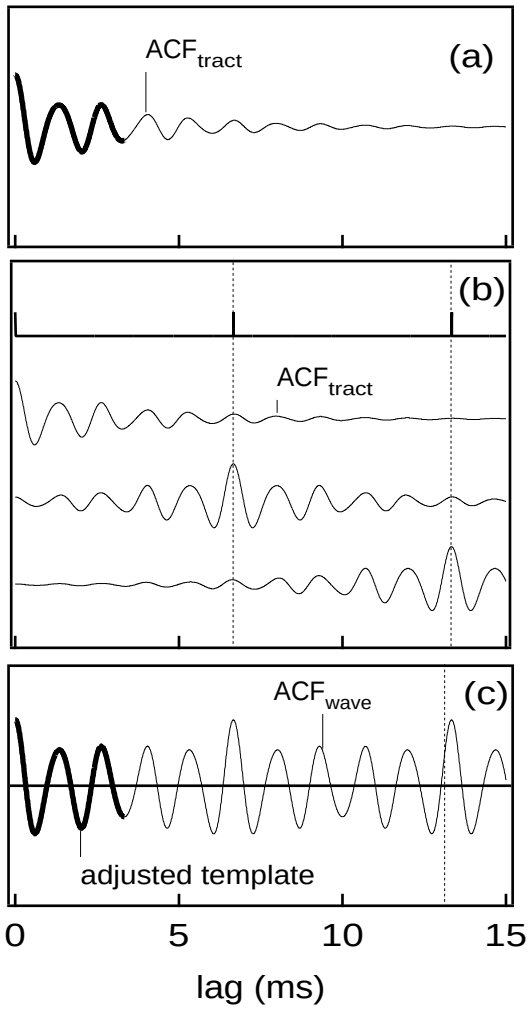

Fig. 5 (a) Autocorrelation of vocal tract transfer function (ACF_{tract}). (b) Illustration of the convolution process by which ACF_{wave} is derived from ACF_{tract} in the case where $F_0 = 150$ Hz. Copies of ACF_{tract} are shifted and added to obtain ACF_{wave} [thin line in (c)]. The vertical dotted lines indicate multiples of the period, 6.67 ms. The thick line in (c) is an adjusted template.

A problem remains. The autocorrelation function of the impulse response of the vocal tract (ACF_{tract}) is not directly observable. Observable is the autocorrelation of the waveform (ACF_{wave}), that is related to (ACF_{tract}) by the following relation:

$$ACF_{wave}(\tau) = ACF_{tract}(\tau) \circ \sum_{k=-\infty}^{\infty} \delta(\tau - kT) \qquad (3)$$

where ∘ represents convolution. ACF_{wave} is the result of convolving ACF_{tract} by a periodic series of delta functions with period T_0. The convolution is illustrated in [Fig. 5(b)]. Copies of (ACF_{tract}) are shifted to multiples of $_0$, and added up. Because of overlap between the shifted functions, ACF_{wave} differs from ACF_{tract}, even in the region $\tau < T_0/2$. The discrepancy depends on F_0.

For this reason, ACF_{wave} cannot make a perfect match to the templates, even if the match 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 ACF_{tract} [exactly as in the convolution illustrated in Fig. 5(b)]. In this way, an accurate match is obtained between the correct template and the observed ACF_{wave} [Fig. 5(c)].

We can thus formulate an autocorrelation version of the missing-feature vowel perception model. The following steps occur when a vowel is recognized: (a) the ACF_{wave} is estimated from the waveform, (b) the F_0 is estimated, (c) the ACF_{tract} templates are adjusted based on the F_0, and (d) they are compared to the ACF_{wave} over the $0 - T_0/2$ range of lags.

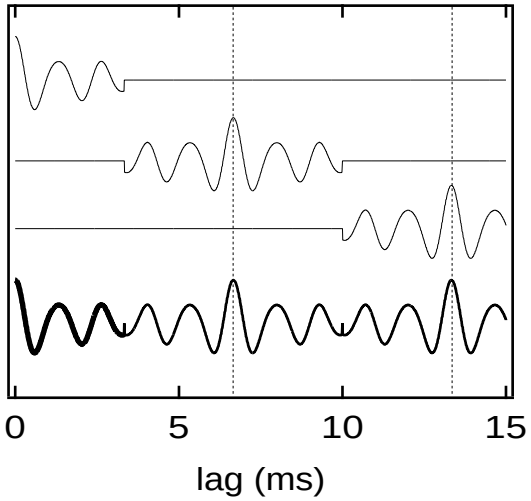

Fig. 6 Ilustration of the hypothetical case of a vocal tract with a transfer function band-limited to lags smaller than 3.33 ms. ACF_{wave} is derived from ACF_{tract} by convolution, but the $\tau < 3.33ms$ portion is unaffected by the convolution and remains equal to ACF_{tract}. Template adjustment is unnecessary in this (hypothetical) case.

The adjustment step would be unnecessary if ACF_tract were limited to $T_0/2$, as illustrated in the top of Fig. 6. In that case, ACF_{wave} and ACF_{tract} would be equal for $\tau < T_0/2$ (thick line in the bottom of Fig. 6). Leaving out

the adjustment stage is equivalent to assuming that such is the case. This is almost equivalent to the simple, smoothed-spectrum matching model discussed in Sect. 1.2, the difference being that in that case the magnitude spectrum was supposed to be band-limited, whereas here it is the squared magnitude spectrum.

The squared magnitude spectrum emphasizes the high-amplitude parts of the spectrum at the expense of others [Fig. 7 (a)]. Formant F1 is well represented (accounting for the ripple that dominates ACF_{tract}), F2 less well, and higher formants hardly at all. The magnitude spectrum used in Sect. 1.2 is a slightly more balanced representation [Fig. 7 (b)]. The log magnitude spectrum represents peaks and valleys in equal detail, whatever their amplitude[Fig. 7 (c)], and its inverse Fourier transform, the cepstrum, is widely used in speech processing. The success of the cepstrum in speech processing applications suggests that the log magnitude spectrum, and cepstrum, might be more useful substrates for pattern matching than the square magnitude spectrum and autocorrelation used in the present model. This weakness of the autocorrelation function is partly compensated for in the model of the next section, where autocorrelation functions are calculated within channels of a basilar-membrane/hair-cell model.

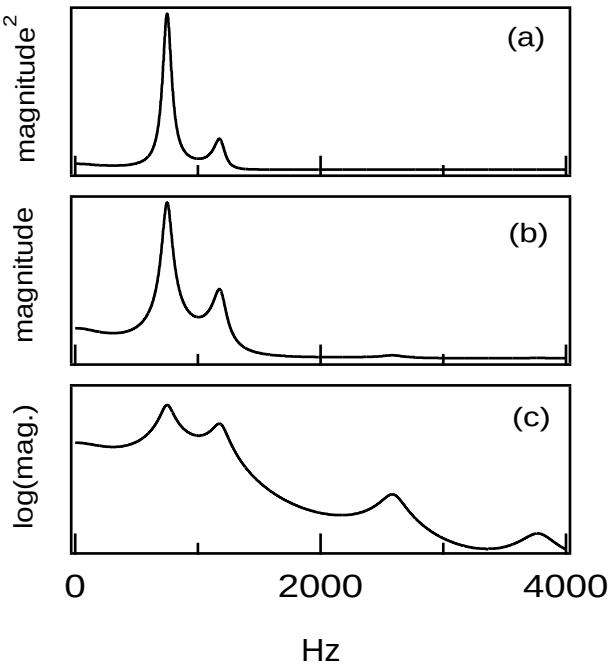

Fig. 7 (a) Squared-magnitude of transfer function of /a/. (b) Magnitude transfer function of /a/. (c) Log-magnitude transfer function of /a/.

An interesting feature of the autocorrelation model is that high-pass filtering in the lag domain can be applied to the autocorrelation function to factor out large-scale variations (such as spectral tilt, etc.). A similar operation in the cepstral domain (high-pass cepstral "liftering") is used effectively for the same purpose in speech recognition applications (Tohkura, 1987).

2.3 Spike-train coincidence version

In Section 2.1 we suggested that the spectral version of the model might be implemented by frequency-domain processing within the auditory system, based a tonotopic representation. Here we describe how the autocorrelation version of the model might be implemented by time-domain processing within the auditory system, based on the temporal structure of nerve fiber discharge patterns.

Autocorrelation of nerve fiber discharge pattern has been suggested as a basis for pitch perception (Licklider, 1951; Meddis and Hewitt, 1991a,b). In the 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. 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 and related models (de Cheveigné, 1998).

The SACF was also proposed as a substrate for vowel identification by Meddis and Hewitt (1992). In their model, vowels were identified by matching the "low-lag" portion of the SACF ($\tau < 4ms$). A similar scheme was used with success by de Cheveigné (1997), also for vowel identification.

Figure 8 (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 10000 Hz. The stimulus was a single impulse response of the vocal tract corresponding to the vowel /a/. The combination of vocal tract and basilar membrane filter has a much longer impulse response than the vocal tract alone, which explains the slow decay of the ACFs in Fig. 8 compared to Fig. 5(b), particularly in channels tuned to low frequencies or near a formant. Due to the rectifying property of the hair-cell model, ACFs are never negative. The non-zero baseline value is due to the fact that the hair-cell model parameters produced a relatively high spontaneous discharge

probability (probability of discharge in the absence of stimulation).

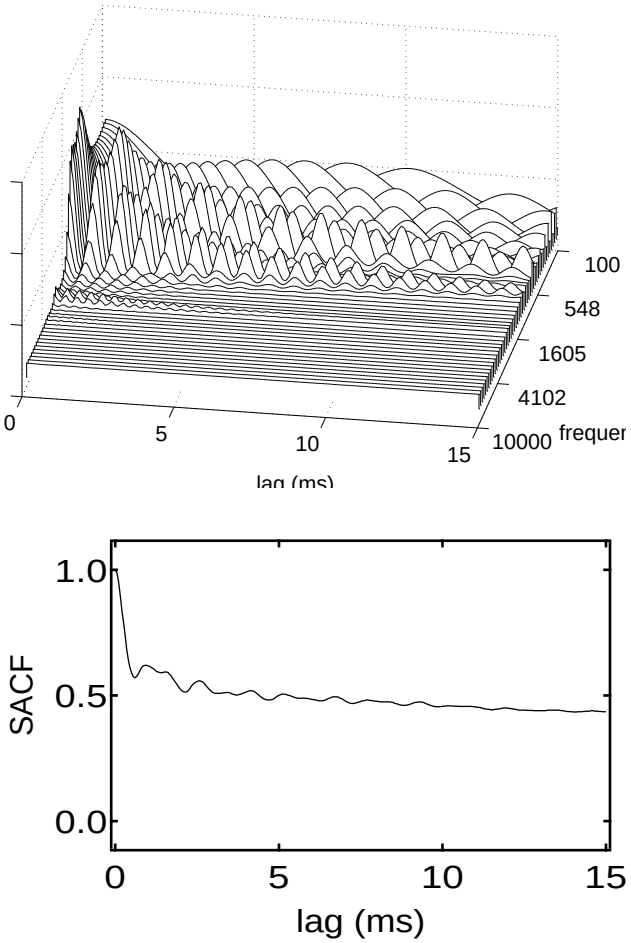

Fig. 8 Top: Autocorrelation of auditory nerve fiber discharge probability as a function of channel frequency, in response to a single impulse of the vowel /a/. Probabilities were produced by a model of peripheral filtering and hair-cell transduction, with 50 channels uniformly spaced between 100 and 10000 Hz on a scale of equivalent rectangular bandwidth (ERB). Bottom: summary autocorrelation function (SACF).

Figure 8 (bottom) shows the corresponding summary autocorrelation function (SACF) (normalized by dividing by the value at zero lag). The SACF decays to a value that is relatively high, due to the summation of non-negative ACFs responding to different frequencies and phases, and to the relatively high spontaneous rate. Compared to the autocorrelation of the waveform [ACF_{tract}, Fig. 5(b)], the SACF lacks the strong ripple at the period of F1. This is due to the saturating properties of the basilar membrane model, that

limit the response at high amplitudes, and therefore equalize contributions of different channels. Compared to ACF_{tract}, the SACF is more affected by F2 and other components, and relatively less by F1.

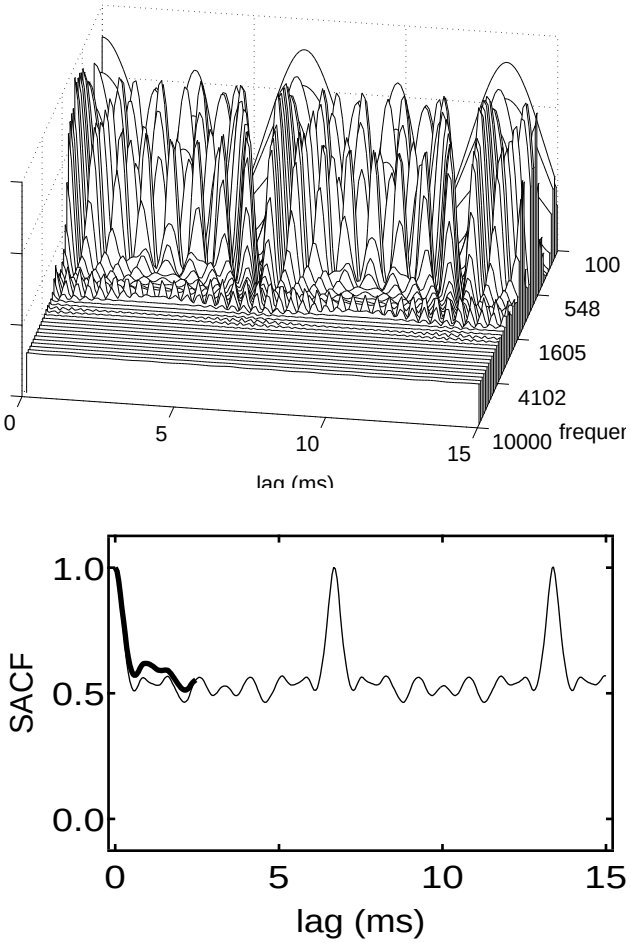

Fig. 9 Same as Fig. 8, in response to vowel /a/ at $F_0 = 150$ Hz.

Figure 9 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 $0 - T_0$, the SACF resembles the SACF obtained in response to the impulse response (thick line). This justifies the choice made by Meddis and Hewitt (1992) to use the low-lag portion of the SACF for their vowel-identification model.

There are nevertheless differences between the two SACFs, as we observed previously for the wave-form based autocorrelation functions. Analysis of these differences is complicated by the presence of the non-linear hair-cell transduction stage. Because of the non-linearity, within-channel ACFs in response to

the periodic stimulus cannot accurately be calculated as a convolution, as in Sect. 2.2. The consequences of this difference need to be worked out.

It is customary to consider the SACF instead of on the full ACF array (Meddis and Hewitt, 1992), but this is not mandatory. Matching can be performed on the 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 array is a richer pattern than the SACF. Another is the possibility to factor out spectral tilt and other transmission channel characteristics (within-channel compression or automatic gain control might serve this purpose). A third is that parts of the partial-SACF array can be weighted differently, to handle "missing features" in the spectral domain (Cooke et al., 1997; Morris et al., 1998). An expanded version of Meddis and Hewitt's (1992) concurrent vowel segregation model could be implemented in this way. The ACF or sub-SACF array is thus a flexible starting point for sophisticated models of identification and segregation.

3 Discussion

3.1 F_0-dependency of vowel perception

F_0 affects any spectral pattern that can reasonably be extracted from the waveform, making the search for F_0-invariant features a difficult goal to attain. Our model sidestepped this problem by making the pattern matching process itself also dependent on F_0. It should be stressed that in doing so we dealt with only one source of F_0-dependent variability. Other sources are variations with F_0 of the shape of glottal pulses, changes in the height of the larynx induced by changes in F_0, the covaration of F_0 range and vocal tract dimensions within a population of speakers, intrinsic F_0 of vowels, etc. (Carlson et al., 1975; Assmann et al., 1982; Rosner et al., 1994; Neary, 1989; Miller, 1989; Hirahara et al., 1992, 1993; Ainsworth, 1971). 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. Although vowel identity changes little with F_0 (Neary, 1989; Slawson, 1967), elevation of F_0 is known to increase error rates in multivowel identifying or matching tasks (Ryalls and Lieberman, 1982; Sundberg et al., 1982; Benkolken et al., 1990; Rosner et al. 1994; Gottfied et al., 1986). This is understandable, as at high F_0s the widely spaced sample points provide less information about the vocal tract, even if they are used optimally as here. Variation in time of F_0 might help by providing information about the derivative of the spectral envelope at each sample point (McAdams and Rodet, 1988). However Marin and

McAdams (1991) looked for such an advantage of "envelope tracing" in a concurrent vowel identification task, and failed to find any. Transitions between vowels and consonants may also be of use, particularly at high F_0s (Sundberg et al., 1982, Strange et al., 1976; Gottfried et al., 1986). In fact, there is evidence that vowels can be identified from the unvoiced parts belonging to the consonant with which they are articulated (Bonneau, 1996).

3.2 Perception in the absence of F_0

The model uses F_0 to adjust templates and/or restrict matching to certain features. That is impossible when a reliable F_0 estimate is unavailable, such as for stimuli that are short or otherwise non-stationary, whispered vowels, etc.. Whispered vowels are nonetheless intelligible, and it is known that vowel identity can be perceived from a single period of the vowel, whereas F_0 estimation requires several periods (Robinson and Patterson, 1995). A model that requires F_0 seems to be in trouble.

Actually, this is less of a problem than it seems. When the vocal tract is stimulated by a single pulse or noise, the short-term spectrum reflects the shape of the transfer function in a relatively unbiased way. The adjustments provided by our model are not required. A bias appears when the excitation is periodic, but in that case F_0 can be reliably estimated, and the "missing-feature" model can step in and make the necessary adjustments. The model would thus operate by default in an F_0-independent mode, and would switch to an F_0-dependent mode as soon as F_0 is available.

Transitions may aid identificaton: Summerfield et al. (1984) remarked that a long-duration vowel waveform sounds vowel-like at onset, but may then loose its identity, and partially regain it at offset (see also Carré and Lancia, 1975).

3.3 Relation to other models

The spectral-domain version of the model is quite similar to the "harmonic sieve" of Duifhuis et al. (1982; Scheffers, 1983). The harmonic sieve was proposed by Duifhuis as a means to select the components of a sound that should enter the calculation of its pitch. Scheffers used it to assign components of a mixed-speech spectrum to each voice. Moore et al. (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 all 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.

The model of Scheffers (1983) and the PEAK model of Assmann and Summerfield (1989) also sampled the excitation pattern at multiples of F_0, but they used it to construct a spectral envelope by interpolation. Features were extracted from peaks and shoulders of this envelope (as determined by zeros in the first and second differential respectively). Interpolation beween sample points can be seen as the result of smoothing the sampled excitation pattern with a triangular window of width $2F_0$. As such it is similar to the spectral smoothing that we discussed in the Introduction, and vulnerable to the same criticism. For example, applied to the spectral envelope of Fig. 2(a) or (b) the PEAK algorithm would find two formants, but applied to that of Fig. 2(c) it would find only one. The strength of that algorithm, according to our analysis, is that it samples the excitation pattern at multiples of F_0. Its weakness is that it then attempts to derive invariant features (formant peaks or shoulders) rather than exploiting the sample points directly. Similar problems are likely to exist for other smoothing schemes, such as the second integration of Rosner and Pickering (1994) or the 3.5-bark integration scheme of Chistovich (1985).

Perception is known to depend on spectral characterics near peaks more than near valleys. This gives an advantage to models that extract formant positions. Ours does not do so implicitely, but this does not necessarily mean that it gives peaks and troughs equal weight. All depends upon the representation. For example, the square magnitude spectrum emphasizes high-amplitude parts of the spectrum, and the same is thus true of the ACF, its Fourier transform. Indeed, the excessive weight of F1 is a weakness of this representation, as cues to F2 are weak, and cues to higher formants almost non-existent. By calculating the ACF independently in several channels, the auditory system may possibly obtain a better-balanced representation, in a sense closer to the cepstrum.

The present model is a way similar to the whole-spectrum model of Bladon (1982). It takes care of the problem of F_0-related mismatch, without running into the problems related to spectral smoothing and integration. Most models can be seen as "bottom-up", in that the signal processing module is responsible for producing the best possible "F_0-invariant" representation to be matched to internal templates. The present model has more of a "top-down" flavor, in that the pattern matching module has the responsibility of producing F_0-dependent templates to ease the match with the representation derived from the waveform. In this sense, it is related to "analysis-by-synthesis" models (Bell et al., 1961), and bears some relation to the ideas of Chistovich (1985).

The proposition that the F_0 of a vowel might be useful for its identification seems to contractict 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., 1995, 1997).

3.4 Application to Speech Recognition

The reasoning of Sect. 2.2, based on the square-magnitude-spectrum / autocorrelation-function Fourier transform pair, can be applied as well to the log-magnitude-spectrum / cepstrum pair. The useful information carried by the set of F_0-spaced samples of the log magnitude spectrum is entirely contained in the part of the cepstrum with lags (or "quefrencies") below $t_0/2 = 1/F_0$.

When the cepstrum is used for speech recognition, it is common to limit it to small quefrencies to avoid the influence of small-scale details of the log-spectral envelope, in particular those due to F_0. Indeed, taking a cepstrum, setting to zero beyond a limit, and taking the transform back to the spectral domain is a process known as "cepstral smoothing". The limit is usually fixed. Our reasoning suggests that things might work better if the limit depended on F_0. What is sure is that values beyond the the Nyquist quefrency $T_0/2$ are meaningless, as far as the spectral envelope is concerned.

As in the case of the autocorrelation function, values of the cepstrum smaller than $T_0/2$ are distorted in a way that depends on F_0. In contrast to the autocorrelation case, it does not seem possible to describe this distortion as a simple convolution, as in Fig. 5. Without a model of the distortion, it is not possible to perform the template adjustment that was suggested for the autocorrelation (except via table-lookup). This question needs more work.

LPC analysis is also based on the autocorrelation function limited to small lags (Rabiner and Shaffer, 1978). The reasoning of Sect. 2.2 tells us that values of the ACF beyond the Nyquist lag are meaningless. The logical implication is that the limit used in the LPC calculation (and thus the LPC order) should be made to depend on F_0. Variable-order LPC is not without problems.

Missing-data techniques handle the effects of F_0 in the recognition phase. Remains the problem of F_0 in the training phase. Cooke et al. (1994) report experiments in which a speech recognizer was trained on incomplete data. Such techniques might be applied to handle the effects of F_0 in the training phase of a speech recognizer. Alternatively, if low-F_0 speech is abundant and representatitive, it might suffice for training.

3.5 Relation to PPS

PPS (pitch-period smoothing) uses F_0 to provide the best possible smoothing (de Cheveigné and Biem, 1998), as does Kawahara's (1997) STRAIGHT. PPS is equivalent to the smoothing of Sect. 1.2, and is the best way of handling

F_0-related structure in a bottom-up fashion. However top-down application of missing-feature techniques (based on the present model) should potentially lead to better results.

3.6 Some related results

These results will be weaved into the discussion of some future version of this report.

- Assmann and Neary (1987) matched synthetic vowel-like stimuli with various shapes to a continuum of synthetic vowels with variable F1.

  In general, the matched F1 was best predicted from the weighted sum of the two harmonics of highest amplitude in the F1 region. The single highest harmonic, the weighted sum of the three highest harmonics, and LPC analysis gave less good predictions.

  In one experiment, F1 was represented by harmonic pairs. Both harmonics had the same amplitude, or either the lower or the higher harmonic was boosted by 9 dB. F_0 was 125 Hz or 250 Hz. Results were similar for both F_0s: the matched F1 corresponded to the weighted sum of the two harmonics, even when they were widely spaced (250 Hz + 500 Hz, 500 Hz + 750 Hz, etc.) and would have been resolved by peripheral filtering.

- Beddor and Hawkins (1990) matched nasal to oral vowels, and found that the F1 of the matching oral vowel was intermediate between the F1 and the "centroid" of the nasal vowel. It was closer to the sharp nominal F1 than to the blunt FN.

  They also matched two-formant vowels with close formants (similar to /o/, /a/, /u/) to single-formant vowels. The frequency of the matched single formant fell between the F1 and centroid of the two-formant vowel, and closer to its F1. The dominance of F1 was greater than for nasal vowels, suggesting that the centroid is more important when formants are wide (FN) and less when they are narrow (F2).

  Finally, they matched two-formant vowels with either wide (150 Hz) or narrow (45 Hz) bandwidths to two-formant vowels with medium bandwidths (75 Hz) (/o/ and /a/). The match for narrow-formant vowels corresponded to a match of centroid frequencies, but the match for wide-formant vowels was different. The difference was attributed to differences in "spectral shape", consequence of the interaction between formants and harmonics (when the formant falls between harmonics, the peak appears flatter and thus matches wide-formant vowels better).

"...where harmonics are sufficiently prominent, their frequency is of prime importance..." "When the lowest frequency spectral prominence is broad, what is important is...the general correspondance in shape of the entire region of spectral prominence..."

- Carlson, Fant, and Granström (1975) determined the F1 boundary between Swedish /i/ and /e/, for F0s between 100 and 160 Hz. The boundary increased (from about 300 to 350 Hz), regularly and with no jumps. It did not correspond to either the "most significant" or the loudest harmonic. The weighted sum of two most prominent partials was the best predictor (but it was not clear how it predicted a shift with F0...).

- Carré and Lancia (1975) remarked that formant information in stationary high-F_0 voices is poor, anf suggested that transients may provide the missing information.

- Bladon (1982) points out that formants are "indeterminate", in particular because when measured (on a spectrogram), at high F_0, they tend to follow a harmonic, and thus vary with F_0 whereas the vowel timbre remains constant. They use this to argue against the adequacy of the formant, as opposed to a "whole spectrum" difference metric.

- Sundberg and Gauffin (1982) state that singers may maintain vowel intelligibility up to 500 Hz for isolated vowels, and 1000 Hz in a CVC context. In soprano singing F_0 may be as high as 1500 Hz.

  They performed an identification experiment with 6 isolated vowels (12 allowed response categories) at F_0s from 260 to 700 Hz. Identification rate fell as F_0 increased.

- Traunmüller (1982) found that single-formant vowels were identified on the basis of that formant's frequency if $F_0 < 350$ Hz, and on the basis of the frequency of the second partial if $F_0 > 350$ Hz.

- Baru (1982) had dogs discriminate between vowels /a/ and /i/ at fundamentals of 120 Hz and 240 Hz. Discrimination was as good at 240 Hz as at 120 Hz, and training at one F_0 transferred to the other. Dogs were also capable of discriminating 120 Hz from 240 Hz, and training from one vowel transferred to the other (does it transfer from one dog to the other?).

- Ainsworth (1975) notes that children's voices are about 1 octave above adult men, and their formant frequencies about 30 % higher.

In an identification task with "h*d" words, a 1-octave rise in F_0 caused a 3-4% increase of perceived formant frequency. Increase was roughly the same for nominal F_0s of 120, 240 and 360 Hz (details are unclear).

- Darwin and Gardner (1985) found evidence that harmonics other than the two largest affect the "effective F1". Effective F1 corresponded well with results of LPC analysis.

- Assmann et al. (1982) found that inclusion of an F_0-based parameter improved discriminant analysis of sets of acoustic parameters of vowels.

- Benolken and Swanson (1990) asked a soprano singer to produce 12 american vowels at F_0s of 262 to 1047 Hz. Identification rate decreased from about 70% at 262 Hz to about 10% at 1047 Hz.

- Neary (1989) notes the covariation of F_0 (over a 1 octave range) and formant frequencies (over a 30 % range) between adult men and children. However he also notes that a) the coupling between source and tract characteristics is very loose (orthogonal to a first approximation), and b) in a given speaker the variation of formant frequencies with F_0 is small.

  In an identification experiment investigating intrinsic (such as F_0) and extrinsic (such as context) normalizaton effects, he found that an octave change in F_0 caused a 7-9% change in F1 and about 1.5% change in F2.

  Neary notes that resolution of individual harmonics poses a problem for both formant-based and "whole-spectrum" theories. He also notes that the 3.5 bark critical separation between formants has "no counterpart in general auditory psychophysics", and should be classed as a "speech mode" effect.

- Ryalls and Lieberman (1982) synthesized nine diphthong vowels of American English at fundamentals of 100, 135, and 250 Hz. Error rates were similar at 100 and 135 Hz, and greater at 250 Hz, even for vowels synthesized with formant frequencies adequate for female voices (based on data of Peterson and Barney, 1952).

  Ryalls and Lieberman attribute poor identification at high F_0 to sparse sampling of the spectrum.

- Klatt (1982) insists on the conflict between the need for filters to be wide-band (to remove differences between male and female speech) and narrow-band (to obtain sufficient spectral resolution).

- Miller (1989) proposes an "auditory-perceptual" model which scales formant frequencies based on the cubic root of F_0.

- Hirahara and Kato (1992) found that perceptual phoneme boundaries between Japanese vowels were much more stable in an (F1-F0) vs (F2-F0) plane than in an F1 vs F2 plane, when F0 varied from 123 to 423 Hz). Clusters for male and female speakers also overlapped better in that space.

  They suggested that the effect of F_0 might result from the dependency on F_0 of the strongest harmonic near a formant, but they did not give enough details to know whether this explanation works for all vowels, at all F_0s.

- Hirahara, Cariani and Delgutte (1996) recorded from the auditory nerve of the cat in response to Japanese vowels /e/ and /i/ (for which F_0 affected the boundary). At low F_0, F1 and F2 were evident in the population rate-place response, and also in temporal patterns of discharge. At high F_0, individual harmonics were evident in both representations.

- Hirahara (1993) found that Japanese /i/-/e/ F1 boundaries were a function of F_0. When F_0 was varied from 100 to 450 Hz, the F1 boundary shifted from about 320 to 540 Hz, but not linearly. It shifted from 320 to 400 Hz when F0 increased from 100 to 150 Hz. It then remained stable at 400 Hz when F0 increased from 150 to 250 Hz. Finally it increased from 400 to 540 Hz when F_0 increased from 250 to 450 Hz. This is not what one would expect if F1 were strongly influenced by the strongest harmonic near F1.

Conclusion

This paper addressed the problem of F_0-invariancy of vowel perception, using missing-feature pattern matching techniques. In voiced vowel production, the transfer function of the vocal tract is sampled at multiples of the F_0. Values between these points are missing. Reliable pattern matching can nevetheless be performed if the missing data are ignored in the matching process. This can be done in the frequency domain, using a spectral representation, or in the time domain, using the autocorrelation function. The autocorrelation version of the model can be implemented physiologically as a two-dimensional array of delay lines and coincidence-counting neurons, that calculate autocorrelation functions within each channel of the peripheral filter bank.

Acknowledgements

Thanks to Minoru Tsuzaki, Hiroaki Kato, Tatsuya Hirahara and Erik McDermott for comments on a previous version.

References

[1] Ainsworth, W. A. (1975). "Intrinsic and extrinsic factors in vowel judgements," in "Auditory analysis and perception of speech," Edited by G. Fant and M. A. A. Tatham, London, Academic Press, 103-113.

[2] 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.

[3] 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.

[4] 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.

[5] Baru, A. V. (1975). "Discrimination of synthesized vowels [a] and [i] with varying parameters (fundamental frequency, intensity, duration and number of formants) in dog," in "Auditory analysis and perception of speech," Edited by G. Fant and M. A. A. Tatham, London, Academic Press, 91-101.

[6] Beddor, P. S., and Hawkins, S. (1990). "The influence of spectral prominence on perceived vowel quality," J. Acoust. Soc. Am. 87, 2684-2704.

[7] 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," JASA 33, 1725-1736.

[8] 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.

[9] Bladon, R. A. W., and Lindblom, B. (1981). "Modeling the judgement of vowel quality differences," JASA 69, 1414-1422.

[10] 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, 95-102.

[11] Bonneau, A. (1996). "Identification of vowels from french stop bursts.", Proc. ESCA Workshop on the auditory basis of speech perception, Keele, 133-136.

[12] 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, London, Academic Press, 83-90.

[13] 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. F. a. M. A. A. Tatham, London, Academic, 55-82.

[14] Chistovich, L. A. (1985). "Central auditory processing of peripheral vowel spectra," J. Acoust. Soc. Am. 77, 789-805.

[15] Cooke, M., Crawford, M., and Green, P. (1994), "Learning to recognise speech from partial descriptions," (tech report?)

[16] Cooke, M., Morris, A., and Green, P. (1997). "Missing data techniques for robust speech recognition.", Proc. ICASSP, 863-866.

[17] 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.

[18] 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.

[19] de Cheveigne, A., McAdams, S., Laroche, J., and Rosenberg, M. (1995). "Identification of concurrent harmonic and inharmonic vowels: A test of the theory of harmonic cancellation and enhancement," J. Acoust. Soc. Am. 97, 3736-3748.

[20] 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.

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

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

[23] de Cheveigné, A., and Biem, A. (1998), "Pitch Synchronous Feature Estimation for Speech Recognition," ATR-HIP (in preparation) technical report, .

[24] Fahey, R. P., Diehl, R. L., and Traunmüller, H. (1996). "Perception of back vowels: effects of varying F1-F0 bark distance," JASA 99, 2350-2357.

[25] Fant, G. (1970). "Acoustic theory of speech production," .

[26] Gottfried, T. L., and Chew, S. L. (1986). "Intelligibility of vowels sung by a countertenor," JASA 79, 124-130.

[27] 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, Tokyo, Ohmsha, 89-112.

[28] Hirahara, T. (1993). "On the role of relative harmonics level around the F1 of high vowel identification.", Proc. ARO abstracts (ISSN 0742-3152), abstract #258, p 65.

[29] Hirahara, T., Cariani, P., and Delgutte, B. (1996). "Representation of low-frequency vowel formants in the auditory nerve.", Proc. ESCA Workshop on the auditory basis of speech perception, 83-86.

[30] Hoemeke, K. A., and Diehl, R. L. (1994). "Perception of vowel height: the role of F1-F0 distance," JASA 96, 661-674.

[31] Kawahara, H. (1997). "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited.", Proc. ICASSP, 1303-1306.

[32] Kewley-Port, D. (1996). "Psychophysical studies of vowel formants.", Proc. ESCA Workshop on the auditory basis of speech perception, Keele, 148-153.

[33] 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.

[34] 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, Amsterdam, Elsevier, 181-196.

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

[36] McAdams, S., and X., R. (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, London, Academic Press, 359-369.

[37] 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.

[38] 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.

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

[40] Miller, J. D. (1989). "Auditory-perceptual interpretation of the vowel," J. Acoust. Soc. Am. 85, 2114-2134.

[41] 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.", Proc. ICASSP, 737-740.

[42] Neary, T. M. (1989). "Static, dynamic, and relational properties in vowel perception," J. Acoust. Soc. Am. 85, 2088-2113.

[43] Peterson, G. E., and Barney, H. L. (1952). "Control methods in a study of the vowels," J. Acoust. Soc. Am. 24, 175-184.

[44] 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.

[45] Ryalls, J. H., and Lieberman, P. (1982). "Fundamental frequency and vowel perception," J. Acoust. Soc. Am. 72, 1631-1634.

[46] Rosner, B. S., and Pickering, J. B. (1994). "Vowel perception and production," Oxford, Oxford University Press.

[47] Slaney, M. (1993), "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Apple Computer technical report, 35.

[48] Slawson, A. W. (1967). "Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency," JASA 43, 87-101.

[49] Strange, W., Verbrugge, R. R., Shankweiler, D. P., and Edman, T. R. (1976). "Consonant environment specifies vowel identity," JASA 60, 213-224.

[50] Strange, W. (1987). "Information for vowels in formant transitions," J. Memory and Language 26, 550-557.

[51] 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. C. a. B. Granström, 223-228.

[52] Tohkura, Y. (1987). "A weighted cepstral distance measure for speech recognition," IEEE Trans. ASSP 35, 1414-1422.

[53] Traunmüller, H. (1982). "Perception of timbre: evidence for spectral resolution bandwidth different from critical band?," in "The representation of speech in the peripheral auditory system," Edited by R. Carlson and B. Granström, Elsevier, 103-108.