On the upper cutoff frequency of the auditory critical-band envelope detectors in the context of speech perception[a)]

Oded Ghitza
Media Signal Processing Research, Agere Systems, Murray Hill, New Jersey 07974

(Received 8 August 2000; revised 20 February 2001; accepted 7 June 2001)

Studies in neurophysiology and in psychophysics provide evidence for the existence of temporal integration mechanisms in the auditory system. These auditory mechanisms may be viewed as ''detectors,'' parametrized by their cutoff frequencies. There is an interest in quantifying those cutoff frequencies by direct psychophysical measurement, in particular for tasks that are related to speech perception. In this study, the inherent difficulties in synthesizing speech signals with prescribed temporal envelope bandwidth *at the output of the listener's cochlea* have been identified. In order to circumvent these difficulties, a dichotic synthesis technique is suggested with interleaving critical-band envelopes. This technique is capable of producing signals which generate cochlear temporal envelopes with prescribed bandwidth. Moreover, for unsmoothed envelopes, the synthetic signal is perceptually indistinguishable from the original. With this technique established, psychophysical experiments have been conducted to quantify the upper cutoff frequency of the auditory critical-band envelope detectors at threshold, using high-quality, wideband speech signals (bandwidth of 7 kHz) as test stimuli. These experiments show that in order to preserve speech quality (i.e., for inaudible distortions), the minimum bandwidth of the envelope information for a given auditory channel is considerably smaller than a critical-band bandwidth (roughly one-half of one critical band). Difficulties encountered in using the dichotic synthesis technique to measure the cutoff frequencies relevant to intelligibility of speech signals with fair quality levels (e.g., above MOS level 3) are also discussed.
[DOI: 10.1121/1.1396325]

PACS numbers: 43.71.Pc, 43.66.Ba, 43.72.Ar [DOS]

I. INTRODUCTION

Studies in neurophysiology and in psychophysics provide evidence for the existence of temporal integration mechanisms in the auditory system (e.g., Eddins and Green, 1995). The neural circuitry that realizes these mechanisms is yet to be understood. At the least, we may view these mechanisms as ''detectors,'' characterized in part by their lower- and upper cutoff frequencies. These cutoff frequencies determine which part of the input information that is present at the auditory-nerve (AN) level is perceptually relevant. Hence, it is important to quantify these frequencies, particularly for tasks that are related to speech perception.

Two recent studies (Drullman *et al.*, 1994 and Chi *et al.*, 1999) seem to provide psychophysically based estimates of the cutoff frequencies of the auditory detectors involved in tasks related to speech intelligibility. These studies are inspired by the apparent ability of the speech transmission index (STI) to predict intelligibility scores for speech recorded in auditorium-like conditions (e.g., Steeneken and Houtgast, 1980). Recall that the STI is computed from the modulation transfer functions (MTFs) of the transmission path between the location of the speech source and that of the microphone. An MTF is specified at a given frequency as the degree to which the original intensity modulations are preserved at the microphone location. In Steeneken and Houtgast, 1980, the MTFs are measured for 7 one-octave-wide noise carriers centered at frequencies that are one octave apart (from 125 to 8000 Hz), with 14 modulation frequencies (0.63 to 12.5 Hz, in one-third-octave steps). [Note that the range of center frequencies covers the frequency range used in speech communication, and that the range of the modulation frequencies covers the time constants of the articulatory mechanisms used by the human speaker.] The high correlation of STI and speech intelligibility scores (Steeneken and Houtgast, 1980), and the fact that STI is based upon MTFs, raises the question whether auditory detectors active in the speech intelligibility task have a cutoff frequency of the order of 12.5 Hz (i.e., the maximum modulation frequency in Steeneken and Houtgast, 1980). In Drullman *et al.* (1994), an attempt was made to assess the amount by which temporal modulations can be reduced without affecting the performance in a phoneme identification task. Results showed that temporal envelope smoothing hardly affect the performance, even for cutoff frequency as low as 16 Hz. In Chi *et al.* (1999), detection thresholds were measured for spectral and temporal MTFs using broadband stimuli with sinusoidally rippled profiles that vary with time. Results showed that temporal MTFs exhibit low-pass characteristics, with cutoff frequencies similar to those of Drullman *et al.* (1994).

A question that emerges at this point is whether the psychophysical data obtained by these experiments, about the bandwidth of temporal MTFs, can also be considered as evidence of the characteristics of the relevant auditory mechanisms (i.e., that they are low-pass in nature, with cutoff frequencies of about 16 Hz). As shown in Sec. II, such a

[a)]This work was done while the author was with Bell Labs, Lucent Technologies.

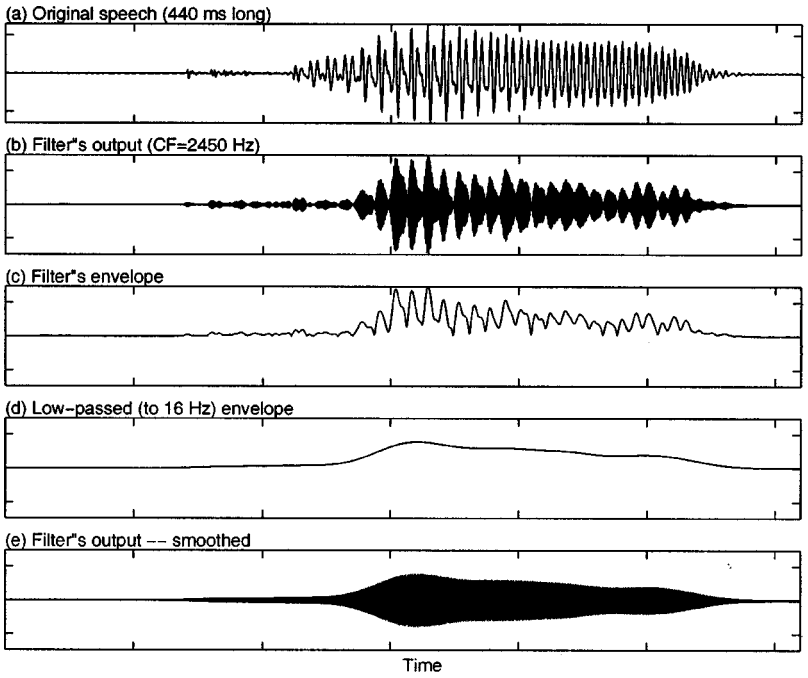

FIG. 1. From top to bottom: (a) a 440-ms-long segment of the original speech $s(t)$; (b) the output signal, $s_i(t)$, of a critical-band filter centered at 2450 Hz; (c) the envelope $a_i(t)$; (d) the smoothed envelope $\tilde{a}_i(t)$ (low-pass filtered to $B=16$ Hz); and (e) the envelope-smoothed critical-band signal $\overline{s_i}(t)$. The ordinate of panels (b) to (e) have the same scale. The ordinate of panel (a) has a different scale.

conclusion is not permissible. This is so because the observed psychophysical performance is, in part, a consequence of using signal-processing techniques which, for a prescribed envelope bandwidth, produce synthetic signals that generate internal auditory representations whose temporal envelopes are wideband signals, with envelope bandwidths as wide as one critical band. Therefore, while performing the psychophysical experiments the human observer was presented with rich temporal envelope information, with a bandwidth much beyond the nominal value prescribed at the input.

In Sec. III, the difficulties inherent in synthesizing speech signals with prescribed temporal envelope bandwidth at the output of the listener's cochlea are identified. In order to circumvent these difficulties, a dichotic[1] synthesis has been suggested with interleaving smoothed critical-band envelopes. This technique has two desired capabilities: (1) it produces synthetic signals which generate cochlear temporal envelopes with prescribed bandwidth, and (2) for unsmoothed envelopes, the synthetic signal is perceptually indistinguishable from the original. With this technique established, psychophysical experiments have been conducted to quantify the upper cutoff frequency of the auditory critical-band envelope detectors at threshold (i.e., in the context of preserving speech quality) using high-quality, wideband speech signals (bandwidth of 7 kHz) as test stimuli (Sec. IV). Finally, in Sec. V, the difficulties encountered in using the dichotic synthesis technique to measure the cutoff frequencies relevant to intelligibility of speech signals with some reasonable level of quality (say, ''fair''—or 3—on the MOS scale[2]) are also discussed.

II. TEMPORAL SMOOTHING AND SPEECH INTELLIGIBILITY

It is widely accepted that a decomposition of the output of a cochlear filter into a temporal envelope and a ''carrier'' may be used to quantify the role of auditory mechanisms in speech perception (e.g., Flanagan, 1980). This is supported by our current understanding of the way the auditory system (the periphery, in particular) operates.

Let $s(t)$ be the original speech signal, and let $s_i(t)$ be a bandlimited signal resulting from filtering $s(t)$ through $h_i(t)$

$$s_i(t)=s(t)*h_i(t). \tag{1}$$

Here, $h_i(t)$ is the impulse response of the ith critical-band filter and the operator $*$ represents convolution. We can express $s_i(t)$ of Eq. (1) as

$$s_i(t)=a_i(t)\cos\phi_i(t), \tag{2}$$

where $a_i(t)$ is the *Hilbert envelope*[3] of $s_i(t)$, $\phi_i(t)$ is the *Hilbert instantaneous phase*[3] of $s_i(t)$, and $\cos\phi_i(t)$ is the *carrier* of $s_i(t)$. We refer to the expression of Eq. (2) as ''the envelope/carrier decomposition'' of $s_i(t)$.

Let $\tilde{a}_i(t)$ be a filtered version of $a_i(t)$, low-passed to some cutoff frequency B. The *envelope-smoothed* critical-band signal is defined as

$$\overline{s_i}(t)=\tilde{a}_i(t)\cos\phi_i(t), \tag{3}$$

and the envelope-smoothed speech signal is defined as

$$\overline{s}(t)=\sum_{i=1}^{N}\overline{s_i}(t)=\sum_{i=1}^{N}\tilde{a}_i(t)\cos\phi_i(t), \tag{4}$$

where N is the number of critical bands.

Figure 1 shows (from top to bottom) (a) a 440-ms-long segment of the original speech $s(t)$; (b) the output signal, $s_i(t)$, of a critical-band filter centered at 2450 Hz; (c) the envelope $a_i(t)$; (d) the smoothed envelope $\tilde{a}_i(t)$, low-pass filtered to $B=16$ Hz, and (e) the envelope-smoothed critical-band signal $\overline{s_i}(t)$.

In Drullman *et al.* (1994), the envelope-smoothed speech of Eq. (4) was used to measure human performance in a phoneme identification task as a function of the cutoff

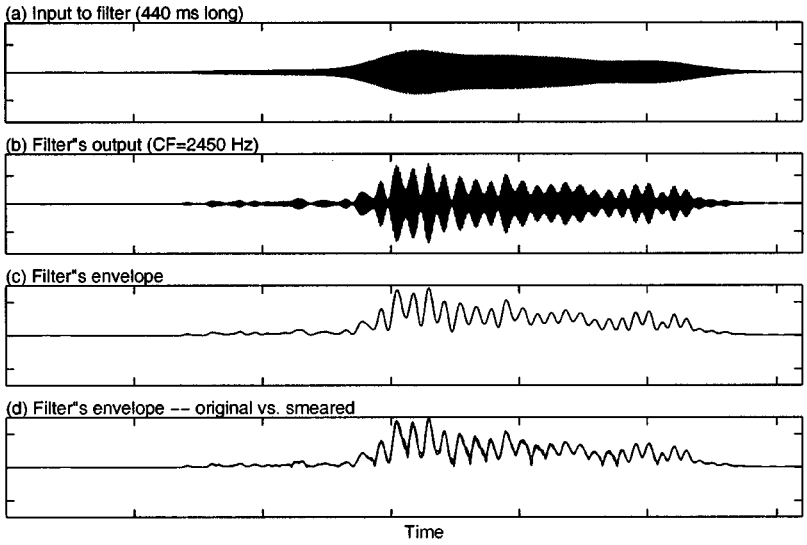

FIG. 2. From top to bottom: (a) Fig. 1(e), redrawn; (b) the output signal of a critical-band filter centered at 2450 Hz, for the input signal shown in (a); (c) the envelope signal of the critical-band signal of (b); and (d) comparison of the envelope signals of Figs. 1(c) and 2(c). Ordinate of all panels have the same scale.

frequency B of a low-pass filter representing the temporal smoothing. Results showed that performance was hardly affected by temporal envelope smoothing characterized by cutoff frequencies higher than 16 Hz.

A question that emerges at this point is whether these findings can be considered as evidence that relevant auditory mechanisms are low-pass in nature, with cutoff frequency of about 16 Hz. This question stems from our current understanding of the relationship between the envelope $a_i(t)$ of the driving signal and the properties of the auditory-nerve firing patterns they stimulate. This understanding is better, in particular, for AN fibers with high characteristic frequencies (CFs),[4] where the synchrony of neural discharges to frequencies near the CF is greatly reduced, due to the physiological limitations of the inner hair cell (IHC) in following the carrier information. At these frequencies, temporal information is preserved by the instantaneous average rate of the neural firings, which is related to the temporal envelope of the underlying driving cochlear signal.[5] Is it correct to assume that, by presenting the listener with the envelope-smoothed signal $\tilde{a}_i(t)\cos\phi_i(t)$, the instantaneous average rate of the corresponding stimulated AN fibers is also smoothed, limiting the bandwidth of the information available to the upper auditory stages to B?

A. The role of interaction between temporal envelope and phase

Such a conclusion would be justified if the processing of the speech signal would result in the signal of Fig. 1(e) *at the output of the listener's cochlear filter*. This, however, is not the case as illustrated in Fig. 2. Figure 2(b) shows the output signal of a critical-band filter, identical to the one used in Fig. 1, for the input signal shown in Fig. 1(e). [For pictorial clarity, Fig. 1(e) is redrawn as Fig. 2(a).] Figure 2(c) shows its envelope. Clearly, these signals [of Figs. 2(b) and (c)] do not look at all like the smooth signals of Figs. 1(e) and (d), respectively. Indeed, they look very much like the original (nonsmoothed) signals of Figs. 1(b) and (c), respectively. [To highlight this point, a comparison of the envelope signals, Fig. 1(c) and Fig. 2(c), is shown in Fig. 2(d).] The implication of this finding is that the envelope-smoothed speech signal $\overline{s(t)}$ of Eq. (4) is inappropriate for the purpose of measuring the cutoff frequency of the auditory envelope detector. This is so because, when listening to $\overline{s(t)}$, the human observer is presented with rich envelope information, much beyond the nominal cutoff frequency of the smoothing filter.

The fact that filtering the smooth signal restores much of the nonsmoothed envelope appears to be somewhat unexpected. However, two theorems, one in the field of signal processing and one in the field of communications, provide analytic support to this finding. These theorems determine that: (1) For a bandlimited signal $s_i(t)=a_i(t)\cos\phi_i(t)$, the envelope signal $a_i(t)$ and the phase signal $\phi_i(t)$ are related (e.g., Voelcker, 1966), and (2) If $\phi(t)$ is a bandlimited signal, and if $\cos\phi(t)$ is the input to a bandpass filter [note that the envelope of the input signal is a constant, i.e., $a_i(t)=1$], then the filter's output has an envelope that is related to $\phi(t)$ (e.g., Rice, 1973). A corollary to these theorems is that if we pass the envelope-smoothed signal $\overline{s_i(t)}=\tilde{a}_i(t)\cos\phi_i(t)$ through a bandpass filter, the bandwidth of the output envelope is larger than the bandwidth of $\tilde{a}_i(t)$ [where the extra information is regenerated from $\phi_i(t)$]. If the bandpass filter represents a cochlear filter, the bandwidth of the temporal envelope information available to the listener is greater than the nominal smoothing cutoff frequency, B!

One clarification is noteworthy. The envelope signal of Fig. 2(c) (representing the envelope at the listener's cochlear output) exhibits both pitch modulations and articulatory modulations. Recall that the articulatory modulations (the main carrier of speech intelligibility) of the input envelope signal were low-pass filtered to B (e.g., 16 Hz). A question arises whether the envelope signal shown in Fig. 2(c) is mainly composed of pitch modulations (i.e., a secondary carrier of speech intelligibility), while the articulatory modulations are bandlimited to B, as intended. To answer this question, recall that the *phase information* of the input signal is unsmoothed, comprising the unsmoothed articulatory modulations and the unsmoothed pitch modulations. It is impossible to use the analytic expressions derived by Rice to isolate the response of the filter to the articulatory modulations from its response to the pitch modulations. (This is so be-

cause of the complexity of these expressions.) Suffice it to say that even though the articulatory information of the input envelope signal was appropriately smoothed (e.g., to 16 Hz), it still exists in its entirety in the input phase signal and, therefore, will be regenerated as part of the envelope signal at the filter's output.

III. DICHOTIC SYNTHESIS WITH INTERLEAVING CHANNELS

For a direct psychophysical measurement of the cutoff frequency of the auditory envelope detector, we have to ensure bandlimited envelope information at the listener's AN. This requirement can be elaborated as follows. Recall that information is conveyed to the AN by a large number of highly overlapped cochlear filters, with a density and location determined by the discrete distribution of the IHCs along the continuous cochlear partition. When the source signal $s(t)$ is passed through this cochlear filter bank, the resulting envelopes change gradually with CF as we move across the filter bank. The signal-processing method we seek should enable us to generate a signal that, when passed through the cochlear filter bank, will result in smoothed envelopes that are the envelopes generated by the source signal $s(t)$, low-pass filtered to the prescribed cutoff frequency B. This requirement, termed "the globally smoothed cochlear envelopes criterion," is formulated in Sec. III A.

In Sec. III B we consider a signal-processing technique based on diotic[6] speech synthesis, using pure cosine carriers. We shall demonstrate that this technique indeed generates smoothed envelopes at the output of the listener's cochlea, but only at the locations that correspond to the frequencies of the cosine carriers. At all other locations, distortions are generated that are perceptually noticeable. In Sec. III C we suggest a signal-processing technique designed to circumvent this problem. The technique is based upon dichotic speech synthesis with interleaving smoothed critical-band envelopes, and is based on the assumption that when the two streams are presented to the left and the right ears, the auditory system produces a single fused image (e.g., Durlach and Colburn, 1978). By using this procedure, perceivable distortions are greatly reduced.

Finally, we note that the present study is limited to measuring the cutoff frequency of the auditory envelope detectors only at the high CF region (i.e., frequencies above 1500 Hz). As mentioned before, ascending information at this frequency range is conveyed mainly via the temporal envelope of the cochlear signals (while the carrier information is lost). The lower frequency range (i.e., below 1500 Hz) was not addressed here since we lack understanding of the post-AN mechanisms that are active at the low CFs (and are sensitive to synchrony).

A. The globally smoothed cochlear envelopes criterion

Let $s(t)$ be processed by a filter bank consisting of the cochlear-shape filters H_1, H_2, and H_x (realized, for example, as gammatone filters, Slaney, 1993), where H_1 and H_2 are one critical band apart, and H_x is located in between H_1 and H_2 (Fig. 3). Let the envelope signals of Fig. 3, $a_1(t)$, $a_2(t)$,

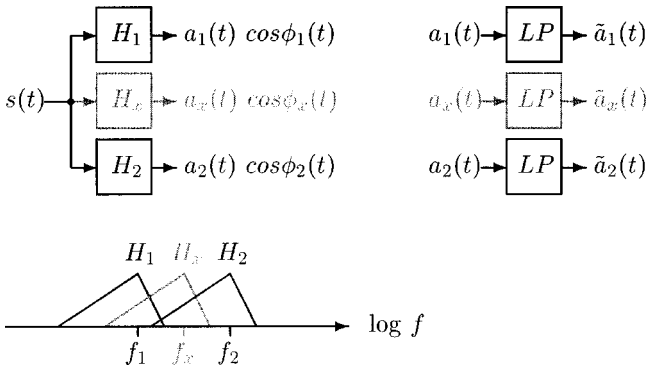

FIG. 3. Passing $s(t)$ through cochlear-shape filters H_1, H_2, and H_x. The spacing between H_1 and H_2 is one critical band. H_x represents one of the many overlapping cochlear filters located in between H_1 and H_2. The envelope signals $a_i(t)$ are temporally smoothed to $\tilde{a}_i(t)$, using a low-pass filter.

and $a_x(t)$, be temporally smoothed to $\tilde{a}_1(t)$, $\tilde{a}_2(t)$, and $\tilde{a}_x(t)$, respectively, and let

$$\tilde{s}(t)=F(\tilde{a}_1(t),\tilde{a}_2(t)), \tag{5}$$

where $F(\cdot,\cdot)$ stands for the desired signal-processing method. Let this $\tilde{s}(t)$ be fed to the filter bank of Fig. 3, as shown in Fig. 4. The resulting output signals, $b_i(t)\cos\psi_i(t)$, $i=1,2,x$, have envelope signals $b_1(t)$, $b_2(t)$, and $b_x(t)$ [and carrier signals $\cos\psi_1(t)$, $\cos\psi_2(t)$, and $\cos\psi_x(t)$]. For filters located at the high-frequency range (say, above 1500 Hz), the desired signal-processing method $F(\cdot,\cdot)$ should be designed to produce $\tilde{s}(t)$ such that

$$b_i(t)=\tilde{a}_i(t), \quad i=1,2,x. \tag{6}$$

Note that the properties of the signal carriers $\cos\psi_i(t)$ are being ignored since, at this frequency range, they are considered irrelevant due to the inability of the inner hair cell to follow the carrier information.

B. Diotic synthesis with pure cosine carriers

Reiterating Eqs. (1) and (2), let

$$s_i(t)=s(t)*h_i(t)=a_i(t)\cos\phi_i(t), \tag{7}$$

where $s(t)$ is the input signal, $h_i(t)$ is the impulse response of a gammatone filter centered at frequency f_i (above 1500 Hz), the operator $*$ represents convolution, and $a_i(t)$ and $\cos\phi_i(t)$ are, respectively, the envelope and the carrier of the filtered signal $s_i(t)$. Motivated by the observation that neural firings of AN fibers originating at this frequency range

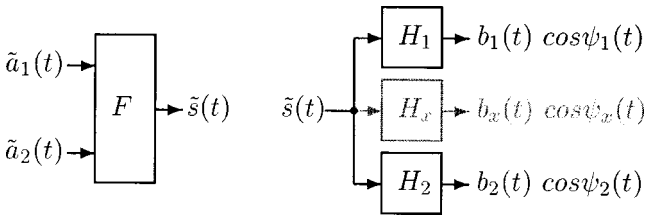

FIG. 4. Passing $\tilde{s}(t)$ through H_1, H_2, and H_x of Fig. 3. The desired signal processing method $F(\cdot,\cdot)$ should be designed to produce $\tilde{s}(t)$, which satisfies Eq. (6).

mainly transmit the envelope information $a_i(t)$, let us consider the signal

$$\hat{s}_i(t) = a_i(t)\cos 2\pi f_i t = a_i(t)\cos \omega_i t, \tag{8}$$

that is, $s_i(t)$, with the original carrier $\cos \phi_i(t)$ of Eq. (7) replaced by a cosine carrier $\cos \omega_i t$. Let $a_i(t)$ be low-pass filtered to $\tilde{a}_i(t)$, and let

$$\tilde{s}_i(t) = \tilde{a}_i(t)\cos \omega_i t. \tag{9}$$

Note that $\tilde{s}_i(t)$ is a bandlimited signal centered at frequency f_i. If $\tilde{s}_i(t)$ is presented to the listener's ear, the resulting envelope signal at the place along the cochlear partition that corresponds to frequency f_i will be the smoothed envelope $\tilde{a}_i(t)$. One possible signal-processing strategy could, therefore, be to generate a signal

$$\tilde{s}(t) = s_{\text{baseband}}(t) + \sum_{i=1}^{N} \tilde{a}_i(t)\cos \omega_i t, \tag{10}$$

where $s_{\text{baseband}}(t)$ represents the low-frequency range (i.e., below 1500 Hz), and $\tilde{a}_i(t)$, $i=1,\ldots,N$ are the smoothed-envelope signals of N gammatone filters equally spaced along the critical-band scale, with a spacing of one critical band, above 1500 Hz.

Let $\tilde{s}(t)$ of Eq. (10) be presented diotically to the listener's ear. The envelope at the output of the listener's cochlear filter located at frequency f_i is (ideally) $\tilde{a}_i(t)$, for each i, $i=1,\ldots,N$. However, the output of a cochlear filter located in between two successive cosine carrier frequencies f_i and f_{i+1} will reflect "beating" of the two modulated cosine carrier signals passing through the filter. This will result in a perceptually noticeable distortion. [Using the terminology of Sec. III A, if $F(\cdot,\cdot)$ is the diotic synthesis technique, i.e., $\tilde{s}(t) = \tilde{a}_1(t)\cos \omega_1 t + \tilde{a}_2(t)\cos \omega_2 t$, then $b_1(t) \cong \tilde{a}_1(t)$ and $b_2(t) \cong \tilde{a}_2(t)$. However, $b_x(t) \neq \tilde{a}_x(t)$, and such will be the case (to a different degree of dissimilarity) for every filter H_x located in between filters H_1 and H_2.]

C. Dichotic synthesis with interleaving critical-band envelopes

1. Principle

To reduce the amount of distortion due to beating, a dichotic synthesis with interleaving critical-band envelopes is proposed. As we shall see, this synthesis procedure is not perfect [i.e., it produces synthetic speech which does not satisfy Eq. (6) in a perfect way]. However, it allows us to circumvent the difficulties encountered in the diotic synthesis procedure and significantly reduce distortions.

Let $\tilde{s}_{\text{odd}}(t)$ and $\tilde{s}_{\text{even}}(t)$ be the summation of the odd components and even components of $\tilde{s}(t)$ of Eq. (10), respectively, i.e.,

$$\tilde{s}_{\text{odd}}(t) = s_{\text{baseband}}(t) + \sum_{i \in \text{odd}} \tilde{a}_i(t)\cos \omega_i t, \tag{11}$$

$$\tilde{s}_{\text{even}}(t) = s_{\text{baseband}}(t) + \sum_{i \in \text{even}} \tilde{a}_i(t)\cos \omega_i t. \tag{12}$$

The distance between two successive cosine carriers in each of these signals is two critical bands, resulting in a reduction of distortion due to carrier beating. When $\tilde{s}_{\text{odd}}(t)$ and $\tilde{s}_{\text{even}}(t)$ are presented to the left and the right ears, respectively, the auditory system produces a single fused image. In Secs. III D and III E, we shall examine the extent to which the fused auditory image achieves the property of Eq. (6).

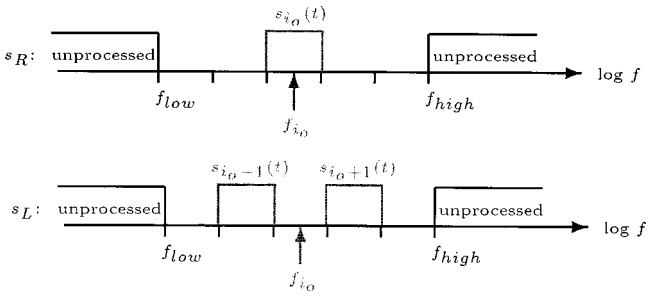

FIG. 5. Dichotic synthesis with interleaving channels. For pictorial clarity, the critical-band spectra are sketched as "flat" spectra.

2. Stimuli for the psychophysical experiments

Let us assume that, for a given input signal $s(t)$, we want to generate a fused auditory image with a range of smoothed-envelope representations that are one critical-band wide and that are centered at frequency f_{i_o}. To achieve this goal, we generate two signals, $\tilde{s}_R(t)$ and $\tilde{s}_L(t)$, as sketched in Fig. 5. More specifically, let the original signal $s(t)$ be divided into three regions: (1) the "low-frequency range," up to frequency f_{low}, denoted as $s_{\text{low}}(t)$; (2) the "high-frequency range," from frequency f_{high}, denoted as $s_{\text{high}}(t)$; and (3) the "middle-frequency range," five successive critical bands wide, located in between frequencies f_{low} and f_{high} and centered at the "target" frequency f_{i_o}. The critical-band signals are $s_i(t) = s(t) * h_i(t) = a_i(t)\cos \phi_i(t)$, where $h_i(t)$ is a gammatone filter centered at frequency f_i, $i = i_o - 2, i_o - 1, i_o, i_o + 1, i_o + 2$. Note that in Figs. 5 and 6 these critical-band spectra are sketched as "flat" spectra, for pictorial clarity.

We define $s_R(t)$ and $s_L(t)$ as

$$s_R(t) = s_{\text{low}}(t) + s_{i_o}(t) + s_{\text{high}}(t), \tag{13}$$

$$s_L(t) = s_{\text{low}}(t) + s_{i_o-1}(t) + s_{i_o+1}(t) + s_{\text{high}}(t). \tag{14}$$

Thus, $s_R(t)$ and $s_L(t)$ are obtained by adding the unprocessed outputs of the filters as illustrated in Fig. 5. Similarly,

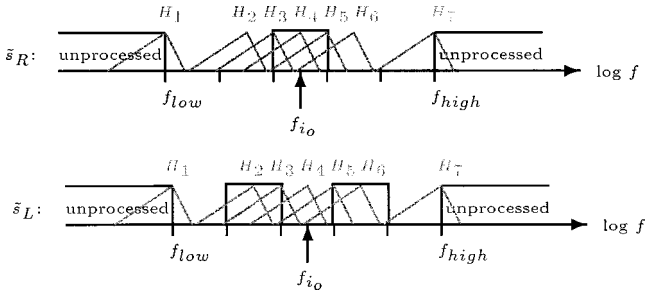

FIG. 6. Overlapping cochlear filters (in gray) superimposed over the spectral representation of $\tilde{s}_R(t)$ (top) and $\tilde{s}_L(t)$ (bottom). For pictorial clarity, the critical-band spectra are sketched as "flat" spectra.

the right- and the left smoothed-envelope signals are defined as

$$\tilde{s}_R(t) = s_{\text{low}}(t) + \tilde{a}_{i_o}(t)\cos\omega_{i_o}t + s_{\text{high}}(t), \tag{15}$$

$$\tilde{s}_L(t) = s_{\text{low}}(t) + \tilde{a}_{i_o-1}(t)\cos\omega_{i_o-1}t + \tilde{a}_{i_o+1}(t)\cos\omega_{i_o+1}t + s_{\text{high}}(t), \tag{16}$$

where $\tilde{a}_i(t)$, $i = i_o-1, i_o, i_o+1$, are the smoothed envelopes of the critical-band signals, and f_i, $i = i_o-1, i_o, i_o+1$, are the center frequencies of the critical bands in the middle frequency range (the gray-colored bands in Fig. 5), respectively. Compared to diotic synthesis, the distance between two successive occupied frequency bands in each of these signals is at least one critical band, resulting in a reduction of distortion due to carrier beating. At CF$=f_{i_o}$ and its one-critical-band neighborhood, the resulting fused auditory image contains smooth-envelope information in accordance with the prescribed bandwidth. This will be demonstrated in the remainder of the section.

D. Properties of the simulated cochlear signals

Figure 6 illustrates the filtering of the signals $\tilde{s}_R(t)$ of Eq. (15) (Fig. 6, top) and $\tilde{s}_L(t)$ of Eq. (16) (Fig. 6, bottom) by a simulated cochlea. In both figures, a sketch of seven (overlapping) cochlear filters is superimposed (in gray) over the spectral description of the signals.

Figure 6, top, illustrates the processing of $\tilde{s}_R(t)$ by the filters. All cochlear filters located to the left of filter H_1 (i.e., filters with lower CFs), and all the filters located to the right of filter H_7 (i.e., filters with higher CFs) will produce envelope signals with unsmoothed temporal structure. Filters H_2 to H_6 will produce temporally smoothed envelopes which are merely filtered versions of $\tilde{a}_{i_o}(t)$, with the response of H_4 being the strongest [and the most similar to $\tilde{a}_{i_o}(t)$]. The responses of filters H_2 and H_6 are negligible, since they are located at the energy gaps of the input signal. The amount of distortion due to beating is negligible since, for any CF, only one occupied frequency band is passing through the corresponding cochlear filter. (This is due to the wide gap, two critical-bands wide, between any adjacent occupied channels.)

Figure 6, bottom, illustrates the processing of $\tilde{s}_L(t)$ by the filters H_1 to H_7 of Fig. 6, top. Since $\tilde{s}_R(t)$ and $\tilde{s}_L(t)$ are identical for $f < f_{\text{low}}$ and for $f > f_{\text{high}}$, so is the response of all cochlear filters located in these frequency ranges. However, the response of cochlear filters in the midfrequency range is different. In contrast to their response to $\tilde{s}_R(t)$, the response of filter H_4 to $\tilde{s}_L(t)$ is the weakest while the envelope signals at the outputs of H_2 and H_6 are the strongest, similar in shape to $\tilde{a}_{i_o-1}(t)$ and $\tilde{a}_{i_o+1}(t)$, respectively [see Fig. 6, bottom, and Eq. (16)]. Also, compared to Fig. 6, top, the gap between adjacent occupied frequency bands is only one critical-band wide, resulting in some distortion due to beating.

Figure 7 shows simulated IHC response at 20 successive CFs to a 70-ms-long segment of the vowel /U/, cut from diphone /m_U/, starting at the transition point of /m/ into /U/. The top section shows the response to $s_R(t)$ and $\tilde{s}_R(t)$; bottom section is for $s_L(t)$ and $\tilde{s}_L(t)$. The channels' CFs (indicated in the upper-left corner of each panel) are equally spaced along the critical-band scale with a spacing of one-fourth critical band, from $f_{\text{low}} = 1722$ Hz to $f_{\text{high}} = 2958$ Hz, i.e., every column (four successive channels) covers one critical band. Each cochlear channel is realized as a gammatone filter, followed by an IHC model.[7] In this example, the target frequency is $f_{i_o} = 2227$ Hz, and the parameters of the dichotic synthesizer are set to $f_{\text{low}} = 1722$ Hz, $f_{\text{high}} = 2958$ Hz, $f_{i_o-1} = 1988$ Hz, $f_{i_o} = 2227$ Hz, and $f_{i_o+1} = 2494$ Hz [see Fig. 5 and Eqs. (13)–(16)]. Each panel in the figure shows the output of the IHC model to the following input signals: Black lines show the output for the signals with unprocessed critical bands, $s_R(t)$ of Eq. (13) (top) and $s_L(t)$ of Eq. (14) (bottom); gray lines show the output for the signals with the envelope-smoothed critical bands, $\tilde{s}_R(t)$ of Eq. (15) (top) and $\tilde{s}_L(t)$ of Eq. (16) (bottom), where a smoothed envelope $\tilde{a}_i(t)$ is the envelope $a_i(t)$, low-pass filtered to 64 Hz. The panel labeled 1722 Hz represents channel H_1 of Fig. 6, panel 2958 Hz represents channel H_7, and panels 1988, 2227, and 2494 Hz represent channels H_2, H_4, and H_6, respectively.

The response shown in Fig. 7 is in accordance with the observations made in Fig. 6. As we see in the top section, the IHCs' response to $s_R(t)$ of Eq. (13) (i.e., black lines) is rich in temporal structure. The overall energy changes with CF, with a stronger response by filters located in occupied frequency regions. The IHCs' response to $\tilde{s}_R(t)$ of Eq. (15) (superimposed gray lines) is rich in temporal structure for CFs below f_{low} and for CFs above f_{high}. However, the response gradually changes with CF, becoming temporally smoothed (and similar to the envelope signal $\tilde{a}_{i_o}(t)$). The output energy peaks at CF$=f_{i_o}$, then slowly decays for filters located at the frequency gap of Fig. 6, top. Note that distortion due to beating is negligible. Analogous behavior is illustrated in the bottom section of Fig. 7. Here, minimum response is produced at CF$=f_{i_o}$ while maximum response is produced at CF values near f_{i_o-1} and f_{i_o+1}. Note also the distortion produced by beating which, for this particular vowel, is most noticeable at CFs in the left energy gap of Fig. 6, bottom (i.e., CF$\approx$1900 Hz).

E. Properties of the fused auditory image

1. Integration of left and right channels

During listening, the subject's response is based upon the information contained in the fused auditory image. The ''low-frequency range'' and the ''high-frequency range'' [$s_{\text{low}}(t)$ and $s_{\text{high}}(t)$ of Eqs. (13)–(16)] are presented to the listener diotically, creating an auditory image with conventional properties. However, the midfrequency range is presented dichotically, with interleaving critical bands. This raises a question about the properties of the resulting fused (internal) auditory image. It is reasonable to assume that information from left and right ears originating at similar CFs will be integrated to generate a fused image. The use of

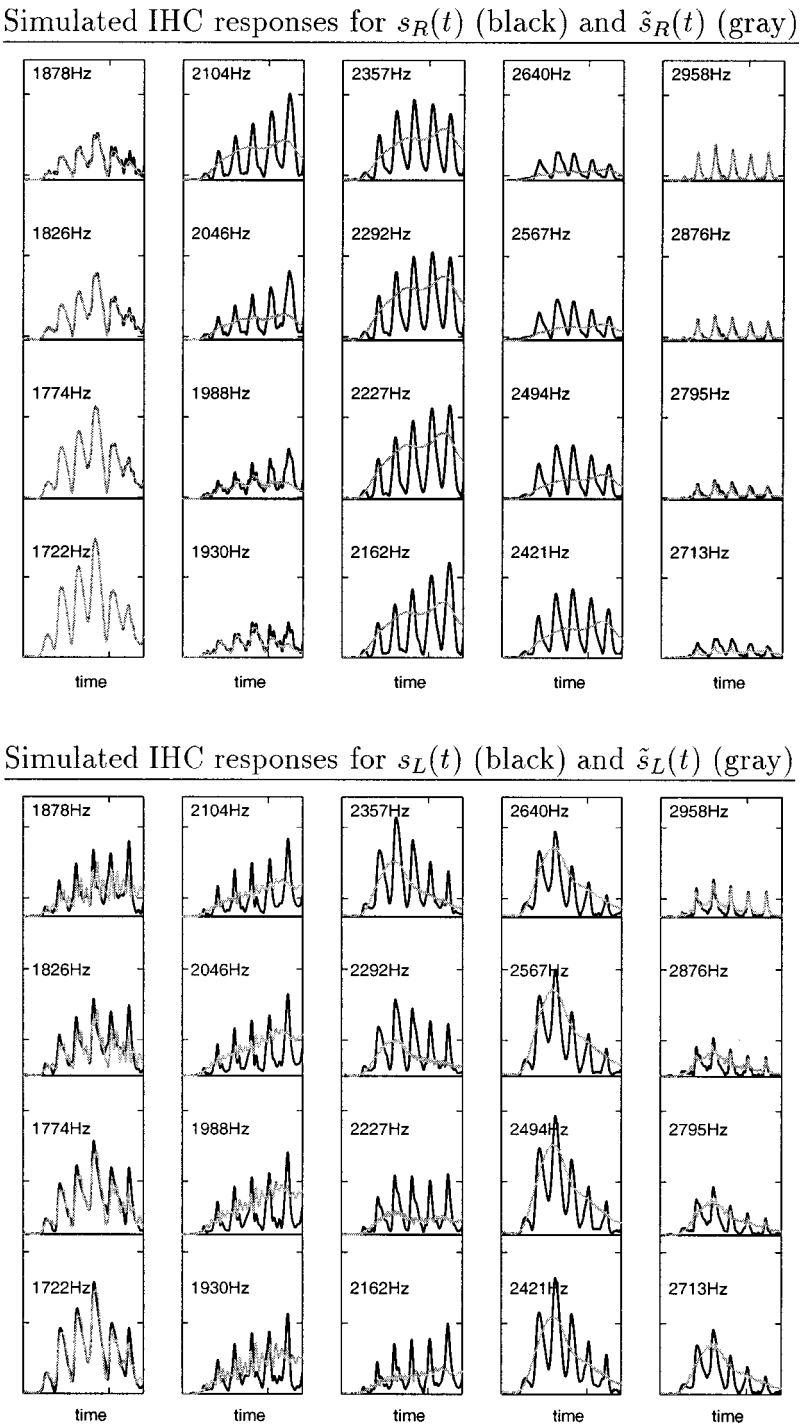

FIG. 7. Simulated IHC response at 20 successive CFs to a dichotically synthesized speech. The figure shows the response to a 70-ms-long segment of the vowel /U/, cut from diphone /m_U/, starting at the transition point of /m/ into /U/. The channels are located one-fourth of one critical band apart, with every column (four successive channels) covers one critical band. Black lines show the output for the input signals with unprocessed critical bands, $s_R(t)$ of Eq. (13) (top) and $s_L(t)$ of Eq. (14) (bottom). Gray lines show the output for the input signals with envelope-smoothed critical bands, $\tilde{s}_R(t)$ of Eq. (15) (top) and $\tilde{s}_L(t)$ of Eq. (16) (bottom), where the envelopes are low-pass filtered to 64 Hz. See the text for details.

dichotic stimulus with interleaving critical bands ensures that, at any CF, when one ear is stimulated, the opposite ear is not. Nevertheless (as illustrated in Figs. 6 and 7), cochlear channels located at the energy gaps of the input signal produce a nonzero output. The proposed synthesis procedure, therefore, only ensures that, at any CF, information from the stimulated ear is stronger than the information from the opposite ear. In Fig. 7, at any given CF, the panel from the top section (say, right ear) is assumed to be combined with the corresponding panel from the bottom section (left ear). In particular, for CFs near f_{i_o}, the signals from the stimulated ear are stronger than the signals from the other ear.

2. Coarse variation of IHC responses with CF

The proposed dichotic synthesis technique produces an inherent distortion due to undersampling (in CF) of the IHC response. Recall that information is conveyed to the AN by a large number of highly overlapped cochlear channels, with a density and location determined by the discrete distribution of the IHCs along the continuous cochlear partition. When a signal with unprocessed critical bands [e.g., $s_R(t)$ or $s_L(t)$] is passed through this cochlear filter bank, the resulting IHC responses change gradually with CF. Passing a signal with envelope-smoothed critical bands [$\tilde{s}_R(t)$ or $\tilde{s}_L(t)$] through

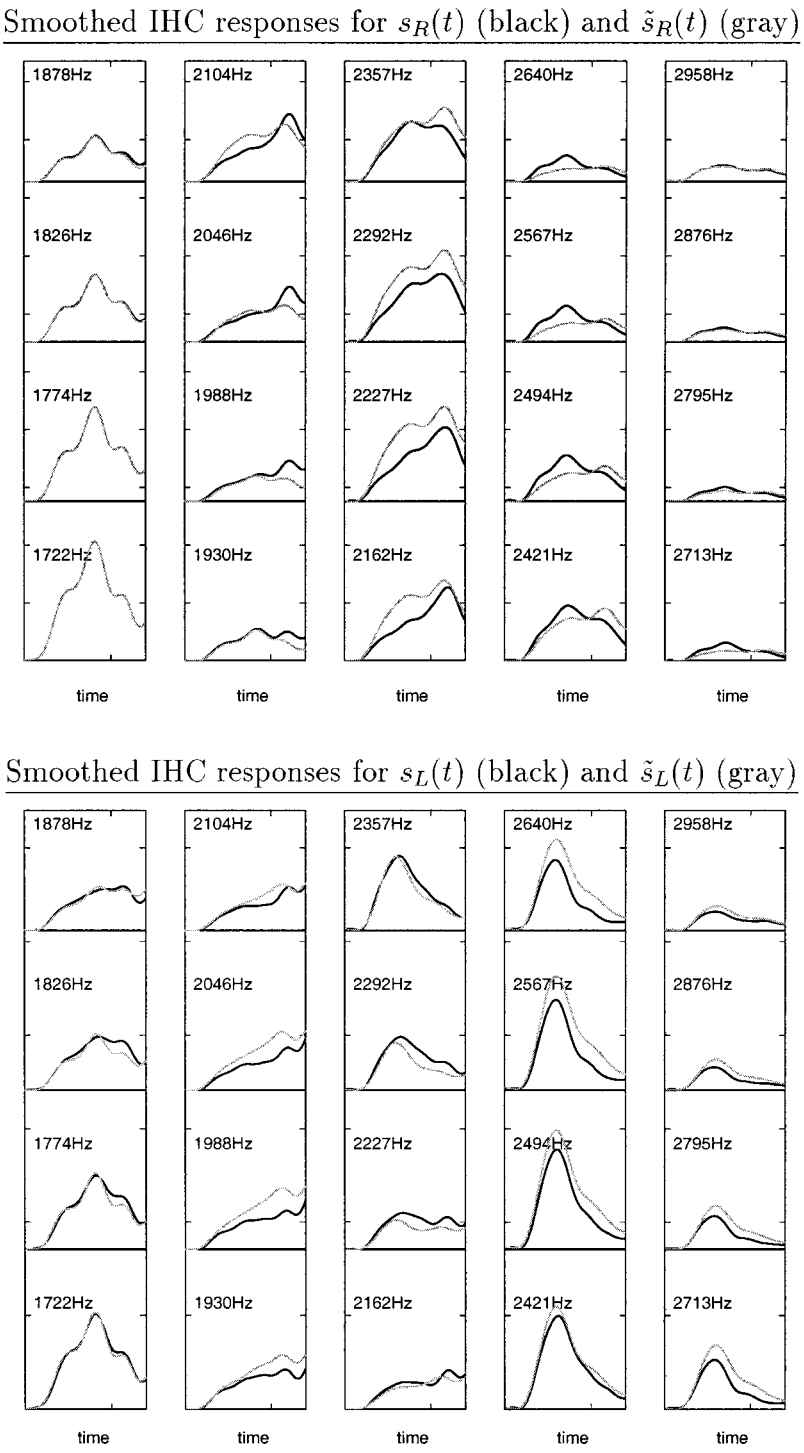

FIG. 8. Illustrating the coarse variation of IHC response with CF, due to the undersampling of the auditory channels (an inherent property of the dichotic synthesis technique). The figure shows the simulated IHC response of Fig. 7 smoothed to 64 Hz, for the input signals with unprocessed critical bands (black), and for the input signals with the envelope-smoothed critical bands (gray). Note the richer variation with CF for the unprocessed input signals (black). Notations are same as in Fig. 7. See the text for details.

the same filter bank will result in much coarser change. This is so because, in synthesizing $\tilde{s}_R(t)$ and $\tilde{s}_L(t)$, pure cosine carriers are used to place a few smoothed-envelope samples (sampled with a frequency resolution of two critical bands) at the appropriate locations along the basilar membrane. This is illustrated in Fig. 8, which is similar to Fig. 7 with the exception that, at each panel, the signals are the corresponding signals of Fig. 7 low-pass filtered to 64 Hz. The figure shows the change in envelope as a function of CF for the input signals with unprocessed critical bands (black) and for the input signals with envelope-smoothed critical bands (gray). With $\tilde{s}_R(t)$ as an input (top section), all overlapping cochlear channels located in the center column are fed with the same amplitude-modulated (AM) signal $\tilde{a}_{i_o}(t)\cos\omega_{i_o}t$, with $f_{i_o}=2227$ Hz. Therefore, the simulated IHC responses of these channels (in gray) are merely filtered versions of $\tilde{a}_{i_o}(t)$, and their similarity to $\tilde{a}_{i_o}(t)$ depends on the frequency response of the corresponding gammatone filter. In contrast, with $s_R(t)$ as an input, the variation in the simulated IHC responses of the corresponding channels (in black) is richer, reflecting the detailed information of the signal with the unprocessed critical bands. Analogous behavior will occur for $s_L(t)$ and $\tilde{s}_L(t)$ as inputs (bottom section). Note that

the coarse variation of the IHC responses with CF limits the extent to which the fused auditory image achieves the property of Eq. (6).

3. Sparse IHC responses for excessive envelope smoothing

Due to the undersampling of the IHC responses (Sec. III E 2) the coarse representation with CF becomes sparse for an excessive envelope smoothing, causing a significant perceivable distortion. If the bandwidth of $\tilde{a}_i(t)$ is B, the bandwidth of the AM signal $\tilde{a}_i(t)\cos\omega_i t$ is $2\times B$. Hence, for $\tilde{s}_{\text{odd}}(t)$ and $\tilde{s}_{\text{even}}(t)$ of Eqs. (11) and (12), each defined as a sum of AM signals for $f>1500$ Hz, the energy gap between two successive occupied frequency bands increases as B decreases. Consequently, more cochlear channels located in between successive cosine carriers will have a weak response, resulting in a sparse fused image. Illustratively, if $B\rightarrow 0$, the upper frequency band of $\tilde{s}_{\text{odd}}(t)$ and $\tilde{s}_{\text{even}}(t)$ becomes a sum of sinusoids. The perceived distortion sounds as an additive monotonic ''musical note.''

4. Spacing between successive cosine carriers

Recall that the dichotic synthesis technique was introduced to reduce perceivable distortions rising from the beating of two modulated cosine carriers passing through a cochlear filter located in between the carriers' frequencies. For the signals $\tilde{s}_{\text{odd}}(t)$ and $\tilde{s}_{\text{even}}(t)$ of Eqs. (11) and (12), the spacing between successive cosine carriers was set to be two critical bands wide. This choice was somewhat arbitrary. Obviously, the greater the spacing is, the smaller the beating-induced distortions are. However, increase in spacing will result in a coarser variation of IHC responses with CF (Sec. III E 2). Analogously, decreasing the spacing, e.g., to reduce sparse envelope representation for small values of B (Sec. III E 3), will reintroduce a perceptible amount of beating-induced distortions. This trade-off between beating-induced distortion and distortions due to sparse envelope representation is inevitable.

IV. DICHOTIC SYNTHESIS AND SPEECH QUALITY—EXPERIMENTS

In this section we use the dichotic synthesis technique to conduct two separate experiments in the context of preserving speech quality. In experiment I (described in Sec. IV B) we examine how speech quality is affected by replacing the carrier information of the critical-band signal by a cosine carrier [i.e., replacing $\cos\phi_{i_o}(t)$ by $\cos\omega_{i_o}t$], while keeping the envelope information untouched. In experiment II (Sec. IV C) we measure how speech quality deteriorates as the envelope bandwidth at the listener's cochlear output is gradually reduced.

A. Database, psychophysical procedure, subjects

The stimuli for the experiments were generated by implementing the dichotic synthesis technique [Eqs. (13)–(16)]. Twelve speech sentences were used, spoken by three female speakers and three male speakers (each speaker contributed two sentences). Since the experiments were conducted in the context of preserving speech quality, wideband speech signals were used, with a bandwidth of 7000 Hz. The speech intensity was set to 75 dB SPL. The stimuli are characterized by the center frequency of the middle frequency range [i.e., f_{i_o} of Eqs. (13)–(16)] and by the processing condition. We used five center frequencies, equally spaced on the critical-band scale and separated by (roughly) two critical bands (1600, 2000, 2500, 3200, and 4000 Hz). We used six processing conditions: one condition representing the signals with *unprocessed critical bands* [where the right and left signals are $s_R(t)$ and $s_L(t)$ of Eqs. (13) and (14), respectively], four conditions representing signals with *envelope-smoothed critical bands* [where the right and left signals are $\tilde{s}_R(t)$ and $\tilde{s}_L(t)$ of Eqs. (15) and (16), respectively], with envelope bandwidths of 512, 256, 128, and 64 Hz, and a *control condition*, termed the *null* condition, where the five successive critical bands centered at f_{i_o} are set to zero.[8]

In both experiments, we used the ABX psychophysical procedure. In this procedure, two sets of stimuli, the ''reference set'' and the ''test set,'' are defined. A stimulus in the reference set has a counterpart in the test set; both stimuli differ only by their processing condition. At each trial, a stimulus from the reference set and its counterpart from the test set are assigned to be the A stimulus and the B stimulus, at random. Then, the X stimulus is randomly chosen to be either the A or the B stimulus. The listener is presented with the A, B, and X stimuli (in this order), and must decide whether X is A or B. In our version, there is no ''repeat'' option. Note that if the listener makes his decisions at random (this may occur if the reference set and the test set are perceptually indistinguishable), the probability of correct decision is 50%.

Five subjects participated in each experiment (same subjects for both experiments). All subjects are well experienced in listening to high-quality audio signals (speech and music).

B. Experiment I—Carrier information

In this experiment we validate the hypothesis that at high CFs the auditory system is insensitive to the carrier information of the critical-band signals and that ascending auditory information in this frequency range is conveyed mainly via the temporal envelope of the cochlear signals. Towards this goal, we measure the probability of correct response in an ABX psychophysical procedure, using a reference set and a test set as defined in Table I. A stimulus in the reference set and its counterpart in the test set differ in the characteristics of the carrier information of the critical-band signals at the middle-frequency range (Fig. 5). As indicated in the middle column of Table I (processing condition), a reference stimulus is comprised of the signals $\tilde{s}_R(t)$ and $\tilde{s}_L(t)$ of Eqs. (15) and (16), respectively, with the envelopes low-pass filtered to 512 Hz (i.e., zero carrier information but full envelope information[9]). The corresponding test stimulus is composed of the signals $s_R(t)$ and $s_L(t)$ of Eqs. (13) and (14), respectively (i.e., containing the full carrier and the full envelope information).

TABLE I. Stimuli for experiment I (Sec. IV B) and experiment II (Sec. IV C). Each entry denoted by * contains 12 sentences, spoken by three female and three male speakers (two sentences each).

	Processing condition		Center frequency f_{i_o}, in Hz				
	Carrier	Envelope bandwidth	1600	2000	2500	3200	4000
Reference	$\cos \omega_{i_o} t$	512 Hz	*	*	*	*	*
Test—Experiment I	$\cos \phi(t)$	full	*	*	*	*	*
	$\cos \omega_{i_o} t$	256 Hz	...	...	*	*	*
Test—Experiment II	$\cos \omega_{i_o} t$	128 Hz	*	*	*	*	*
	$\cos \omega_{i_o} t$	64 Hz	*	*	...	...	...
Test—Control	null	null	*	*	*	*	*

C. Experiment II—Envelope bandwidth

In this experiment we measure the upper cutoff frequency of the auditory critical-band envelope detector, in terms of the minimal bandwidth of the critical-band envelope that ensures transparent speech quality. Towards this goal, we measure the probability of correct response in an ABX psychophysical procedure, using a reference set and a test set as defined in Table I. A reference stimulus and the corresponding test stimulus are composed of the signals $\tilde{s}_R(t)$ and $\tilde{s}_L(t)$ of Eqs. (15) and (16), respectively. They differ only in the bandwidth of the critical-band envelopes, with the bandwidth of a reference stimulus being 512 Hz. In the test set, only two smoothing conditions were used at each center frequency (to reduce the overall number of trials, and hence the experimental load on the subjects). For $f_{i_o} = 1600$ Hz and $f_{i_o} = 2000$ Hz, the envelope bandwidths were 64 and 128 Hz. (Note that the bandwidth of critical bands located at these center frequencies are 180 and 250 Hz, respectively.) For $f_{i_o} = 2500$ Hz, $f_{i_o} = 3200$ Hz, and $f_{i_o} = 4000$ Hz, the envelope bandwidths were 128 and 256 Hz (where the corresponding bandwidth of critical bands are 300, 360, and 440 Hz).

D. Results

In conducting the experiment, all test stimuli of experiment I, experiment II, and the control experiment were combined into one set ([5 center frequencies]×[4 test processing conditions]×[12 sentences]=240 sentences–see Table I). These sentences were randomly shuffled, then divided into four groups of 60 sentences each. The counterpart reference stimuli were arranged in the same order. Each subject participated in four sessions (a group of 60 sentences per session), lasting about 10 min each ([60 ABX trials]×[3 sentences]×[≈3 seconds]=≈600 seconds).

The results are presented in Fig. 9. Each panel represents performance at the center frequency specified at the upper-right corner of the panel. The bandwidth of a critical band[10] centered at that frequency is also indicated in parentheses. The abscissa of each panel indicates the processing condition of the test set stimuli. The entry $s_i(t)$ represents the condition with unprocessed critical bands (experiment I), the entries 256, 128, and 64 Hz represent the conditions with envelope-smoothed critical bands (experiment II), and the entry *null* represents the control experiment. (We chose to display all conditions in the same panel since a test set, in all experiments, is always contrasted with the same reference set—see Table I.) The ordinate is the probability of correct identification of the identity of the X stimuli (during the ABX procedure), in percent. The proportion of correct response for each subject was computed from 12 binary responses (one binary response for each sentence in the experiment). Each entry shows the mean and the standard deviation

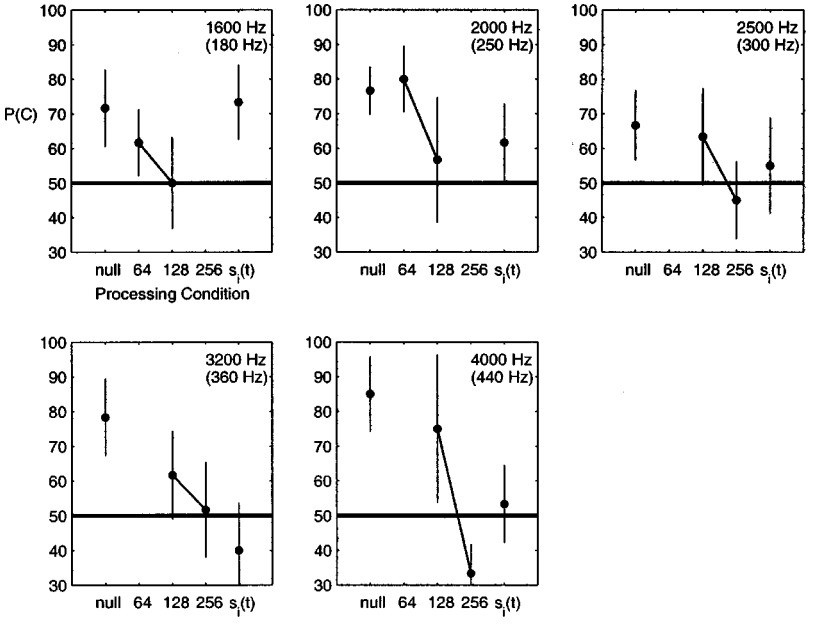

FIG. 9. Probability of correct response as a function of processing condition, with the center frequency as a parameter. Center frequencies are specified at the upper-right corner of the panel (the bandwidth of the corresponding critical bands is also indicated, in parentheses). The abscissa of each panel indicates the processing condition of the test set stimuli. The ordinate is the probability of correct identification of the identity of the X stimuli (during the ABX procedure), in percent. Each entry shows the mean percentage of correct response and the standard deviation among the five subjects. See the text for details.

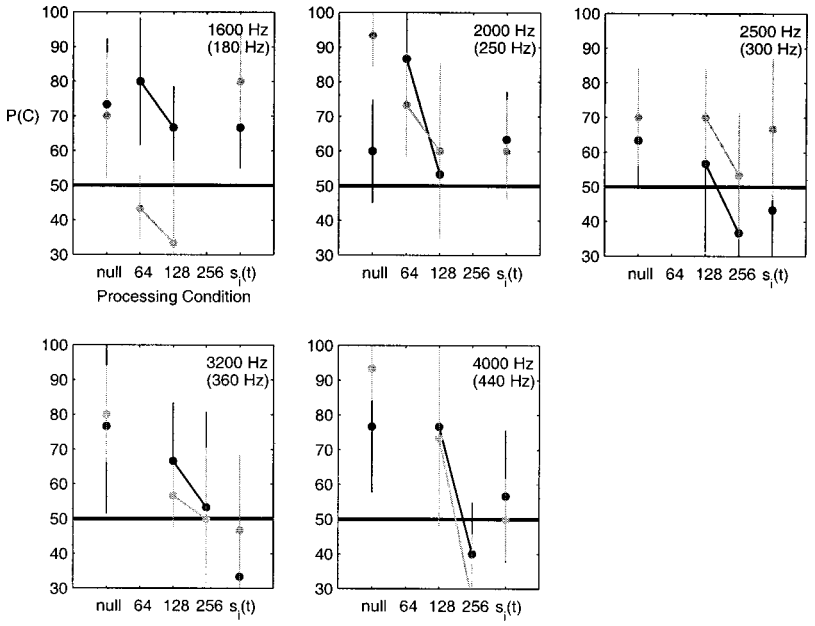

FIG. 10. Experimental results of Fig. 9, broken into two groups according to speaker gender, male speakers in black, female speakers in gray. Differences may be attributed to the interaction between the spectral contents of the stimulus (location of formants, pitch) and the center frequency under consideration.

of these five numbers. (A simple analysis of variance demonstrated that the interaction between subject and processing condition was not significant, so that it is legitimate to pool results from the five subjects.)

The control experiment (indicated *null* on the abscissa) confirms the assumption that a removal of a frequency band five-critical-bands wide results in a perceivable degradation in quality. This is so because for all center frequencies we considered, the mean probability of correct response is significantly above 50%.

For experiment I [indicated as $s_i(t)$ on the abscissa], the mean probability of correct response is about 50% for the higher center frequencies (i.e., 2500, 3200, and 4000 Hz). As the center frequency decreases, the mean probability of correct response increases (62% for $f_{i_o}=2000\,\mathrm{Hz}$, and 74% for $f_{i_o}=1600\,\mathrm{Hz}$). This result confirms the hypothesis that at high center frequencies (above $\approx 1800\,\mathrm{Hz}$) the auditory system is insensitive to the temporal details of the carrier information, and that the full carrier $\cos\phi(t)$ can be replaced with a cosine carrier $\cos\omega t$.

For experiment II (indicated as 64, 128, and 256 Hz on the abscissa), at higher center frequencies (i.e., 2500, 3200, and 4000 Hz) the mean probability of correct response is about 50% for an envelope bandwidth of 256 Hz.[11] For the other two center frequencies (1600 and 2000 Hz), a 50% mean probability of correct response is measured for an envelope bandwidth of 128 Hz. Note that these bandwidth values are considerably smaller than the bandwidth of the critical bands centered at the corresponding center frequencies (indicated in the upper-right corner, in parentheses), and are roughly one-half of one critical band.

Finally, Fig. 10 shows the experimental results of Fig. 9, broken into two groups according to speaker gender, male speakers in black, female speakers in gray. (Obviously, the number of observations per entry per subject is now only six.) The figure shows that at most center frequencies and for most processing conditions, performance is not affected much by the speaker gender. Differences may be attributed to the interaction between the spectral contents of the stimulus (location of formants, pitch) and the center frequency under consideration.

V. DICHOTIC SYNTHESIS AND SPEECH INTELLIGIBILITY

In Sec. IV, the dichotic synthesis technique was used to measure the cutoff frequencies of the auditory envelope detectors at threshold (i.e., the cutoff frequencies which maintain the quality of the original speech). A question arises whether the technique can also be used to measure the cutoff frequencies in the context of speech intelligibility, for speech signals that maintain some reasonable level of speech quality (say, above MOS level 3). In the following, it will be argued that speech stimuli produced by dichotic synthesis for intelligibility-related experiments are of poor quality, with MOS readings well below 3.

Suppose that we want to repeat the phoneme identification experiment reported by Drullman *et al.* (1994), by using a dichotically synthesized speech, with temporal envelopes that are low-pass filtered to a cutoff frequency B. Which values of B are reasonable for such an experiment? Expressing temporal envelope information in terms of the amplitude-modulation spectrum, two kinds of modulations may be considered as information carriers of speech intelligibility—the articulatory modulations and the pitch modulations. Of these, the pitch modulations convey only a limited amount of phonemic information (this is so because for speech signals, the salient mechanism for pitch perception is based on resolved harmonics at the lower frequency range[12]). The major carriers of phonemic information are, therefore, the articulatory modulations. [Indeed, the STI method is aimed at measuring these MTFs (Steeneken and Houtgast, 1980).] Hence, the B values for a phoneme identification experiment should be on the order of a few tens of Hz, determined by the mechanical properties of the articulators. Recall the properties of the speech signals generated by the dichotic synthesis technique

(Secs. III D and III E). For an appropriate spacing between successive cosine carriers (Sec. III E 4), and for B values of a few tens of Hz, the resulting speech stimuli generate fused auditory images that are too sparse (Sec. III E 3), and suffer severe degradation in speech quality (to MOS levels well below 3) due mainly to an overriding monotonic tonal accent. The speech signals produced by the dichotic synthesis technique are, therefore, inadequate for experiments intended to measure intelligibility-related Bs while maintaining fair quality levels. The appropriate signal-processing method is yet to be found.

VI. DISCUSSION

This study was motivated by the need to quantify the minimum amount of information, at the auditory-nerve level, that is necessary for maintaining human performance in tasks related to speech perception (e.g., threshold measurements for speech quality, phoneme classification for speech intelligibility). Such data are needed, for example, for a quantitative formulation of a perception-based distance measure between speech segments (e.g., Ghitza and Sondhi, 1997). The study was restricted to the frequency range above 1500 Hz, where the information conveyed by the auditory nerve is mainly the temporal envelopes of the critical-band signals. From the outset, it was assumed that these envelopes are processed by distinct, albeit unknown, auditory detectors characterized by their upper cutoff frequencies which, in turn, determine the perceptually relevant information of the envelope signals in terms of their effective bandwidth. The main contribution of this study is the establishment of a framework that allows the direct psychophysical measurement of this bandwidth, using speech signals as the test stimuli.

Measuring the perceptually relevant content of temporal envelopes was the subject of numerous studies, most of which were aimed at measuring the amplitude-modulations spectra using threshold-of-detection criteria. These studies (e.g., Viemeister, 1979; Dau *et al.*, 1997a, 1997b, 1999; Kohlrausch *et al.*, 2000) used nonspeech signals as test stimuli—mostly signals with a bandwidth of one critical band.[13] The present study extends the scope of previous studies by providing threshold measurements of the cochlear temporal envelope bandwidth (which may be regarded as the bandwidth of the amplitude-modulation spectrum) for speech signals, hence providing an estimate of the threshold bandwidth of a target auditory channel *while all other channels are active simultaneously*.

In order to conduct these experiments, a signal-processing framework had to be formulated that would be capable of producing speech signals with appropriate temporal envelope properties. As was shown in Sec. II, if the envelope of a critical-band signal is temporally smoothed while the instantaneous phase information remains untouched (e.g., Drullman *et al.*, 1994), the resulting synthetic speech signal evokes cochlear envelope signals that are not necessarily smoothed. This rather counterintuitive behavior (which is theoretically founded, as discussed in Sec. II A) suggests that a different criterion should be used for signal synthesis, such that the resulting speech signal will evoke temporal envelopes with a prescribed bandwidth *at the output of the listener's cochlea* (Sec. III A). Such a signal-processing technique is yet to be found. However, in Sec. III C, an approximate solution has been introduced based upon dichotic speech synthesis with interleaving smoothed critical-band envelopes.[14]

With this technique established, psychophysical measurements were conducted using high-quality, wideband, speech signals (bandwidth of 7 kHz) as the test stimuli. The measurements show that in order to maintain the quality of the original speech signal (1) there is no need to preserve the detailed timing information of the critical-band signal (experiment I, Sec. IV B); (2) the perceptually relevant information in this frequency range is mainly the temporal envelope of this signal, and (3) the minimum bandwidth of the temporal envelope of the critical-band signal is, roughly, one-half of one critical-band (experiment II, Sec. IV C). These results are in line with the widely accepted observation that at higher center frequencies, due to the physiological limitations of the inner hair cells to follow detailed timing information, neural firings at the auditory nerve mainly represent the temporal envelope information of the critical-band signal.

The data obtained here can be compared to previously published data only qualitatively, because of the marked difference in the underlying frameworks. As discussed by others (e.g., Dau *et al.*, 1999; Kohlrausch *et al.*, 2000), a reliable measurement of amplitude-modulation spectra can be obtained when the stimulus bandwidth is sufficiently narrower than the critical band of the target auditory channel. Previous studies that meet this requirement provide tight estimates of the envelope bandwidth at threshold, since the measurements for the target auditory channel are obtained with zero external stimulation of all other channels. In contrast, the measurements in the present study are taken with all auditory channel simultaneously active (the test stimuli are wideband speech signals), allowing interaction across channels (e.g., due to spread of masking). A qualitative comparison shows that estimates of envelope bandwidths obtained in this study are indeed lower than those published earlier. For example, for an auditory channel at CF of 3000 Hz, the estimate of the envelope bandwidth using a cosine carrier is roughly one critical band (i.e., about 350 Hz, Kohlrausch *et al.*, 2000). For speech stimuli at similar CFs, the envelope bandwidth is about 250 Hz (Fig. 9).

The methodology presented in this study provides a framework for the design of transparent coding systems[15] with a substantial information reduction (due to the use of fixed cosine carriers, modulated by smoothed critical-band envelopes, Ghitza and Kroon, 2000). One desirable property of this coding paradigm is that it performs equally well for speech, noisy speech, music signals, etc. This is so since the coding paradigm is based solely on the properties of the auditory system and does not assume any specific properties of the input source.

Finally, the dichotic synthesis technique is inadequate for the purpose of measuring the cutoff frequencies relevant to intelligibility of speech signals with fair quality levels (say, above MOS=3). Recall that the main information car-

riers of speech intelligibility are the articulatory modulations (e.g., Sec. V). Following a reasoning similar to the one used in measuring the cutoff frequencies at threshold, the appropriate speech stimuli should satisfy the criterion of generating temporal envelopes with *smoothed articulatory modulations* at the output of the listener's cochlea. In view of the discussion in Sec. II, a speech signal produced by smoothing the envelope signal alone (while keeping the original instantaneous phase information untouched) is inadequate because it will regenerate, at the cochlear output, most of the original envelope information, including the articulatory modulations and the pitch modulations. Indeed, the dichotic synthesis technique is capable of producing speech stimuli that generate cochlear temporal envelopes with smoothed articulatory modulations as desired. Alas, the quality of these signals is well below MOS=3 (Sec. V). We still lack the knowledge of how to synthesize speech stimuli which simultaneously satisfy both requirements (i.e., cochlear temporal envelopes with smoothed articulatory modulations *and* a prescribed level of speech quality).

ACKNOWLEDGMENTS

I wish to thank M. M. Sondhi and Y. Shoham for stimulating discussions throughout this work, and S. Colburn and two anonymous reviewers for reviewing earlier versions of the paper.

[1]Signals presented to left and right ears are different.

[2]The Mean-Opinion-Score, or MOS, is a test which is widely used to assess quality of speech coders. It is a subjective test that can be categorized as a rating procedure. Subjects are presented, once, with a speech sentence and are requested to score its quality using a scale of five grades. The grades (and their numerical aliases) are Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1). The MOS is the mean score, averaged over the database and the subjects.

[3]The *Hilbert envelope* and the *Hilbert instantaneous phase* are defined as follows: Let $z_i(t)$ be the *analytic signal* of $s_i(t)$, i.e., $z_i(t)=s_i(t)+j\hat{s}_i(t)$, where $\hat{s}_i(t)$ is the Hilbert transform of $s_i(t)$. We express $s_i(t)$ in terms of $z_i(t)$ as $s_i(t)=\Re(z_i(t))=a_i(t)\cos\phi_i(t)$, where $a_i(t)=\sqrt{s_i^2(t)+\hat{s}_i^2(t)}$ is the *envelope* of $s_i(t)$, and $\phi_i(t)=\arctan[\hat{s}_i(t)]/[s_i(t)]$ is the *instantaneous phase* of $s_i(t)$.

[4]CF, for *Characteristic Frequency*, indicates the place of origin of a nerve fiber along the basilar membrane in frequency units.

[5]Obviously, there is no distinct boundary between the low-CF and high-CF AN regions. Rather, the change in properties is gradual. Our working hypothesis is that the region of transition is around 1500 Hz.

[6]The same signal is presented to both ears.

[7]The IHC model is comprised of a half-wave rectifier, followed by a low-pass filter with the amplitude transfer function $\|H(f)\|=1/\sqrt{(1+(f/600)^2)(1+(f/3000)^2)}$, reflecting the synchrony roll-off in AN firings (e.g., Johnson, 1980).

[8]The null condition is for control purposes, to validate the assumption that a removal of a frequency band five-critical-bands wide indeed causes perceivable degradation in quality.

[9]Note that the bandwidth of the critical band centered at the highest center frequency considered in this experiment (i.e., $f_{i_o}=4000$ Hz) is about 440 Hz.

[10]We follow the ERB definition of a critical band, according to Moore and Glasberg (1983).

[11]Note that at center frequency of 4000 Hz the mean probability of correct response, for an envelope bandwidth of 256 Hz, is about 33%. This indicates that the two conditions are being distinguished somehow, but that the response is consistently incorrect.

[12]Recall the existence of two competing mechanisms for pitch perception. One is based upon resolved harmonics and, for speech signals in particular, operates at the lower frequency range (say, below 1500 Hz);, the other is based on temporal envelope periodicities and operates at the higher frequency range. When both mechanisms are active (as in the case of speech signals) the salient mechanism is the former one (e.g., Goldstein, 2000).

[13]The study by Drullman *et al.* (1994) belongs to a different category since it used a threshold criterion related to speech intelligibility (i.e., percent correct in a phoneme classification task). Obviously, Drullman *et al.* had to use speech signals as test stimuli.

[14]See Secs. III D and III E for a discussion on the properties and the shortcomings of this approximate solution.

[15]That is, at the receiving end, the system produces speech signals that are perceptually indistinguishable from the original speech.

Chi, T., Gao, Y., Guyton, M. C., Ru, P., and Shamma, S. (**1999**). "Spectro-temporal modulation transfer functions and speech intelligibility," J. Acoust. Soc. Am. **106**, 2719–2732.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997a**). "Modeling auditory processing of amplitude modulation. I. Detection and masking with narrow-band carriers," J. Acoust. Soc. Am. **102**, 2892–2905.

Dau, T., Kollmeier, B., and Kohlrausch, A. (**1997b**). "Modeling auditory processing of amplitude modulation. II. Spectral and temporal integration," J. Acoust. Soc. Am. **102**, 2906–2919.

Dau, T., Verhey, J., and Kohlrausch, A. (**1999**). "Intrinsic envelope fluctuations and modulation-detection thresholds for narrow-band noise carriers," J. Acoust. Soc. Am. **106**, 2752–2760.

Drullman, R., Festen, J. M., and Plomp, R. (**1994**). "Effect of temporal envelope smearing on speech reception," J. Acoust. Soc. Am. **95**, 1053–1064.

Durlach, I. N., and Colburn, S. (**1978**). "Binaural phenomena," in *Handbook of Perception, Volume IV: Hearing*, edited by E. C. Carterette and M. P. Friedman (Academic, New York), pp. 365–466.

Eddins, D. A., and Green, D. M. (**1995**). "Temporal integration and temporal resolution," in *Hearing*, edited by B. C. J. Moore (Academic, New York), pp. 207–242.

Flanagan, J. L. (**1980**). "Parametric coding of speech spectra," J. Acoust. Soc. Am. **68**, 412–430.

Ghitza, O., and Kroon, P. (**2000**). "Dichotic presentation of interleaving critical-band envelopes: An application to multi-descriptive coding," in *Proceedings of the IEEE Workshop on Speech Coding*, Delavan, Wisconsin (September), pp. 72–74.

Ghitza, O., and Sondhi, M. M. (**1997**). "On the perceptual distance between speech segments," J. Acoust. Soc. Am. **101**, 522–529.

Goldstein, J. L. (**2000**). "Pitch perception," in *Encyclopedia of Psychology*, edited by A. E. Kazdin (American Psychological Association, Washington, D.C.), Vol. VI, pp. 201–210.

Johnson, D. H. (**1980**). "The relationship between spike rate and synchrony in responses of auditory-nerve fibers to single tones," J. Acoust. Soc. Am. **68**, 1115–1122.

Kohlrausch, A., Fassel, R., and Dau, T., (**2000**). "The influence of carrier level and frequency on modulation and beat-detection thresholds for sinusoidal carriers," J. Acoust. Soc. Am. **108**, 723–734.

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

Rice, S. O. (**1973**). "Distortion produced by band limitation of an FM wave," Bell Syst. Tech. J. **52**, 605–626.

Slaney, M. (**1993**). "An efficient implementation of the Patterson-Holdsworth auditory filter bank," Technical Report 33, Apple Computer.

Steeneken, H. J. M., and Houtgast, T. (**1980**). "A physical method for measuring speech-transmission quality," J. Acoust. Soc. Am. **67**, 318–326.

Viemeister, N. F. (**1979**). "Temporal modulation transfer functions based upon modulation thresholds," J. Acoust. Soc. Am. **66**, 1364–1380.

Voelcker, H. B. (**1966**). "Towards a unified theory of modulation. I. Phase-envelope relationships," Proc. IEEE **54**(3), 340–354.