Robust estimation of speech in noisy backgrounds based on aspects of the auditory process[a)]

John H. L. Hansen and Srinivas Nandkumar[b)]
Robust Speech Processing Laboratory, Department of Electrical Engineering, Box 90291, Duke University, Durham, North Carolina 27708-0291

(Received 12 July 1994; revised 20 January 1995; accepted 6 February 1995)

A new approach to speech enhancement is proposed where constraints based on aspects of the auditory process augment an iterative enhancement framework. The basic enhancement framework is based on a previously developed dual-channel scenario using a two-step iterative Wiener filtering algorithm. Constraints across broad speech sections and over iterations are then experimentally developed on a novel auditory representation derived by transforming the speech magnitude spectrum. The spectral transformations are based on modeling aspects of the human auditory process which include critical band filtering, intensity-to-loudness conversion, and lateral inhibition. The auditory transformations and perceptual based constraints are shown to result in a new set of auditory constrained and enhanced linear prediction (ACE-LP) parameters. The ACE-LP based speech spectrum is then incorporated into the iterative Wiener filtering framework. The improvements due to auditory constraints are demonstrated in several areas. The proposed auditory representation is shown to result in improved spectral characterization in background noise. The auditory constrained iterative enhancement (ACE-II) algorithm is shown to result in improved quality over all sections of enhanced speech. Adaptation of auditory based constraints to changing spectral characteristics over broad classes of speech is another novel aspect of the proposed algorithm. The consistency of speech quality improvement for the ACE-II algorithm is illustrated over time and across all phonemes classified over a large set of phonetically balanced sentences from the TIMIT database. This study demonstrates the application of auditory based perceptual properties of a human listener to speech enhancement in noise, resulting in improved and consistent speech quality over all regions of speech.

PACS numbers: 43.72.Ew

INTRODUCTION

Enhancement of speech in the presence of additive continuous broadband noise remains a challenging task, especially in moderate to high noise levels (signal-to-noise ratios of -10 to 5 dB). Several reasons contribute to task complexity. First, broadband noise overlaps the speech signal both in time and frequency domains, and local noise characteristics cannot be determined exactly in either domain from the noisy speech signal. Second, speech is a highly varying signal both in terms of time and frequency characteristics, and the amount of speech distortion due to background noise varies across both time and frequency. In perceptual terms, the affect of broadband noise on different speech classes is not uniform. Most traditional enhancement algorithms are limited in terms of suppressing noise (improving SNR) and improving perceptual quality at the same time across all speech classes.

A speech enhancement algorithm can be termed successful if it accomplishes two tasks, (i) suppressing the perceivable background noise, and (ii) preserving or enhance perceived signal quality. Additionally, it is also desirable to improve intelligibility, and improve the performance of other speech processing systems (e.g., coding or recognition in noise). Traditional speech enhancement algorithms are based on optimizing mathematical criteria, which in general are not well correlated with speech perception. In general, these have not been as successful in preserving or improving quality in all regions of speech, especially transitional and unvoiced. In fact, many speech enhancement algorithms introduce additional speech distortion while suppressing noise, which can increase listener annoyance and cause preprocessing for speech coding and automatic recognition systems to be unreliable. Recently, enhancement algorithms which augment mathematical criteria with perceptual criteria have shown reasonable consistency in speech quality enhancement (Hansen and Clements, 1991; Nandkumar and Hansen, 1992, 1994; Cheng and O'Shaughnessy, 1991). Perceptual criteria can involve aspects of both speech production or speech audition. Use of perceptual criteria has also been shown to aid in reducing annoying artifacts or speech correlated distortion in the enhanced speech (Hansen and Clements, 1991; Nandkumar and Hansen, 1992, 1994, 1995; Peterson and Boll, 1981).

In an earlier study, we developed a dual-channel iterative speech enhancement framework augmented with constraints developed on the auditory based mel-cepstral param-

[a)]This work was supported in part by National Science Foundation Grant No. NSF-IRI-90-10536.

[b)]Dr. Nandkumar was with the Dept. of Electrical Engineering, Duke Univ. when this work was performed. He has since joined Martin Marietta Labs, Baltimore, MD.

 0001-4966/95/97(6)/3833/17/$6.00

eters (Nandkumar and Hansen, 1992) (ACE-I). However, critical band frequency analysis is only one aspect of the complex processing performed in the human auditory system. Researchers in auditory neurophysiology have proposed several auditory functional representations for speech, which incorporate peripheral and central auditory processing phenomena (Jennison *et al.*, 1991; Yang *et al.*, 1992). Other studies have also considered psychoacoustic models for speech recognition (Zwicker *et al.*, 1979). In order to further exploit the auditory based aspects of speech, several auditory modeling schemes have also been proposed as front-ends for automatic speech recognition (Cohen, 1989; Ghitza, 1986; Hermansky, 1990; Hunt and Lefebvre, 1986; Seneff, 1986). These techniques augment or replace traditional linear spectral parameters with parametric representations based on aspects of neurophysiological and psychoacoustic features. However, few techniques which incorporate such auditory models have been proposed for the purpose of speech enhancement in noisy backgrounds. The main reason for this is that most auditory models decompose the speech signal into complex auditory neural representations, from which it is difficult to resynthesize enhanced speech. Earlier studies, such as spectral subtraction in the loudness domain (Peterson and Boll, 1981) and adaptive filtering based on a degraded speech spectrum convolved with a frequency-dependent lateral inhibition function (Cheng and O'Shaughnessy, 1991), have attempted to augment traditional enhancement frameworks with an auditory processing property. In this study, a previously developed dual-channel Wiener filtering framework (Nandkumar and Hansen, 1992) is augmented with constraints based on several aspects of the auditory process. A new set of parameters termed auditory constrained and enhanced linear predictive (ACE-LP) parameters are derived based on aspects of auditory modeling which include critical band filtering, intensity-to-loudness transformation, and lateral inhibition. The enhanced speech is then reconstructed using ACE-LP parameters within the iterative Wiener filtering framework. The paper is organized as follows. In Sec. I, the iterative enhancement framework is described. A brief overview of auditory processing and applications to speech engineering problems is presented in Sec. II. The new auditory process constrained algorithm is developed and presented in Secs. III and IV. Results and conclusions follow in Secs. V and VI.

The tools used to assess speech quality improvement in this study are informal listening tests and objective quality measures. Objective speech quality measures used in this study are the spectral distortion based Itakura–Saito log-likelihood measure and the auditory perception based weighted spectral slope (Klatt) measure. For additive noise and speech coder distortions, these objective measures have been shown to be well correlated with perceived quality as measured by subjective tests such as the DAM (Quackenbush *et al.*, 1988). It should be emphasized that objective measures cannot, and should not, replace subjective testing for enhancement algorithm evaluation. In general, a balance should exist between subjective and objective methods for speech enhancement evaluation. While no formal listener evaluation was conducted, extensive informal comparisons were made throughout this study to confirm general directions for quality improvement, as seen using objective speech quality measures. It should also be noted that spectral based distortion measures, while useful for assessment of analysis-by-synthesis speech coders, may not be capable of representing the complete level of quality with respect to certain enhancement or vocoder artifacts and other distortions such as intelligibility loss and glottal source based distortions. It is strongly suggested that the use of objective speech quality measures as performance indicators for enhancement be confirmed with either informal or formal listener evaluations. Here, extensive informal listening tests are performed at each development step to ensure that such distortions are not introduced during each phase of enhancement.

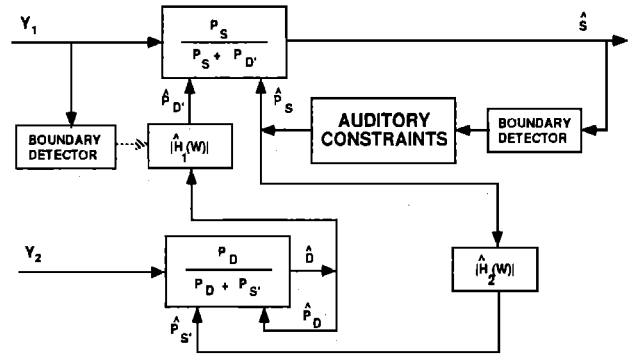


FIG. 1. The auditory constrained enhancement framework.

I. THE ITERATIVE ENHANCEMENT FRAMEWORK

The iterative enhancement framework developed in an earlier study (Nandkumar and Hansen, 1992, 1995) in a dual-microphone scenario is used in all simulations. It is noted that the auditory process based constraints presented in later sections can also be used in single-channel iterative filtering scenarios. In fact, the dual-channel framework is an extension of the single-channel constrained iterative framework developed by Hansen and Clements (1991). Unconstrained iterative Wiener filtering was originally considered by Lim and Oppenheim (1978) for an autoregressive (AR) speech model, and later generalized to an autoregressive-moving average (ARMA) speech model by Musicus (1979).

The dual-channel observations y_1 and y_2 can be expressed in the frequency domain as

$$Y_1(\omega)=S(\omega)+H_1(\omega)D(\omega)=S(\omega)+D'(\omega),$$
$$Y_2(\omega)=D(\omega)+H_2(\omega)S(\omega)=D(\omega)+S'(\omega). \quad (1)$$

Here, $H_1(\omega)$ and $H_2(\omega)$ represent the frequency-dependent correlation functions. The assumptions made in this scenario are that speech and noise are uncorrelated, $s(t)$ and $s'(t)$ are samples from a short-time stationary AR Gaussian process, and $d(t)$, $d'(t)$ are samples from a slowly varying Gaussian process. In this scenario, a two-step iterative dual-channel Wiener filtering solution can be derived as shown in Fig. 1. The enhancement solution is based on estimation of noise from the second channel in a MMSE sense, followed by estimating speech from the primary observation, using a par-

ticular interpretation of the iterative EM algorithm. The noise and speech spectra are updated at each iteration using the current estimates of speech and noise. The speech estimate using frequency-domain Wiener filtering at iteration i is given by

$$\{\hat{S}(\omega)\}_i = \left[\frac{\{P_S(\omega)\}_i}{\{P_S(\omega)\}_i + \{P_{D'}(\omega)\}_i}\right] Y_1(\omega). \quad (2)$$

The noise estimate at iteration i is also obtained by a Wiener filter operation, given by

$$\{\hat{D}(\omega)\}_i = \left[\frac{\{P_D(\omega)\}_i}{\{P_D(\omega)\}_i + \{P_{S'}(\omega)\}_i}\right] Y_2(\omega). \quad (3)$$

The speech spectra in the above equations are estimated at each iteration using the AR model assumption. The noise spectra are estimated using FFT based magnitude transforms. Estimates of the magnitude spectra of $H_1(\omega)$ and $H_2(\omega)$ obtained from noise-only and speech-only sections aid in transforming the speech and noise spectra between the primary and reference channels. At the first iteration, speech and noise spectra are estimated from the noisy observations.

It has been shown in a single-channel case that iterative Wiener filtering suppresses noise sufficiently after three to four iterations, but produces unnatural sounding speech due to narrow formant bandwidths (Hansen and Clements, 1991). Moreover, quality enhancement is inadequate in unvoiced and transitional regions of speech. In order to improve performance, auditory based transforms are derived and incorporated into the speech spectrum estimation process. Auditory process based constraints are then developed to allow the iterative process to converge to improved quality across all sections while simultaneously keeping residual noise at a minimum.

II. AN OVERVIEW OF AUDITORY PROCESSING

Various models have been developed by researchers in an attempt to describe the peripheral and central processing that occurs in the auditory system. Some of the more analytically tractable auditory models have been used in applications such as analysis, synthesis, and automatic recognition of speech signals. However, it is noted that application and performance evaluation of complex, nonlinear auditory models is a difficult task, and is experimental in nature for most cases. Yang *et al.* (1992) propose an analytical framework in order to model the transformations that acoustic signals undergo during peripheral and central auditory processing stages. They also define auditory models from a biophysical point of view as involving three stages—analysis, transduction, and reduction. This characterization of auditory modeling provides a functional view of the underlying phenomena. The analysis stage involves the relationship between the basilar membrane displacements and the amplitude and frequency content of the sound stimulus. The cochlear mechanism is seen to segregate incoming frequencies into different spatial displacements along the length of the basilar membrane. This process can also be viewed as applying a parallel bank of bandpass filters on the incoming signal. Several studies in psychoacoustics, which relate acoustic signals to what the listener perceives, have experimentally determined the center frequencies and bandwidths of such bandpass filters termed critical band filters (Zwicker, 1961). Broadly speaking, cochlear filtering is seen to be on a logarithmic frequency scale which becomes progressively linear for frequencies below 800 Hz (Scharf, 1970; Zwicker and Terhardt, 1980). In many auditory based models, spectral intensity is perceived as the sum of intensities of the critical bands. The critical band intensities are then either amplitude warped logarithmically or raised to a noninteger power which transforms the intensities into perceived loudness. Examples of applications which use the above transformations are as follows. Hermansky (1990) performs critical band analysis on the nonlinear *Bark* scale, intensity-to-loudness transformation, and lower-order linear prediction on the resulting auditory representation, resulting in a set of perceptual linear prediction parameters which have been successfully used in speech analysis and speech recognition (Hermansky, 1990). Davis and Mermelstein (1980) proposed derivation of the cepstral parameters based on critical band energies on a *mel* frequency scale, which have been successfully used for speech recognition, and have also been applied in an auditory constrained enhancement (ACE-I) algorithm (Nandkumar and Hansen, 1992, 1995). Critical band filtering based on psychoacoustic data has also been used to develop perceptually relevant objective measures of speech quality (Klatt, 1982).

The transduction stage which follows the analysis stage of peripheral auditory processing involves transduction of the mechanical motion along the basilar membrane of the cochlea into electrical firings along an array of auditory-nerve fibers. Again, several studies have modeled these transduction steps for applications in speech analysis and recognition. Yang *et al.* (1992) model this stage to consist of a time derivative representing the fluid–cilia coupling in the cochlea, an instantaneous nonlinearity to represent the nonlinear channels for ionic current flow into hair cells along the cochlea, and a low-pass filter with a short-time constant to represent the ionic flow through the hair cells which results in a temporally smoothed slowly varying signal in each critical band. Seneff (1986) develops an auditory based front-end for speech recognition based on critical band filtering and a hair-cell/synapse model which involves nonlinearities such as dynamic range compression and half-wave rectification on the time-domain signals in each critical band, short-term adaptation to represent the current flow, and a low-pass filter to represent the partial loss of synchrony with increasing frequency. The output of such a model which represents auditory-nerve firing rates in each critical band is then used for speech recognition. Cohen (1989) uses an auditory model which incorporates critical band filtering, loudness transformation, and a short-term adaptive mechanism relating stimulus intensities to auditory-nerve firing rates, as a front-end to a large vocabulary recognition system. Ghitza (1986, 1988) proposes a complex model of the auditory periphery which represents the intensity to neural firing rate transformation along with a neural feedback mechanism. Application of Ghitza's model to speech recognition has shown improved recognition rates in the presence of wideband noise.

TABLE I. Critical band center frequencies F_{0k} and bandwidths BW_k for filter number k spanning a frequency range of 4 kHz.

Critical band specifications								
Num. k	F_{0k}	BW_k	Num. k	F_{0k}	BW_k	Num. k	F_{0k}	BW_k
1	50.0	70.0	9	617.4	86.0	17	1610.7	183.5
2	120.0	70.0	10	703.4	95.3	18	1794.2	199.8
3	190.0	70.0	11	798.7	105.4	19	1993.9	217.2
4	260.0	70.0	12	904.1	116.3	20	2211.1	235.6
5	330.0	70.0	13	1020.4	127.9	21	2446.7	255.3
6	400.0	70.0	14	1148.3	140.4	22	2702.0	276.1
7	470.0	70.0	15	1288.7	153.8	23	2978.0	298.1
8	540.0	77.3	16	1442.5	168.2	24	3276.2	321.5
						25	3597.6	346.1

Finally, the reduction stage which occurs in the auditory nerve can be seen as enhancement of spectral characteristics of the sound-pressure wave before it is conveyed to the central auditory system. An important aspect of the reduction stage is lateral inhibition which occurs due to a biological neural network acting on the auditory-nerve responses. Lateral inhibition networks can be found in all sensory systems such as vision, touch, and the auditory system. The essence of lateral inhibition is to sharpen spatial and temporal stimulus variations. In audition, it has been seen that the lateral inhibition network produces a spectral profile by rapidly detecting spatial discontinuities in the auditory-nerve response (Shamma, 1985; Yang *et al.*, 1992). In general, lateral inhibition in several studies has been modeled as a frequency-dependent function with an excitatory area flanked by two inhibitory areas on either side, which is convolved with an auditory representation of the speech spectrum. Yang *et al.* (1992) model lateral inhibition by a spatial derivative, followed by thresholding using a half-wave rectifier and short-time integration to obtain a short-time auditory equivalent of the magnitude speech spectrum which has been successfully integrated into automatic recognition schemes. Ifukube and White (1987) use a three-range lateral inhibition function convolved with critical band energies in a cochlear implant design and evaluation. Similar lateral inhibition functions have been used to analyze and recognize vowels (Grocholewski and Krenz, 1992). A frequency-dependent lateral inhibition function has also been used for spectral sharpening

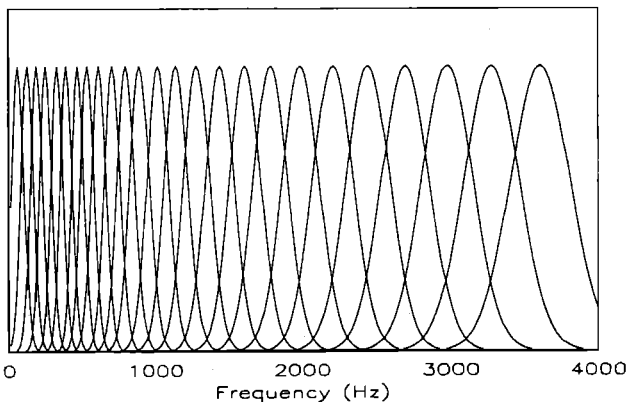


FIG. 2. A parallel bank of filters motivated by critical band auditory analysis using specifications shown in Table I and Eq. (4).

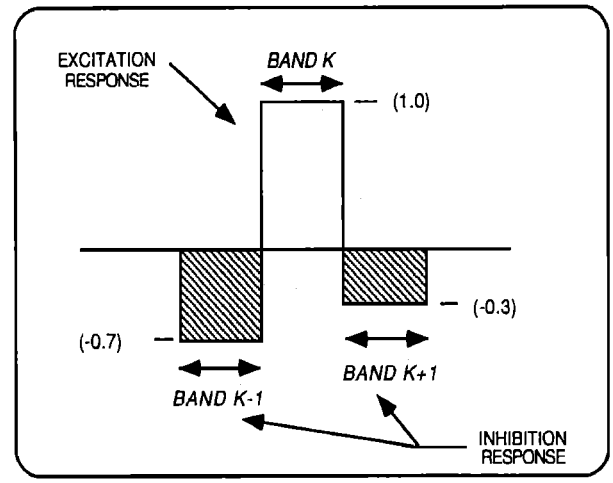


FIG. 3. A functional representation of lateral neural inhibition.

in an adaptive filtering based speech enhancement scheme in the presence of broadband noise (Cheng and O'Shaughnessy, 1991).

A novel scheme which integrates aspects of the above complex stages of auditory processing into the frequency-domain dual-channel enhancement framework described in Sec. I will be discussed and evaluated in the following sections.

III. INTEGRATION OF AUDITORY PROCESSING IN THE ITERATIVE FRAMEWORK

As described in the previous section, the auditory process involves a series of complex and nonlinear temporal and spatial transformations on the speech stimulus. The scheme proposed in this study is to integrate suitable aspects of the complex auditory transformations in order to obtain a more meaningful auditory representation in an iterative enhancement framework. The following sections describe three major auditory transformations and the way they are incorporated in the speech enhancement scenario under consideration.

A. Filters based on psychoacoustic critical band data

Extensive studies in psychoacoustics have been performed in order to measure the ear's critical bandwidths and spacing in the frequency domain (Zwicker *et al.*, 1957; Patterson, 1976; Scharf, 1970; Zwicker and Terhardt, 1980; Zwislocki, 1965). The critical band mechanism is seen to discriminate between sound energy within a single critical band and energy outside the band, thus allowing the auditory system to treat subcritical stimuli alike with respect to auditory phenomena such as masking, loudness, and harmonic discrimination (Scharf, 1970). Masking is one of the phenomena that has led to the determination of critical band shapes and frequencies. Masking can occur in two cases, frequency masking where a lower frequency sound generally masks a higher frequency one, and temporal masking where sounds delayed with respect to one another cause masking of one or both sounds. It is also seen that when two competing sounds occur in a critical band range, the sound with the higher energy masks the second (Scharf, 1970). Critical

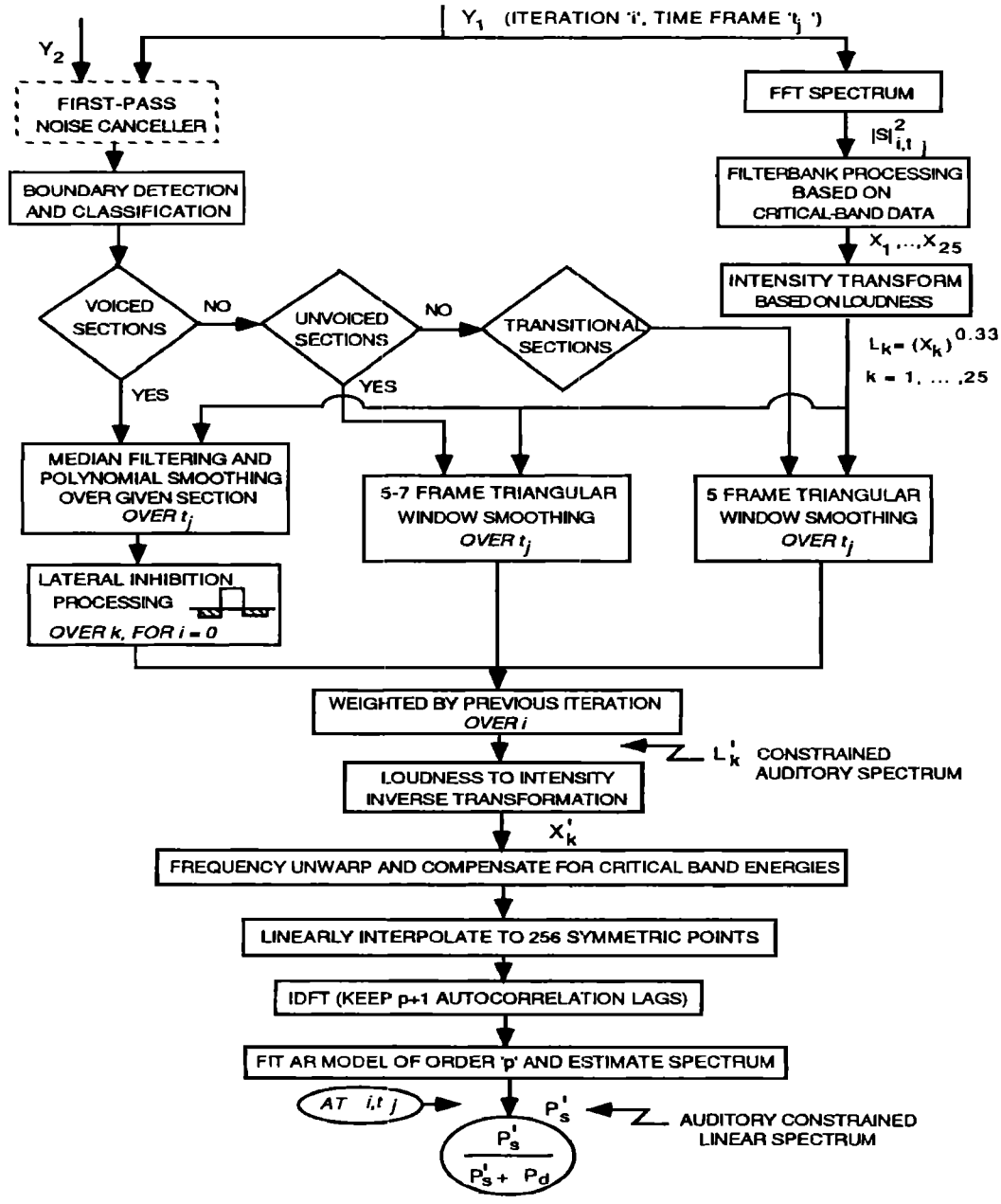


FIG. 4. A detailed flowchart of the proposed auditory processing and speech-specific constraints.

bandwidths are experimentally determined by the fact that a band of noise kept at a constant sound-pressure level while its bandwidth is increased is heard with equal loudness until critical bandwidth is reached (Scharf, 1970). The shape of critical band filters can be determined experimentally by using broadband low-pass or high-pass noise to mask a tone. One such experiment with wideband noise stimuli and tones led to Gaussian critical band filter shapes and a set of center frequencies and bandwidths (Patterson, 1976). These critical band filter specifications may be approximate for conversational speech. However, they have been successfully used to derive a perceptually relevant objective speech quality measure (Klatt, 1982; Quackenbush *et al.*, 1988) (referred to as the weighted spectral slope or Klatt measure). A parallel bank of filters which were motivated by critical band specifications in these previous studies were used to obtain an integrated auditory representation into the enhancement algorithm. The complete set of 25 critical band motivated filter bandwidths and center frequencies is given in Table I for a frequency range of 4 kHz.

The Gaussian approximation for the shapes of the critical band filters can be expressed in terms of center frequency and bandwidth (Patterson, 1976), and is given by

$$B_k(j)=\exp\left[-1.5\left(\frac{j-F_{0k}}{BW_k}\right)^2\right] \quad \text{for } k=1,2,\ldots,25, \quad (4)$$

where $j \in (a_k, b_k)$. Here, a_k and b_k are lower and upper limits of the frequency range of the kth critical filter. The critical band filters as described above are illustrated in Fig. 2.

All processing in the enhancement framework is performed on overlapping time frames. The first transformation

on a given time frame (jth frame) at the ith iteration is to obtain the linear magnitude spectrum $|S_i(\omega_j)|^2$. The 25 point critical band representation of the magnitude spectrum at the ith iteration is then given by

$$X_{k,i}=\sum_{j=a_k}^{b_k} B_k(j)|S_i(\omega_j)|^2 \quad \text{for } k=1,2,\ldots,25. \tag{5}$$

Thus the critical band transformation is applied in the frequency domain, and therefore incorporates aspects of cochlear filtering in the analysis stage of the auditory process. The critical band energies $X_{k,i}$ will also be referred to as critical band intensities.

B. Intensity transformation based on loudness

The next stage in the auditory process is transduction where a nonlinearity is applied to the critical band representation followed by smoothing across time. The nonlinearity in this case is assumed to be a transformation from the critical band intensity $X_{k,i}$ to a perceived loudness sensation. This transformation is given by the psychophysical power law of loudness sensation postulated by Stevens (1955). Hence the loudness based transformation of the critical band intensities at iteration i can be written as

$$L_{k,i}=(X_{k,i})^{1/3} \quad \text{for } k=1,2,\ldots,25. \tag{6}$$

Next, in order to obtain a slowly varying signal in each critical band, smoothing over time frames is performed based on a broad speech classification technique presented in Hansen (1991), as well as in Nandkumar and Hansen (1995) for the ACE-I system. The reasoning behind smoothing over broad classes (voiced, transitional, and unvoiced) is to constrain the auditory processing to match speech production properties. An example of a speech production property is that for most voiced sounds, spectral characteristics are stationary over the entire section rather than just a short-time frame. The specific smoothing constraints on the auditory representation $L_{k,i}$ results in an auditory constrained spectrum $L'_{k,i}$, and are discussed in detail in Sec. IV.

C. Lateral neural inhibition

The final stage of the auditory process incorporated in this study is based on lateral neural inhibition. Lateral inhibition is a sensory phenomena which, in audition, acts upon the auditory-nerve responses in order to obtain and convey a spectral representation to the central auditory system (Yang *et al.*, 1992). Lateral inhibition is a concept based upon the fact that the response of a neuron can be affected by the response of adjacent neurons, spatially and temporally. In general, lateral inhibition is seen to sharpen spatial input patterns to highlight edges and peaks, and in some cases sharpen temporal input changes (Shamma, 1985). The model for a lateral inhibition function used in this study is similar to the one used in Grocholewski and Krenz (1992), where the excitatory and the inhibitory ranges each span one critical band as illustrated in Fig. 3.

The function of lateral inhibition is applied on the auditory representation $L'_{k,i}$ by a discrete convolution operation. The resulting auditory constrained spectrum is given by

$$L'^{(\mathrm{LI})}_{k,i}(m)=\sum_{n=1}^{25} L'_{k,i}(n)H(m-n) \quad \text{for } m=1,2,\ldots,25, \tag{7}$$

where $H(j)$ is the functional representation of lateral inhibition given by

$$\begin{aligned} H(j)&=-0.7, \quad j=-1,\\ &=1.0, \quad j=0,\\ &=-0.3, \quad j=1,\\ &=0.0, \quad \text{elsewhere.} \end{aligned} \tag{8}$$

The lateral inhibition stage is implemented only on $L_{k,0}$, which is the auditory representation of degraded speech at the start of the iterative enhancement process. This step was decided based on the observation that spectral sharpening due to lateral inhibition adds to spectral distortion as the iterations proceed. However, the spectral estimate of noisy speech is shown to be a better starting point for iterative enhancement if the auditory constrained spectrum includes the lateral neural inhibition stage. This is due to the sharpening of spectral features by lateral inhibition which results in an enhanced spectral estimate at the beginning of the iterative procedure, especially when wideband noise has the effect of suppressing spectral peaks. In addition, lateral inhibition is seen to be effective in the auditory constrained spectrum as an initial spectral estimate only for sections classified as voiced. Spectral enhancement was not significant over transitional and unvoiced sections of the noisy utterance. In fact, further spectral distortion was observed over some unvoiced sections. Hence the lateral inhibition stage is enabled only during sections classified as voiced.

IV. AUDITORY CONSTRAINTS ACROSS SPEECH SECTIONS

Smoothing constraints over time and iteration are discussed in this section based on the auditory representation $L_{k,i}$. The constraints are directed by a broad classification of speech into voiced, transitional, and unvoiced sections. The classification is performed by a boundary detector developed in Hansen (1991, 1994) and integrated into the enhancement algorithm in a manner similar to that in the ACE-I algorithm (Nandkumar and Hansen, 1995). A detailed flowchart of the proposed auditory constraints is illustrated in Fig. 4.

A. Constraints on an auditory spectral representation

Smoothing constraints on the auditory representation are applied over time in order to obtain a slowly varying time signal for each auditory critical band response. Speech sections classified as "voiced" are in general steady state and hence the smoothing is performed over an entire given section. Smoothing in "voiced" sections is of the form of three frame median filtering in order to remove single frame outliers followed by polynomial smoothing using a least-

squares-fitting technique. Speech sections classified as "unvoiced" may not have smooth variations over the entire section. Hence a triangular weighted smoothing is applied over five frames for $L_{1,i}...L_{20,i}$, the auditory representation up to approximately 2200 Hz, and over seven frames for the auditory representation at higher frequencies, $L_{21,i}...L_{25,i}$. The smoothing over five to seven frames corresponds to smoothing over approximately 35–50 ms of the input speech signal. A slowly varying signal over 35–50 ms is a reasonable assumption over unvoiced sections of speech. An example of smoothing over five frames for the kth auditory representation at time frame t_j is given by

$$L'_{k,i}(t_j)=[L_{k,i}(t_j-2)+2L_{k,i}(t_j-1)+3L_{k,i}(t_j)+2L_{k,i}(t_j+1)+L_{k,i}(t_j+2)]/9. \quad (9)$$

Speech sections classified as "transitional" are transient regions, and a five frame triangular weighted smoothing is maintained over all $L_{k,i}$, and seen to result in an improved spectral representation.

Constraints are also applied over iterations on $L_{k,i}$ in order to obtain smoother transitions across iterations and allow the iterative enhancement algorithm to achieve mproved noise suppression along with minimum possible spectral distortion at the same iteration. The above constraint was experimentally derived with the help of objective quality measures, vocal-tract spectra, and informal listening tests. The constraints across iterations which achieved the best quality are of the form

$$L'_{k,i}(t_j)=\left(\frac{4L_{k,i}(t_j)+L_{k,i-1}(t_j)}{5.0}\right), \quad (10)$$

where the auditory representation from the current iteration is fractionally weighted by the auditory representation from the previous iteration. A detailed flowchart of the auditory processing and the speech-specific constraints is shown in Fig. 4.

B. Auditory constrained linear prediction based speech spectrum

The auditory transformation in frequency and magnitude, and speech-specific constraints across frequency, time, and iteration, result in a spectral representation termed constrained auditory spectrum. Next, the 25-point, frequency warped constrained auditory spectrum must undergo further transformations in order to obtain a linear vocal-tract spectrum for use in the primary channel Wiener filter of the proposed dual-channel enhancement framework. The following transformations lead to such an auditory constrained linear speech spectrum. First, an inverse transform converts the auditory constrained spectrum to an approximation of the intensity magnitudes from critical bands as follows:

$$X'_{k,i}=(L'_{k,i})^3 \quad \text{for } k=1,2,...,25. \quad (11)$$

Next, the constrained critical band representation $X'_{k,i}$ is unwarped along the frequency scale (each energy value is placed at the center frequency of the respective critical band), and compensated in energy for the critical band integration, given by

$$|S'_{F_{0k},i}|^2=\frac{X'_{k,i}}{\sum_{j=a_k}^{b_k}B_k(j)} \quad \text{for } k=1,2,...,25, \quad (12)$$

where a_k, b_k, and B_k are as given in Eq. (4). The resulting constrained energy spectrum is linearly interpolated from 25 points to 128 points. Next, the following steps are performed to obtain a smooth vocal-tract spectrum. An inverse discrete Fourier transform (IDFT) of the interpolated constrained spectrum extended to 256 even symmetric points results in the autocorrelation coefficients $r'_{k,i}$ given at time frame t_j by

$$r'_{k,i}(t_j)=\text{IDFT}[|S'_{k,i}(t_j)|^2]. \quad (13)$$

Applying the Durbin recursion on the first 11 autocorrelation coefficients $r'_{k,i}$ results in ten linear prediction parameters which will be termed auditory constrained and enhanced linear prediction (ACE-LP) parameters. Next, an auditory constrained smooth linear spectrum is given by

$$\hat{P}'_S(\omega_k,i,t_j)=\left|\frac{g_{i,t_j}}{1-\sum_{l=1}^{10}a'_{i,t_j}(l)e^{-jl\omega_k}}\right|^2, \quad (14)$$

where g_{i,t_j} is a gain term and a'_{i,t_j} are the ACE-LP parameters for time frame t_j at the ith iteration. The auditory constrained smooth linear spectrum is then included in the primary channel Wiener filter of the proposed dual-channel enhancement framework, and the resulting enhancement algorithm is termed auditory constrained enhancement-II (ACE-II). Linear vocal-tract spectra of a single representative time frame for four different speech sounds are shown in Figs. 5 and 6. Each spectra is illustrated for four conditions—original noise-free, degraded (SNR=5.0 dB), ACE-LP on the degraded speech as a starting point for iterative enhancement, and enhanced speech after three iterations of ACE-II. The ACE-LP spectral representations are seen to be better starting points than the degraded speech spectrum for all speech sounds shown (a vowel /aa/ and a nasal /n/ in Fig. 5(a) and (b), an unvoiced stop /p/ and a fricative /s/ in Fig. 6(a) and (b). Furthermore, the ACE-II algorithm is seen to bring the enhanced spectra closer to the original spectral variations for all speech sounds shown, especially well in higher energy regions of the frequency spectra. Results regarding the quality of speech enhanced by ACE-II are presented in the following section.

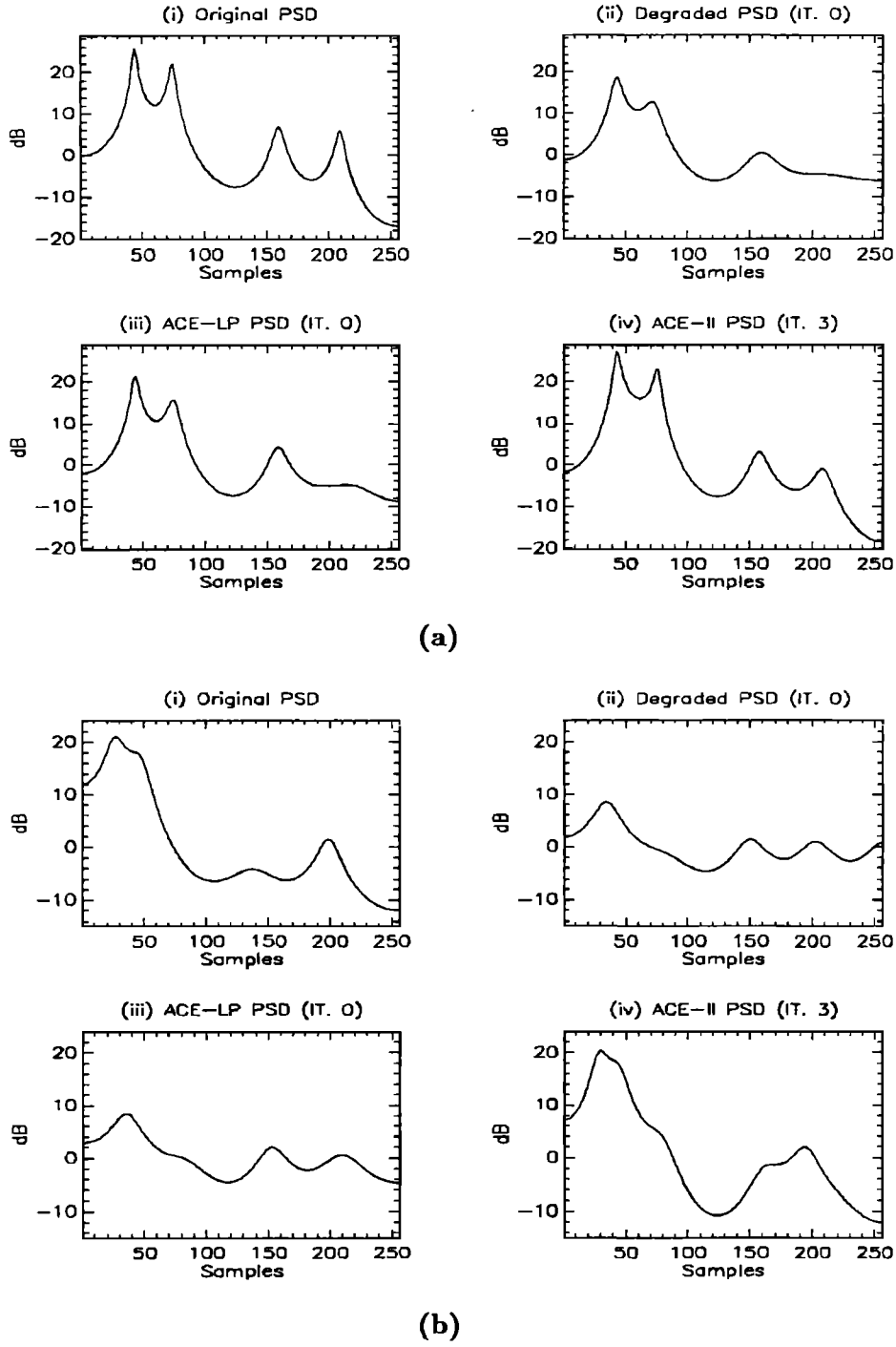


FIG. 5. Linear vocal-tract spectra for (a) a vowel section /aa/ and (b) a nasal section /n/, for four cases: (i) original undegraded, (ii) degraded, SNR=5.0 dB, (iii) ACE-LP processed and input to first iteration, (iv) ACE-II enhanced after three iterations.

V. RESULTS FOR THE ACE-II SYSTEM

The ACE-II enhancement system was developed by simulating aspects of the auditory process, resulting in a novel auditory spectral representation which is incorporated along with perceptual based constraints into the iterative enhancement algorithm. The quality of speech enhanced by ACE-II is assessed using the Itakura–Saito (IS) log-likelihood objective quality measure and the weighted spectral slope (Klatt) measure. The weighted spectral slope measure (Klatt, 1982; Quackenbush *et al.*, 1988) is a perceptually relevant measure based on critical band analysis of the speech spectrum. In this study, the Klatt measure was implemented with the same critical band specifications as the one used for the ACE-II algorithm. Hence the auditory perception based Klatt measure, along with the IS measure which is a good measure of spectral distortion, provides a relevant way to assess quality improvement for the ACE-II algorithm.

In all cases, signal-to-noise ratio (SNR) is defined simply as the ratio of the noise-free speech energy to the degrading noise energy (the two energy values are assumed to be known in an experimental scenario where speech and noise

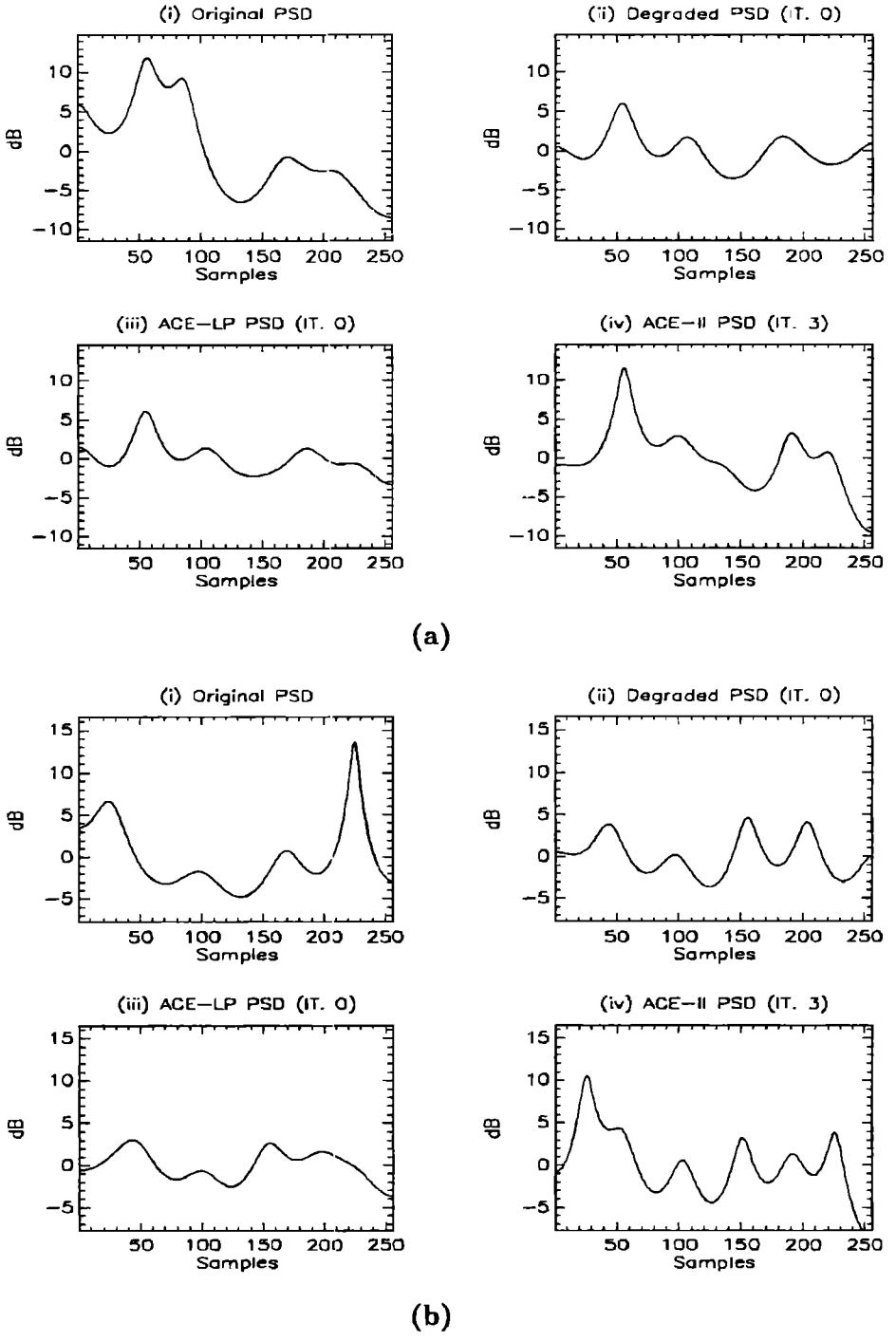


FIG. 6. Linear vocal-tract spectra for (a) an unvoiced stop /p/ and (b) a fricative section /s/, for four cases: (i) original undegraded. (ii) degraded. SNR=5.0 dB, (iii) ACE-LP processed and input to first iteration, (iv) ACE-II enhanced after three iterations.

are digitally mixed at controlled SNR levels). The dual-channel iterative framework has been shown to be robust in varying cross-talk levels for the ACE-II algorithm (Nandkumar and Hansen, 1992, 1994; Nandkumar, 1993). In this study, we shall focus on evaluating the new auditory constraint technique at a fixed cross-talk level of zero. So, for all results discussed in this section, the best cross-talk condition (zero cross-talk) is assumed. The correlation functions $H_1(\omega)$ and $H_2(\omega)$ are estimated during noise-only and speech-only conditions in a dual-channel experimental setup. Further details about experimental simulation of the dual-channel scenario can be found in Nandkumar and Hansen (1992; 1995) and Nandkumar (1993). Evaluation for a single utterance and a single noise condition will first be presented, followed by evaluation over a large set of utterances and varying noise conditions.

A. Evaluation for a single utterance, single noise condition

Enhancement performance is illustrated using several tools. First, objective quality measures classified over different speech classes are shown. Time waveforms and time ver-

TABLE II. IS measures for ACE-II over five iterations, for AWGN, SNR=5.0 dB. Optimum perceived objective quality is indicated by a ◇.

Sound type	Itakura–Saito likelihood measure (across iterations)						
	Degraded	No. 1	No. 2	No. 3	No. 4	No. 5	No. frames
Silence	3.42	1.96	1.54	◇1.38	1.57	2.61	51
Vowel	5.55	4.13	2.15	◇0.76	1.05	4.43	158
Nasal	3.53	2.34	1.74	◇1.40	1.68	3.65	13
Stop	3.33	2.37	1.95	1.67	◇1.60	2.05	72
Fricative	1.41	1.21	◇1.19	1.29	1.65	2.47	39
Liquids and glides	3.04	1.92	0.95	◇0.52	1.14	8.95	21
Voiced+unvoiced	4.23	3.11	1.88	◇1.06	1.29	3.89	303
Total	4.11	2.94	1.83	◇1.10	1.33	3.71	354

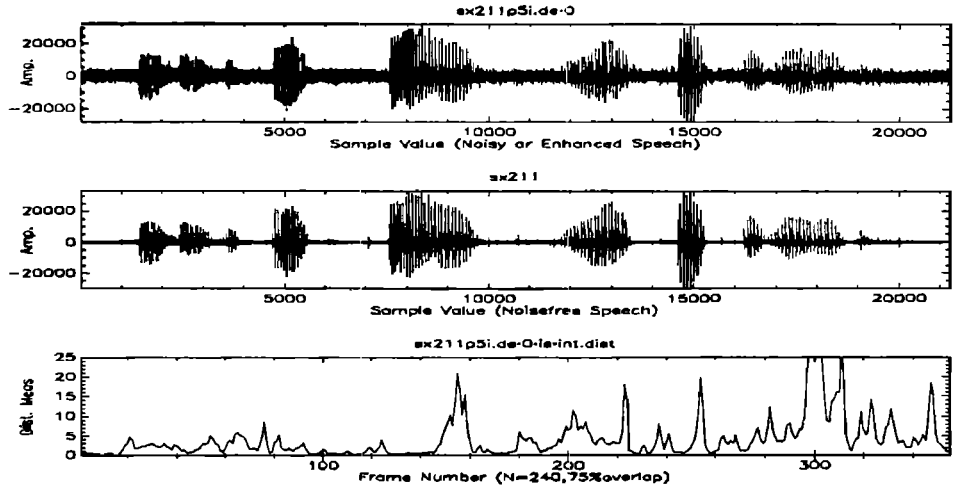
sus frequency vocal tract spectra are also used to illustrate performance. In many cases, ACE-II performance will be compared with the unconstrained dual-channel iterative framework. All objective measures are calculated with respect to the original noise-free utterance. Hence an objective measure closer to zero implies lesser spectral distortion or lesser deviation from the original. The improvement in speech quality is illustrated using the IS measure and the Klatt measure classified over different speech classes for the degraded utterance and five iterations of ACE-II in Tables II and III, respectively, for a single utterance.

The degrading noise is additive white Gaussian noise (AWGN) at a SNR of 5 dB. The IS measure and the Klatt measure both show significant and consistent improvement over all speech classes. However, the IS measure is seen to have a larger variance (that is, the distance between the enhanced and degraded measures is larger) than the Klatt measure. Quality improvement for ACE-II using IS measure is seen to be similar to that of ACE-I which was illustrated in Nandkumar and Hansen (1992, 1995). In addition, ACE-II is seen to result in improved spectral distortion for fricatives, which are noiselike and difficult to enhance. Quality improvement is seen to be the best at iteration 3 for most speech classes. The perceptually relevant Klatt measure in Table III also shows significant improvement especially for vowels, nasals, liquids, and glides. Optimal speech quality in this case is seen to be achieved at both iteration 3 and 4 for the different speech classes. There seems to be very little perceptual difference between the overall speech quality at these two iterations. Informal listening tests confirm the above results and provide a choice of the third or fourth as terminating iterations. Residual noise at the fourth iteration seemed to be lesser than at the third iteration, and hence preferred more during informal listening tests.

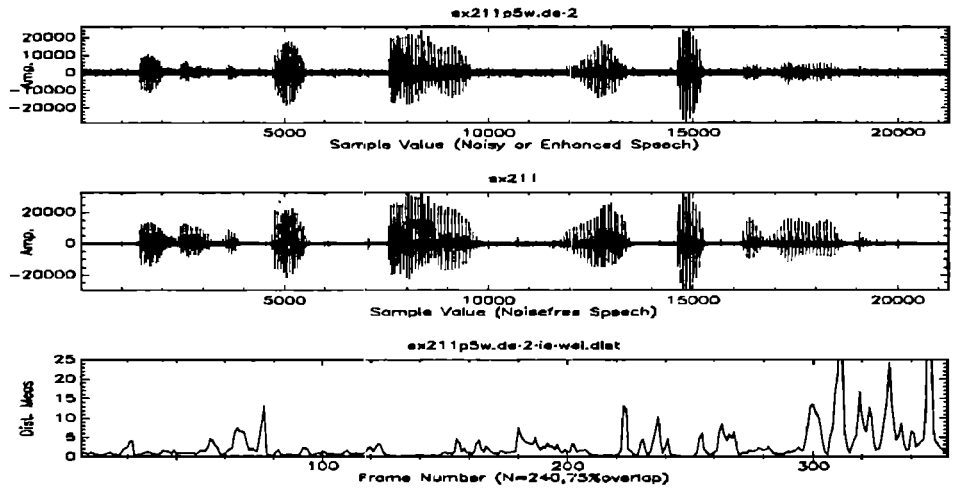
Next, time waveforms and frame-to-frame IS measures are presented for speech degraded with AWGN, SNR=5.0 dB in Fig. 7(a), and speech enhanced from iteration 2 of the unconstrained enhancement framework in Fig. 7(b), and from iteration 3 of ACE-II in Fig. 7(c). Spectral distortion due to AWGN is seen to affect low-energy regions of speech more than steady-state high-energy sections as seen in Fig. 7(a). Again, ACE-II is seen to successfully suppress noise, and result in consistent quality improvement across the entire utterance [Fig. 7(c)]. It can be concluded from the frame-to-frame results presented for ACE-I in Nandkumar and Hansen (1992; 1995) and Nandkumar (1993) and for ACE-II above that auditory based constraints contribute to significant improvement in quality especially for low-energy unvoiced and transitional regions of speech, when compared to an unconstrained dual-channel iterative enhancement framework. Time versus frequency based vocal-tract spectra for the word *players* are also shown for the original, degraded, and enhanced from iteration 4 of the unconstrained technique, and enhanced from iteration 4 of the ACE-II algorithm, in Fig. 8(a)–(d). Time versus frequency spectra in this case illustrate the improvement in spectral degradation across frequency, and for the different sounds in the given word. The ACE-II

TABLE III. Weighted spectral slope (Klatt) measures for ACE-II over five iterations, for AWGN, SNR=5.0 dB. Optimum perceived objective quality is indicated by a ◇.

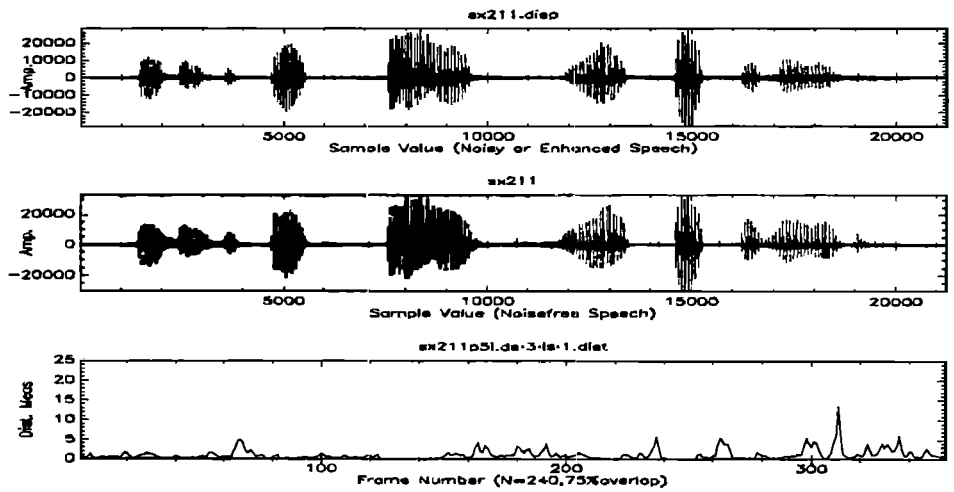
Sound type	Weighted spectral slope measure (across iterations)						
	Degraded	No. 1	No. 2	No. 3	No. 4	No. 5	No. frames
Silence	3.91	4.33	2.86	◇2.72	2.76	2.92	51
Vowel	2.33	2.30	1.66	1.48	◇1.45	2.77	158
Nasal	2.51	2.12	1.86	1.74	◇1.66	1.83	13
Stop	2.75	2.63	◇2.56	2.60	2.75	2.94	72
Fricative	3.01	2.50	2.97	◇2.39	2.53	2.82	39
Liquids and glides	3.47	2.56	1.95	1.88	◇1.54	4.55	21
Voiced+unvoiced	2.60	2.41	2.07	◇1.88	1.91	2.90	303
Total	2.79	2.69	2.18	◇1.99	2.03	2.90	354



(a) Average Measure = 4.11



(b) Average Measure = 2.89



(c) Average Measure = 1.10

FIG. 7. Time waveforms and frame-to-frame IS measures between original noise-free utterance and processed utterances: (a) degraded with AWGN, SNR=5.0 dB, (b) iteration No. 2 of unconstrained enhancement, (c) iteration No. 3 of the ACE-II technique.

technique is seen to improve some of the highly distorted-spectral characteristics, while the unconstrained technique shows further distortion at the same iteration. ACE-II is also seen to improve high-frequency charac-teristics, especially for the trailing fricative, without introducing further distortion in other spectral regions. Further evaluations on a large set of speakers, sentences, signal-to-noise ratios, and noise cases will be discussed in the following section.

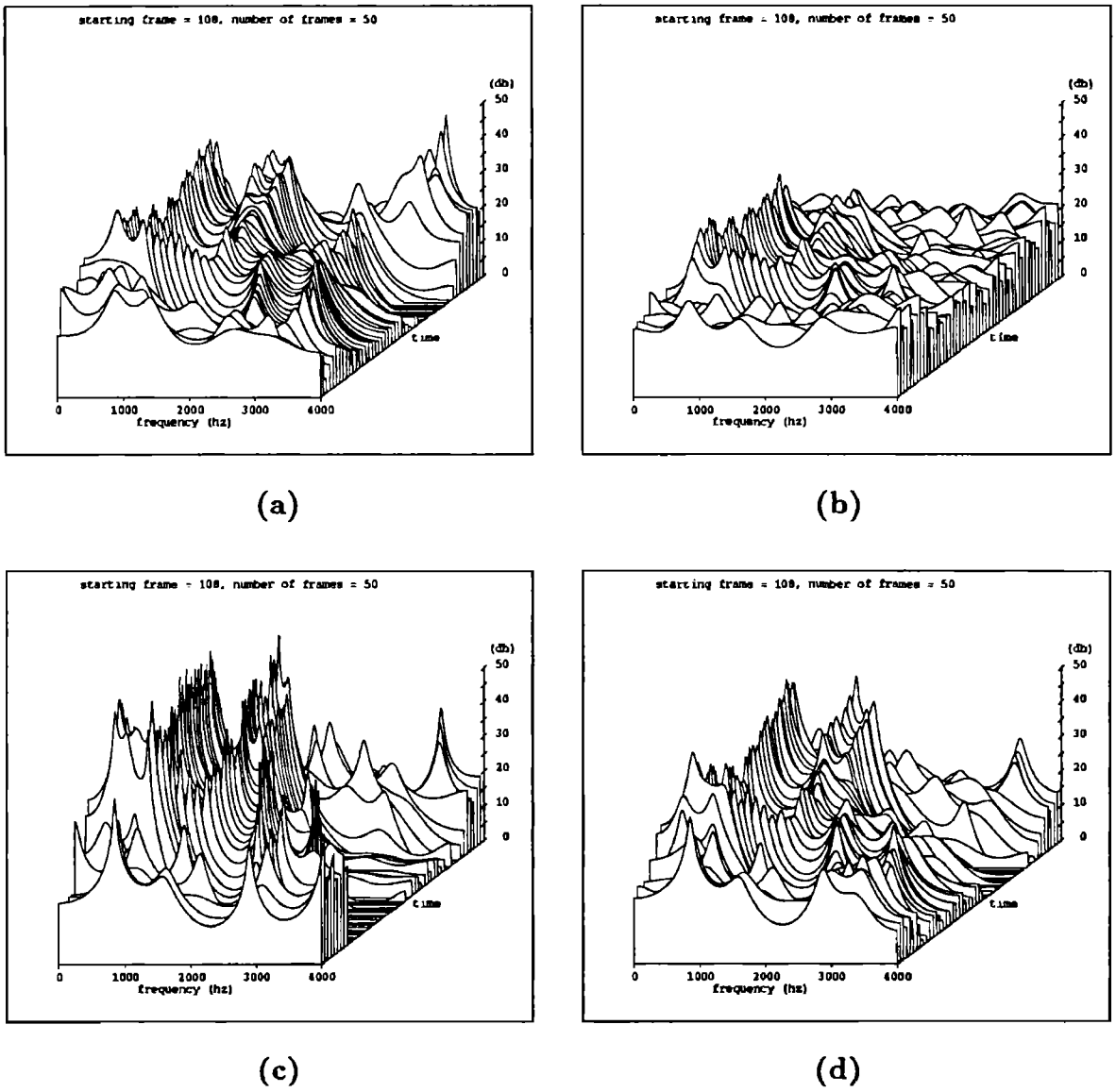


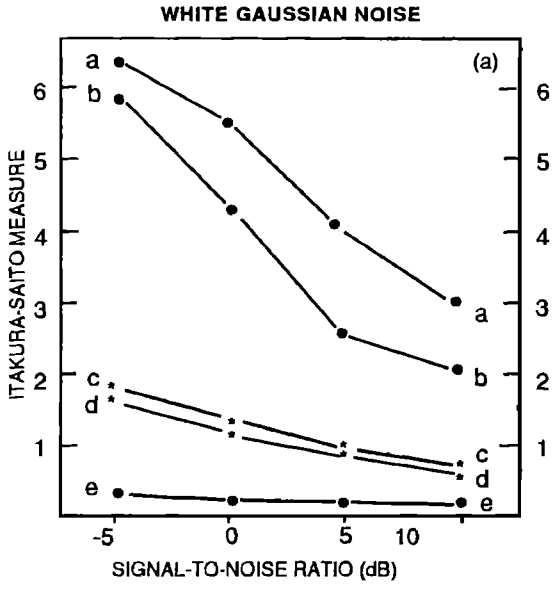
FIG. 8. Time versus frequency vocal-tract responses for the word *players* for (a) original noise-free, (b) degraded, SNR=5.0 dB (AWGN), (c) iteration No. 4 using dual-channel unconstrained Wiener filtering, and (d) iteration No. 4, using the ACE-II algorithm.

B. Evaluation over the TIMIT database with different noise conditions

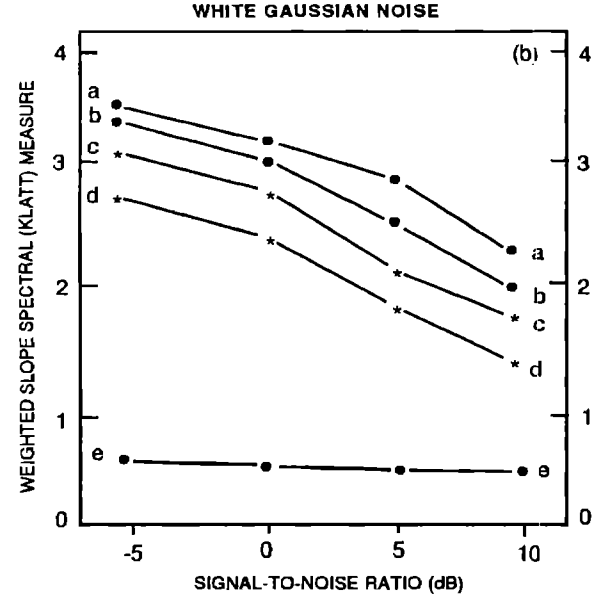
In this section, a more detailed evaluation of the ACE-II system is presented based on a large set of utterances for a wide range of signal-to-noise ratios, and for different noise cases such as slowly varying colored noise. The purpose is to show that improvement for a single noisy speech condition can be extended to larger noise and speaker populations. The primary means of demonstrating quality of the proposed enhancement technique are objective quality measures. Visualization of speech waveforms and spectra over time and informal listening tests have also been used to support observed speech quality measures. Objective quality measures have been shown to possess fair to good correlation with subjective perceived quality (Quackenbush *et al.*, 1988), and have been used extensively in the evaluation of speech coding (Hansen and Nandkumar, 1992; Quackenbush *et al.*, 1988) and speech enhancement systems (Hansen and Clements, 1985, 1989; Hansen, 1991). In this study, objective measures have not only been used as a global measure over utterances, but also visualized over time frames and over classified phonemes and speech classes, in order to demonstrate a more in-depth understanding of enhanced quality.

1. White, Gaussian noise

Performance evaluation for an additive white Gaussian noise (AWGN) distortion will be discussed in this section. First, performance based on global Itakura–Saito (IS) measures for a sentence from the TIMIT database (NIST, 1988) degraded with AWGN at different SNR levels was determined. All processing was done on the following sentence spoken by a male speaker: "*Only the best players enjoy popularity.*" Figure 9(a) illustrates global IS measures versus SNR which range from −5 to 10 dB for the following five cases: (a) the degraded original speech, (b) dual-channel unconstrained Wiener filtering, (c) the ACE-I algorithm, (d) the ACE-II algorithm, and (e) the theoretical limit. The theoreti-



a) Original Degraded.
b) Dual-Channel Unconstrained Wiener Filter.
c) Proposed ACE-I Dual-Channel Algorithm.
d) Proposed ACE-II Dual-Channel Algorithm.
e) Theoretical Limit (using undegraded spectrum).



a) Original Degraded.
b) Dual-Channel Unconstrained Wiener Filter.
c) Proposed ACE-I Dual-Channel Algorithm.
d) Proposed ACE-II Dual-Channel Algorithm.
e) Theoretical Limit (using undegraded spectrum).

FIG. 9. White Gaussian noise performance: global (a) Itakura-Saito and (b) Klatt measures for a single utterance versus SNR.

cal limit represents best possible enhancement (IS measure closest to 0) that could be obtained using exact noise-free speech parameters, generally not available in practice, during dual-channel enhancement. The ACE-I technique with constraints on mel-cepstral parameters (Nandkumar and Hansen, 1992; Nandkumar, 1993) and the proposed ACE-II algorithm

TABLE IV. A comparison of objective speech quality measures for degraded, unconstrained dual-channel Wiener filtering, and the proposed ACE-I and ACE-II systems, for white Gaussian noise, SNR=5.0 dB. $|\hat{\rho}|$ is the average correlation coefficient between objective and subjective speech quality as measured by composite acceptability of the diagnostic acceptability measure (Quackenbush *et al.*, 1988).

	Objective speech quality measures		
	Itakura–Saito	log area ratio	weighted spectral slope
$\|\hat{\rho}\|$	0.59	0.62	0.74
Degraded original	4.11	8.48	2.79
Unconstrained	2.89	7.41	2.54
ACE-I	1.08	5.78	2.10
ACE-II	1.04	3.66	1.95

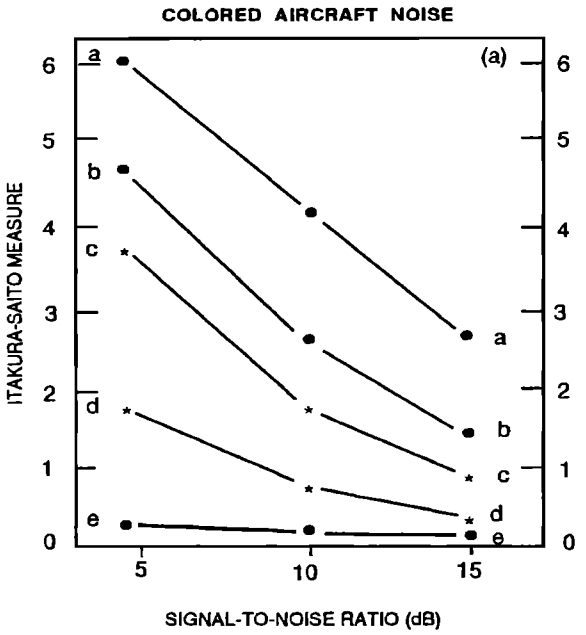
are seen to perform better and are more consistent as SNR decreases when compared to dual-channel unconstrained Wiener filtering. However, the measured spectral distortion of the enhanced speech is still seen to increase as SNR decreases. Overall quality of ACE-II vs SNR is seen to be comparable with that of ACE-I, with a slight improvement for ACE-II at lower SNR levels. The overall weighted spectral slope (Klatt) measures are also illustrated for ACE-II in Fig. 9(b). ACE-II is seen to improve overall quality as compared to unconstrained dual-channel Wiener filtering and the ACE-I algorithm for all tested SNR levels. One advantage of the general class of Wiener filtering approaches is that no "musical tone" artifacts are present after processing as can be observed in spectral subtraction techniques (Lim and Oppenheim, 1978; Hansen, 1988). Informal listening tests using a collection of TIMIT sentences have confirmed improvement in speech quality, with no additional artifacts being introduced after enhancement processing using ACE-I and ACE-II, for both male and female speakers.

Similar improvement in overall quality is seen for the log-area-ratio measures for both ACE-I and ACE-II. A comparison of the three objective speech quality measures for unconstrained dual-channel Wiener filtering, and the ACE-I and ACE-II systems, is shown in Table IV. The correlation between each objective quality measure and subjective quality as measured by composite acceptability of the diagnostic acceptability measure (Quackenbush *et al.*, 1988) is also shown.

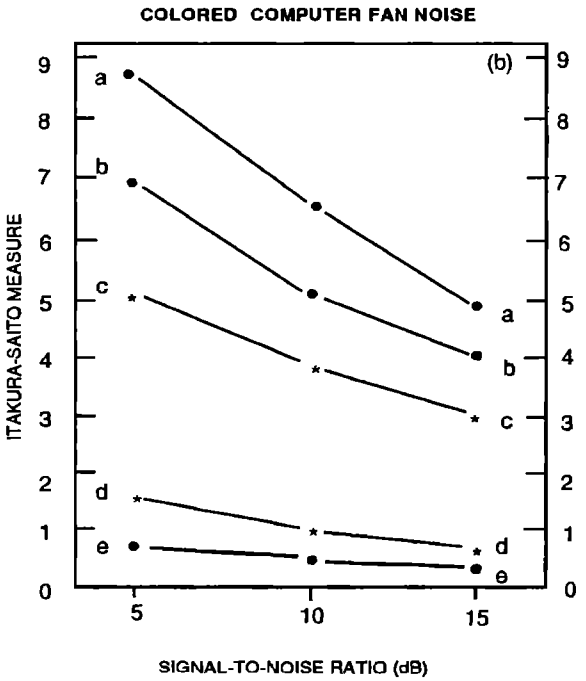
2. Nonstationary colored noise

Next, performance for slowly varying colored aircraft cockpit noise[1] and noise from the cooling fan of a workstation (computer fan noise) is obtained for the enhancement systems under consideration. Samples of these colored noise cases were obtained from actual noise recordings.

Enhancement performance is illustrated by means of global quality measures over a single utterance versus SNR, for aircraft noise distortion in Fig. 10(a) and for computer fan noise distortion in Fig. 10(b). Best overall objective speech quality was obtained for both colored noise cases at the third iteration. It can be seen from Fig. 10(a) that overall quality improvement for the proposed ACE-II algorithm for aircraft cockpit noise is comparable to that obtained for white noise, especially at higher SNR values. It is noted that



a) Original Degraded.
b) Dual-Channel Unconstrained Wiener Filter.
c) Proposed ACE-I Dual-Channel Algorithm.
d) Proposed ACE-II Dual-Channel Algorithm.
e) Theoretical Limit (using undegraded spectrum).



a) Original Degraded.
b) Dual-Channel Unconstrained Wiener Filter.
c) Proposed ACE-I Dual-Channel Algorithm.
d) Proposed ACE-II Dual-Channel Algorithm.
e) Theoretical Limit (using undegraded spectrum).

FIG. 10. Colored noise performance: global Itakura–Saito measures for a single utterance versus SNR for (a) aircraft cockpit noise and (b) computer fan noise.

TABLE V. IS measures over speech classes for speech degraded by aircraft cockpit noise at SNR=10 dB, and enhanced using unconstrained Wiener filtering, ACE-I, and ACE-II.

Sound type	Aircraft cockpit noise Itakura–Saito likelihood measure				
	Degraded	Unconstrained	ACE-I	ACE-II	No. frames
Silence	2.42	6.24	4.10	1.06	51
Vowel	0.18	0.09	0.16	0.08	158
Nasal	0.58	0.68	0.69	0.39	13
Stop	3.77	2.17	1.66	0.94	72
Fricative	27.05	8.34	6.49	3.52	39
Liquids and glides	0.34	0.12	0.87	0.20	21
Voiced+unvoiced	4.52	1.67	1.40	0.75	303
Total	4.22	2.33	1.79	0.80	354

the ACE-I algorithm performs better than unconstrained Wiener filtering, and the proposed ACE-II algorithm significantly outperforms both ACE-I and unconstrained Wiener filtering for all SNR levels shown. Overall IS quality measures in Fig. 10(b) indicate very high distortion for the tested speech utterance degraded with computer fan noise. The measures also indicate poor enhancement quality for unconstrained Wiener filtering and the ACE-I algorithm. However, ACE-II again shows constant improvement over unconstrained Wiener filtering and ACE-I, achieving quality levels comparable to the white Gaussian noise and aircraft noise cases. Informal listening tests confirm the quality improvement demonstrated by global objective measures. In summary, it is seen that ACE-II with auditory process based constraints provides excellent enhancement quality for colored noise cases when compared to unconstrained dual-channel Wiener filtering and the ACE-I algorithm. One of the reasons ACE-II performs better than ACE-I in colored noise cases could be that the auditory constraints for ACE-II were derived independent of noise type, whereas the auditory constraints for ACE-I were derived based on the behavior of mel-cepstral parameters during iterative enhancement for a white Gaussian noise distortion.

Objective measures of speech quality can be grouped into broad speech classes in order to illustrate improvement for each class. Frame-to-frame Itakura–Saito measures are classified over speech classes and shown for a degraded utterance (SNR=10 dB), unconstrained dual-channel Wiener

TABLE VI. IS measures over speech classes for speech degraded by computer fan noise at SNR=10 dB, and enhanced using unconstrained Wiener filtering, ACE-I, and ACE-II.

Sound type	Computer fan noise Itakura–Saito likelihood measure				
	Degraded	Unconstrained	ACE-I	ACE-II	No. frames
Silence	15.12	15.69	16.64	2.47	51
Vowel	0.09	0.04	0.30	0.06	158
Nasal	1.02	0.21	1.03	0.20	13
Stop	12.77	11.57	6.27	1.95	72
Fricative	14.88	3.11	1.06	1.82	39
Liquids and glides	0.06	0.02	0.53	0.05	21
Voiced+unvoiced	5.04	3.18	1.86	0.74	303
Total	6.50	4.98	3.99	0.99	354

TABLE VII. Itakura–Saito quality measures across phonemes for a set of 100 speech utterances for ACE-II compared to degraded (AWGN, SNR=5 dB).

OBJECTIVE SPEECH QUALITY ACROSS AMERICAN PHONEMES								
hline *Ph.*		DEG	ACE-II	# *Fr*	*Ph.*		DEG	ACE-II	# *Fr*
CONSONANTS – nasals					*CONSONANTS – unvoiced stops*				
/m/	me	7.986	2.970	683	/p/	pan	2.733	1.242	508
/n/	no	9.710	3.652	1153	/t/	tan	1.770	1.068	542
/ng/	sing	9.621	3.618	159	/k/	key	2.689	1.191	559
/nx/	many	7.475	1.286	77	*CONSONANTS – voiced stops*				
/em/	problem	6.197	3.514	33	/b/	be	2.877	0.982	135
/en/	traction	10.376	3.118	135	/d/	dawn	1.453	0.844	186
/eng/	greasing	4.261	1.200	18	/g/	give	2.788	1.175	142
CONSONANTS – unvoiced fricatives					*CONSONANTS – closure stops*				
/s/	sip	0.815	1.184	1433	/tcl/	it pays	1.983	1.428	999
/th/	thing	1.110	1.201	203	/kcl/	pockets	2.166	1.647	655
/f/	fan	1.032	1.173	796	/bcl/	to buy	4.117	2.270	399
/sh/	show	1.471	1.535	673	/dcl/	sandwich	3.693	2.219	636
CONSONANTS – voiced fricatives					/gcl/	iguanas	3.240	1.774	241
/z/	zip	1.025	1.352	1054	/pcl/	accomplish	1.628	1.271	779
/zh/	garage	1.277	1.314	66	*CONSONANTS – glottal stop, flap*				
/dh/	that	3.852	1.463	270	/q/	allow	5.103	1.941	661
/v/	van	4.387	1.497	273	/dx/	put in	5.546	1.031	142
CONSONANTS – affricates					*CONSONANTS – unvoiced whisper*				
/jh/	joke	1.526	2.125	263	/hh/	had	4.345	2.052	143
/ch/	chop	1.791	2.628	336	*CONSONANTS – voiced whisper*				
					/hv/	you have	8.101	1.459	103
VOWELS – front					*DIPHTHONGS*				
/ih/	hid	2.403	0.720	947	/ay/	hide	2.641	0.619	1033
/eh/	head	2.998	0.603	856	/oy/	coin	5.595	1.102	171
/ae/	had	2.407	0.464	977	/ey/	pain	1.733	0.670	725
/ux/	to buy	2.962	0.850	636	/ow/	code	3.453	0.976	660
VOWELS – mid					/aw/	pout	3.093	0.629	288
/aa/	odd	4.503	0.986	1339	/iy/	new	2.278	1.311	1220
/er/	earth	9.798	2.174	562	*SEMIVOWELS – liquids*				
/ah/	up	3.518	0.693	625	/r/	ran	12.257	2.590	747
/ao/	all	6.682	1.528	750	/l/	lawn	5.326	1.496	1079
VOWELS – back					/el/	chemicals	6.261	2.008	356
/uw/	boot	5.068	1.280	197	*SEMIVOWELS – glides*				
/uh/	foot	3.401	0.531	116	/w/	wet	7.290	2.507	289
VOWELS – front schwa					/y/	you	2.417	1.107	318
/ix/	heed	3.974	1.426	1043	*Silence*				
VOWELS – back schwa					/# /	extended	1.736	0.871	5087
/ax/	a ton	4.872	1.190	628	/pau/	pause	3.017	1.825	175
VOWELS – retroflexed schwa					/epi/	epenthetic	4.656	3.437	98
/axr/	after	12.011	2.539	594					
VOWELS – voiceless schwa					**Overall**		3.677	1.394	36006
/ax-h/	sub	3.201	2.722	35	**Overall - /#/**		3.996	1.480	30919

filtering, ACE-I, and ACE-II, for aircraft cockpit noise distortion in Table V and computer cooling fan noise in Table VI.

Unlike the case of white Gaussian noise, both colored noise cases do not show significant distortion for the vowels, nasals, liquids, and glides. This is as expected since energy distribution over frequency for colored noise cases is similar to that of the above sonorant classes (that is, the lower frequencies have high energies while the higher frequencies have decreasing energies). However, spectral distortion for low-energy stops, fricatives, and silence regions is seen to be very high. Unconstrained dual-channel Wiener filtering is seen to retain sonorant quality while the highly distorted stops and fricatives show little improvement. ACE-I is seen to do a better job with stops and fricatives, while slightly distorting sonorants. ACE-II not only gives good overall quality, but also improves quality over all speech classes, which is especially significant over stops and fricatives as compared to unconstrained Wiener filtering and ACE-I.

3. Performance classification over individual phonemes

In the previous sections, performance improvement has been demonstrated in the form of global objective quality measures and frame-to-frame quality measures classified over broad speech classes. Phonetic labeling of the TIMIT sentences enables further classification of quality measures over individual American English phonemes. Such classifi-

cation would provide important information about the effects of noise and enhancement processing on individual acoustic–phonetic units of speech. This information could be useful not only for possible further improvement of enhancement quality but also in the development of postenhancement speech processing systems. One hundred sentences (72 male speakers and 28 female speakers) from the TIMIT database were chosen, degraded with AWGN at a SNR of 5 dB and a cross-talk level of $-\infty$ dB, and enhanced using ACE-II. Performance evaluation over individual phonemes was automated using acoustic–phonetic boundary information available for each sentence in the TIMIT database.

The set of phonemes with degraded and enhanced IS measures are shown for ACE-II in Table VII for AWGN degradation (SNR=5 dB). Good quality improvement is observed over virtually all classified phonemes for the proposed ACE-II algorithm. This is encouraging for phonemes which were particularly sensitive to noise such as glottal stops, voiced stops, nasals, liquids, and glides. Voiced fricatives on the average showed some improvement, whereas unvoiced fricatives and affricates which form only 9% of all processed phoneme frames show little improvement. The proposed ACE-II technique is seen to maintain or slightly degrade quality for all unvoiced fricatives and affricates. In an earlier study, it was seen that ACE-I (Nandkumar and Hansen, 1992, 1995; Nandkumar, 1993) further distorts unvoiced fricatives and affricates. This tendency has been corrected by ACE-II due to constraints over time and iteration being applied over all sections classified as unvoiced and transitional, and due to the enhanced ACE-LP spectrum as a starting point for the iterative algorithm. All nasals are seen to be severely degraded, and ACE-II is seen to do a good job in improving quality. ACE-II performs well for severely degraded phonemes belonging to vowels, diphthongs, stops, and whispers. Significant quality improvement was noted for highly distorted phonemes such as /oy/, /l/, /axr/, /dx/. Overall, with quality improvement for 52 of the 58 classified phonemes, and maintaining the distortion in the remaining 6 phonemes, speech quality over the entire 100 sentence set is consistently improved.

It is known that auditory fatigue arising from temporary modifications in hearing is caused by exposure to noise (Sorin and Thouin-Daniel, 1983). Moreover, large distortions during low-energy sounds of speech can be perceptually annoying or detrimental to postprocessing system performance (such as speech recognition or coding). The consistency of noise suppression (determined via informal listening tests) and consistent improvement in speech quality over phonemes and over a large set of speech suggest a possible reduction in listener fatigue and improved perceptual quality for the proposed ACE-II algorithm.

VI. CONCLUSIONS

This study extends the idea of augmenting mathematical criteria based speech enhancement with perceptual properties, by incorporating aspects of the peripheral auditory process into constrained iterative enhancement. The framework was chosen to be a frequency-domain dual-channel scenario, but similar performance has been noted for single-channel scenarios where noise characteristics are updated during nonspeech frames (noise is assumed to be stationary or slowly varying). The dual-channel enhancement framework is a previously developed two-step iterative Wiener filtering scheme. Peripheral auditory processing and lateral inhibition are simulated resulting in a unique spectral representation of speech. Constraints based on broad speech classes are developed on the auditory representation, which in turn are incorporated into the iterative enhancement scheme. The system was termed ACE-II, in order to differentiate between an earlier algorithm developed by us (ACE-I), where for the first time, auditory properties in the form of mel-cepstral parameters were integrated into an iterative enhancement framework. The ACE-II algorithm is seen to achieve superior levels of quality and noise suppression in four iterations. An advantage of ACE-II is that at each iteration a set of enhanced speech parameters are available for use along with the enhanced speech. In the case of ACE-II, the auditory representation can be transformed into the cepstral domain or spectral parametric domain for speech coding or speech recognition in noise.

The speech quality improvement due to ACE-II has been demonstrated using several novel techniques. Objective speech quality measures such as the spectral distortion based Itakura–Saito measure and the perceptually based Klatt measure have been used in different ways to illustrate quality. Objective measures over individual frames of speech along with time waveforms show the consistency of ACE-II over all sections of speech. Global quality measures show significant improvement for ACE-II over a -5- to 10-dB range of SNR for white Gaussian and colored noise cases. Time versus frequency vocal-tract spectra are also shown to demonstrate restoration of spectral features over the two domains. Finally, performance is shown to be consistent for a large set of TIMIT sentences (male and female speakers), and over classified American English phonemes. Speech quality improvement is also confirmed by informal listening tests. Several comments may be in order concerning enhanced speech quality. For white Gaussian noise, at very low SNR levels, residual noise is perceivable, though signal distortion is minimal, and speech quality is maintained. However, at SNR levels greater than or equal to 5 dB, residual noise is barely perceivable and speech quality is maintained or improved in some sections. The enhanced speech sounds very clear with ACE-II achieving consistent levels of enhancement by restoring low-energy high-frequency spectral features. ACE-II results in efficient quality improvement especially during unvoiced sounds, resulting in improved overall quality both for white and colored noise cases. While quality improvement is observed for white Gaussian, aircraft cockpit, and computer cooling fan noise sources, ACE-II produces more significant levels of improvement versus ACE-I for colored noise sources [see quality measures in Fig. 9(b) vs Fig. 10(a) and (b)], than for white Gaussian. The reasons for this are that (i) for white Gaussian noise, quality loss is not as severe as for aircraft cockpit and computer fan noise for the same SNR level. Therefore this is not as much "room" from an objective quality measure point of view for enhancement of speech in AWGN. (ii) Since ACE-II constraints are auditory

based, their impact is greater for lower frequency distortions where critical band filters are more closely spaced. In conclusion, the speech enhancement objective of achieving effective noise suppression, while maintaining or improving perceived quality, is brought closer to reality with the proposed auditory constrained iterative enhancement scheme.

ACKNOWLEDGMENTS

This work was supported in part by grants from the National Science Foundation, Grant No. IRI-9010536 and The Whitaker Foundation. The authors wish to express their appreciation to the referees for their conscientious review of the manuscript.

[1]Recorded from the interior of a Lockheed C130 aircraft cockpit at an altitude of 25 000 feet; see Hansen and Clements (1991).

Cheng, Y. M., and O'Shaughnessy, D. (**1991**). "Speech enhancement based conceptually on auditory evidence," IEEE Trans. Signal Process. **39**, 1943–1954.

Cohen, J. R. (**1989**). "Application of an auditory model to speech recognition," J. Acoust. Soc. Am. **85**, 2623–2629.

Davis, S., and Mermelstein, P. (**1980**). "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE Trans. Acoust. Speech Signal Process. **ASSP-28**, 357–366.

Ghitza, O. (**1986**). "Auditory nerve representation as a front-end for speech recognition in a noisy environment," Comput. Speech Lang. **1**, 109–130.

Ghitza, O. (**1988**). "Auditory neural feedback as a basis for speech processing," *Proceedings of the 1988 IEEE ICASSP* (IEEE, New York), pp. 91–94.

Grocholewski, S., and Krenz, R. (**1992**). "Lateral inhibition in vowel processing," in *Signal Processing VI: Theories and Applications* (Elsevier, New York), Vol. 1, pp. 299–302.

Hansen, J. H. L. (**1988**). "Analysis and compensation of stressec and noisy speech with application to robust automatic recognition," Ph.D. thesis, Georgia Institute of Technology, Atlanta, Georgia.

Hansen, J. H. L. (**1991**). "A new speech enhancement algorithm employing acoustic endpoint detection and morphological based spectral constraints," *Proceedings of the 1991 IEEE ICASSP*, Toronto, Canada (IEEE, New York), pp. 901–904.

Hansen, J. H. L. (**1994**). "Morphological constrained enhancement with adaptive cepstral compensation (MCE-ACC) for speech recognition in noise and Lombard effect," IEEE Trans. Speech Audio Process. **2**, 598–614.

Hansen, J. H. L., and Clements, M. A. (**1985**). "Objective quality measures applied to enhanced speech," J. Acoust. Soc. Am. Suppl. 1 **78**, S8.

Hansen, J. H. L., and Clements, M. (**1989**). "Stress compensation and noise reduction algorithms for robust speech recognition," *Proceedings of the 1989 IEEE ICASSP*, Glasgow, Scotland (IEEE, New York), pp. 266–269.

Hansen, J. H. L., and Clements, M. (**1991**). "Constrained iterative speech enhancement with application to speech recognition," IEEE Trans. Signal Process. **39**, 795–805.

Hansen, J. H. L., and Nandkumar, S. (**1992**). "Speech quality assessment of a real-time RPE-LTP vocoder," in *Signal Processing VI: Theories and Applications* (Elsevier, New York), Vol. 1, pp. 515–518.

Hermansky, H. (**1990**). "Perceptual linear predictive PLP analysis of speech," J. Acoust. Soc. Am. **87**, 1738–1752.

Hunt, M. J., and Lefebvre, C. (**1986**). "Speech recognition using a cochlear model," *Proceedings of the 1986 IEEE ICASSP*, Tokyo, Japan (IEEE, New York), pp. 1979–1982.

Ifukube, T., and White, R. L. (**1987**). "A speech processor with lateral inhibition for an eight channel cochlear implant and its evaluation," IEEE Trans. Biomed. Eng. **34**, 876–882.

Jennison, R. L., Greenberg, S., Kluender, K. R., and Rhode, W. S. (**1991**). "A composite model of the auditory periphery for the processing of speech based on the filter functions of single auditory-nerve fibers," J. Acoust. Soc. Am. **90**, 773–786.

Klatt, D. (**1982**). "Prediction of perceived phonetic distance from critical-band spectra: A first step," *Proceedings of the 1982 IEEE ICASSP*, Paris, France, May (IEEE, New York), pp. 1278–1281.

Lim, J., and Oppenheim, A. (**1978**). "All-pole modeling of degraded speech," IEEE Trans. Acoust. Speech Signal Process **ASSP-26**, 197–210.

Musicus, B. R. (**1979**). "An iterative technique for maximum likelihood parameter estimation on noisy data," M.S. thesis, Massachusetts Institute of Technology, Cambridge, MA.

Nandkumar, S. (**1993**). "Dual-channel iterative speech enhancement with constraints based on the auditory process," Ph.D. thesis, Duke University, Durham, NC.

Nandkumar, S., and Hansen, J. H. L. (**1992**). "Dual-channel speech enhancement with auditory spectrum based constraints," *Proceedings of the 1992 IEEE ICASSP*, San Francisco, CA, March (IEEE, New York), pp. 297–300.

Nandkumar, S., and Hansen, J. H. L. (**1994**). "Speech enhancement based on a new set of auditory constrained parameters," *Proceedings of the 1994 IEEE ICASSP*, Adelaide, Australia, April (IEEE, New York), Vol. I, pp. 001–004.

Nandkumar, S., and Hansen, J. H. L. (**1995**). "Dual-channel iterative speech enhancement with constraints on an auditory based spectrum," IEEE Trans. Speech Audio Proc. **SA-3**, 22–34.

NIST. (**1988**). "Getting started with the DARPA TIMIT CD-ROM: An acoustic phonetic continuous speech database," National Institute of Standards and Technology, Gaithersburg, Maryland, prototype as of December.

Patterson, R. D. (**1976**). "Auditory filter shapes derived with noise stimuli," J. Acoust. Soc. Am. **59**, 640–654.

Peterson, T. L., and Boll, S. F. (**1981**). "Acoustic noise suppression in the context of a perceptual model," *Proceedings of the 1981 IEEE ICASSP*, Atlanta, GA, March (IEEE, New York), pp. 1086–1088.

Quackenbush, S., Barnwell, T., and Clements, M. (**1988**). *Objective Measures of Speech Quality* (Prentice-Hall, New York).

Scharf, B. (**1970**). *Foundations of Modern Auditory Theory* (Academic, New York), Vol. 1, Chap. Critical Bands, pp. 157–201.

Seneff, S. (**1986**). "A computational model for the peripheral auditory system: Application to speech recognition research," *Proceedings of the 1986 IEEE ICASSP*, Tokyo, Japan (IEEE, New York), pp. 1983–1986.

Shamma, S. A. (**1985**). "Speech processing in the auditory system II: Lateral inhibition and the central processing of speech evoked activity in the auditory nerve," J. Acoust. Soc. Am. **78**, 1622–1632.

Sorin, C., and Thouin-Daniel, C. (**1983**). "Effects of auditory fatigue on speech intelligibility and lexical decision in noise," J. Acoust. Soc. Am. **74**, 456–466.

Stevens, S. S. (**1955**). "The measurement of loudness," J. Acoust. Soc. Am. **27**, 815–829.

Yang, X., Wang, K., and Shamma, S. A. (**1992**). "Auditory representations of acoustic signals," IEEE Trans. Inf. Theory **38**, 824–839.

Zwicker, E. (**1961**). "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," J. Acoust. Soc. Am. **33**, 248.

Zwicker, E., Flottorp, G., and Stevens, S. S. (**1957**). "Critical band width in loudness summation," J. Acoust. Soc. Am. **29**, 548–557.

Zwicker, E., and Terhardt, E. (**1980**). "Analytical expressions for critical band rate and critical bandwidth as a function of frequency," J. Acoust. Soc. Am. **68**, 1523–1525.

Zwicker, E., Terhardt, E., and Paulus, E. (**1979**). "Automatic speech recognition using psychoacoustic models," J. Acoust. Soc. Am. **65**, 487–498.

Zwislocki, J. (**1965**). *Analysis of Some Auditory Characteristics* (Wiley, New York).