Speech Enhancement Using a Constrained Iterative Sinusoidal Model

Jesper Jensen and John H. L. Hansen, *Senior Member, IEEE*

***Abstract*—This paper presents a sinusoidal model based algorithm for enhancement of speech degraded by additive broad-band noise. In order to ensure speech-like characteristics observed in clean speech, smoothness constraints are imposed on the model parameters using a spectral envelope surface (SES) smoothing procedure. Algorithm evaluation is performed using speech signals degraded by additive white Gaussian noise. Distortion as measured by objective speech quality scores showed a 34%–41% reduction over a SNR range of 5-to-20 dB. Objective and subjective evaluations also show considerable improvement over traditional spectral subtraction and Wiener filtering based schemes. Finally, in a subjective AB preference test, where enhanced signals were coded with the G729 codec, the proposed scheme was preferred over the traditional enhancement schemes tested for SNR's in the range of 5 to 20 dB.**

***Index Terms*—Sinusoidal speech model, speech and noise, speech enhancement, speech quality.**

I. Introduction

In general, the need exists for digital voice communications or automatic speech recognition systems to perform reliably in noisy environments. For example in hands-free operation of cellular phones in vehicles, the speech signal to be transmitted may be contaminated by reverberation and background noise. In many cases, these systems work well in nearly noise-free conditions, while their performance deteriorates rapidly in noisy conditions. Therefore, development of preprocessing algorithms for reducing background degradation in speech signals is of current interest.

In the past, a number of single-microphone speech enhancement algorithms have been proposed. A number of these are discussed in overview studies by Lim and Oppenheim [15], Ephraim [3], and Hansen [7]. These include variants of spectral subtraction [2], methods based on all-pole modeling [8], [15], subspace model based methods [5], [14], schemes based on hidden Markov models [4], [25], and algorithms that exploit masking effects, [26], [27]. Some methods have focused on iterative speech modeling schemes that employ speech production-like spectral constraints (e.g., Auto-LSP [8], [10]), auditory based constraints (e.g., ACE-I,II [19], [11]), and noise dependent constraints (e.g., [9], [20]). These methods however assume a linear all-pole based speech model. Although successful for speech coding (e.g., [16]), and speech signal modification (e.g., [24]), the sinusoidal model has not received the same level of attention in a speech enhancement context [1], [17], [22], [23].

In this paper, we propose a sinusoidal model based algorithm for enhancement of speech degraded by additive broadband noise. Adopting a similar idea as that used in [8], for an all-pole model, the present algorithm exploits the notion that during the speech production process, the vocal tract transfer function varies continuously and relatively slowly with time. Furthermore, in most voiced speech regions, the fundamental frequency (F_0) varies relatively slowly as well. The aim of the proposed enhancement scheme is to improve the quality of the enhanced speech signal, by exploiting this knowledge of the signals origin to apply speech production constraints in the enhancement process.

This paper is structured as follows. Section II introduces the signal model and notation. In Section III each step of the enhancement algorithm is considered in detail. Algorithm performance using objective and subjective testing with signals degraded with additive white Gaussian noise (AWGN) are presented in Section IV. The evaluation is based on an objective speech quality measure as well as subjective listener preference tests. Finally, in Section V, we summarize our study and identify directions for future research.

II. Signal Model and Notation

In this paper, we assume that speech is corrupted by additive broadband noise as follows:

$$\mathbf{x} = \mathbf{s} + \mathbf{n}$$

where $\mathbf{x}, \mathbf{s}$ and $\mathbf{n}$ denote the noisy speech, clean signal, and noise component, respectively. Furthermore, the noise is assumed stationary, such that the estimate of the noise spectrum in silence regions is still valid during speech activity. Finally, it is assumed that the noise is uncorrelated with the speech signal. The noise level is measured with the global SNR defined as

$$\text{SNR [dB]} = 10\log_{10}(\mathbf{s}^T\mathbf{s}/\mathbf{n}^T\mathbf{n})$$

where T denotes vector transposition.

Manuscript received April 25, 2000; revised June 28, 2001. This work was conducted at CSLR during a 1999/2000 visiting internship, with financial support provided by the Danish Government and CPK. This work was also supported in part by SPAWAR under Grant N66001-92-0092. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Hsiao-Chuan Wang.

J. Jensen was with Center for PersonKommunikation (CPK), Aalborg University, Denmark. He is now with the Information and Communication Theory Group, Delft University of Technology, Delft, The Netherlands (e-mail: j.jensen@its.tudelft.nl).

J. H. L. Hansen is with Center for Spoken Language Research (CSLR), Robust Speech Processing Laboratory (RSPL), University of Colorado, Boulder, CO 80309-0594 USA (e-mail: john.hansen@colorado.edu; http://cslr.colorado.edu/rspl/).

Publisher Item Identifier S 1063-6676(01)08236-0.

The enhanced speech signal is modeled as a sum of sinusoids on a frame-by-frame basis as

$$\hat{\mathbf{s}}_m^{(i)}(n) = \sum_{k=1}^{K_m} a_{k,m}^{(i)} \cos\left(\omega_{k,m}^{(i)} n + \phi_{k,m}^{(i)}\right), \quad n \in [-N, \ldots, 0, \ldots, N] \quad (1)$$

where $\hat{\mathbf{s}}_m$ is the enhanced speech, K_m is the number of sinusoids used; and $a_{k,m}, \omega_{k,m}$, and $\phi_{k,m}$ are kth sinusoidal amplitude, angular frequency, and phase, respectively, for frame m. Finally, since the proposed algorithm is iterative, the superscript (i) is used to denote the iteration number, (e.g., $\mathbf{a}_m^{(i)}$ represents the amplitude vector for frame m at iteration i).

III. Enhancement Algorithm

A. Algorithm Overview

It is well-known that most clean voiced speech segments can be represented accurately with the model from (1) such that amplitudes and frequencies are relatively smooth functions of time (provided frequent parameter updates, e.g., every 10 ms). However, when estimated from a noisy signal, using peak-picking of the FFT magnitude spectrum as described in [16, p. 143], the sinusoidal parameters show a much more unstructured behavior. The aim of the proposed algorithm is to obtain an enhanced signal, where the amplitudes and frequencies evolve smoothly with time, as is expected in a clean voiced speech signal.

In Fig. 1, the proposed algorithm is outlined. First, the noisy speech signal $\mathbf{x}$ is divided into overlapping analysis frames. Initial sinusoidal parameters are estimated for each frame. Amplitudes and frequencies (in voiced regions) are refined iteratively, while the initial phase values are not modified. Finally, enhanced frames are synthesized using estimated and constrained smoothed parameters for the Lth iteration in an overlap-and-add manner in order to generate the enhanced speech signal.

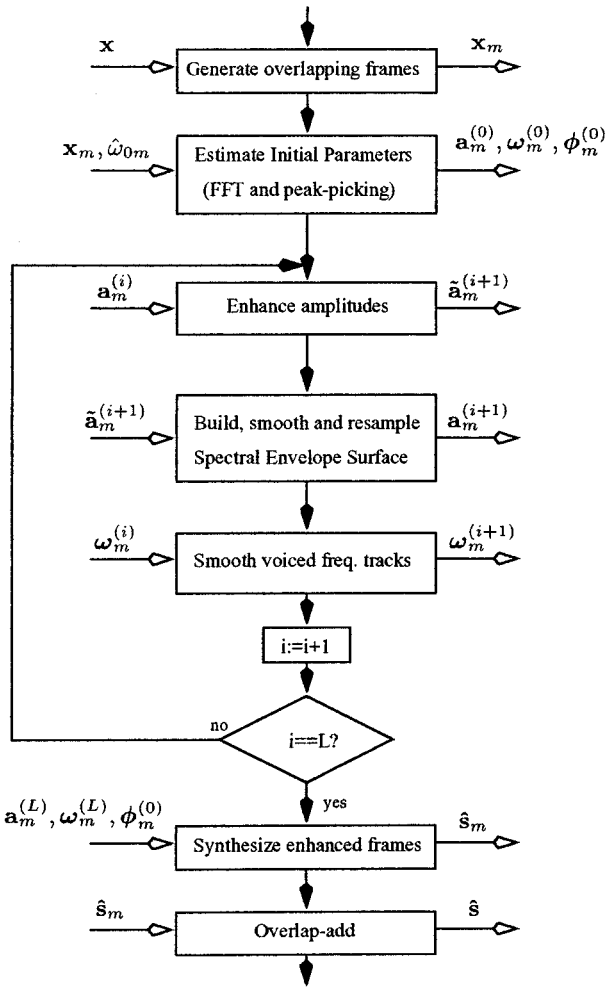

Fig. 1. Block diagram of the enhancement algorithm. Black arrows: algorithm flow and white arrows: parameter input/output.

B. Estimation of Initial Sinusoidal Model Parameters

Each of the noisy signal frames are Hanning windowed, zero-padded, and transformed with a high-order FFT. In noise-only regions, an estimate of the noise amplitude spectrum is calculated as the average FFT magnitude spectrum across consecutive analysis signal frames. The updating of this smoothed, high-order FFT magnitude spectrum is terminated in regions where speech is present.

In signal frames with speech (and noise), the goal is to represent the speech signal with the sinusoidal model in (1). The first step toward this goal is to select all spectral peaks of each FFT magnitude spectrum. These peaks represent candidate triplets $(a_{k,m}, \omega_{k,m}, \phi_{k,m})$ from which the speech signal relevant triplets are selected.

For voiced speech frames, the relevant signal peaks are mainly due to periodicity of the speech signal, while other peaks are related to analysis window side lobes or noise. Using a rough F_0 estimate $\hat{\omega}_{0m}$ in frame m, the frequency axis is divided into nonoverlapping bands of the form $[(1/2)\hat{\omega}_{0m}; (3/2)\hat{\omega}_{0m}), [(3/2)\hat{\omega}_{0m}; (5/2)\hat{\omega}_{0m}), \ldots$, and a peak from each band is selected. For each of these frequency bands an SNR is estimated as

$$\mathrm{SNR}_{k,m} = \frac{\bar{a}_{k,m}^2}{\bar{N}_{k,m}^2} - 1$$

where $\bar{a}_{k,m}$ denotes the highest amplitude in frequency band number k, and $\bar{N}_{k,m}$ denotes the value of the noise amplitude spectrum sampled at the frequency bin associated with $\bar{a}_{k,m}$. In bands with SNR above a prespecified value, T_{SNR} (in the reported simulations a value of $T_{\mathrm{SNR}} = 10$ dB was used), the highest peak is selected. In all other bands, the peak closest to $k\hat{\omega}_{0m}$ is selected, where $k\hat{\omega}_{0m}$ denotes the middle of the frequency band in question. Using this peak-picking strategy in low SNR bands generally works better than selecting the highest peak in all bands.

While for unvoiced frames it is more difficult to decide which peaks best describe the clean speech signal, it is well-known that relatively many sinusoidal components are needed to represent noise-like speech sounds (e.g., unvoiced fricatives) well [16]. Thus, one reasonable approach might be to keep all the candidate peaks for the later processing stages. However, in the presence of noise, some voiced frames will typically be mis-classified as unvoiced, and experiments have shown that the side lobe peaks between harmonic peaks in these mis-classified

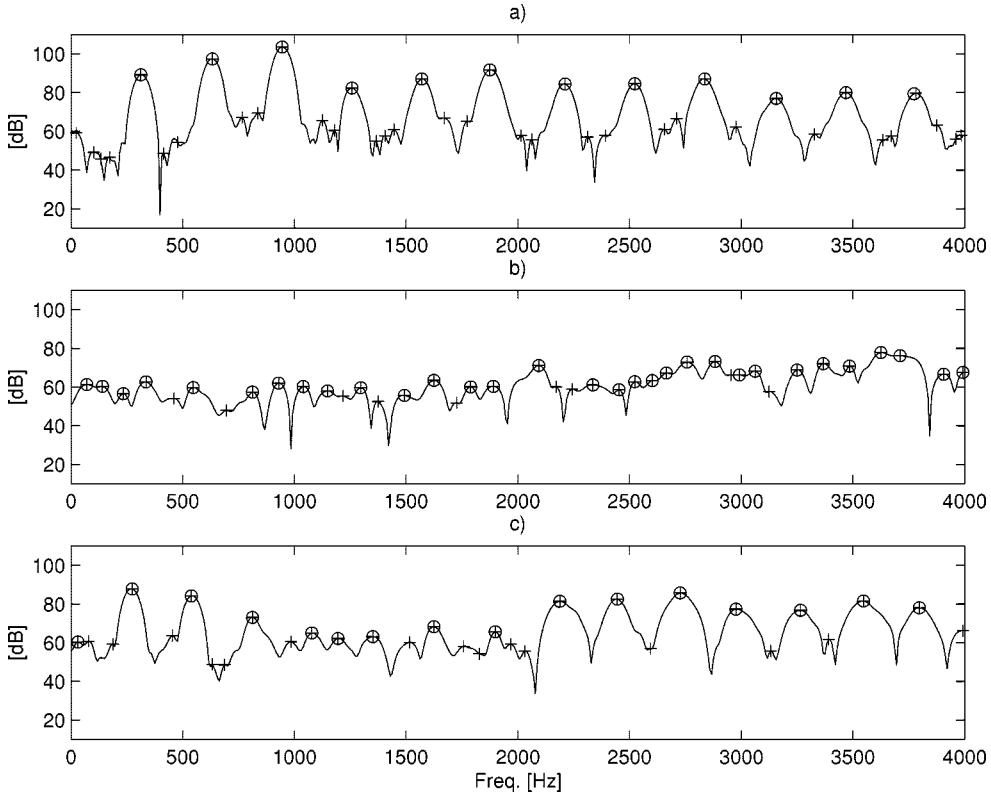

Fig. 2. FFT Magnitude spectra with candidate peaks ('+') and selected peaks ('∘'). (a) Voiced frame /ae/ in "had" ($\hat{F}_0 = 314$ Hz), (b) Unvoiced frame /s/ in "she", (c) Misclassified voiced frame /i/ in "she".

voiced frames should be discarded to avoid audible artifacts in the enhanced signal (We mention in passing that mis-classified voiced frames constitute more than 90% of the observed voicing errors). In order to increase robustness against mis-classified voiced frames, the following sidelobe detection scheme is applied. Side lobe peaks are detected by calculating the slopes of the lines between each peak and its two neighbors. If neighbor peaks have higher amplitude and steeper slopes than a pre-specified value, T_{unv} (simulations have shown a value of $T_{unv} = 0.07$ dB/Hz to work well), the peak represents a side lobe and should be discarded. The sidelobe detection scheme succeeds in discarding most of the sidelobe peaks in mis-classified voiced frames, while the perceptual quality of the truly unvoiced frames remains essentially unchanged. In principle, sidelobe detection could also be used for voiced frames in combination with the F_0 based scheme described above. However, this combined approach did not lead to improved performance, and was thus abandoned.

Fig. 2 illustrates the peak-picking procedure for a female speech signal sampled at 8 kHz and degraded with additive white Gaussian noise at a global SNR of 20 dB. Fig. 2(a) shows the FFT magnitude spectrum for a voiced signal frame (/ae/ in "had"). The peaks of the magnitude spectrum are marked with '+'; these represent potential sinusoidal components. In addition, the peaks selected in the peak-picking procedure for voiced speech described above are marked with '∘'. Clearly, all peaks representing harmonics have been picked, while all other peaks have been discarded. Fig. 2(b) illustrates the peak-picking procedure for an unvoiced signal frame (/s/ in "she"). Here, the peaks are generally spaced less structured compared to the voiced case. For illustrative purposes, we deliberately misclassified a voiced frame (/i/ in "she") as unvoiced in Fig. 2(c), and the unvoiced peak-picking procedure was applied. From this figure it is clear that using the unvoiced peak-picking procedure here causes most of the harmonic peaks to be picked, while most of the sidelobe peaks between the harmonics have been discarded. Thus, the peak-picking scheme in unvoiced frames is relatively robust toward voicing errors.

Using this peak-picking approach with clean speech signals results in modeled signals of high perceptual quality, nearly indistinguishable from the originals.

C. Enhancement and Amplitude Smoothing Constraints

The procedure for enhancement of sinusoidal amplitudes consists of two steps. The first step aims at reducing the noise, while the second step ensures that amplitudes evolve smoothly with time.

In the first step, enhanced amplitudes are estimated using a weighted average between amplitude values from the previous iteration and their Wiener filtered counterparts,

$$\tilde{a}_{k,m}^{(i+1)} = W_w H_{k,m}^{(i)} a_{k,m}^{(i)} + (1 - W_w) a_{k,m}^{(i)}, \quad \text{where}$$
$$H_{k,m}^{(i)} = \frac{\left(a_{k,m}^{(i)}\right)^2}{\left(a_{k,m}^{(i)}\right)^2 + N_{k,m}^2} \tag{2}$$

where

$H_{k,m}^{(i)}$ value of the Wiener filter at frequency $\omega_{k,m}$;
$a_{k,m}^{(i)}$ estimated sinusoidal amplitude at iteration (i);

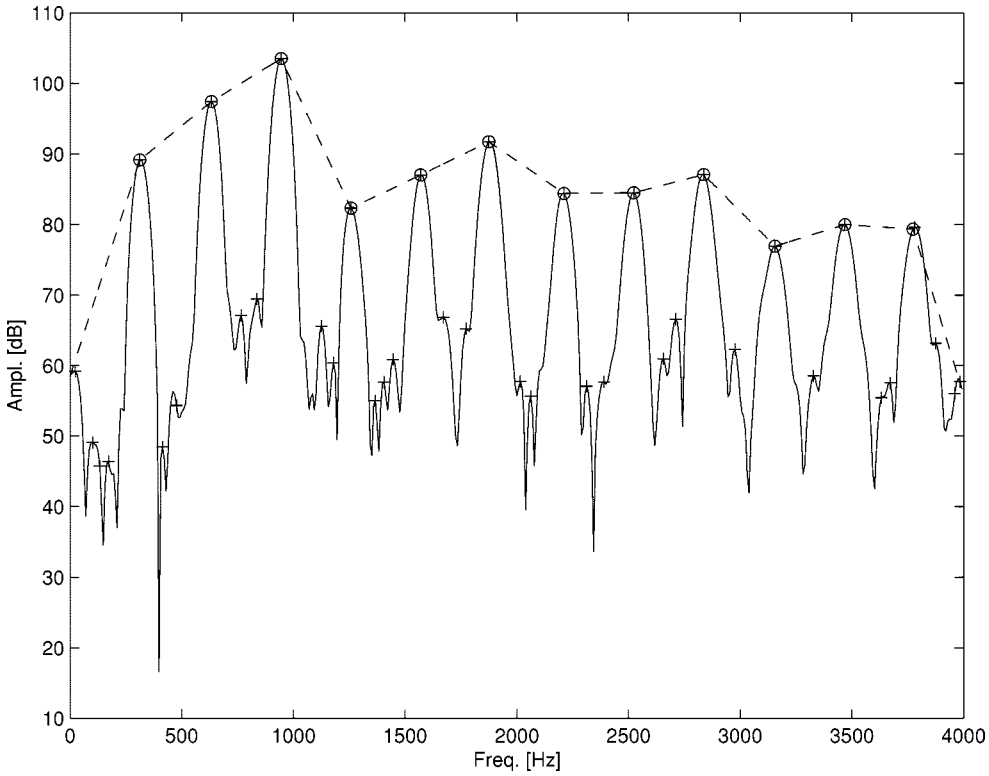

Fig. 3. Spectral envelope approximation (——) for clean, voiced, speech segment obtained by linear interpolation between spectral peaks in the log-amplitude domain.

$N_{k,m}$ noise amplitude spectrum sampled at the frequency bin corresponding to $\omega_{k,m}^{(i)}$.

The weight factor W_w controls the amount of noise reduction at each iteration.

In the second step, enhanced amplitude values are used to obtain a vocal tract spectral envelope approximation for frame m by linear interpolation between the points

$$\left(\omega_{k,m}^{(i)}, \log\left(\tilde{a}_{k,m}^{(i+1)}\right)\right), \quad k \in [1, K_m].$$

As suggested in [16, p. 143], interpolation is performed in the log-amplitude domain. Fig. 3 shows an example of a vocal tract envelope approximation. Spectral envelope approximations for consecutive frames constitute what we call the spectral envelope surface (SES), which reflects a grid of $S_{j,m}^{(i)}$ points; see Fig. 4 for an example. Since the vocal tract transfer function presumably varies continuously and relatively slowly with time, and the SES is a spectral interpretation of this evolution, the SES should be a smooth surface.

For this reason, a smoothing procedure is applied to the estimated SES, by calculating a weighted average between each of the points on the SES and the eight neighboring points

$$\hat{S}_{j,m}^{(i)} = W_s S_{j,m}^{(i)} + \frac{1}{8}(1 - W_s) \times \left(\sum_{p=-1}^{1} \sum_{q=-1}^{1} S_{j+p,m+q}^{(i)} - S_{j,m}^{(i)} \right). \quad (3)$$

The weight factor W_s is adjusted close to $1/9$ in low SNR regions of SES and close to 1 in high SNR regions. With this scheme, amplitudes at high SNR are not greatly modified, while low SNR amplitudes can change more during the smoothing process. After smoothing of the SES, enhanced and smoothed amplitude values $\mathbf{a}_{k,m}^{(i+1)}$ are obtained by resampling the smoothed SES at points corresponding to the frequencies $\omega_{k,m}^{(i+1)}$. We note that the amplitude smoothing procedure (3) makes use of a one frame look-ahead per iteration.

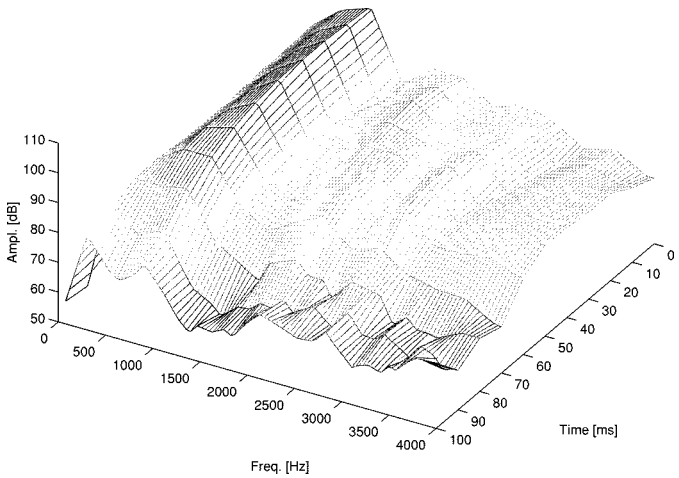

Fig. 4. SES obtained from consecutive spectral envelope approximations of clean voiced speech segments.

D. Smoothing of Frequency Tracks

In noisy conditions, only a rough fundamental frequency (F_0) estimate can be expected. This issue coupled with the peak-picking procedure described above, results in rough frequency tracks for voiced regions, particularly at higher

frequencies. However, clean voiced speech can be represented with the model in (1), such that the frequency tracks evolve smoothly with time. For this reason, a smoothing procedure is applied to the sinusoidal frequencies in voiced regions. Smoothed frequencies are determined by linking each frequency to a frequency in the previous and following frames, and calculating a weighted average

$$\omega_{k,m}^{(i+1)} = W_f \omega_{k,m}^{(i)} + \frac{1}{2}(1 - W_f)\left(\omega_{k,m-1}^{(i)} + \omega_{k,m+1}^{(i)}\right). \quad (4)$$

This smoothed frequency value is used, unless the relative frequency change from frame to frame exceeds a threshold T_f (informal evaluations showed that $T_f = 20\%$ performed well). In this case, a pitch halving/doubling error may have occurred, and no smoothing is performed, i.e., $\omega_{k,m}^{(i+1)} = \omega_{k,m}^{(i)}$. The procedure for smoothing of frequency tracks, shown in (4), requires a look-ahead of one frame per iteration.

Fig. 5 illustrates the frequency smoothing scheme. Fig. 5(a) shows a clean female speech signal: "She had your dark suit in greasy wash water all year". This signal is degraded with AWGN, $\mathrm{SNR} = 10$ dB, and voicing and F_0 is estimated, Fig. 5(b). Fig. 5(c) shows the frequencies picked in the peak-picking procedure for each frame of the noisy signal. Clearly, the peak-picking procedure results in rough frequency tracks, especially at high frequencies. Fig. 5(d) shows the result of the frequency smoothing scheme after termination of the iterative process. Obviously, the frequency smoothing scheme results in smoothly evolving frequency tracks.

E. Termination Criterion

An open question in this, as well as other constrained iterative schemes, is when to stop the iterative process. This is a trade-off between noise reduction and speech processing artifacts. If the iterations are too few, the enhanced signal is noisier than necessary, and if too many iterations are performed, parts of the speech signal could have artifacts introduced.

In order to develop a termination criterion, enhanced speech signals were generated after each iteration. The objective speech quality of each of the enhanced signals was estimated by using the symmetric log-likelihood ratio (LLR) quality measure defined in ([21, p. 49]), which has been shown to correlate fairly well with subjective quality [21].

Initially, the optimum terminating iteration number (in terms of LLR) varied both with input SNR and individual speaker. However, further evaluations showed that with a proper selection of the weight factors W_w, W_s and W_f defined above, optimum enhanced signals could be achieved after $L = 7$ iterations, almost independent of speaker and input AWGN level. Furthermore, in cases where the optimum LLR was reached after $L = 6$ or 8 iterations, the optimum point was usually broad, so the performance loss using $L = 7$ iterations was insignificant.

F. Signal Resynthesis

After terminating the iterations, enhanced signal frames are generated by inserting the enhanced/smoothed sinusoidal amplitudes and frequencies with the original noisy phase values into (1). The enhanced speech is then synthesized by overlap-adding the enhanced frames using a triangular synthesis window (i.e., the last step of the flow diagram in Fig. 1).

IV. Algorithm Evaluation

For the proposed enhancement scheme, voicing decisions were made based on frame energy and zero-crossing rate. In voiced frames, fundamental frequency estimates were obtained using an improved correlation based pitch estimator. The analysis frame length used in the enhancement algorithm was 200 samples (25 ms) and new frames were selected every 80 samples (10 ms). In the procedures for estimating the noise spectrum and initial model parameters an FFT order of 1024 was used. The triangular window for overlap-add synthesis had a length of 160 samples (20 ms). In all simulations, an estimate of the noise amplitude spectrum was calculated from a noise-only region preceding each test signal. To be more specific, the noise spectrum estimate was calculated as the average FFT magnitude spectrum across the first 6 analysis frames.

A. Objective Evaluation

Four test sentences, two female and two male, sampled at 8 kHz were randomly selected from the TIMIT database. In these test signals, speech was present 93–95% of the time. The test sentences were degraded with AWGN at global SNR levels of 20, 15, 10 and 5 dB, and enhanced using $L = 7$ iterations of the proposed scheme.

For comparison, enhanced signals were generated with the spectral subtraction method in [2] using halfwave rectification of enhanced spectral magnitudes, and magnitude averaging across three signal frames. The spectral subtraction scheme made use of signal frames with a length of 160 samples (20 ms), and an overlap of 80 samples (10 ms) between consecutive frames. In addition, enhanced signals were generated with the unconstrained, iterative Wiener filtering approach in [15], where speech spectral envelopes were represented with LPC models of order 10, while the noise amplitude spectrum was represented with an LPC model of order 2. Up to ten enhancement iterations were performed for each signal, and the best enhanced signal in terms of LLR selected. We note that this selection procedure gives an enhancement performance, which can generally not be achieved in practice, since the optimum iteration number is signal dependent and not known in advance [15].

For objective quality assessment of sentences, the average symmetric LLR value was calculated from frames of length 240 samples (30 ms) taken with an overlap of 75%.

Fig. 6 compares the objective speech quality of the enhanced signals. Fig. 6(a), (b), and (c) show average LLR values from unvoiced regions, voiced regions, and nonsilence regions (i.e., unvoiced, voiced and transitional speech regions with silence removed), respectively. Fig. 6(a) shows that spectral subtraction and unconstrained Wiener filtering perform poorly in unvoiced segments, while some improvement can be observed with the proposed method. In voiced segments, Fig. 6(b), all methods improve speech quality, but the proposed method has the largest degree of improvement. Fig. 6(c) shows that, in general, better

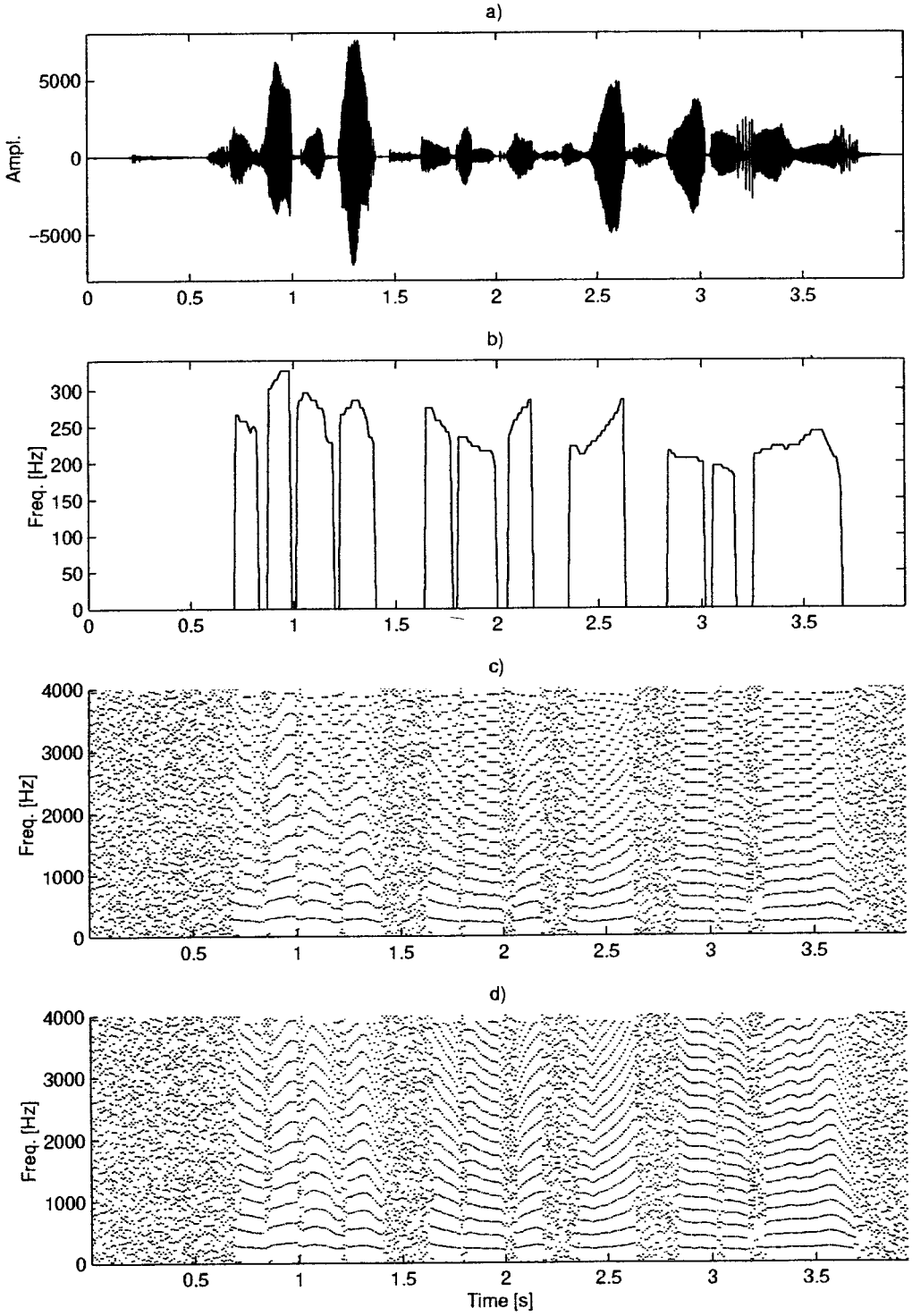

Fig. 5. Smoothing of frequency tracks. (a) Noise-free female signal. (b) F_0 and voicing estimates based on noisy signal with global $\mathrm{SNR} = 10$ dB. (c) Frequencies picked during peak-picking procedure. (d) Frequencies after iterative enhancement.

overall performance can be obtained with the proposed method. Furthermore, there is little performance difference in using F_0 and voicing estimates based on noisy signals instead of clean signals. In particular, the average LLR with noisy F_0 estimates is only 1% higher than with clean F_0 estimates.

Fig. 7 shows spectrograms of a clean female signal, the signal degraded with AWGN, $\mathrm{SNR} = 10$ dB, and the signal enhanced with the proposed method, respectively. It is clear that the proposed method has nice advantages in obtaining, and ensuring speech-like periodic structure during voiced-speech sections (e.g., compare spectral tracks between 1.3–1.7 s).

B. Informal Listening

In general, signals enhanced with the proposed method have high subjective quality in voiced regions. However, at times, enhanced voiced regions are slightly reverberant, especially for

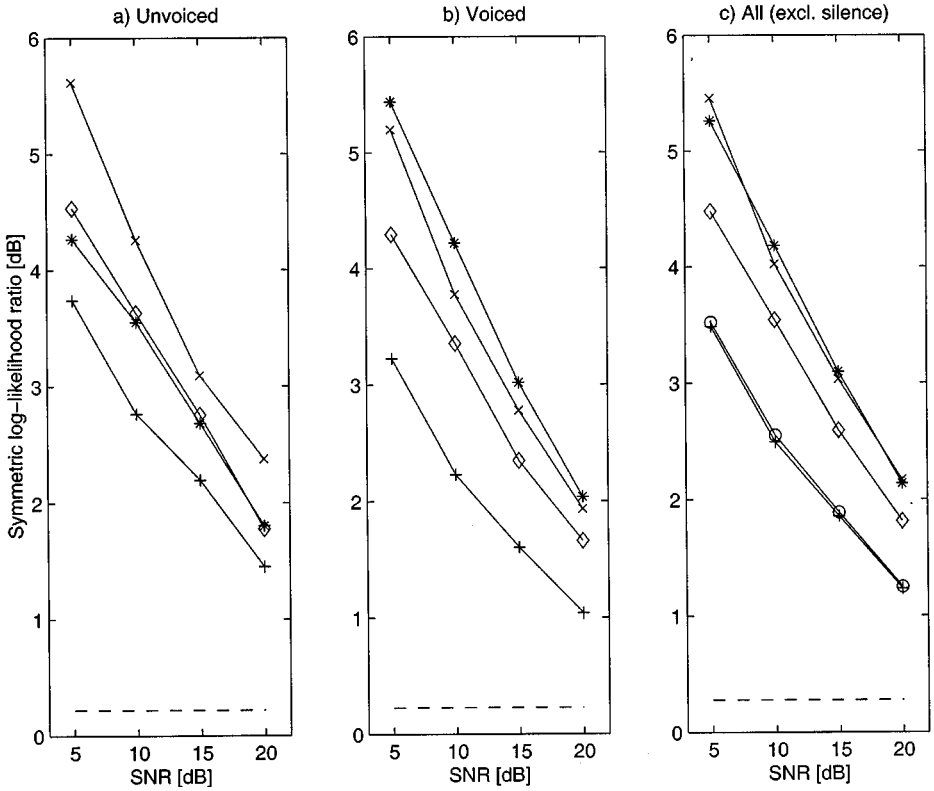

Fig. 6. Enhancement performance in terms of LLR. (a) Unvoiced segments, (b) voiced segments, and (c) all segments (excluding silence). Noisy signals: $*$. Spectral Subtraction: $\times$. Unconstrained Wiener: $\diamond$. Proposed method with F_0 and voicing from clean signals: $+$. Proposed method with F_0 and voicing from noisy signals: $\circ$. Clean signals represented with the sinusoidal model (1): —.

high pitch female speakers. This reverberance may partly be due to the noisy phases used in the resynthesis process.

In unvoiced regions, the enhanced signal seems to have lower subjective quality; in particular, stops tend to sound slightly "muffled." The reason may be that the local SNR in unvoiced regions is typically much lower than in voiced regions. Moreover, the sinusoidal model in (1) is not well-suited for representing speech segments with rapid amplitude changes, when model parameters are based on FFT peak picking [13].

In order to study the influence of inaccurate voicing and F_0 information on perceived signal quality, noisy signals were enhanced using F_0 and voicing estimates from clean and noisy signals, respectively. Informal listening confirms the objective test results from Fig. 6(c), that the algorithm is not particularly sensitive to accurate F_0 and voicing information, (i.e., a rough F_0-contour is adequate for near-optimum performance). However, the algorithm at times is sensitive to bursts of F_0 halving/doubling errors; sometimes, such bursts cause audible artifacts in the enhanced speech signal. This does not occur often, since only one series of 10–15 consecutive frames containing pitch errors was observed in the approximately 5600 frames employed in the objective evaluation. Also, such errors almost always occurred for high dynamic pitch range female speakers. Further effort in pitch tracking error detection would be useful in eliminating this issue.

C. AB-Comparison Test With G729-Coded Speech

Often, the performance of a speech enhancement algorithm as a front-end for a speech coding algorithm is of importance. Other studies have shown that enhancing speech in combination with speech coding methods, such as CELP, can result in a measureable improvement in speech quality [18]. It has also been demonstrated that changing noise and language conditions seriously impacts voice coders such as the GSM (RPE-LTP) coder [12]. For this reason, an informal AB-preference test with coded speech was conducted. Three different clean signals from different speakers, two male and one female, were degraded with AWGN at global SNR levels of 20, 15, 10, and 5 dB. The noisy signals were then enhanced with the proposed scheme using F_0 and voicing estimates from the noisy signals. Subsequently, the enhanced signals were encoded and decoded with the G729 8 kbit/s CS-ACELP speech codec [6]. Ten subjects were asked to compare these signals to coded versions of the unprocessed noisy signals, signals enhanced with spectral subtraction, and signals enhanced with unconstrained Wiener filtering. The signals were presented in random order over headphones, and the specific processing of the A and B signal was unknown to the subjects. The purpose here was to determine if the enhancement method could provide a degree of improvement over other enhancement methods, and to establish listener preference for unprocessed versus processed speech signals. This second point is important because speech coding algorithms generally assume clean input speech, and react differently when noise corrupts the assumed speech signal model. In contrast, while speech enhancement methods can reduce the background noise prior to voice coding [18], the resulting processed speech at times can also differ from an assumed clean speech model. From a general perspective in listener testing for voice communications,

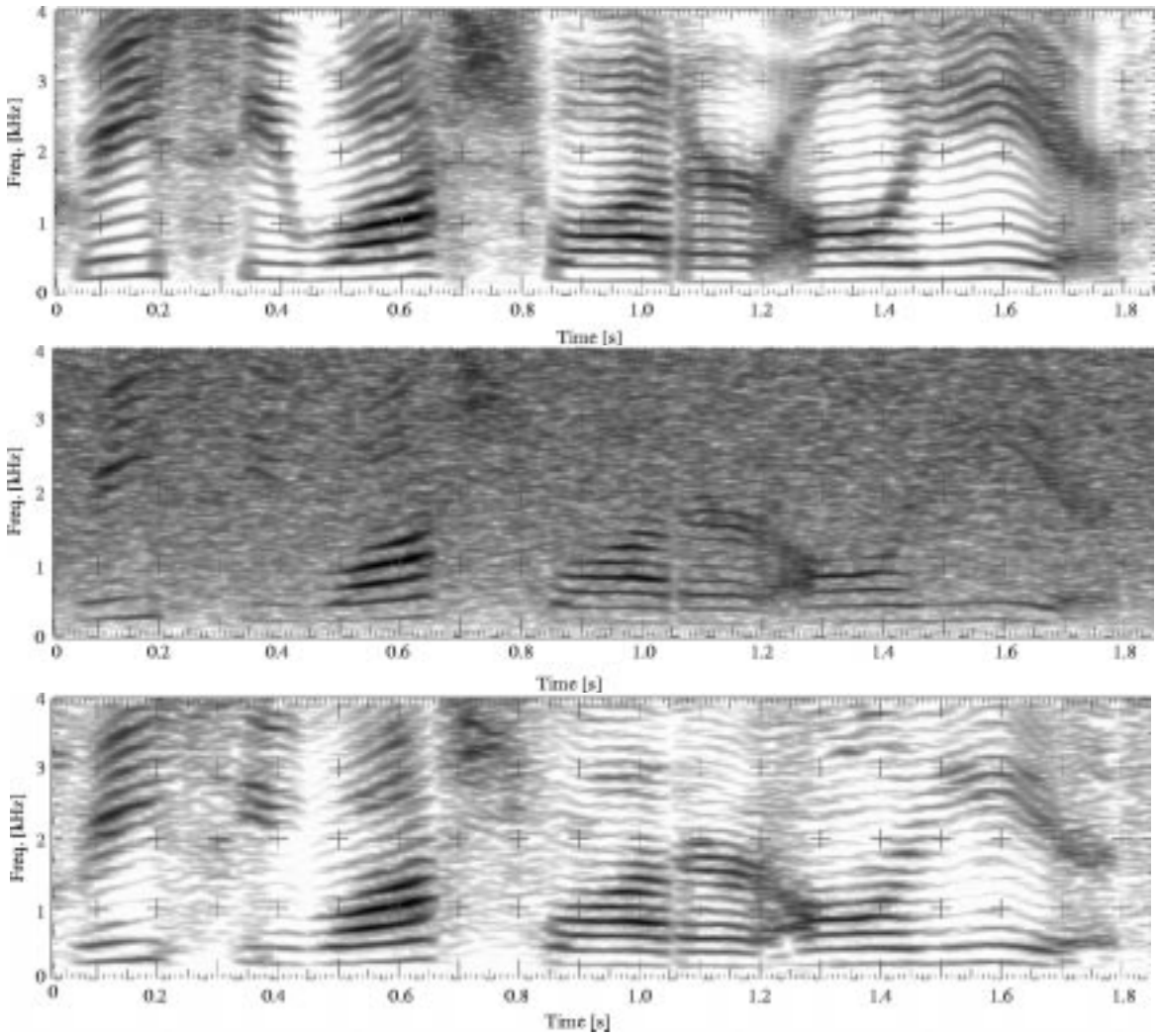

Fig. 7. Top: Spectrogram of the clean female speech signal: "dirty wash water all year". Middle: Spectrogram of the speech signal corrupted by additive white Gaussian noise, global $\mathrm{SNR} = 10$ dB. Bottom: Spectrogram of the enhanced speech signal using the proposed method.

listeners often concentrate on longer sustained periods of noise between speaker phrases. While it is important to suppress the noise between longer speech segments over time, it is also critical for ease in voice communication understanding that the resulting speech quality across individual phones be as high as possible.

In Table I, we summarize results that show the preference toward the combined proposed enhancement scheme plus CS-ACELP coder for different test conditions. A total of 30 tests were performed for each condition, where listeners were asked to choose between (a) the proposed enhancement algorithm plus CS-ACELP coder combination versus original noisy speech plus CS-ACELP coding, (b) the proposed enhancement algorithm plus CS-ACELP coder combination versus traditional spectral subtraction plus CS-ACELP coder, and (c) the proposed enhancement algorithm plus CS-ACELP coder combination versus traditional unconstrained Wiener filtering plus CS-ACELP coder. From this table it is clear that the proposed enhancement scheme plus CS-ACELP coder combination is generally preferred over CS-ACELP coded speech signals with either spectral subtraction or unconstrained Wiener filtering front-ends, or nonenhanced (noisy) signals with CS-ACELP coding alone.

The coded/enhanced signals can be described in general terms as follows. For the nonenhanced, noisy signals, the CS-ACELP codec tends to reduce the noise energy. However the residual noise component in the coded signals have a 'hissing' characteristic and is generally more disturbing than the original noise. Signals enhanced with spectral subtraction have a high level of 'musical noise'. The codec tends to reproduce this residual noise fairly faithfully, though some slight changes in the residual noise occurs resulting in coded signals of poor subjective quality. The Wiener filtering scheme reduces the noise energy without changing the character of the remaining noise noticeable. When coded, these signals suffer from the same 'hissing' noise as with nonenhanced, coded signals, although at a much lower energy level. For signals enhanced with the proposed scheme, the coding process makes the signals sound slightly more 'muffled'. However, in some cases the CS-ACELP codec attenuates some of the artifacts

TABLE I
AB PREFERENCE TEST WITH A COMPARISON FOR: (a) THE PROPOSED ENHANCEMENT SCHEME PLUS CS-ACELP CODER VERSUS THE ORIGINAL NOISY SIGNAL PLUS CS-ACELP CODER, (b) THE PROPOSED ENHANCEMENT SCHEME PLUS CS-ACELP CODER VERSUS SPECTRAL SUBTRACTION PLUS CS-ACELP CODER, AND (c) THE PROPOSED ENHANCEMENT SCHEME PLUS CS-ACELP CODER VERSUS UNCONSTRAINED WIENER FILTERING PLUS CS-ACELP CODER

AB Listener Preference Test			
SNR [dB]	(a) Noisy Signal	(b) Spectral Subtraction	(c) Unconstrained Wiener Filter
20	29/30	30/30	30/30
15	30/30	30/30	30/30
10	30/30	30/30	30/30
5	29/30	30/30	26/30

introduced by the proposed enhancement scheme, e.g., the reverberant artifacts in female speech.

V. CONCLUSIONS AND DISCUSSION

An iterative sinusoidal model-based scheme has been proposed for enhancement of speech degraded by additive broad-band noise. Smoothness constraints were imposed on the sinusoidal amplitudes and frequencies (in voiced regions) in order to ensure a parameter behavior similar to that observed in clean speech. A performance evaluation was conducted on speech from male and female speakers using objective speech quality measures and a subjective AB preference test. It was shown that the proposed method is effective in producing speech-like sinusoidal trajectories in low and high-frequency voiced speech domains. A measureable improvement in objective speech quality was observed compared to more traditional speech enhancement methods such as spectral subtraction and unconstrained Wiener filtering. It should be noted that the improvement in an objective speech quality measure is only useful if the measure is well correlated with the noise sources of interest, and that while there was a 34–41% reduction in distortion using the LLR measure, further evaluations would be necessary to establish performance for other types of noise. This quality improvement was also seen for the case of a subjective AB-comparison test with G729-coded signals.

In terms of extensions to this work, further performance improvement in unvoiced speech regions is clearly a topic for continued research. This could, for example, be performed by introducing an alternative signal model in these regions. In addition, the current implementation of the proposed algorithm introduces an algorithmic delay of approximately 90 ms. This delay is related to the smoothing procedure applied to model amplitudes and frequencies, since the algorithm needs to maintain a look-ahead frame of speech to ensure speech-like structure across iterations. However, due to the nonstationarity of speech we find it reasonable to assume that incidents in the speech signal 90 ms ahead in time should not influence the processing of the present signal frame. For this reason, we expect that the algorithmic delay can be reduced without sacrificing the quality of the enhanced speech signal. The study of other smoothing procedures with smaller delay is a topic for future research. Furthermore, alternatives to reducing the computationally requirements of the parameter smoothing scheme are currently an area of interest.

REFERENCES

[1] D. V. Anderson and M. A. Clements, "Audio signal noise reduction using multi-resolution sinusoidal modeling," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1999, pp. 805–808.

[2] S. F. Boll, "Suppression of acoustic noise in speech using spectral subtraction," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-27, pp. 113–120, Apr. 1979.

[3] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, pp. 1526–1555, Oct. 1992.

[4] Y. Ephraim, D. Malah, and B.-H. Juang, "On the application of hidden Markov models for enhancing noisy speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 37, pp. 1846–1856, Dec. 1989.

[5] Y. Ephraim and H. L. Van Trees, "A signal subspace approach for speech enhancement," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 251–266, July 1995.

[6] *Coding of Speech at 8 kbit/s Using Conjugate-Structure Algebraic-Code-Excited Linear-Predictive (CS-ACELP) Coding*.

[7] J. H. L. Hansen, "Speech enhancement," in *Encyclopedia of Electrical and Electronics Engineering*: Wiley, 1999, vol. 20, pp. 159–175.

[8] J. H. L. Hansen and M. A. Clements, "Constrained iterative speech enhancement with application to speech recognition," *IEEE Trans. Signal Processing*, vol. 39, pp. 795–805, Apr. 1991.

[9] J. H. L. Hansen and L. Arslan, "Markov model based phoneme class partitioning for improved constrained iterative speech enhancement," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 98–104, Jan. 1995.

[10] J. H. L. Hansen and L. M. Arslan, "Robust feature-estimation and objective quality assessment for noisy speech recognition using the credit card corpus," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 169–184, May 1995.

[11] J. H. L. Hansen and S. Nandkumar, "Robust estimation of speech in noisy backgrounds based on aspects of the auditory process," *J. Acoust. Soc. Amer.*, vol. 97, no. 7, pp. 3833–3849, June 1995.

[12] ——, "Objective quality assessment and the RPE-LTP vocoder in different noise and language conditions," *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 609–627, Jan. 1995.

[13] J. Jensen, S. H. Jensen, and E. Hansen, "Exponential sinusoidal modeling of transitional speech segments," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, pp. 473–476, 1999.

[14] S. H. Jensen, P. C. Hansen, S. D. Hansen, and J. A. Sorensen, "Reduction of broad-band noise in speech by truncated QSVD," *IEEE Trans. Speech, Audio Processing*, vol. 3, pp. 439–448, Nov. 1995.

[15] J. S. Lim and A. V. Oppenheim, "All-pole modeling of degraded speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-26, no. 3, pp. 197–210, 1978.

[16] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, Eds. Amsterdam, The Netherlands: Elsevier, 1995, ch. 4.

[17] D. Morgan, B. George, L. Lee, and S. Kay, "Cochannel speaker separation by harmonic enhancement and suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 407–424, Sept. 1997.

[18] S. Nandkumar, J. H. L. Hansen, and R. J. Stets, "A new dual-channel speech enhancement technique with application to CELP coding in noise," in *Int. Conf. Spoken Language Processing*, vol. 1, Banff, AB, Canada, Oct. 1992, pp. 527–530.

[19] S. Nandkumar and J. H. L. Hansen, "Dual-channel iterative speech enhancement with constraints based on an auditory spectrum," *IEEE Trans. Speech Audio Processing*, vol. 3, pp. 22–34, Jan. 1995.

[20] B. Pellom and J. H. L. Hansen, "An improved (Auto:I,LSP:T) constrained iterative speech enhancement algorithm for colored noise environments," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 573–579, Nov. 1998.

[21] S. R. Quackenbush, T. P. Barnwell III, and M. A. Clements, *Objective Measures of Speech Quality*. Englewood Cliffs, NJ: Prentice-Hall, 1988.

[22] T. F. Quatieri and R. G. Danisewicz, "An approach to co-channel talker interference suppression using a sinusoidal model for speech," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 38, pp. 56–69, Jan. 1990.

[23] T. F. Quatieri and R. J. McAulay, "Noise reduction using a soft-decision sine-wave vector quantizer," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1990, pp. 821–824.

[24] ——, "Shape invariant time-scale and pitch modification of speech," *IEEE Trans. Signal Processing*, vol. 40, pp. 497–510, Mar. 1992.

[25] H. Sameti, H. Sheikhzadeh, L. Deng, and R. L. Brennan, "HMM-based strategies for enhancement of speech signals embedded in nonstationary noise," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 445–455, Sept. 1998.

[26] D. E. Tsoukalas, J. N. Mourjopoulos, and G. Kokkinakis, "Speech enhancement based on audible noise suppression," *IEEE Trans. Speech Audio Processing*, vol. 5, pp. 497–513, Nov. 1997.

[27] N. Virag, "Single channel speech enhancement based on masking properties of the human auditory system," *IEEE Trans. Speech Audio Processing*, vol. 7, pp. 497–513, Mar. 1999.

Jesper Jensen received the M.Sc. and Ph.D. degrees from Aalborg University, Aalborg, Denmark, in 1996 and in 2000, respectively, both in electrical engineering.

From 1996 to 2001, he was a Researcher and Assistant Research Professor with the Center for PersonKommunikation (CPK), Aalborg University, Denmark. In 1999, he was a Visiting Researcher at the Center for Spoken Language Research, University of Colorado, Boulder. Currently, he is a Postdoctoral Researcher at the Delft University of Technology, Delft, The Netherlands. His main research interests are digital speech and audio signal processing, including coding, synthesis, and enhancement.

John H. L. Hansen (S'81-M'82-SM'93) was born in Plainfield, NJ. He received the B.S.E.E. degree with highest honors from Rutgers University, New Brunswick, NJ, in 1982 and the M.S. and Ph.D. degrees in electrical engineering from Georgia Institute of Technology, Atlanta, in 1983 and 1988.

He is presently Associate Professor in the Departments of Speech, Language, and Hearing Sciences, and Electrical and Computer Engineering at the University of Colorado, Boulder. In 1988, he established and has since directed the Robust Speech Processing Laboratory (RSPL). He serves as Associate Director for The Center for Spoken Language Research (CSLR), and directs the research activities of RSPL at CSLR. He was a Faculty Member with the Departments of Electrical and Biomedical Engineering, Duke University, for 11 years before joining the University of Colorado in 1999. His research interests span the areas of digital speech processing, analysis and modeling of speech and speaker traits, speech pathology, speech enhancement and feature estimation in noise, robust speech recognition with current emphasis on robust recognition and training methods for phrase spotting in noise, accent, stress, and Lombard effect, and speech feature enhancement in hands-free environments for human–computer interaction. He has served as a Technical Consultant to industry and the U.S. Government, including ATT Bell Labs, IBM, Sparta, Signalscape, BAE Systems, ASEC, VeriVoice, and DoD in the areas of voice communications, wireless telephony, robust speech recognition, and forensic speech/speaker analysis. He is the author of more than 130 journal and conference papers in the field of speech processing and communications, coauthor of the textbook *Discrete-Time Processing of Speech Signals* (New York: IEEE Press, 2000), and lead author of the report "The impact of speech under "stress" on military speech technology," (NATO RTO-TR-10, 2000, ISBN: 92-837-1027-4).

He was an invited tutorial speaker for IEEE ICASSP'95 and the 1995 ESCA-NATO Speech Under Stress Research Workshop, Lisbon, Portugal. He has served as Technical Advisor to U.S. Delegate for NATO (IST/TG-01: Research Study Group on Speech Processing, 1996–1998), Chairman for the IEEE Communications and Signal Processing Society of North Carolina (1992–1994), Advisor for the Duke University IEEE Student Branch (1990–1997), Tutorials Chair for IEEE ICASSP'96, Associate Editor for IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING (1992–1998), and Associate Editor for IEEE SIGNAL PROCESSING LETTERS (1998–2000). He has also served as guest editor of the October 1994 special issue on Robust Speech Recognition for the IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING. He was the recipient of a Whitaker Foundation Biomedical Research Award, an NSF Research Initiation Award, and has been named a Lilly Foundation Teaching Fellow for "contributions to the advancement of engineering education." He will be serving as General Chair for the International Conference on Spoken Language Processing in September 2002.