SPEECH ENHANCEMENT USING A STATE-BASED TRANSFORM MODEL

Michael E. Deisher *Andreas S. Spanias*

Department of Electrical Engineering
Telecommunications Research Center
Arizona State University
Tempe, Arizona 85287-7206 USA
Phone: (602)965-1837 / Fax: (602)965-8325
E-mail: deisher@asu.edu and spanias@asu.edu

ABSTRACT

An analysis/synthesis technique based on harmonic sinusoidal modeling of speech is used to develop a new hidden Markov model (HMM) based speech enhancement algorithm. State sequence estimation is done using a standard HMM-based approach. State-based enhancement is carried out by assuming a harmonic model for speech, i.e., by representing each block of speech as a sum of sine waves in terms of a set of amplitudes, phases, and harmonically related frequencies. Given the maximum a-posteriori probability (MAP) state sequence, the amplitudes, phases, voicing, and fundamental frequency are estimated. Simulation results are presented, comparing the performance of the proposed algorithm to that of a standard HMM-based approach. The proposed method was found to reduce the structured residual noise normally associated with HMM-based algorithms.

1. INTRODUCTION

Speech communication in mobile environments is often difficult due to high-energy ambient noise. The reduction in speech *quality* due to noise is known to cause listener fatigue. Moreover, speech *intelligibility* is severely reduced when low-energy, perceptually important, unvoiced speech sounds are masked by high-energy noise. Speech enhancement algorithms attempt to improve these perceptual aspects of degraded speech.

In addition to improving speech quality or intelligibility for the human listener, speech enhancement preprocessors can improve the performance of other speech processing algorithms. For example, the accuracy of speech recognition algorithms used for "hands-free" dialing of mobile cellular telephones is severely reduced when speech is corrupted by background noise. In this situation, a speech enhancement preprocessor can be added to improve recognition accuracy. In addition, speech compression algorithms used in digital cellular telephones perform poorly in noisy environments, especially when coding at low bit-rates. A speech enhancement preprocessor can be employed in this case to decrease loss in the coder. In mobile applications such as these, speech enhancement algorithms capable of adapting to variable noise environments are especially useful.

The sinusoidal model represents speech by the set of equidistant-in-frequency sinusoids that capture most of the signal energy. This model has received attention in speech coding applications because it provides an accurate, compact representation for speech [1]. Sinusoidal analysis/synthesis has also been successfully applied to co-channel talker interference suppression [2]. The model is of interest for the enhancement of speech corrupted by additive noise for several reasons. In cases where the interference has a fairly flat spectrum, the model tends to select the components with the highest SNR. These components can be estimated more accurately than those "buried in noise." In addition, the model naturally rejects interference whose energy is concentrated between the frequencies of the selected components.

Recently, a family of speech enhancement algorithms based on pattern matching via Gaussian autoregressive hidden Markov models (HMMs) was developed by Ephraim [3, 4]. This approach is attractive in mobile applications because (a) it is well suited to non-stationary noise, (b) it requires only one microphone, (c) it may be trained for a wide range of noise conditions, and (d) it does not suffer from the "musical" residual noise problem associated with other approaches such as spectral subtraction. This paper describes the application of an HMM-based speech enhancement algorithm that uses a harmonic analysis/synthesis model for speech. Our approach differs from other HMM-based approaches in the mechanism for state-based signal estimation. In [3]–[5], Wiener filtering is used; whereas, in [6] the minimum mean square error (MMSE) affine estimator is used. In both cases implementation is done in the frequency domain. Frequency-domain Wiener filtering is accomplished under the assumption that speech and noise covariance matrices are nearly circulant. The MMSE affine estimator is applied to the frequency-domain outputs of an "analysis resonator filterbank." In both cases, all frequency components are generally retained. The method proposed here, however, applies the state-based estimator only to components selected by the sinusoidal model. Therefore, only magnitudes and phases of sinusoids at multiples of the fundamental pitch (voiced or mixed cases) or at equally spaced frequencies (unvoiced case) are estimated. Noise at the chosen frequencies is attenuated; whereas, noise between the chosen frequencies is eliminated.

The rest of the paper is organized as follows. The sinusoidal model and its application to noisy speech are described in Section 2. Section 3 contains a short summary of the HMM-based speech enhancement approach. The proposed speech enhancement system is described in Section 4. Section 5 presents the experimental results and is followed by concluding remarks in Section 6.

1058-6393/95 $4.00 © 1995 IEEE

2. THE SINUSOIDAL MODEL

The sinusoidal model of speech is described in [1]. The n^{th} sample of speech within a 10–20 mS block is represented by

$$y(n) = \sum_{\kappa=1}^{K} A_\kappa \cos(\omega_\kappa n + \phi_\kappa) \qquad (1)$$

where K is the number of sinusoids used in the representation and A_κ and ϕ_κ are the amplitude and phase associated with the κ^{th} frequency track ω_κ. The parameters, $\{A_\kappa, \omega_\kappa, \phi_\kappa\}$, provide a representation for one block of speech (denoted $\mathbf{y}$).

A more compact sinusoidal representation of speech is obtained by considering voiced and unvoiced segments separately. Completely voiced (strongly periodic) speech is represented by a set of harmonics $\{\kappa\omega_0\}$ with the inverse of the pitch period used as the fundamental. The amplitudes $\{A_\kappa\}$ and phases $\{\phi_\kappa\}$ can be measured from the short-time Fourier transform (STFT). Unvoiced speech is represented by a set of equidistant-in-frequency sinusoids (usually no more than 100 Hz apart). Amplitudes in this case can again be measured from the STFT but phases could be randomized with virtually no perceptual loss in unvoiced speech synthesis. Mixed segments can be represented by using a harmonic model for components below a certain voicing-adaptive cutoff frequency (no less than 1500 Hz) and an equidistant-frequency model with random phases for the sinusoidal components above that frequency. In that sense, the sinusoidal model looks similar to mixed-excitation LPC synthesis of the type proposed by McRee and Barnwell [7].

When the harmonic model is applied in the presence of additive noise, $v(n)$, the parameters, $\{A_\kappa, \omega_0, \phi_\kappa\}$, must be estimated from noisy observations, $z(n) = y(n) + v(n)$. Pitch determination in noise is a difficult problem. However, the synthesis model in a speech enhancement application need not be as compact as the one used in sinusoidal transform coding (STC). In fact, a rich parametric set *must* be chosen to ensure good quality speech and avoid "vocoding" noise. Therefore, the dependence on an accurate measurement of the pitch may be weakened by choosing several additional sinusoidal components surrounding the harmonics, e.g., in the range $[\kappa(\omega_0 - \Delta_{\omega_0}), \kappa(\omega_0 + \Delta_{\omega_0})]$.

Amplitudes and phases must also be determined from noisy measurements. Let $V(k)$, $Y(k)$, and $Z(k)$ be the k^{th} discrete Fourier transform (DFT) components of the noise, speech, and noisy speech, respectively. One possible estimate of $Y(k) = \frac{1}{2}A_\kappa \angle\phi_\kappa$ is the MMSE estimate,

$$\widehat{Y(k)} = E\{Y(k)|Z(k)\}. \qquad (2)$$

To compute this estimate, the following assumptions are made. Speech and noise are initially assumed to be stationary random processes. This assumption is necessary only for the derivation of the estimator and will be relaxed in Section 3 by applying a state-based signal model. $V(k)$ and $Y(k)$ are assumed to have jointly Gaussian probability densities. This is motivated by the "central limit theorem for strongly mixing (weakly dependent) processes" [8], considering that DFT components are simply weighted sums of the samples of a random process. Since DFT components

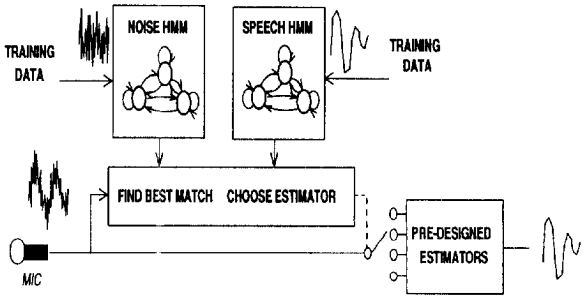

Figure 1. HMM-based speech enhancement system.

are uncorrelated (w.r.t. frequency index k), the assumption that $V(k)$ and $Y(k)$ are jointly Gaussian is equivalent to assuming that they are independent (w.r.t. k). Finally, $\mathbf{y}$ and $\mathbf{v}$ are assumed to be independent with zero mean. Under these assumptions, the MMSE estimate given by (2) can be computed as

$$\widehat{Y(k)} = \frac{S_{yy}(k)}{S_{yy}(k) + S_{vv}(k)} Z(k) = H(k)Z(k) \qquad (3)$$

where $S_{yy}(k)$ and $S_{vv}(k)$ are the power spectral densities (PSDs) of the speech and noise, respectively. This is equivalent to the time domain MMSE estimator ($\hat{\mathbf{y}} = E\{\mathbf{y}|\mathbf{z}\}$) used in [4] where the covariance matrices of speech and noise were approximated with circulant matrices.

Equation (3) is not directly applicable because 1) neither speech nor noise is, in general, stationary and 2) the PSDs of speech and noise are not usually known in advance. Fortunately, the HMM-based speech enhancement approach provides a way to circumvent these restrictions.

3. HMM-BASED ENHANCEMENT APPROACH

Hidden Markov models have proven to be quite useful in modeling the probability density functions of speech parameters in voice recognition. HMMs are also useful models for a wide variety of other time-varying signals including many classes of noise sources [3]. Hidden Markov models represent signals over fixed time-intervals (τ) as the output of a single random source chosen from a finite collection of stationary, discrete random sources. Every τ seconds, a new source is selected from the set according to a first-order Markov chain. Therefore, quasi-stationary signals are modeled as piecewise stationary. A hidden Markov model is specified in terms of its output probability densities $\{b_j(\mathbf{o})\}$, state transition probabilities $\{a_{ij}\}$, and initial state probabilities, $\{\pi_j\}$. In practice, the densities $\{b_j\}$, may characterize the signal itself (as in [3]–[5]) or a set of parameters that describe the signal, such as filterbank outputs (as in [6]). The HMM-based speech enhancement approach is to model both speech and noise with HMMs. Given noisy observations, the most likely state sequence is determined and a state-based estimator is applied.

The HMM-based approach involves two stages, namely, state sequence estimation and state-based enhancement. First, two ergodic HMMs (i.e., HMMs where all state transitions are allowed) are trained; one (λ_y) with clean speech and the other (λ_v) with noise alone. The two models may

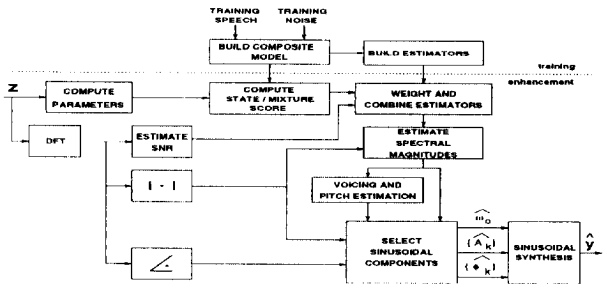

Figure 2. Block diagram of proposed system.

then be combined to form a composite HMM (λ_c) for noisy speech under the assumptions that noise and speech are independent and additive. During state sequence estimation, noisy speech z is divided into blocks of 20-30 mS duration, $Z = \{\mathbf{z}_0, \mathbf{z}_1, \mathbf{z}_2, ...\}$. The most likely composite state sequence $\hat{X} = \{\hat{x}_0, \hat{x}_1, \hat{x}_2, ...\}$ given the noisy observations is found by computing $\arg\max_{\hat{X}_t}(\Pr\{\hat{\mathrm{X}}_t, \mathrm{Z}_t|\lambda_c\})$ for each block index t via the Viterbi algorithm, where $\hat{X}_t = \{\hat{x}_0, \hat{x}_1, ..., \hat{x}_t\}$ and $Z_t = \{\mathbf{z}_0, \mathbf{z}_1, ..., \mathbf{z}_t\}$. During state-based enhancement, each block of clean speech $\mathbf{y}_t$ is estimated given the present block of noisy speech $\mathbf{z}_t$, the most likely composite state $\hat{x}_t$, and the model λ_c. Possible estimators include the MAP estimator of [3], the MMSE estimator of [4], the affine MMSE estimator of [6], or any other appropriate estimator.

The estimator (2) applied to the harmonic components in the previous section fits well into the HMM-based speech enhancement scheme. The discrete random sources associated with the hidden Markov model are stationary. Therefore, if state transitions occur every N_s samples and the estimator is applied to length N_s blocks (between transitions), then the assumption of stationarity in the derivation of (3) is not violated. To compute (3), the power spectral densities of speech and noise are required. If the HMM is designed so that parameters used to describe the output densities also parameterize the PSD of the state output, then (3) may be computed from the HMM parameters given knowledge of the correct speech and noise state. Gaussian autoregressive HMMs (AR-HMMs) meet this requirement since their state output densities are parameterized by linear prediction coefficients [9].

4. PROPOSED SYSTEM

The proposed speech enhancement system applies the HMM-based method to estimate harmonic synthesis parameters for speech. From one point of view, this approach attempts to make the sinusoidal model more robust by using a state-based estimator to find its parameters. From another point of view, this approach attempts to remove noise remaining at the output of an HMM-based enhancement system by capitalizing on the noise suppression properties of the sinusoidal model. A block diagram of the proposed system is shown in Figure 2.

4.1. Training

The system uses ergodic, continuous density, Gaussian mixture AR-HMMs. Two sets of training data are collected, one consisting of speech recorded under quiet conditions, and another consisting of noise from the expected operating environment. An AR-HMM representing clean speech and another representing noise are trained using the segmental k-means method and the Baum-Welch re-estimation procedure [9]. Both AR-HMMs have state output densities of the form

$$b_j(\mathbf{y}) = \sum_m c_{jm} b_{jm}(\mathbf{y}) \tag{4}$$

where c_{jm} are mixture weights associated with the j^{th} state output density,

$$b_{jm}(\mathbf{y}) = (\frac{2\pi\sigma^2}{N_o})^{-\frac{N_o}{2}} e^{-\frac{N_o}{2\sigma^2}\delta(\mathbf{y};\mathbf{a}_{jm})}, \tag{5}$$

$N_o = (1 - P_o)N_s$, N_s is the number of samples of the training data (block length) used in the LPC analysis, P_o is the percent overlap of the analysis blocks, $\mathbf{a}_{jm}$ is a vector of linear prediction coefficients,

$$\delta(\mathbf{y}; a_{jm}) = r_{a_{jm}}(0) r_y(0) + 2 \sum_{i=1}^{p} r_{a_{jm}}(i) r_y(i), \tag{6}$$

$$r_y(i) = \frac{1}{N_s} \sum_{n=0}^{N_s - i - 1} y(n) y(n+i). \tag{7}$$

and

$$r_{a_{jm}}(i) = \sum_{n=0}^{p-i} a_{jm}(n) a_{jm}(n+i), \tag{8}$$

with $\mathbf{a}_{jm}(0) = 1$. The $p + 3$ parameters, σ^2, $\mathbf{r}_{a_{jm}}$, and N_o completely describe the density in equation (5).

The composite (noisy speech) model λ_c is constructed by using the assumptions of statistical independence and additivity of the speech and noise. The new state output densities are obtained by convolving each speech state output density with every noise state output density and approximating the result with an AR Gaussian density of sufficient order. The new transition probability matrix is $A_c = A_y \otimes A_v$ where $\otimes$ is the Kronecker product, and A_y and A_v are the transition probability matrices of the clean speech and noise models, respectively. If the speech model has N_y states with M_y mixtures per state and the noise model has N_v states with M_v mixtures per state then the composite model will have $N_y N_v$ states with $M_y M_v$ mixtures per state.

At the time the composite model is constructed, the PSDs associated with each speech and noise mixture are estimated using the linear prediction coefficients. During enhancement, the PSDs are needed to compute the estimator H in equation (3).

4.2. Enhancement

In the enhancement stage, the autocorrelation sequence is first computed from a block of input samples according to (7). Each mixture is then assigned a score,

$$\varphi_{jm}(t) = \left[\max_{i=1,...,N} \delta_i(t-1) a_{ij}\right] b_j^m(\mathbf{z}_t). \tag{9}$$

This is the probability that the most likely path to state j was taken and the observation $\mathbf{z}_t$ was generated according to the m^{th} mixture density. In (9),

$$b_j^m(\mathbf{z}_t) = \frac{c_{jm} b_{jm}(\mathbf{z}_t)}{b_j(\mathbf{z}_t)}, \tag{10}$$

$\delta_i(t)$ is the probability of the most likely path to state i at time t, a_{ij} are the state transition probabilities for the composite model, and b_{jm} is the AR Gaussian density inside the summation in (4). The expression in brackets is computed using the Viterbi algorithm. Note that only one iteration of the Viterbi algorithm is performed per block of input.

At this point, an estimate of the clean speech could be computed using (3) for the mixture with the highest score. However, a "soft-decision" approach [4] is taken instead. Spectral components are estimated by applying a weighted combination of estimators to each DFT component of the input, i.e.,

$$\hat{Y}_{hmm}(k) = \left(\sum_j \sum_m \rho_{jm} H_{jm}(k) \right) Z(k) \qquad (11)$$

where

$$H_{jm}(k) = \frac{S'_{yy,jm}(k)}{S'_{yy,jm}(k) + \gamma_{jm}^{-1} S'_{vv,jm}(k)} \qquad (12)$$

Here, the PSDs have been normalized to unit energy and γ_{jm} is a signal-to-noise ratio. The value of γ_{jm} may be adjusted if the SNR is determined to be significantly different than the one obtained during training. The weights, $\{\rho_{jm}\}$, are computed from the mixture scores using $\rho_{jm}(t) = \varphi_{jm}(t)/[\sum_{j,m} \varphi_{jm}(t)]$.

The frequency-domain output of the HMM-based algorithm (denoted $\hat{\mathbf{Y}}_{hmm}$) is used to determine the parameters for harmonic synthesis. Once the fundamental pitch, ω_0, has been determined, the amplitudes of the harmonic model are chosen to be $\hat{A}_\kappa = 2|\hat{Y}_{hmm}(k)|$ where k is the index of the component closest in frequency to $\kappa\omega_0$. The phases for the harmonic model are $\hat{\phi}_\kappa = \angle \hat{Y}_{hmm}(k)$. This is equivalent to using the measured phase, $\angle Z(k)$, since ρ_{jm} and H_{jm} are real. As mentioned in Section 2, several components surrounding each harmonic are also selected in order to reduce the dependence on accurate pitch measurement and to decrease vocoding noise. The number of side components must not be excessive, however. Otherwise, noise energy between harmonics will not be suppressed. The frequency domain output of the harmonic model is denoted $\hat{\mathbf{Y}}_{sm}$. After harmonic synthesis, the gain of the reconstructed signal is adjusted by multiplying with $g = \| \hat{\mathbf{Y}}_{hmm} \|_2 / \| \hat{\mathbf{Y}}_{sm} \|_2$ to produce the clean speech estimate, $\hat{\mathbf{y}}_t$.

4.3. Pitch Determination

The fundamental pitch ω_0 is estimated using an analysis-by-synthesis technique similar to that of [10]. However, instead of minimizing the mean squared error between speech synthesized with a peak-picked spectrum and speech synthesized with harmonic components, this algorithm attempts to maximize the energy of the reconstructed signal. A block diagram of the procedure is shown in Figure 3. The magnitude squared of each estimated DFT component is computed. Since peaks near the center of the first formant tend to dominate, a smoothed version of the magnitude squared spectrum is subtracted. Negative components are set to zero. This processed spectrum is denoted $\tilde{\mathbf{Y}}_{hmm}$. The fundamental frequency is allowed to take on a discrete set of values between $\omega_{0\min}$ and $\omega_{0\max}$. For each pitch candidate, ω_0, the components of $\tilde{\mathbf{Y}}_{hmm}$ closest in frequency to $\{\kappa\omega_0\}$ are chosen. The L largest chosen components are summed to yield the score $\xi(\omega)$. The value of ω that maximizes $\xi(\omega)$ may be chosen as the pitch. However, it may be desirable under noisy conditions to apply smoothing to the pitch via the Viterbi algorithm. In this case, the set of pitch candidates is then reduced to the locations of the top M peaks of $\xi(\omega)$. A two-frame delay is introduced and the pitch included in the most likely pitch trajectory is chosen. The score of the most likely pitch trajectory can be compared to a threshold to determine whether the speech is voiced or unvoiced. If the block of speech is classified as unvoiced, components spaced 70 Hz apart are used for synthesis.

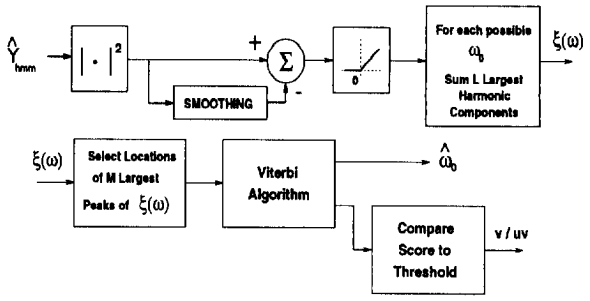

Figure 3. Block diagram of pitch detector.

5. SIMULATION RESULTS

The proposed speech enhancement system was evaluated using speech from the TIMIT database [11] and noise from the NOISEX-92 database [12]. One hundred sentences (approximately 5 minutes total) were selected for training the speech model. Fifty were chosen from the "si" natural phonetic sentences (25 male, 25 female) and fifty were chosen from the "sx" phonetically compact sentences (25 male, 25 female). Speech was modeled by an 8-state, 4-mixture, 12th-order AR-HMM. The broadband noise "06" was used to train the noise model. Noise was modeled by a 1-state, single mixture, 4th-order AR-HMM. The composite model was a 8-state, 4-mixture, 16th-order AR-HMM. The algorithm was tested using 60 sentences from the TIMIT database. Thirty were taken from the "si" sentences (15 male, 15 female) and thirty were taken from the "sx" sentences (15 male, 15 female). Noise "06" from the NOISEX-92 was added at various SNRs.

Training and testing were carried out at each of the input SNRs shown in Table 1. The results shown are averaged over the entire testing set. The columns of this table and the tables that come after it show input signal-to-noise ratio (SNR_{in}), change in signal-to-noise ratio at the output ($\Delta\text{SNR} = \text{SNR}_{out} - \text{SNR}_{in}$), change in segmental signal-to-noise ratio at the output ($\Delta\text{SEGSNR} = \text{SEGSNR}_{out} - \text{SEGSNR}_{in}$), and relative change in modified Itakura-Saito distortion ($\%\Delta\rho = \frac{\rho(\hat{y},y) - \rho(z,y)}{\rho(z,y)} \times 100\%$). Here, SNR is defined as

$$SNR = 10 \log_{10} \frac{\sum_n y^2(n)}{\sum_n (y(n) - \hat{y}(n))^2} \qquad (13)$$

where the sums are taken over all length N_s non-overlapping blocks of clean speech whose energy is at least -40dB. The

Table 1. Performance of the proposed speech enhancement system.

$SNR_{in}(dB)$	$\Delta SNR(dB)$	$\Delta SEGSNR(dB)$	$\%\Delta\rho$
20.0	+0.21	+1.36	-93.39
15.0	+0.78	+2.64	-95.62
10.0	+2.04	+4.57	-97.34
5.0	+3.53	+6.97	-97.93
0.0	+5.66	+9.93	-98.74
-5.0	+8.47	+13.19	-98.26
-10.0	+11.39	+17.05	-99.15

Table 2. Performance of the standard HMM-based speech enhancement system.

$SNR_{in}(dB)$	$\Delta SNR(dB)$	$\Delta SEGSNR(dB)$	$\%\Delta\rho$
20.0	+0.02	+1.26	-93.88
15.0	+0.77	+2.57	-95.62
10.0	+1.73	+4.35	-97.88
5.0	+3.05	+6.61	-97.92
0.0	+5.02	+9.55	-98.73
-5.0	+7.63	+13.00	-98.94
-10.0	+10.97	+16.89	-99.14

modified Itakura-Saito (spectral) distance ρ is defined as

$$\rho(\hat{\mathbf{y}}, \mathbf{y}) = \frac{\delta(\hat{\mathbf{y}}; \mathbf{a}(\mathbf{y}))}{\sigma_y^2(\mathbf{y})} - \log \frac{\sigma_{\hat{y}}^2(\hat{\mathbf{y}})}{\sigma_y^2(\mathbf{y})} - 1 \qquad (14)$$

where $\delta(\cdot;\cdot)$ is defined in equation (6), $\mathbf{a}(\mathbf{y})$ indicates the linear prediction coefficients computed from $\mathbf{y}$, and $\sigma^2(\mathbf{y})$ indicates the linear prediction residual computed from $\mathbf{y}$. SEGSNR is defined as

$$\text{SEGSNR} = \frac{1}{N_b} \sum_i 10 \log_{10} \frac{\sum_{n=iN_s}^{iN_s+N_s-1} y^2(n)}{\sum_{n=iN_s}^{iN_s+N_s-1} (y(n) - \hat{y}(n))^2} \qquad (15)$$

where N_b is the number of non-overlapping blocks of length N_s whose energy is greater than -40dB and the sum is over all such blocks (indexed by i).

The results reported in Table 1 are considered preliminary since the fundamental pitch was estimated from the *clean* speech rather than the noisy speech. However, these results illustrate the potential of the proposed algorithm. Comparison of Tables 1 and 2 reveals that the total and segmental SNRs were improved in every case. The relative spectral distance was improved in half the cases. The improvement with respect to objective criteria is small. However, subjective improvement was quite noticeable.

Although the HMM-based speech enhancement algorithm does not produce musical noise artifacts, it is known to produce "a low-level structured residual noise," especially for female speakers [4]. The proposed algorithm noticeably attenuates this residual noise. It is particularly effective for female speakers since the harmonics of female speakers are spaced further apart.

6. CONCLUSIONS

A new HMM-based speech enhancement algorithm has been proposed. Enhancement is accomplished by incorporating a spectral estimator into the harmonic sinusoidal speech model. The performance of the proposed algorithm was evaluated in terms of average increase in total SNR, segmental SNR, and spectral distance. These results were compared with those of the standard HMM-based approach. The proposed algorithm was found to reduce the low-level residual noise produced by the standard HMM-based algorithm.

Acknowledgments

This work has been supported by the Mobile Research Council of Intel Corporation. The authors would like to thank Brian Mears of Intel Corporation for his comments and suggestions.

REFERENCES

[1] R.J. McAulay and T.F. Quatieri, "Low-Rate Speech Coding Based on the Sinusoidal Model", *Advances in Speech Signal Processing*", S. Furui and M. M. Sondhi ed., pp. 165–208, Marcel Dekker, 1992.

[2] T.F. Quatieri and R.G. Danisewicz, "An Approach to Co-Channel Talker Interference Suppression Using a Sinusoidal Model for Speech", *IEEE Trans. ASSP*, Vol. 38, pp. 56–69, January 1990.

[3] Y. Ephraim, D. Malah, and B.H. Juang, "On the Application of Hidden Markov Models for Enhancing Noisy Speech", *IEEE Trans. on Acoustics, Speech and Signal Processing*, Vol.37, pp. 1846–1856, December 1989.

[4] Y. Ephraim, "A Minimum Mean Square Error Approach for Speech Enhancement", *Proc. ICASSP-90*, pp. 829–832, May 1990.

[5] H. Sheikhzadeh, H. Sameti, L. Deng, and R.L. Brennan, "Comparative Performance of Spectral Subtraction and HMM-Based Speech Enhancement Strategies with Application to Hearing Aid Design", *Proc. ICASSP-94*, pp. I.13–I.16, May 1994.

[6] L. Gagnon, "A state-based noise reduction approach for non-stationary interference", *Speech Communication*, Vol. 12, pp. 213–219, July 1993.

[7] A. McCree and T. Barnwell III, "A New Mixed Excitation LPC Vocoder", *Proc. ICASSP-91*, pp. 593–596, May 1991.

[8] P. Hall and C.C. Heyde, *Martingale Limit Theory and its Application*, Academic Press, 1980.

[9] L.R. Rabiner, "A Tutorial on Hiden Markov Models and Selected Applications in Speech Recognition", *Proceedings of the IEEE*, Vol. 77, pp. 257–286, February 1989.

[10] R.J. McAulay and T.F. Quatieri,, "Pitch Estimation and Voicing Detection Based on a Sinusoidal Speech Model", *Proc. ICASSP-90*, pp. 249–252, May 1990.

[11] J.S. Garofolo, "Getting Started with the DARPA TIMIT CD-ROM: An Acoustic Phonetic Continuous Speech Database", "National Institute of Standards and Technology", December 1988.

[12] A. Varga *et. al.*, "The NOISEX-92 Study on the Effect of Additive Noise on Automatic Speech Recognition", "DRA Speech Research Unit (UK) and TNO Institute for Perception (Netherlands)", June 1992.