HMM-BASED SPEECH ENHANCEMENT USING HARMONIC MODELING

Michael E. Deisher[1] *Andreas S. Spanias*[2]

[1]Intel Corporation, Hillsboro, OR, USA
[2]Arizona State University, Tempe, AZ, USA

ABSTRACT

This paper describes a technique for reduction of non-stationary noise in electronic voice communication systems. Removal of noise is needed in many such systems, particularly those deployed in harsh mobile or otherwise dynamic acoustic environments. The proposed method employs state-based statistical models of both speech and noise, and is thus capable of tracking variations in noise during sustained speech. This work extends the hidden Markov model (HMM) based minimum mean square error (MMSE) estimator to incorporate a ternary voicing state, and applies it to a harmonic representation of voiced speech. Noise reduction during voiced sounds is thereby improved. Performance is evaluated using speech and noise from standard databases. The extended algorithm is demonstrated to improve speech quality as measured by informal preference tests and objective measures, to preserve speech intelligibility as measured by informal Diagnostic Rhyme Tests, and to improve the performance of a low bit-rate speech coder and a speech recognition system when used as a pre-processor.

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 speech is 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 typically 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 [1], an MMSE speech enhancement approach using hidden Markov models (HMMs) for both speech and noise sources was proposed. It has been recognized as one of the most promising methods to date. In particular, the fact that this method uses a noise model that can capture the dynamic behavior of the short-term noise statistics is particularly significant. This paper extends the HMM-based MMSE estimator to incorporate a ternary voicing state and uses a harmonic representation for improved noise suppression during voiced speech.

2. HMM-BASED SPEECH ENHANCEMENT

The following notation is used. Let y and v be zero mean, statistically independent speech and noise processes, respectively. Corresponding non-overlapping blocks of time samples are denoted $\mathbf{y}_t$ and $\mathbf{v}_t$ where t is a time index. Speech and noise are assumed to be additively combined and their sum is denoted $\mathbf{z}_t$. The upper case symbols $\mathbf{Y}_t$ and $\mathbf{Z}_t$ represent the discrete Fourier transforms of $\mathbf{y}_t$ and $\mathbf{z}_t$, respectively. The DFT components of $\mathbf{Y}_t$ and $\mathbf{Z}_t$ are $Y_t(k)$ and $Z_t(k)$ where $k = 0, ..., K-1$.

In [1], Ephraim derived a MMSE estimate of the clean speech DFT component,

$$E\{Y_t(k)|\mathbf{Z}^{\tau_e}, \lambda^z\} = \sum_{j\in\mathcal{N}} \sum_{m\in\mathcal{M}} P(x_t = j, u_t = m|\mathbf{Z}^{\tau_e}, \lambda^z) \cdot E\{Y_t(k)|x_t = j, u_t = m, \mathbf{z}_t, \lambda^z\} \quad (1)$$

where λ^z is an N-state, M-mixture autoregressive hidden Markov model (ARHMM [2]) of noisy speech, $\mathcal{M} = \{0, ..., M-1\}$, $\mathcal{N} = \{0, ..., N-1\}$, x_t is the HMM state at time t, u_t is the HMM mixture at time t, $\mathbf{Z}^{\tau_e} = \mathbf{z}_0 \ldots \mathbf{z}_{t+\tau_e-1}$, and the estimation delay in blocks, $\tau_e \geq 0$, determines the number of future observations included. The joint state/mixture probability, $P(x_t = j, u_t = m|\mathbf{Z}^{\tau_e}, \lambda^z)$, is computed using the well-known HMM "forward-backward" recursion [2]. The conditional mean within the summation in (1) is calculated, assuming circulant speech and noise covariance matrices ($\mathbf{C}^y_{jm}$ and $\mathbf{C}^v_{jm}$), as

$$E\{Y_t(k)|x_t = j, u_t = m, \mathbf{z}_t, \lambda^z\} \approx \frac{\tilde{S}^{yy}_{jm}(k)}{\tilde{S}^{yy}_{jm}(k) + \tilde{S}^{vv}_{jm}(k)} Z_t(k) \quad (2)$$

where $\tilde{\mathbf{S}}^{yy}_{jm} = \mathbf{F}\mathbf{C}^{y\ T}_{jm}[1\ 0 \cdots 0]^T$, $\tilde{\mathbf{S}}^{vv}_{jm} = \mathbf{F}\mathbf{C}^{v\ T}_{jm}[1\ 0 \cdots 0]^T$, and $\mathbf{F}$ is the DFT matrix.

Providing that the HMM λ^z is well-trained, the HMM-based MMSE estimator outperforms other algorithms (e.g., spectral subtraction) in terms of SNR improvement and subjective quality [3, 4]. In addition, the method can cope with noise variations between pauses in speech when

the number of noise states is greater than one. However, at SNRs below 15 dB, the processed speech exhibits a "low-level, structured residual noise," particularly for high-pitched speakers [1].

3. EXTENSION OF THE HMM-BASED MMSE ESTIMATOR

The residual noise associated with the HMM-based MMSE estimator is most perceptible in voiced segments of speech. Therefore, we propose an extension of this method that incorporates voicing and pitch information in order to improve voiced regions. Let s_t be the voicing state of $\mathbf{y}_t$, taking on values $\mathcal{V}$ (voiced), $\mathcal{UV}$ (unvoiced), or $\mathcal{NS}$ (non-speech). Define $\Omega_k(\rho)$ as the event that the frequency corresponding to the k^{th} DFT bin lies within $(\rho/2)\omega_0$ radians of a multiple of the fundamental frequency ω_0 when speech is voiced. The parameter, $0 \leq \rho \leq 1$, is chosen such that speech energy contained in DFT bins whose center frequencies are not within $(\rho/2)\omega_0$ of a harmonic is negligible. The MMSE estimator may be written in terms of s_t and $\Omega_k(\rho)$ as

$$E\{Y_t(k)|\mathbf{Z}^{\tau_e},\lambda^z\} = \sum_{j\in\mathcal{N}}\sum_{m\in\mathcal{M}} P(x_t=j,u_t=m|\mathbf{Z}^{\tau_e},\lambda^z) \cdot \Big[p_{uv} \cdot E\{Y_t(k)|x_t=j,u_t=m,s_t=\mathcal{UV},\mathbf{z}_t,\lambda^z\} + p_v \cdot P(\Omega_k(\rho),u_t=m,s_t=\mathcal{V},\mathbf{z}_t,\lambda^z) \cdot E\{Y_t(k)|x_t=j,u_t=m,s_t=\mathcal{V},\Omega_k(\rho),\mathbf{z}_t,\lambda^z\}\Big] \tag{3}$$

where $p_{uv} = P(s_t = \mathcal{UV}|x_t = j, u_t = m, \mathbf{z}_t, \lambda^z)$ and $p_v = P(s_t = \mathcal{V}|x_t = j, u_t = m, \mathbf{z}_t, \lambda^z)$. Equation (3) may be further simplified by approximating the expected values in the second and third terms of the sum within the square brackets using the expected value that is uninformed of voicing. This results in

$$E\{Y_t(k)|\mathbf{z}_t,\lambda^z\} \approx \sum_{j\in\mathcal{N}}\sum_{m\in\mathcal{M}} \Gamma(j,m,k,\rho,\mathbf{z}_t) \cdot P(x_t=j,u_t=m|\mathbf{Z}^{\tau_e},\lambda^z) E\{Y_t(k)|x_t=j,u_t=m,\mathbf{z}_t,\lambda^z\} \tag{4}$$

where

$$\Gamma(j,m,k,\rho,\mathbf{z}_t) = p_v \cdot P(\Omega_k(\rho)|x_t=j,u_t=m,s_t=\mathcal{V},\mathbf{z}_t,\lambda^z) + p_{uv} \tag{5}$$

The extended estimator differs from that of [1] by addition of the factor $\Gamma(j,m,k,\rho,\mathbf{z}_t)$. An alternative formulation is obtained by rewriting (5) in terms of the *a posteriori* probability of speech.

$$\Gamma(j,m,k,\rho,\mathbf{z}_t) = (P(\Omega_k(\rho)|x_t=j,u_t=m,s_t=\mathcal{V},\mathbf{z}_t,\lambda^z) - 1) \cdot p_v + p_s \tag{6}$$

where $p_s = P(s_t = \overline{\mathcal{NS}}|x_t = j, u_t = m, \mathbf{z}_t, \lambda^z)$. To implement (6) two simplifications are made. The first simplifying approximation is justified by observing the behavior of the original HMM-based algorithm. When speech is not present, the products in (1) are almost always very close to zero for well-trained models. Therefore, assuming that implicit speech detection is accomplished reasonably well,

$$\widetilde{\Gamma}(j,m,k,\rho,\mathbf{z}_t) = 1 + [P(\Omega_k(\rho)|x_t=j,u_t=m,s_t=\mathcal{V},\mathbf{z}_t,\lambda^z) - 1] \cdot p_v \tag{7}$$

Table 1. Total SNR for F-16 aircraft noise.

Input SNR	MMSE Algorithm SNR (dB) Mean	StdDev	Min	Max	
10	13.22	0.52	11.80	14.21	
5	9.70	0.63	7.90	10.84	
0	6.36	0.69	4.62	7.56	
Input SNR	**Extended MMSE Method SNR (dB) Mean**	**StdDev**	**Min**	**Max**	ρ
10	13.84	0.70	12.14	15.24	0.89
5	10.14	0.75	8.16	11.53	0.79
0	6.63	0.74	4.78	8.05	0.70

provides approximately the same results. The second simplifying approximation uses average harmonic and voicing probabilities instead of those conditioned on state / mixture occupancy.

$$\widehat{\Gamma}(j,m,k,\rho,\mathbf{z}_t) = 1 + [P(\Omega_k(\rho)|s_t=\mathcal{V},\mathbf{z}_t,\lambda^z) - 1] \cdot P(s_t=\mathcal{V}|\mathbf{z}_t,\lambda^z) \tag{8}$$

The final approximation (8) is much simpler to implement than (7) since it requires that harmonic status and voicing probabilities be calculated once per block instead of $N^y M^y N^v M^v$ times per block. Furthermore, the simplification in (8) is intuitively appealing since (8) and (7) are equivalent if voicing and harmonic status are independent of the hidden Markov model state and mixture. Since ARHMMs essentially model the average spectral shape of speech, the dependence of voicing and harmonic status upon HMM state and mixture is at most mild. Therefore the second simplification is not unreasonable. For implementation, voicing probability and harmonic status are computed using a pitch detection algorithm such as that described in [5]. In addition, zero-padding is used to increase the size of the DFTs so that the regions defined by ρ are sufficiently well separated. The value of ρ is chosen empirically to maximize the output SNR over the training set.

4. SELECTED RESULTS

Three noise reduction simulation experiments were conducted to evaluate the performance of the extended estimator. The speech models used in all the simulations were 8-state, 5-mixture ARHMMs of order $p^y = 12$. Noise models used in the experiments were trained with noise from the NOISEX-92 database. Noisy speech was created by artificially adding speech and noise. In the first experiment, noisy speech processed using the original MMSE estimator and the extended estimator were compared in terms of SNR and listener preference. Second, the intelligibility of noisy speech processed with a low bit-rate coder was measured using the Diagnostic Rhyme Test (DRT) [8]. The effect of noise reduction on intelligibility was investigated by comparing DRT results when the extended estimator and the spectral subtraction algorithm [6] were used as pre-processors. Finally, the original and extended estimators were tested as pre-processors for automatic speech recognition.

Table 1 shows the output SNR when noisy speech is processed using the original and extended MMSE estima-

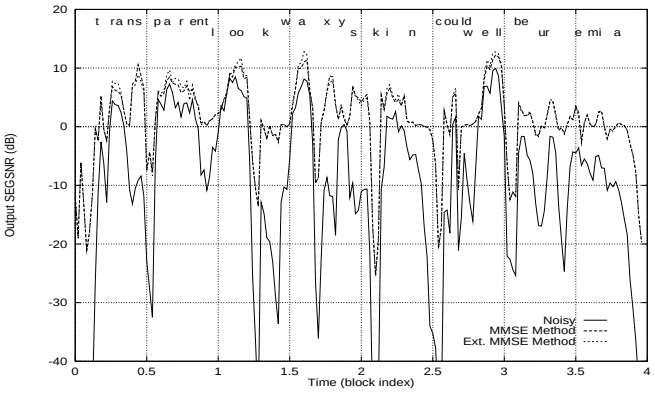

Figure 1. Segmental SNR at 0 dB input SNR for a female speaker.

tors. The noise model was trained using F-16 cockpit noise from the NOISEX-92 database at levels corresponding to 0 dB, 5 dB, and 10 dB average SNR when speech is present ($N^v = 3$, $M^v = 3$, $p^v = 4$). Noisy speech models were created by combining each of the noise models with a speech model trained using sentences from 100 speakers in the TIMIT database. For testing, noise was added to sentences from another 60 speakers from TIMIT and the noisy files were processed. The mean, minimum, maximum and standard deviation of the output SNR is shown for each input SNR. The harmonic widening parameter ρ used in the extended estimator is also shown in the table.

These simulations were repeated using Lynx cockpit noise ($N^v = 3$, $M^v = 3$, $p^v = 20$), operations room noise ($N^v = 5$, $M^v = 5$, $p^v = 12$), and "speech-shaped" noise ($N^v = 1$, $M^v = 3$, $p^v = 4$) from the NOISEX-92 database. In each case, a similar increase in output SNR with respect to the original algorithm was measured. Increased SNR is a good indication of the performance of noise reduction algorithms based on waveform matching in the mean squared error sense. Improvement in SNR for the type of interference involved here was almost always accompanied by perceptual improvement. Although the improvement in average output SNR was small (less than one dB), there was a noticeable improvement in speech quality due to better noise suppression during voiced regions. This is further illustrated by Figure 1. The figure shows the SNR per block for speech-shaped noise added at 0 dB SNR to a sentence spoken by a female speaker. The text of the sentence is shown at the top of the plot. As much as 2 dB improvement is shown in Figure 1 during voiced segments. In particular, /æ/, /e/, /u/, and /l/ are most improved. The degradation in segmental SNR during /s/ is due to erroneous voicing decisions. The degradation is small relative to the overall improvement of /s/ and is not audibly perceptible.

An "A/B" preference test was conducted with a panel of fifteen listeners. Three female/male pairs of sentences from the test set were selected at random. Each pair was concatenated, corrupted with noise, and processed using the original MMSE estimator, the extended estimator, and spectral subtraction. The processed and unprocessed records were presented to the listeners in randomly-ordered pairs. Table 2 shows the percentage of the trials $\hat{\mu}$ in which listeners preferred speech processed using the extended MMSE estimator over that processed by the other algorithms or the noisy speech. An estimate of the standard deviation $\sigma_{\hat{\mu}} = \sqrt{\mu(1-\mu)/J}$ is given in the rightmost column. J is the number of trials. In every case the majority of the listeners preferred speech processed using the extended estimator. In most cases the 95% confidence interval $[\hat{\mu} - 1.96\sigma_{\hat{\mu}}, \hat{\mu} + 1.96\sigma_{\hat{\mu}}]$ lies entirely above the 50% mark.

Table 2. Percentage of listeners that preferred speech from the extended MMSE algorithm.

Speech-Shaped Noise	SNR	% Pref ($\hat{\mu}$)	$\sigma_{\hat{\mu}}$
Unprocessed	5dB	71.4	8.5
MMSE	5dB	85.7	6.6
Spectral Subtraction	5dB	75.0	8.2
Unprocessed	0dB	71.4	8.5
MMSE	0dB	71.4	8.5
Spectral Subtraction	0dB	67.9	8.8
Operations Room Noise	**SNR**	**% Pref ($\hat{\mu}$)**	**$\sigma_{\hat{\mu}}$**
Unprocessed	5dB	67.9	8.8
MMSE	5dB	85.7	6.6
Spectral Subtraction	5dB	64.3	9.1
Unprocessed	0dB	67.9	8.8
MMSE	0dB	64.3	9.1
Spectral Subtraction	0dB	53.6	9.4

Table 3. Adjusted DRT Scores.

Speech-Shaped Noise			
SNR	Noisy	SSUB	EMMSE
0.0	25.62	34.98	37.44
5.0	58.13	56.40	60.34
Operations Room Noise			
SNR	Noisy	SSUB	EMMSE
0.0	47.78	48.03	56.16
5.0	62.07	63.05	56.90

The listeners were interviewed following completion of the test. Most found the speech produced by the HMM-based algorithms to be most comfortable to listen to because the background noise was strongly attenuated. However, listeners generally disliked the rough residual noise of the original algorithm and that remaining in the speech from the extended estimator. The listeners were sharply divided in their opinion of the so-called "musical" residual noise produced by the spectral subtraction method. Some found it very objectionable while others described it as quite tolerable. In some cases, listeners chose the noisy speech simply because it was free of processing artifacts.

In addition to the experiments using speech from the TIMIT database, several simulations were carried out on the DRT word lists available from Dynastat, Inc. In this case, the goal was to evaluate the extended estimator as

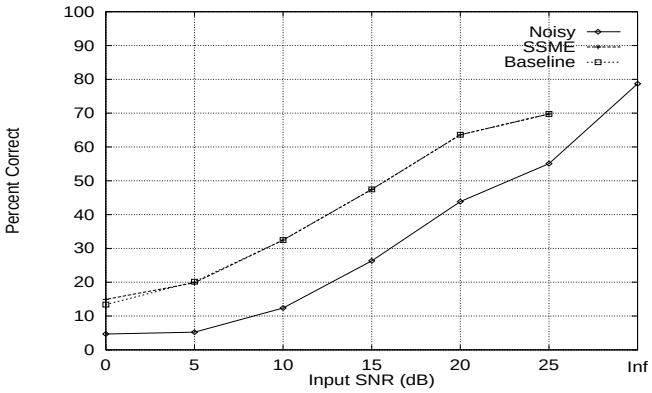

Figure 2. Alphabet recognition performance.

a pre-processor for low bit-rate speech compression. Low bit-rate speech coders often exhibit artifacts and poor performance when deployed in harsh acoustic environments. Noise reduction pre-processing is aimed at preventing these artifacts and restoring coder performance. The speech coder used in these experiments was the 2400 bps sinusoidal transform coder (STC) developed at MIT Lincoln Labs [7].

The DRT described in [8] was carried out with a panel of thirteen listeners. Two sets of test data were prepared: one corrupted with speech-shaped noise and another with operations room noise. Each set consisted of 4 six-minute recordings. The initial 90 seconds of each was used to train an eight state speech model. Noise was added to the remainder of each recording. The noise level was 0 dB for two of the recordings and 5 dB for the remaining two. Each pair contained speech from one adult female and one adult male. The third 90-second segment of each recording was then processed using spectral subtraction. The extended estimator was used to process the final 90 seconds of each record. The resulting speech files were coded and decoded using the 2.4 Kbps STC. The original MMSE estimator was not evaluated due to the limited size of the data set and because it differs from the extended estimator mainly during vowels. Recognition of vowels is not tested by the DRT.

Figure 3 shows the intelligibility scores (adjusted for random guessing) obtained by informal DRT evaluation of the output of the STC decoder. Seven listeners were used in each case. One of the thirteen listeners participated in both tests. At 0 dB input SNR, STC speech with extended MMSE pre-processing is more intelligible than both noisy speech and speech processed with spectral subtraction prior to coding. Similar results were obtained at 5 dB SNR for speech-shaped noise. However, for operations room noise at 5 dB SNR, speech processed by the extended method is less intelligible. In this case, the background noise was completely removed but portions of some initial stop consonants (e.g., /b/) were removed along with the noise. This suggests that it may be important to weaken the suppression rule in applications where intelligibility is paramount. However, considering the small number of listeners involved in the test and the fact that most of the listeners had no prior experience with tests such as the DRT, these results should be interpreted cautiously.

Finally, the extended estimator was evaluated as a preprocessor to the HMM-based isolated alphabet recognizer described in [9]. The performance of the recognizer was evaluated using the OGI Telephone Speech Database of Spelled and Spoken Names. The training set consisted of the letters of the alphabet uttered once by fifty speakers (25 male, 25 female). An embedded training procedure was used whereby context-independent phoneme models were first trained and used to initialize the context-dependent phoneme models. The test set consisted of the letters of the alphabet spoken by an additional set of 25 male and 25 female speakers. White Gaussian noise was added to the test set at several SNRs. Figure 2 shows the recognition accuracy with and without pre-processing. A 20% increase in the percentage of correct recognitions was achieved over most of the noise levels. The fact that the original and extended MMSE estimators performed about the same is not surprising since recognition features model spectral shape rather than fine spectral structure.

5. REMARKS

This paper has outlined an improved HMM-based speech enhancement algorithm. Results were presented in terms of SNR improvement, informal listener preference scores, informal DRT scores when the algorithm is used as a preprocessor for a low bit-rate speech coder, and recognition performance when the algorithm is used as a preprocessor for an alphabet recognition system.

REFERENCES

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

[2] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proc. IEEE*, Vol. 77, pp. 257–286, February 1989.

[3] H. Sheikhzadeh, *et al.*, "Comparative Performance of Spectral Subtraction and HMM-Based Speech Enhancement Strategies with Application to Hearing Aid Design", *Proc. ICASSP*, pp. I.13–I.16, May 1994.

[4] M.E. Deisher and A.S. Spanias, "Speech Enhancement Using a State-Based Transform Model", *Proc. Asilomar Conference on Signals, Systems, and Computers*, pp. 1242–1246, Nov. 1994.

[5] M.E. Deisher, "State-Based Noise Reduction Using the Sinusoidal Speech Model", Ph.D. Dissertation, ASU, May 1996.

[6] S.F. Boll, "Suppression of Acoustic Noise in Speech Using Spectral Subtraction", *IEEE Trans. ASSP*, pp. 113–120, April 1979.

[7] R.J. McAulay and T.F. Quatieri, "The Sinusoidal Transform Coder at 2400 b/s", *Proc. MILCOM*, pp. 15.6.1–15.6.3, 1992.

[8] W.D. Voiers, "Evaluating Processed Speech Using the Diagnostic Rhyme Test", *Speech Technology*, pp. 30–39, Jan. 1983.

[9] P.C. Loizou, "Robust Speaker Independent Recognition of a Confusable Vocabulary", PhD Thesis, ASU, May 1995.