TRACKING SPEECH-PRESENCE UNCERTAINTY TO IMPROVE SPEECH ENHANCEMENT IN NON-STATIONARY NOISE ENVIRONMENTS

David Malah[*], *Richard V. Cox, and Anthony J. Accardi*

AT&T Labs - Research, Florham Park, NJ 07932

ABSTRACT

Speech enhancement algorithms which are based on estimating the short-time spectral amplitude of the clean speech have better performance when a soft-decision gain modification, depending on the a priori probability of speech absence, is used. In reported works a fixed probability, q, is assumed. Since speech is non-stationary and may not be present in every frequency bin when voiced, we propose a method for estimating distinct values of q for different bins which are tracked in time. The estimation is based on a decision-theoretic approach for setting a threshold in each bin followed by short-time averaging. The estimated q's are used to control both the gain and the update of the estimated noise spectrum during speech presence in a modified MMSE log-spectral amplitude estimator. Subjective tests resulted in higher scores than for the IS-127 standard enhancement algorithm, when pre-processing noisy speech for a coding application.

1. INTRODUCTION

In view of the steady increase in mobile voice communication system usage and applications, there is a renewed interest in single microphone input speech enhancement algorithms. Of particular interest is the use of such algorithms as pre-processors for low bit-rate speech coders, since such coders are very sensitive to background noise.

Although many speech enhancement algorithms have been developed in the last two decades, e.g., [5] [6] [2] [3] [8] [7], improvements are still sought. In particular, effective suppression of non-stationary noise, like moving vehicle noise, is of special importance.

Most of the common enhancement techniques, such as those cited above, operate in the frequency domain. These techniques apply a frequency-dependent gain function to the spectral components of the noisy speech, in an attempt to attenuate the noisier components to a greater degree.

The noise suppression properties of the above enhancement algorithms have been shown to improve when a *soft-decision* based modification of the gain function, which depends on the probability of speech absence, is used [6] [2] [3] [8]. To implement such a gain modification function, one needs to give a value to the *a priori* probability of speech absence in each spectral component of the noisy signal.

In all the works known to the authors, a fixed probability, q, is assumed for all frequency components and all the analyzed input frames. Note, however, that even if a voice activity detector (VAD) is used, so it is known in advance that an input frame contains speech, it is not necessarily best to use $q = 0$ (i.e., no gain modification). This is because voiced speech can be considered quasi-harmonic and hence speech energy may not be present in every spectral component of the analyzed input signal.

If the same value of q is used for all frequency components, this value should reflect the average number of spectral components that do not contain speech (using a discrete frequency representation). Since speech is non-stationary, this number varies in time. Furthermore, instead of assigning the same value of q to all frequency bins, we should allow for a different value in each bin and track it in time.

In this work, we first propose a method for estimating a fixed q for each analyzed frame. We then obtain distinct values of q for each frequency bin in each frame. We propose to use these q's also to control the update of the estimated noise spectrum when speech is present. The above propositions were examined in the context of the MMSE Log-Spectral Amplitude (LSA) estimator [3] speech enhancement algorithm. Subjective tests were performed to examine the performance of the proposed algorithm in pre-processing noisy speech for coding applications.

2. SOFT-DECISION GAIN MODIFICATION

Assuming an additive noise model, the noisy signal $y(n)$ is given by $x(n) + d(n)$, where $x(n)$ is the clean speech signal which is assumed to be independent of the noise $d(n)$. A short-time Fourier analysis is applied to the input signal by computing the DFT of overlapping windowed frames. In the frequency domain we have $Y_k = X_k + D_k$, where, $X_k = A_k \exp(j\varphi_k)$, and $Y_k = R_k \exp(j\theta_k)$, with k denoting the frequency bin index. It is assumed that the DFT coefficients of both the speech and the noise are independent Gaussian random variables.

Let C_k be some function of the short-time spectral am-

[*] This work was performed while on leave from the Dept. of Electrical Eng., Technion – Israel Institute of Technology, Haifa 32000, Israel

plitude A_k of the clean speech in the k-th bin (e.g., A_k, $\log A_k$, A_k^2). Taking into account speech-presence uncertainty, the MMSE estimator of C_k is [6]:

$$\begin{aligned}\tilde{C}_k &= E\{C_k|Y_k, H_1^k\}P(H_1^k|Y_k) \\ &+ E\{C_k|Y_k, H_0^k\}P(H_0^k|Y_k),\end{aligned} \quad (1)$$

where, H_0^k: Speech absent; H_1^k: Speech present (k-th bin).

Since the second term is zero [6], we have:

$$\tilde{C}_k = E\{C_k|Y_k, H_1^k\}P(H_1^k|Y_k). \quad (2)$$

$P(H_1^k|Y_k)$ is thus the 'soft-decision' modification of the optimal estimator under the signal presence hypothesis.

Applying Bayes' rule, one obtains [6] [2]:

$$P(H_1^k|Y_k) = \frac{\Lambda(k)}{1 + \Lambda(k)} \triangleq G_M(k), \quad (3)$$

where,

$$\Lambda(k) \triangleq \mu_k \frac{p(Y_k|H_1^k)}{p(Y_k|H_0^k)};\ \mu_k \triangleq \frac{P(H_1^k)}{P(H_0^k)} = \frac{1-q_k}{q_k}. \quad (4)$$

$\Lambda(k)$ is a likelihood ratio and q_k denotes the *a priori* probability of speech absence in the k-th bin. $\tilde{C}_k$ is then used to find an estimate of the clean speech amplitude A_k.

3. MMSE-LSA AND MM-LSA ESTIMATORS

Based on the results reported in [3] we prefer using the MMSE-LSA estimator over the MMSE-STSA ($C_k = A_k$) estimator [2] as the basic enhancement algorithm. In this case $C_k = \log A_k$ and the amplitude estimator has the form:

$$\begin{aligned}\tilde{A}_{LSA} &= \exp[E\{\log A_k|Y_k, H_1^k\}G_M(k)] \\ &\triangleq [G_{LSA}(k)R_k]^{G_M(k)},\end{aligned} \quad (5)$$

where $G_M(k)$ is the gain modification defined in (3). Because the soft-decision modification in (5) is not multiplicative and, in [3], it did not result in a meaningful improvement over using $G_{LSA}(k)$ alone, we have chosen to use the following *multiplicatively-modified* LSA (MM-LSA) estimator:

$$\hat{A}_L = G_M(k)G_{LSA}(k)R_k \triangleq G_L(k)R_k. \quad (6)$$

Under the above assumptions on the speech and noise, the gain function $G_{LSA}(k)$ is derived in [3] to be:

$$G_{LSA}(\xi_k, \gamma_k) = \frac{\xi_k}{1+\xi_k} \exp\left(\frac{1}{2}\int_{v_k}^{\infty} \frac{e^{-t}}{t} dt\right), \quad (7)$$

where,

$$\begin{aligned} v_k &\triangleq \frac{\xi_k}{1+\xi_k}\gamma_k\ ; & \gamma_k &\triangleq \frac{R_k^2}{\lambda_d(k)} \\ \xi_k &\triangleq \frac{\eta_k}{1-q_k}\ ; & \eta_k &\triangleq \frac{\lambda_x(k)}{\lambda_d(k)} \\ \lambda_x(k) &\triangleq E\{|X_k|^2\}\ ; & \lambda_d(k) &\triangleq E\{|D_k|^2\}. \end{aligned}$$

In [3], γ_k is called the *a posteriori* SNR for bin k, η_k is called the *a priori* SNR, and q_k is the prior probability of speech absence discussed earlier.

With the above definitions, the expression for $\Lambda(k)$ in (3) is given by [2]:

$$\Lambda(k) = \mu_k \frac{\exp(v_k)}{1+\xi_k}\Big|_{\xi_k = \eta_k/(1-q_k)} \quad (8)$$

In order to evaluate these gain functions, one needs to first estimate the noise power spectrum λ_d. This is usually done during periods of speech absence as determined by a VAD. The estimated noise spectrum and the squared input amplitude R_k^2 provide an estimate for the *a posteriori* SNR. In [2] and [3], a decision-directed approach to estimate the *a priori* SNR, $\eta_k(l)$, in each frame l is used.

An important property of both the MMSE-STSA [2] and the MMSE-LSA [3] enhancement algorithms is that they are able to eliminate 'musical noise' [1] in the enhanced signal, which plagues most other frequency-domain algorithms. This can be attributed to the decision-directed estimation method for the *a priori* SNR [1]. It is recommended in [1] to use a lower limit η_{MIN} for the estimated η_k (we used values in the range between 0.1 to 0.2). A weighting factor α, in that estimator, controls a tradeoff between noise reduction and signal distortion [2] [1]. We typically used values like 0.91 (condition L2 in Section 6) and 0.95, whereas for aggressive enhancement we used $\alpha = 0.98$ (condition L1).

4. ESTIMATION OF PRIOR PROBABILITIES

The key issue in this work is the estimation of the priors q_k needed in both (7) and (8). Our first goal was to estimate a fixed q for each frame that contains speech.

Trying to base a decision about speech absence in a particular frequency bin, k, by comparing the estimated *a priori* SNR, $\hat{\eta}_k$, to a threshold value, was not fruitful. Hence, we turned our attention to the *a posteriori* SNR, γ_k. Under our above assumptions the pdf of γ_k, for a given value of η_k, is given by [2]:

$$p(\gamma_k) = \frac{1}{1+\eta_k} \exp\left(-\frac{\gamma_k}{1+\eta_k}\right)\ ;\ \gamma_k \geq 0. \quad (9)$$

To decide whether speech is present in the k-th bin we consider the following composite hypothesis testing problem:

$$\begin{aligned} \mathcal{H}_0 :&\quad \eta_k \geq \eta_{\min} \quad \text{(speech present in } k\text{-th bin)} \\ \mathcal{H}_A :&\quad \eta_k < \eta_{\min} \quad \text{(speech absent in } k\text{-th bin)} \end{aligned}$$

We have chosen the *null hypothesis* $\mathcal{H}_0$ as stated above since its rejection when true is more grave than the alternative error of accepting when false.

Since η_k parameterizes the pdf of γ_k, as shown in (9), γ_k can be used as a statistic for this decision problem. In particular, since the likelihood ratio $p(\gamma_k \mid \eta_k = \eta_k^a < \eta_{\min})/p(\gamma_k \mid \eta_k = \eta_{\min})$ is a monotone function, it can

be shown [4] that a good decision rule for this problem is equivalent to the Neyman-Pearson decision rule for the following hypothesis test between simple hypotheses: $\mathcal{H}_0'$: $\eta_k = \eta_{\min}$ and $\mathcal{H}_A' : \eta_k = \eta_k^a < \eta_{min}$. This gives the test:

$$\gamma_k \underset{\mathcal{H}_A}{\overset{\mathcal{H}_0}{\gtrless}} \gamma_{\text{TH}}, \tag{10}$$

where, γ_{TH} is set to satisfy a desired significance level (or size [4]) α_0 of the test (i.e., the probability of rejecting $\mathcal{H}_0$ when true is α_0). From (9), this leads to:

$$\gamma_{\text{TH}} = (1+\eta_{\min}) \log\left(\frac{1}{1-\alpha_0}\right). \tag{11}$$

It is of interest to note that according to [4] this is a *uniformly most powerful test* for the above posed problem.

Let M be the number of positive frequency bins under consideration, and let $N_q(l)$ be the number of bins in frame l, out of M bins, for which the test in (10) results in rejection of hypothesis $\mathcal{H}_0$. Letting $r_q(l) \stackrel{\triangle}{=} N_q(l)/M$, the proposed estimate for $q(l)$ is formed from $r_q(l)$ by:

$$\hat{q}(l) \;=\; \alpha_q \hat{q}(l-1) \;+\; (1 \;-\; \alpha_q) r_q(l). \tag{12}$$

The smoothing in (12) is performed only for frames which contain speech (using a VAD). The setting of the parameters in the above scheme was done on the basis of informal listening tests. An improvement in performance was noticed with $\gamma_{\text{TH}} = 0.8$ in (10) and $\alpha_q = 0.95$ in (12).

A better gain-modification could be expected if we allow different q's in different bins. Let $I_k(l)$ be an index function that denotes the result of the test in (10), in the k-th bin of frame l (i.e., $I_k(l) = 1$ if $\mathcal{H}_0$ is rejected, and $I_k(l) = 0$ if it is accepted). We suggest the following estimator for q_k:

$$\tilde{q}_k(l) \;=\; \alpha_q \tilde{q}_k(l-1) \;+\; (1 \;-\; \alpha_q) I_k(l). \tag{13}$$

The same settings for γ_{TH} and α_q above are appropriate here also. Then, averaging $\tilde{q}_k(l)$ over k results in the $\hat{q}(l)$ of (12).

Note that the availability of a separate estimate of q in each bin can be used for controlling the update of the estimated noise spectrum when speech is present.

Before we turn to noise spectrum adaptation, we would like to mention a puzzling result. We found, and as yet do not have a good explanation for it, that estimated values obtained for q_k by substituting a fixed $q = 0.5$ in the expression for $P(H_0^k|Y_k) = 1 - P(H_1^k|Y_k)$ (see (3)) were close in value and similar in behavior to $\tilde{q}_k(l)$ above (note that the resulting estimate is *not* $1 - G_M(k)$). Because of its simple functional form, we used this estimate in the subjective experiments described in section 6.

5. NOISE SPECTRUM ADAPTATION

A critical component in any frequency domain enhancement algorithm is the estimation of the noise power spectrum $\lambda_d(k)$. A common technique is to use a VAD and update the estimated noise spectrum during periods of speech absence in the input signal. Recursive smoothing of R_k^2 is typically used. As in [8], we found that the mean value, $\bar{\gamma}$, of γ_k (averaged over all frequency bins in a given frame), is useful for indicating voice activity in that frame. For stationary noise and under the assumption of independent DFT coefficients, $\bar{\gamma}$ is approximately normal with mean 1 and variance $1/M$ (for sufficiently large M). Thus, by comparing $\bar{\gamma}$ to a suitable threshold, one can obtain quite a reliable VAD. Typical values we used for this threshold lie in the range between 1.3 and 2. This also allows for some increase in noise level during a speech utterance without causing a wrong VAD decision when the utterance terminates.

Since spectral changes may also occur during periods of speech absence, we found it useful to control the estimated noise spectrum update-rate by using a dynamic smoothing factor in its recursive adaptation, as follows:

$$\hat{\lambda}_d(k,l) = \alpha_d(l)\hat{\lambda}_d(k,l-1) + (1-\alpha_d(l))R_k^2, \tag{14}$$

where, $\alpha_d(l) \stackrel{\triangle}{=} 1 - F_d|\bar{\gamma}(l-1) - 1|$. F_d is a constant (e.g., 0.2), and α_d is constrained to be positive and within a limited range, such as 0.8 to 0.98. The idea here is that when the noise spectrum is changing faster than the current adaptation rate, this will be reflected in a larger deviation of $\bar{\gamma}$ from its expected value of 1, decreasing the value of α_d.

When the change in the noise spectrum is relatively fast during speech presence, the performance of the enhancement algorithm can be adversely affected. Using the short-term estimates of the q_k's we have the ability to control the update of the estimated noise spectrum. We do this by modifying the update factor α_d in (14) as follows:

$$\alpha_d(k,l) \stackrel{\triangle}{=} 1 - F_d|\bar{\gamma}_{\mathcal{K}}(l-1) - 1|\tilde{q}_k(l), \;\; k \in \mathcal{K}, \tag{15}$$

where, $\mathcal{K}$ denotes the set of frequency bins for which the update is performed, so that $\bar{\gamma}_{\mathcal{K}}$ is the mean of γ_k over all $k \in \mathcal{K}$. We considered two ways of selecting the set $\mathcal{K}$. One is to select all those bins for which $\tilde{q}_k$ is larger than a threshold value. The other way, which we adopted, is to consider only those bins that have a relatively low value of γ_k (e.g., ≤ 4.) This may result in more bins in $\mathcal{K}$, but since α_d in (15) is controlled by $\tilde{q}_k$, it could be an advantage.

The adaptation of $\hat{\lambda}_d$ described above not only directly improves the performance of the enhancement algorithm in non-stationary noise environments, but also improves the performance of the VAD itself, as it also depends on the estimated power spectrum of the noise. This further improves the performance of the enhancement algorithm. It should be noted that over-estimation of λ_d by a factor between 1.2 to 1.4 was found helpful and is recommended (the different threshold values should be adjusted accordingly).

In addition, by setting a lower limit to the gain values in each frame one obtains a uniform noise level in noise-only frames with almost no noise structuring. Noise structuring

during noise-only frames can be completely eliminated if a fixed attenuation factor is applied to the whole frame. The fixed gain used in such frames should then be determined so that the noise level remains approximately the same as in frames containing speech. This can be achieved by setting the gain at a fixed fraction of the mean value (in frequency) of the gains applied in the last frame containing speech. Alternatively, the mean of the gains applied in the bins in set $\mathcal{K}$ can be used. As the number of consecutive noise-only frames increases it is useful to allow this gain to decay (but not to zero, to avoid a switching effect when speech begins).

6. EXPERIMENTAL RESULTS AND DISCUSSION

To evaluate the performance of the proposed MM-LSA technique, two formal subjective experiments were conducted. In the first experiment, summarized in Table 1, the MM-LSA was used as a pre-processor for two different speech coders, denoted WI and LD. The 4 kb/s waveform interpolation coder (WI) is a parametric coder and the 16 kb/s G.728 LD-CELP (LD) is a high-quality waveform coder. Two different tunings of the MM-LSA algorithm were used. Because the WI coder is very sensitive to background noise, a more aggressive noise attenuation was applied (L1), resulting in less noise, but a greater amount of speech distortion. A less aggressive tuning (L2) was made for LD-CELP because of its higher quality. L2 tuning without coding was also tested.

The following observations can be made: (i) L2 tuning of the MM-LSA technique resulted in an improvement to the original speech for every condition (including clean speech) and was statistically significant for all 5 noisy background conditions. (ii) L2 tuning improved G.728 performance for every condition and was significant in 4 of 6 conditions. (iii) L1 tuning improved the performance of 4 kb/s WI in 3 of 6 conditions, all of which were significant. In our opinion now, the L1 tuning was too aggressive.

In the second test, summarized in Table 2, the performance of the MM-LSA technique was measured for use with the 7.4 kb/s IS-641 speech coder used in North American TDMA digital PCS. It was compared against no enhancement and using the technique created for the IS-127 standard used for North American CDMA digital PCS.

In this test, the L2 setting of the MM-LSA technique improved IS-641 performance for all 6 conditions and the difference was significant for the 3 car noise conditions. The MM-LSA technique outperformed the IS-127 technique in 5 of the 6 conditions with the differences being significant for the worst cases (10 dB SNR).

In these two tests, the MM-LSA technique showed that when properly tuned it could result in significant improvements in speech coder performance for many of the worst background noise conditions. At the same time, it did not damage the perceived quality of clean speech. Further testing (not presented here due to lack of space) also leads to these conclusions.

Cond.	Clean	Babble		Car Noise		Heli.
		15 dB	20 dB	10 dB	20 dB	5 dB
None	4.16	3.43	3.70	2.91	3.69	2.62
L2	4.26	3.62	3.86	3.09	3.84	3.08
WI	3.24	2.65	2.95	2.02	2.90	1.69
L1+WI	3.14	2.58	2.91	2.22	3.08	2.27
LD	4.01	3.39	3.65	2.92	3.67	2.73
L2+LD	4.19	3.54	3.74	3.06	3.78	3.09

Table 1: MOS scores for the MM-LSA enhancement technique for different conditions (see notation in text) and background noises. Significant difference = 0.14.

Enh.	Clean	Babble		Car Noise		
		10 dB	20 dB	10 dB	15 dB	20 dB
None	4.09	3.08	3.74	2.72	3.17	3.48
L2	4.12	3.21	3.79	3.19	3.58	3.79
127	4.11	2.93	3.75	2.93	3.45	3.80

Table 2: MOS scores for the MM-LSA (L2 tuning) and the IS-127 (' 127') enhancement techniques, followed by coding with the 7.4 kb/s IS-641 coder, for different noise types and intensities. The score for clean coded speech is included (' None'). Significant difference = 0.16.

7. REFERENCES

[1] O. Cappé, "Elimination of the Musical Noise Phenomenon with the Ephraim and Malah Noise Suppressor", IEEE Trans. Speech and Audio Proc., vol. 2, pp. 345–349, 1994.

[2] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator", IEEE Trans. ASSP, vol. 32, pp. 1109–1121, 1984.

[3] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Log-Spectral Amplitude Estimator", IEEE Trans. ASSP, vol. 33, pp. 443–445, 1985.

[4] T. S. Ferguson, *Mathematical Statistics - A decision Theoretic Approach*, Academic Press, Inc., 1967.

[5] S. Lim and A.V. Oppenheim, "Enhancement and Bandwidth Compression of Noisy Speech", Proc. IEEE, vol. 67, pp. 1586-1604, Dec. 1979.

[6] R. J. McAulay and M. L. Malpass, "Speech Enhancement Using a Soft-Decision Noise Suppression Filter", IEEE Trans. ASSP, vol. 28, pp. 137–145, 1980.

[7] P. Scalart and J. Vieira Filho, "Speech Enhancement Based on A Priori Signal to Noise Estimation", ICASSP, 1996.

[8] J. Yang, "Frequency Domain Noise Suppression Approaches in Mobile Telephone Systems", ICASSP, pp. II-363 – II-366, 1993.