Decision Directed Constrained Iterative Speech Enhancement

Amit Das , John H. L. Hansen

Center for Robust Speech Systems (CRSS)
Erik Jonsson School of Engineering and Computer Science
University of Texas at Dallas
Richardson, Texas 75083-0688, U.S.A.
amit.das@colorado.edu, john.hansen@utdallas.edu

Abstract

Earlier studies have shown that degradation due to environmental background noise is non-uniform across various phoneme classes of speech. In this study, we present an improved formulation of single channel constrained iterative speech enhancement (AutoLSP) that follows a *rover* based paradigm. The new approach overcomes some of the drawbacks observed earlier in the baseline AutoLSP system. First, it eliminates the sensitivity to proper determination of the terminating iteration. Second, it employs a phone level non-uniform enhancement approach which significantly improves perceptual quality of the overall uterrance. Third, audible noise components are suppressed by incorporating an auditory masked threshold (AMT) framework. The proposed algorithm is evaluated using Itakura-Saito (IS) objective quality measure over four noise sources and two SNR levels. Comparative evaluations with other baseline systems (AutoLSP, log-MMSE) reveal that the new algorithm exhibits consistent quality improvement for each noise case over all phoneme classes in the TIMIT corpus. Reduction in IS distance over degraded speech is observed in the range of 35.09-46.88%. The Rover scheme outperforms AutoLSP and log-MMSE by 9.21% and 11.19 % respectively using IS scores.

Index Terms: Speech Enhancement, Rover AutoLSP, Vector Quantization, Auditory Masked Threshold

1. Introduction

The objective of any speech enhancement algorithm is to improve the quality and intelligibility of speech degraded in adverse noisy environments. Many algorithms based on a mathematical framework do not take into account improvement in quality of the processed speech from a psychoacoustical sense. Any enhancement scheme can be considered successful if it i) suppresses perceptual background noise and in addition ii) either preserves or enhances perceived speech quality. Although earlier enhancement approaches such as spectral subtraction [4], and other statistical model based approaches like iterative sequential maximum aposteriori (MAP) autoregressive parameter estimators [3] and short time spectral amplitude minimum mean square error (STSA MMSE) estimators [5] have been successful in suppressing noise, they have also introduced processing artifacts and musical noise. The effects of these limitations can degrade the performance of signal processing algorithms used for speech recognition or speaker identification systems that are primarily built for clean speech inputs. Therefore, it becomes imperative to minimize the impact of noise by building front-end robust speech enhancement algorithms.

This work was funded by grants from CDC under contract no. 1541372, and by the University of Texas at Dallas under project EMMITT

2. Baseline Enhancement Algorithm

In this study, we concentrate on the constrained iterative speech enhancement method popularly known as AutoLSP. Originally formulated by Hansen and Clements, a detailed work on it is present in [1, 2]. Briefly, it belongs to a family of single channel speech enhancement schemes based on a two step iterative sequential maximum aposteriori (MAP) estimation of clean speech waveform $\vec{S_0}$ and all-pole autoregressive (AR) speech model parameters ($\vec{a}$, gain G). An order of 10 was considered for AR models used in this study. Two assumptions were made here. First, the unknown parameters $\vec{S_0}$, $\vec{a}$ and G were assumed to be random with apriori Gaussian probability density functions. Next, noise in a given short time frame was assumed stationary with clean speech and noise being statistically independent. A sub-optimal solution to the estimation problem was solved using a sequential two step MAP approach. In the first step, the AR parameters are obtained from the knowledge of clean speech estimate $\vec{S}_{0,i-1}$ at the $(i-1)^{th}$ iteration. In the second step, a new clean speech estimate $\vec{S}_{0,i}$ is obtained by applying a non-causal filter to $\vec{S}_{0,i-1}$. These two steps are iteratively carried out until a terminating iteration is reached. This is summarized as follows:

$$\text{Step 1: MAX } p(\vec{a_i}|\vec{S}_{0,i-1}, \vec{Y_0}; G, \vec{S_I}) \text{ to give } \vec{a_i} \quad (1)$$

$$\text{Step 2: MAX } p(\vec{S_{0,i}}|\vec{a_i}, \vec{Y_0}; G, \vec{S_I}) \text{ to give } \vec{S_{0,i}} \quad (2)$$

In AutoLSP, between (1) and (2), constraints were applied to autocorrelation lags and line spectrum pairs such that the AR model is stable and possess more speech-like characteristics than the traditional spectral subtraction or Lim-Oppenheim Wiener filtering [3] scheme.

3. Algorithm Issues

There are certain drawbacks present in the baseline AutoLSP system. First, it is sensitive to the terminating iteration. The last iteration, which on average is the third or fourth iteration, is purely empirical in nature as there is no specific convergence criterion to determine the optimal terminating iteration for each utterance. While some utterances are best enhanced by the second iteration, there are others that need as many as seven iterations. It is also dependent on the type of noise and the level of degradation. For example, speech degraded by highway noise needs fewer iterations (typically 2) than those that are degraded by additive white Gaussian noise (typically 3 to 7).

The second issue is that while noise suppression for high en-

ergy sections (vowels) of speech is significant, it is sometimes overly suppressed for low energy sections (fricatives, stops) at the terminating iteration, resulting in the introduction of processing artifacts. These artifacts have a pronounced effect on the perceived quality for the entire utterance. Obviously, the number of iterations can be reduced to minimize the artifacts. However, this will leave noise under suppressed for most high energy sections which does not alleviate the problem.

Third, there is usually some level of audible residual noise in the enhanced speech due to errors caused during estimation of the model parameters and noise spectrum.

This paper addresses these issues in the following way. We introduce a Rover based mechanism that exploits an inventory built from different enhancement levels of the baseline system. Using a decision directed approach, the best enhanced frames for different phoneme classes are selected from this inventory and used for reconstruction of the enhanced speech. This removes the dependency of iteration on noise type, noise level and phonetic structure. To further improve the subjective intelligibility of speech, audible noise components can be suppressed using AMT developed originally by Tsoukalas *et al.* [7]. The remainder of the paper is outlined as follows. Sec.4 describes the decision directed approach of the Rover mechanism and the AMT framework. Sec.5 summarizes the experimental evaluations, and Sec.6 presents the conclusions.

4. Algorithm Formulation

4.1. Enhancement Step

A Rover inventory of enhanced frames is created aprior to the decision making step. The enhancement is implemented using the baseline AutoLSP system that iteratively enhances the degraded speech using a non-causal Wiener filtering technique. The Wiener filter can be parameterized by the noise over suppression factor (α) and the exponent term (β). These can be varied to effect different enhancement levels at each iteration (i). The baseline system uses only one set of parameters for 2-3 iterations. However, the Rover system uses 6 iterations, $i \in \{1, 2, ..., 6\}$, of all combinations of $\alpha \in \{1\}$ and $\beta \in \{0.5, 1.0, 1.5, 2.0\}$. Hence, for every degraded utterance 24 different AutoLSP enhanced utterances are produced. The Rover decision making scheme utilizes this (α, β, i) space to pick the best set of enhanced frames. The Wiener filter is represented by,

$$H(\omega) = \left(\frac{\hat{P}_s^i(\omega)}{\hat{P}_s^i(\omega) + \alpha \hat{P}_n(\omega)} \right)^{\beta} \quad (3)$$

where $\hat{P}_s^i(\omega)$ is the apriori power spectrum estimate of speech at the i^{th} iteration and $\hat{P}_n(\omega)$ is the noise power spectrum estimate.

4.2. Codebook Construction

Since it is assumed that the gender of the speaker is known prior to enhancement, vector quantized (VQ) gender based codebooks are used to classify each short-time frame into one of eight broad phoneme classes (vowels, semivowels, nasals, fricatives, affricates, stops, closures, silence). Phoneme classification is critical in the Rover scheme because it employs a set of class dependent search constraints discussed in the next section. 600 TIMIT sentences from the training set were used to prepare noisy codebooks for the noise types and noise levels used in this study. Using a 30 ms frame size with a 75 % overlap, each short time frame was parameterized using 10 dimensional linear predictor cepstral coefficients (LPCC) and log gain coefficient derived from the AR model parameters ([10], pp.376, eq 6.44). The codebooks, of size 512, were constructed using the LBG algorithm [11] and optimized in a minimum mean square error (MMSE) sense. The distance between the test vectors and codebook entries was defined using a cepstral projection measure. This distance metric uses the property that noise corrupted cepstral vectors are less sensitive to angle perturbation and is given by,

$$d(\vec{C}_r, \vec{C}_t) = |\vec{C}_t| - \frac{\vec{C}_t^T \vec{C}_r}{|\vec{C}_r|}. \quad (4)$$

Phoneme classification errors occured mostly due to silence regions being misclassifed as closures on a single frame within a sequence of frames. The errors were corrected by considering the classifications of the leading and trailing frames.

4.3. Decision Making Step

4.3.1. Itakura-Saito(IS) Inventory

We employ a decision directed approach using the IS distance metric [9] because it bears a high correlation with the subjective quality of speech [9]. From an inventory of AutoLSP enhanced corpus of 600 TIMIT sentences, the IS distances between clean and enhanced speech, and degraded and enhanced speech (d_δ) are recorded per phoneme class (δ) where $\delta \in \{$vowel, semivowel, nasal, affricate, fricative, stop, closure, silence$\}$. Using this knowledge, the median and standard deviation per phoneme class ($\mu_\delta, \sigma_\delta$) are determined. These parameters are used in generating the bounds on the search space described in the next section.

4.3.2. Decision Directed Strategy

An effective decision directed strategy is critical to the reconstruction of the overall enhanced speech. The strategy essentially attempts to choose the best set of enhanced frames from the Rover inventory using the knowledge of IS distances (d_δ) obtained from the training set. The degraded speech can be modeled as,

$$\mathbf{y} = \mathbf{x} + \mathbf{n} \quad (5)$$

where **y**, **x**, **n** represent the degraded speech, clean speech and additive noise respectively. Let Γ, of size $N_F \times N_I \times N_\beta$, denote the set of enhanced frames in the AutoLSP inventory for the input degraded speech **y** where N_F is the number of frames per iteration, N_I is the number of iterations per Wiener filter and N_β is the number of Wiener filters. Here, N_I = 6 and N_β = 4 (Sec.4.1) is used. For the degraded speech, the IS distances (d_Γ) over Γ is determined.

If frames $(k, ..., k + n - 1)$ over a particular filter and iteration (β,i) belonging to a single phoneme class δ from time $t_k, t_{k+1}, \ldots, t_{k+n-1}$ are represented by,

$$\gamma^{\delta}_{(\beta,i)} = \{\gamma_k, \gamma_{k+1}, \ldots, \gamma_{k+n-1}\} \quad (6)$$

then the goal is to find $\gamma^{\delta}_{(\beta^\star,i^\star)}$ such that the average d_{IS} is minimized over $\gamma^{\delta}_{(\beta,i)}$ $\forall\, \beta, i$. Given this foundation, the following steps illustrates the decision directed strategy:

1. Using VQ codebooks, find a broad phoneme class (δ) for the next sequence of frames in $[t_k, t_{k+n-1}]$.
2. Evaluate the IS distances,

$$d_{IS}(\mathbf{y}_j, \hat{\mathbf{x}}_j),\ j = k, \ldots, k+n-1, \ \text{ over } \gamma^{\delta}_{(\beta,i)}\ \forall \beta, i \quad (7)$$

where $\hat{\mathbf{x}}_j$ is the enhanced speech in time $[t_k, t_{k+n-1}]$. Set m = 1 to

choose the initial search space given in the next step.
3. Based on δ, choose a search space S_m given by

$$S_m = \{d_\delta : max(0, \mu_\delta - m\omega_1\sigma_\delta) \leq d_\delta \leq \mu_\delta + m\omega_2\sigma_\delta\}. \quad (8)$$

Hence, the search space is bounded by those IS distances in the IS inventory d_δ that lie between the upper and lower bounds (Sec.4.3.1). Here, ω_1, ω_2 represents the weights on the standard deviation. A value of 0.1 is used for both ω_1, ω_2.
4. Based on $d_{IS}(\mathbf{y}_j, \hat{\mathbf{x}}_j)$ and S_m, the best iteration and filter $(\beta^\star, i^\star)$ combination is the one that gives the maximum mean IS distance for $m \leq 3$ and minimum mean IS distance for $m > 3$. This is given by the following equations:

$$(\beta^\star, i^\star) = \arg\max_{\beta,i} \frac{1}{n} \sum_{j=k}^{k+n-1} d_{IS}(\mathbf{y}_j, \hat{\mathbf{x}}_j | S_m, \gamma^\delta_{(\beta,i)}) \; m \leq 3 \quad (9)$$

$$(\beta^\star, i^\star) = \arg\min_{\beta,i} \frac{1}{n} \sum_{j=k}^{k+n-1} d_{IS}(\mathbf{y}_j, \hat{\mathbf{x}}_j | S_m, \gamma^\delta_{(\beta,i)}) \; m > 3 \quad (10)$$

5. Next, determine whether to continue searching or proceed for reconstruction:

(a) If $(\beta^\star, i^\star)$ exists, choose $\gamma^\delta_{(\beta^\star, i^\star)}$ for reconstrunction of enhanced speech and proceed to Step 1.

(b) Else, continue searching by increasing search space size. Set m = m + 1 and go to Step 3.

We feel this strategy improves the performance for the following reasons. First, selection during the first three searches are biased for choosing frames at a maximum IS distance from the degraded speech since frames with lower distances are expected to retain more noise, whereas those closer to the upper bound are expected to be more noise free. On not finding $(\beta^\star, i^\star)$ in S_m, the search space size is increased to accomodate more frames. However, if $m > 3$, then it is likely that the information in the noise free frames are lost due to the presence of overwhelming amount of artifacts. Hence, the selection procedure is *reversed*. Noisy frames near the lower bound are chosen over noise free frames near the upper bound. The core idea behind this has been in finding a trade-off between noise and artifacts. Second, contiguous sequence of frames are selected instead of individual frames over a given class of phoneme sequence. This is done in order to reduce artifacts and impose a level of naturalness to allow a reasonable rate for the speech spectrum to be allowed to change. However, selection of silence regions can be broken into individual frames because limited spectral variations are expected. Selection of silence frames is restricted to no more than 3 contiguous frames at a time. Third, a broad class phoneme level enhancement approach is employed since search spaces are phoneme class dependent. Fig.1 depicts the initial search spaces for vowels and stops degraded by flat communications channel noise at 5 dB.

4.4. Auditory Masked Threshold (AMT)

The basic idea behind psychoacoustic enhancement technique is to suppress those spectral components that contribute to audible noise to an extent that they just become inaudible. A widely used technique to estimate these components is the determination of the Auditory Masked Threshold (AMT) [7] from the enhanced speech. In the proposed framework, an improved technique [8] is incorporated to estimate the AMT (for the case of normal hearing listeners only)

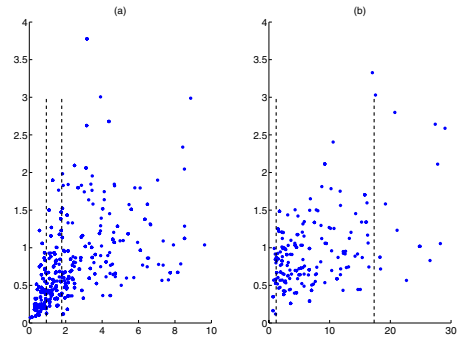

Figure 1: *Initial (m=1) search space (within the vertical lines) for (a) vowels and (b) stops. X-axis: IS(degraded,enhanced). Y-axis: IS(clean,enhanced).*

using the Equivalent Rectangular Bandwidth (ERB). The ERB represents an auditory filterbank model. The steps for calculating the AMT are :

1. Calculate the total energy in each bandpass auditory filter (ERB).
2. Compute the excitation pattern (E) by summing up the power of each signal component with the filter weighting function that is given by the ROEX(p) model, which is described as, $$W(g) = (1 + pg)e^{-pg} \quad (11)$$ where W is the filter shape, p and g are filter parameters. The normalized distance of the signal component (f) from the center frequency (f_c) of the bandpass filter involved is described as $$g = \left(\frac{|f - f_c|}{f_c}.\right) \quad (12)$$
3. Compare the excitation pattern with the absolute threshold of hearing (ATH) to estimate the AMT $$AMT = \max(ATH, E). \quad (13)$$

5. Results

A set of 192 TIMIT test sentences consisting of over 70000 frames was used for objective evaluations using four noise types - flat communications channel noise (FLN), sun cooling fan noise (SUN), in-vehicle wind noise (BL4), large crowd noise (LCR) - and at two noise levels (0, 5dB). Itakura-Saito distance was used as the objective measure to evaluate the results. Fig.2 shows a comparison of the reduction in IS distances of Rover AutoLSP, AutoLSP and log-MMSE enhancement schemes. With the exception of BL4 noise at 0 dB, the Rover scheme consistently exhibited lower IS measures than AutoLSP and log-MMSE over all noise types and SNR levels. The average relative improvement in performance over AutoLSP and log-MMSE for all cases considered in this study was 9.21 % and 11.19 % respectively. Across noise types, the highest percentage reduction in IS distance was observed for FLN noise (48.62% at 0dB) and the lowest for BL4 noise (25.04%). This is not an anomaly since the levels of degradation at the same SNR (0dB) was the lowest for BL4 (IS = 2.24) compared to other noise sources (e.g FLN, IS = 4.21).

Reduction in IS measure was compared across all phoneme classes listed in TIMIT. Fig.3 shows the results for flat communications channel noise at 5dB. For each phoneme class, the Rover scheme always performed better than the baseline systems. However, the performance was only marginally better than AutoLSP

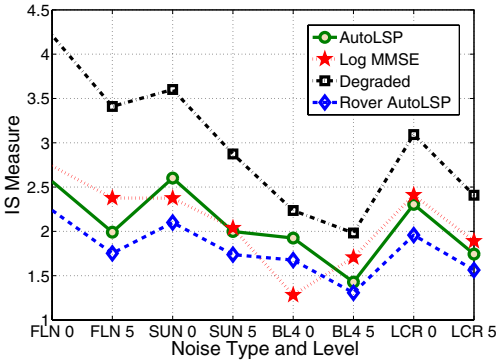

Figure 2: *IS measures across FLN, SUN, BL4, LCR noise types at 0dB and 5dB levels for different enhancement schemes*

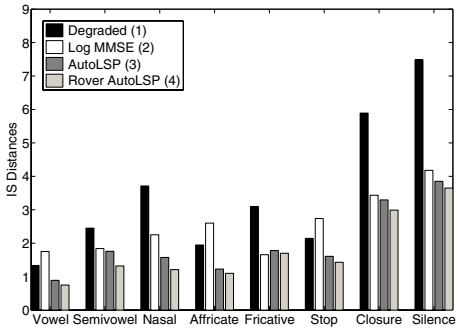

Figure 3: *IS measures across phoneme classes at FLN 5dB for different enhancement schemes*

for affricates, fricatives and stops. Another test on Rover AutoLSP was carried out where clean speech ("*Swing your arms as high as you can*") was used to estimate the best Wiener filter and iteration $(\beta^\star, i^\star)$ for every sequence of frames per phoneme class followed by an evaluation of the mean IS distance. This was compared with the IS distances obtained from the decision directed approach. We noted an average IS difference of 0.0839 for vowels, 0.0784 for semivowels, 0.3713 for nasals, 0.5535 for fricatives, 0.6314 for stops, 1.1519 for closures and 1.3957 for silence sections.

Fig. 4 reports the frame-by-frame IS measure for the degraded speech ("*Don't carry an oily rag like that*"), AutoLSP enhanced speech and Rover AutoLSP enhanced speech. Although the mean IS measure for Rover AutoLSP is lower than AutoLSP, sharp peaks are observed which are possibly the result of codebook phoneme class misclassification or due to a wrong decision in the decision directed strategy. These peaks do not degrade the overall perceptual quality of the enhanced speech since there is usually a reduction in the standard deviation of IS measure suggesting uniform levels of improved quality across the entire utterance. Enhancing further using AMT technique on Rover AutoLSP can siginificantly reduce residual audible noise. Since the IS distance is not a suitable metric to gauge the performance of perceptual based AMT enhancements, we are in the process of carrying out formal subjective listener tests such as those reported in [8]. Most of the computational resource required for the algorithm was used in the preparation the noisy codebooks and IS distance inventories from the trained set. This is not a limitation since these can be prepared offline. However during run-time, Rover search and VQ classification involves moderate to

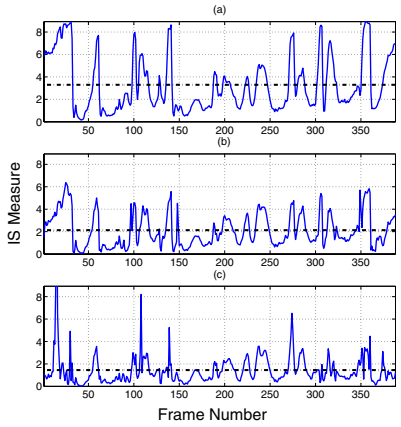

Figure 4: *IS measures across frames (a) Degraded speech (b) AutoLSP enhanced speech (c) Rover AutoLSP enhanced speech. (Legend: Solid line indicates frame IS, Bold dashed line indicates overall mean IS)*

high complexity depending on the size of the Rover inventory generated per degraded test utterance.

6. Conclusions

In this paper, a Rover based scheme was proposed as an improvement over the baseline AutoLSP system that uses a decision directed approach to select the best set of enhanced frame sequences from an inventory consisting of speech at different levels of enhancement. Broad class phone classification was performed using vector quantized codebooks. This approach overcomes some of the shortcomings of the AutoLSP system. It removes the dependency of the terminating iteration and employs phoneme class dependent enhancement. Objective quality evaluations were carried out for four different noise types at two SNR levels. It was shown that Rover AutoLSP consistently improved performace over AutoLSP and log-MMSE enhancement algorithms. The new scheme can be used as a front-end system for non real-time applications.

7. References

[1] J.H.L.Hansen,M.A.Clements,"Constrained iterative speech enhancement with application to speech recognition," IEEE Trans. Sig Proc.,pp. 795-805, Apr 1991.

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

[3] J.S.Lim,A.V.Oppenheim,"All Pole Modeling of Degraded Speech," IEEE ASSP, pp.197-210, Jun 1978.

[4] S.F.Boll,"Suppression of acoustic noise in speech using spectral subtraction," IEEE ASSP, pp.113-120, Apr 1979.

[5] Y.Ephraim,D.Malah,"Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator," IEEE ASSP, pp. 1109-1121, 1984.

[6] Y.Ephraim,"Speech enhancement using a minimum mean square logspectral amplitude estimator", IEEE ASSP, pp 443-445, Apr 1985.

[7] D.E.Tsoukalas,J.N.Mourjoupoulos,G.Kokkinakis,"Speech enhancement based on audible noise suppression," IEEE Trans. Speech & Audio Proc., pp.497-514, Nov 1997.

[8] A.Natarajan,J.H.L.Hansen,K.H.Arehart,J.Rossi-Katz, "Perceptual Based Speech Enhancement for Normal-Hearing & Hearing-Impaired Individuals," pp.1425-1428, Eurospeech 2003

[9] S.R.Quackenbush,T.P.Barnwell,M.A.Clements,"Objective Measures of Speech Quality," Prentice-Hall, NJ, 1988

[10] J.Deller,J.H.L.Hansen,J.Proakis,*Discrete Time Processing of Speech Signals*, Prentice-Hall Publishers, NY,2000

[11] Y.Linde,A.Buzo,R.M.Gray,"An algorithm for vector quantizer design," IEEE Trans Comm, pp.84-95, Jan 1980.