Class Constrained ROVER Based Speech Enhancement

Amit Das, John H. L. Hansen

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

A phoneme class based speech enhancement algorithm is proposed that is derived from the family of constrained iterative enhancement schemes. The algorithm is a Rover based solution that overcomes three limitations of the iterative scheme. It removes the dependency of the terminating iteration, employs direct phoneme class constraints, and achieves suppression of audible noise. In the Rover scheme, the degraded utterance is partitioned into segments based on class, and class specific constraints are applied on each segment using a hard decision method. To alleviate the effect of hard decision errors, a GMM based maximum likelihood (ML) soft decision method is also introduced. Performance evaluation is done using Itakura-Saito, segSNR, and PESQ metrics for four noise types at two SNRs. It is shown that the proposed algorithm outperforms other baseline algorithms like Auto-LSP and log-MMSE for all noise types and achieved a greater degree of consistency in improving quality for most phoneme classes.

Index Terms: constrained iterative speech enhancement, Auto-LSP, Rover, auditory masked threshold, Gaussian mixture models

1. Introduction

The impact of environmental background noise has wide variations across different phoneme classes of speech because each phoneme class has distinct acoustical properties. This variation can be attributed to two primary reasons. First, the phoneme class is characterized by its frequency content and the degree of stationarity of the vocal tract configuration. Second, the bandwidth of the noise influences the level of degradation of the class. For example, narrowband low pass noise like in-vehicle wind noise degrades low frequency vowels more than high frequency fricatives. Therefore, it is important to employ class specific selective enhancement schemes.

Earlier studies have been successful in developing class based enhancement algorithms. McAulay and Malpass [2] adopted a two-state soft-decision ML algorithm in which speech was classified into equally likely binary (silence and non-silence) states. The resultant clean speech ML spectral envelope estimator was a sum of the products of individual MMSE envelope estimators, given the noisy signal and knowledge of the state, and corresponding a posteriori probabilities of the states given the noisy signal. In another study, Hansen and Arslan [4] used hidden Markov models (HMM) to create 13 phoneme class models. ML scores for \( p(\mathbf{x} | \lambda_i) \), \( i = 1, 2, \ldots, 13 \), were obtained where \( \mathbf{x} \) represents the observation vector from noisy speech, and \( \lambda_i \) is the noisy speech HMM model for phoneme class \( i \). The difference of the top two scores was weighted by the inverse of a cost function to evaluate a confidence measure. Enhancement was done selectively based on this measure.

While these methods have used knowledge of the state or class of speech, they have limited benefit of a single enhanced utterance with the introduction of processing artifacts or musical noise. In [4], a constrained iterative speech enhancement model, popularly known as Auto-LSP, is employed which was originally formulated by Hansen and Clements [1]. This scheme is based on a two step iterative sequential maximum a posteriori (MAP) procedure where the 10th order all-pol speech parameters \( (\mathbf{a}, \text{gain } G) \) and clean speech estimates \( (\mathbf{s}) \) are found in two sequential MAP steps in the presence of uncorrelated additive background noise \( \mathbf{y} \). The two MAP steps are carried out until a terminating iteration is reached. This can be summarized as,

\[
\begin{align*}
\text{Step 1:} & \quad \max p(\mathbf{a} | \mathbf{y}, G; \mathbf{s}) \text{ to give } \mathbf{a} \\
\text{Step 2:} & \quad \max p(\mathbf{s} | \mathbf{a}, \mathbf{y}, G; \mathbf{s}) \text{ to give } \mathbf{s},
\end{align*}
\]

where the suffix \( \gamma \) denotes iteration number, and \( \mathbf{s} \) is assumed to be a known initial condition. Constraints were applied to autocorrelation lags and line spectrum pairs between the two MAP steps such that the AR model is stable and possesses more speech-like characteristics. Assuming Gaussian distributed unknowns, it can be shown that the solution to Eq.(2) is equivalent to the MMSE estimate obtained by filtering \( \mathbf{y} \) with,

\[
H(\omega) = \left( \frac{\hat{P}_c^\gamma(\omega)}{P_e(\omega) + \alpha P_n(\omega)} \right)^\beta,
\]

where \( \hat{P}_c^\gamma(\omega) \) is the LP based a priori power spectrum estimate of the clean speech at the \( \gamma \)th iteration and \( P_e(\omega) \) is the noise power spectrum estimate. Since this filter is parameterized by the noise over suppression factor \( (\alpha) \), the exponent term \( (\beta) \) and the iteration \( (\gamma) \), it can be represented as \( H(\theta) \) where \( \theta = (\alpha, \beta, \gamma) \) after dropping the frequency term \( \omega \) for ease of representation.

2. System limitations

There are three limitations present in the Auto-LSP system:

- The terminating iteration, typically the third or the fourth iteration, is based on empirical observations. The last iteration depends on the utterance, type of noise and level

This work was supported by RADC under contract number A40104 and by the University of Texas at Dallas under project EMMITT.
of degradation. More iterations are required for utterances degraded by broadband noise (AWGN) than narrowband noise (in-vehicle wind noise).

- The knowledge of phoneme class is not taken into account. While noise suppression for high energy sections (vowels) of speech is significant, it is sometimes overly suppressed for low energy sections (fricatives, stops), resulting in the introduction of processing artifacts. An immediate step to reduce artifacts is the reduction of number of iterations. However, this will leave noise under suppressed for most high energy sections making the enhanced speech sound noisy 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 overcomes these inherent issues by introducing Rover-based hard and soft decision enhancement algorithms. The remainder of the paper is outlined as follows. The Rover solutions are described in Section 3, followed by evaluations in Section 4, and conclusions summarized in Section 5.

3. Proposed algorithm

In the Rover formulation, the enhancement space is viewed as a three dimensional Auto-LSM system governed by $\theta = (\alpha, \beta, \gamma)$ as in Eq.(3). This space should ensure sufficient minimal to maximal noise suppression for a wide range of noise types and levels. A degraded utterance is enhanced within this enhancement space generating multiple enhanced utterances. In this paper, $\alpha$, $\beta$, and $\gamma$ values lie in the range of 0.25-6, 0.25-4, and 1-6 respectively.

3.1. Phoneme classification

Assuming that the gender of the speaker is known prior to enhancement, a set of 12 dimensional linear predictor cepstral coefficient (LPCC) codebooks are used to classify each 30ms frame belonging to one of 8 broad phoneme classes $\delta$ where $\delta \in \{\text{vowels, semivowels, nasals, affricates, fricatives, stops, closures, silence}\}$. Phoneme classification is critical in the Rover scheme because it defines the boundaries of the search constraints used in the decision making step addressed in the next section. A total of 600 TIMIT degraded utterances are enhanced using the Auto-LSM filter $H(1, 1, 1)$ to prepare class-based codebooks. $H(1, 1, 1)$ is used since it ensures the improvement in recognition rate for all phoneme classes over a wide range of noise types. The distance between the codebook entries ($\vec{C}_r$) and test vectors ($\vec{C}_t$) is defined using a cepstral projection measure since it uses the property that noise corrupted cepstral vectors are less sensitive to angle perturbation. This distance is given by,

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

3.2. Decision phase

The decision phase is the most critical step in the Rover scheme since it utilizes effective hard and soft decisions to select the class based segments from the enhancement space. These decisions are based on a search space evaluated from the Itakura-Saito (IS) distortion [6] objective measure since it possess a high correlation with the subjective quality of speech [6]. Enhanced speech can be obtained from these selected class based segments by aligning them and using overlap add method for reconstruction. The decision phase is explained in the following steps.

In the degraded speech, if the set of frames with frame indices $\{k, k+1, ..., k+n-1\}$ belonging to a class $\delta$ generated from filter $\theta$ is represented by,

$$\chi^\delta(\theta) = \{\chi_k^\delta(\theta), \chi_{k+1}^\delta(\theta), ..., \chi_{k+n-1}^\delta(\theta)\},$$  

then the goal is to find the set of frames $\chi^\delta(\theta^*)$ obtained from filter $\theta^*$ such that IS distortion between clean and enhanced speech can be minimized. This is given as follows:

$$\arg\min_{\theta} E|d(P_s(\omega), P_e(\omega))|\chi^\delta(\theta)|,$$  

where the expectation is taken over all frames $k, k+1, ..., k+n-1$ belonging to the same time segment of class $\delta$.

3.2.1. Hard decision

Based on the foregoing foundation, the following steps illustrate the hard decision strategy:

(a) Using LBG codebooks, class $\delta$ is found for the sequence of frames in $\{k, k+1, ..., k+n-1\}$.

(b) Since clean speech is unknown, the IS distortions are evaluated from degraded speech using,

$$d_j(P_g(\omega), P_e(\omega)), j = k, k+1, ..., k+n-1,$$  

over $\chi^\delta(\theta) \neq \theta$. The search space step $m$ is set to 1 to choose the initial search space as given in the next step.

(c) Based on $\delta$, a search space $S_m$ is obtained from,

$$S_m = \{D_h : \max(0, \mu_\delta - m\omega_3\sigma_3) \leq D_h \leq \mu_\delta + m\omega_3\sigma_3\}.$$  

Hence, the search space is bounded by those IS distortions that lie between the upper and lower bounds determined by class dependent parameters $\mu_\delta$ (median) and $\sigma_3$ (standard deviation). Here, $\omega_1$ and $\omega_2$ are set to 0.1 and represent the backward and forward weights respectively on $\sigma_3$. $\mu_\delta$ and $\sigma_3$ are determined from the training corpus (600 TIMIT sentences).

(d) Considering only those segments $\chi^\delta(\theta)$ whose $d_j(P_g(\omega), P_e(\omega))$ lie in the search space $S_m$, the best filter parameter $\theta^*$ 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:

$$\theta^* = \begin{cases} \arg\max_{\theta} E|d(P_g(\omega), P_e(\omega))|S_m, \chi(\theta), m \leq 3 \\ \arg\min_{\theta} E|d(P_g(\omega), P_e(\omega))|S_m, \chi^\delta(\theta), m > 3 \end{cases},$$  

where the expectation is taken over all IS distortions for frames in $\chi^\delta(\theta)$. It is assumed that each value of $\theta$ is an equally likely candidate for selection. However, in practice, non-uniform weight assignment based on class can give better results.

(e) Next, it is determined whether to continue searching or proceed for reconstruction:

(f) If $\theta^*$ exists, $\chi^\delta_{\theta^*}$ is selected for reconstruction of the enhanced speech followed by proceeding to Step (a) (or by terminating the algorithm if there are no more frames left in the
utterance).

(eii) Else, searching continues by increasing search space step size by one, i.e., \( m = m + 1 \), and then proceeding to Step (c).

3.2.2. Soft decision

Although the hard decision phase attempts to select the best sequence of enhanced frames for a particular phoneme class, there are inconsistencies present in the decision mostly in low energy frames. For example, an undesirable selection of \( \theta^* \) for fricatives may not only suppress noise but also obliterate its acoustical properties. To alleviate this problem, a Gaussian mixture model (GMM) based constrained ML soft decision method is proposed. In this method, instead of finding a scalar value of the parameter \( \theta \), the problem is extended to find a vector parameter \( \vec{\theta}^* \). The LPCCs obtained from \( \chi^\delta \) are weighted to obtain a new LPCC vector. This, in effect, is similar to the procedure of replacing a single filter in the enhancement space with a series of linearly weighted filters. The clean speech phoneme class GMMs are obtained from the training set using 12 dimensional LPCC vectors using the Expectation Maximization (EM) algorithm. The number of mixtures for the GMM models are determined from the size of the broad phoneme class.

In the soft decision method, a predetermined number of filter parameters \( \theta^i \), where \( i = 1, 2, \ldots, N_k \), are chosen based on the number of individual phonemes belonging to the class \( \delta \). Steps (a)-(c) in the soft decision method are identical to the hard decision method. In addition, the following steps are also followed:

(d) If \( n(m, \delta) \) represents the total number of filter parameters \( \theta^* \) to be found at the \( m^{th} \) search step for speech class \( \delta \) and if it is assumed that initial condition \( n(0, \delta) = 0 \) (i.e., no \( \theta^* \) has been found prior to any search step), then the required number of filter parameters to find at the \( m^{th} \) search step is given by,

\[
n(m, \delta) = \min(N_k - \sum_{k=1}^{m-1} n(k, \delta), 4),
\]

where the second argument of \( \min(\cdot) \) operator indicates that the number of filter parameters is restricted to 4 in the current search space even if there are more than 4 to be found.

(e) This is similar to Eq.(9) in the hard decision method except that \( n(m, \delta) \) filters are selected instead of a single filter. This is given as,

\[
\vec{\theta}^*_{n(m, \delta)} = \begin{array}{c}
\arg \max_{\theta^i_{n(m, \delta)}} E[d(P_y(\omega), P_\delta(\omega)|S_m, \chi^\delta(\theta^i)] \quad m \leq 3 \\
\arg \min_{\theta^i_{n(m, \delta)}} E[d(P_y(\omega), P_\delta(\omega)|S_m, \chi^\delta(\theta^i)] \quad m > 3
\end{array}
\]

where \( \vec{\theta}^*_{n(m, \delta)} \) is the top \( n(m, \delta) \) filter parameters for \( \theta^* \).

(f) Next, it is determined whether to continue searching or not. (f1) If \( n(m, \delta) \) values of \( \theta^* \) exist and \( N_k - \sum_{m} n(m, \delta) = 0 \) (i.e., all the top \( \vec{\theta}^*_N \) are found), then choose frames \( \chi^\delta(\vec{\theta}^*_N) \) and use them for reconstruction of enhanced speech using the following GMM weighting procedure before proceeding to Step (a). GMM weights, \( \phi_i \), \( i = 1, 2, \ldots, N_k \), are assigned to the LPCC vectors obtained from the selected sequence of enhanced frames \( \chi^\delta(\vec{\theta}^*_i) \). If \( \vec{x}(\theta^i) \) represents the corresponding LPCCs of \( \chi^\delta(\vec{\theta}^*_i) \), then the soft decision method follows,

\[
\vec{x} = \sum_{i=1}^{N_k} \phi_i \vec{x}(\theta^i),
\]

where, the term \( \phi_i \) is defined by,

\[
\phi_i = \frac{p(\vec{x}(\theta^i)|\lambda_\delta)}{\sum_{\lambda_\delta} p(\vec{x}(\theta^i)|\lambda_\delta)},
\]

where, \( \lambda_\delta \) is the clean speech GMM model for class \( \delta \) and \( p(\vec{x}(\theta^i)|\lambda_\delta) \) is the ML score of the model \( \lambda_\delta \) for the set of feature vectors \( \vec{x}(\theta^i) \). Assuming independence between frames in each segment and using an M-component GMM, the term \( p(\vec{x}(\theta^i)|\lambda_\delta) \) can be further written as,

\[
p(\vec{x}(\theta^i)|\lambda_\delta) = \prod_{k=r}^{k+n-1} \sum_{m=1}^{M} p(\vec{x}(\theta^i)|\lambda_\delta, m)p(m|\lambda_\delta),
\]

since \( \vec{x}(\theta^i) = [\vec{x}_k(\theta^i), \vec{x}_{k+1}(\theta^i), \ldots, \vec{x}_{k+n-1}(\theta^i)] \) and \( \vec{x}_k(\theta^i) \) is the LPCC vector for the frame index \( k \) obtained from filter parameter \( \theta^i \). The individual component density for the \( m^{th} \) mixture is the second term in Eq.(14). If this is denoted by simply \( c(\vec{x}_k) \) and if \( \vec{x}(\theta^i) = \vec{x}_k \), then

\[
c(\vec{x}_k) = \omega(D, m) \exp \left\{ -\frac{1}{2} (\vec{x}_k - \mu_m)^T \Sigma_m^{-1} (\vec{x}_k - \mu_m) \right\},
\]

where, \( \omega(D, m) = \frac{1}{(2\pi)^{D/2}\Sigma_m^{-1/2}} \) and the mean vector and diagonal covariance matrix is given by \( \mu_m \) and \( \Sigma_m \) respectively.

(f2) Else if \( N_k - \sum_{m} n(m, \delta) > 0 \) (not all top values are found yet), searching continues by increasing search space step size, i.e., \( m = m + 1 \), and proceeding to Step (c) as in hard decision.

Speech reconstructed from the Rover solutions usually possess some audible residual noise which is reduced using our auditory masking threshold technique based on equivalent rectangular bandwidths (AMT-ERB) [7]. This is a psychoacoustical enhancement algorithm that suppresses those spectral components that contribute to audible noise to an extent that they just become inaudible.

4. Results

A set of 192 phonetically balanced test utterances from the TIMIT corpus is used for objective quality evaluations. The corpus is sampled at 8kHz and comprised of roughly 69000 frames (240 samples per frame, 75% overlap). Each utterance is degraded with four noise types - flat communications channel noise (FLN), sun cooling fan noise (SUN), in-vehicle wind noise (BL4), large crowd noise (LCR) - at SNRs of 0 and 5 dB. The quality of enhanced speech is assessed using objective speech quality measures such as the Itakura-Saito (IS) distortion (lower the better), segmental SNR (SegSNR) and Perceptual Evaluation of Speech Quality (PESQ) (higher the better). In this section, the hard and soft decision algorithms are referred to as HROV and SROV respectively.

In Fig.1, a summary of the reduction in average IS distortion for all phoneme classes considered in this study is shown for the case of FLN noise at 5dB SNR. With the exception of closures and fricatives, HROV and SROV outperform log-MMSE and Auto-LSF over all other phoneme classes. For example, highest levels of relative improvement over log-MMSE and Auto-LSF are observed for vowels, semivowels (liquids and glides) and silence regions. Although log-MMSE had the best performance for fricatives, there is considerable loss in quality (in comparison to degraded speech) for affricates and vowels by about 45.7% and 12.43% respectively. Also, the relative improvement of log-MMSE is marginal for closures since IS distortion is higher by approximately only 2% and 0.5%
over HROV and SROV respectively which is not significant to the average human listener. This can be mostly attributed to the misclassification of phoneme class or during the selection of the filter parameter $\theta$. When comparing degraded speech with HROV/SROV enhanced speech, high levels of improvement is observed for all phoneme classes with the exception of affricates. However, HROV/SROV exhibited better performance for affricates over the baselines. In summary, the important point to consider here is that the the $Rover$ schemes show higher consistency in improving speech quality for most phoneme classes compared to log-MMSE or Auto-LSP.

In Table 1, the results of all three objective measures are summarized for the case of FLN noise at 0dB and 5dB SNR levels and compared with those of log-MMSE and Auto-LSP schemes. The results presented here are averaged over the entire utterance instead of phoneme classes. An average improvement of approximately 38% over degraded speech is observed in IS distortion with SROV performing the best. In SegSNR results, the improvement from degraded speech varied from 7.44 dB to 9.19 dB for HROV/SROV. However, HROV has the highest improvement over degraded speech by 9.18dB at 0 dB SNR and 7.67dB at 5 dB SNR. The increment in SROV is about 0.2 dB and 0.23 dB below HROV. The contrasting performances in IS and SegSNR results for HROV and SROV is mostly due to the ability of SROV to recover from poor decisions or parameter estimates. However, this comes with a system trade-off that in SROV some amount of artifacts are incorporated during the weighting procedure and hence the slight degradation in SNR. Therefore, there is a trade-off between noise suppression and processing artifacts when using the HROV and SROV methods. PESQ results further emphasize the effectiveness of HROV/SROV over the baselines.

In another experiment, AMT is engaged in all enhancement schemes discussed in this study and PESQ results over all noise types at an SNR of 5dB are tabulated in Table 2. Since the AMT approach [7] followed in this study requires the prior knowledge of clean speech estimate that can be obtained from any of the enhancement schemes, the results for AMT enhancement using degraded speech is not possible. The PESQ results indicate that SROV+AMT performs the best across all noise types resulting in improved levels of speech quality. Among noise types, highest percentage of improvement in PESQ scores is observed for FLN and lowest for BL4. This is not an anomaly since the levels of degradation at the same SNR was the lowest for BL4 (2.45) and highest for FLN (1.81). The complexity of the $Rover$ solutions depend on the size of the enhancement space governed by $\theta$. However, this limitation can be overcome since the enhancement space is class dependent and only selected regions can be effectively used for enhancement without a large compromise on performance.

![Figure 1: Average Itakura-Saito distortion across phoneme classes degraded by FLN noise at an SNR of 0dB over HROV and SROV respectively which is not significant to the average human listener. This can be mostly attributed to the misclassification of phoneme class or during the selection of the filter parameter $\theta$.](image)

Table 1: Itakura-Saito (IS), SegSNR, and PESQ results in FLN noise at SNRs of 0dB and 5dB

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>IS</th>
<th>SegSNR</th>
<th>PESQ</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degraded</td>
<td>5.44</td>
<td>2.67</td>
<td>-7.80</td>
</tr>
<tr>
<td>log-MMSE</td>
<td>2.47</td>
<td>2.72</td>
<td>-0.78</td>
</tr>
<tr>
<td>Auto-LSP</td>
<td>2.35</td>
<td>1.88</td>
<td>-0.21</td>
</tr>
<tr>
<td>HROV</td>
<td>2.22</td>
<td>1.71</td>
<td>3.02</td>
</tr>
<tr>
<td>SROV</td>
<td>2.15</td>
<td>1.65</td>
<td>3.96</td>
</tr>
</tbody>
</table>

Table 2: Performance summary for enhancement schemes with and without AMT using PESQ results in FLN, SUN, LCR and BL4 noises at 5dB SNR

5. Conclusions

In this paper, the problem of enhancing noisy speech using a $Rover$ algorithm was proposed. Hard and soft $Rover$ based solutions are introduced to overcome the limitations of Auto-LSP. It generates multiple enhanced utterances for a given noisy utterance. The noisy utterance is partitioned into class based segments. While a constrained hard decision is used to select one segment from the set of multiple enhanced utterances in every partition, the soft decision linearly combines multiple segments by applying a GMM weighting procedure. Auditory masked threshold based enhancement scheme was integrated within the new algorithms. It was demonstrated that the hard and soft $Rover$ solutions improved speech quality with higher consistency compared to Auto-LSP and log-MMSE enhancement algorithms.

6. References


