Environmentally Aware Voice Activity Detector

Abhijeet Sangwan, Nitish Krishnamurthy, and John H.L. Hansen
Center for Robust Speech Systems (CRSS), University of Texas at Dallas (UTD), Richardson, Texas, U.S.A
{abhijeet.sangwan,nkm052000,john.hansen}@utdallas.edu

Abstract
Traditional voice activity detectors (VADs) tend to be deaf to the acoustical background noise, as they (i) utilize a single operating point for all SNRs (signal-to-noise ratios) and noise types, and (ii) attempt to learn the background noise model online from finite data length. In this paper, we address the aforementioned issues by designing an environmentally aware (EA) VAD. The EA VAD scheme builds prior offline knowledge of commonly encountered acoustical backgrounds, and also combines the recently proposed competitive Neyman-Pearson (CNP) VAD with a SVM (support vector machine) based noise classifier. In operation, the EA VAD obtains accurate noise models of the acoustical background by employing the noise classifier and its prior knowledge of the noise type, and thereafter uses this information to set the best operating point and initialization parameters for the CNP VAD. The superior performance of the EA VAD scheme over the standard AMR (adaptive multi-rate) VADs in low SNR is confirmed in a simulation study, where speech and noise data were drawn from the SWITCHBOARD and NOISEX databases. We report an absolute improvement of 10-15% in detection rates over AMR VADs in low SNR for different noise types.

Index Terms: noise modeling, voice activity detector, environmental sniffing

1. Introduction
Voice activity detectors (VADs) identify the speech and nonspeech parts of a noisy speech signal. The ability to identify speech only regions in an utterance is useful in speech communication systems such as mobile telephony, voice over internet protocol (VoIP), and speech coding [1, 2]. On the other hand, identifying noise only regions assists in boosting performance of speech enhancement and ASR (automatic speech recognition) [3, 4].

Most VAD algorithms are packaged with tunable parameters that impact their performance. It is not always possible to predict the working environment of the VAD, designers are forced to choose an operating point (which is a unique combination of VAD parameter values) that yield a reasonably good performance across a wide variety of noises and different SNRs (signal-to-noise ratios). Unfortunately, choosing a generic operating point is clearly sub-optimal. In order to mitigate the effect of this design difficulty, most VAD algorithms try to learn their operating environment, that is, some statistical characterization of the background acoustic noise that is essential for the operation of the VAD. The learning of background acoustics is generally done online by assuming that the first few frames (generally corresponding to 100-200ms of data) are noise only wherein the necessary statistics are estimated from this data. Thereafter, the subsequent speech-pause decisions are used to update the statistical parameters estimated earlier in order to track its evolution with the progression of the noisy utterance. While this always improves the performance of the VAD, it also poses the problem of learning statistics of the noise signal using extremely short data lengths. This is most critical in poor SNR conditions and non-stationary noise types where the most basic discriminating features between speech and noise (such as energy) become ambiguous, making the operation highly sensitive to initial conditions.

While the search for a VAD with ideal initialization and adaptation of operating parameters continues, such a solution remains elusive. In this paper, we explore an alternate paradigm towards building a robust VAD by attributing intelligence and awareness to the algorithm. We propose the use of noise classification and modeling strategies to build prior offline knowledge of the most commonly encountered acoustical backgrounds. In this manner, the problem of estimating noise statistics from finite data length is modified into a search problem where the short initial segment of data is used for searching the acoustical noise space for a reasonable match. Thereafter, the offline knowledge of that noise type is used to set the optimal operating point for the VAD.

Environmental noise classification is a topic of research for a wide range of fields. The benefits of obtaining useful knowledge of the acoustical background are seen in diverse applications such as mobile telephone, hearing aids and noise event recognition [5, 6, 7]. For example, an environmental sniffing framework proposed by Akbacak and Hansen has shown improvement in performance of speech recognition systems in car environment [3]. Furthermore, Krishnamurthy and Hansen have used the environmental sniffing framework to improve the performance of speech enhancement systems by providing an accurate estimate of the noise update rate required for a given environment [4]. Moreover, many features and classifiers have been employed for the task of noise classification. Maleh, Samouelian, and Khabal have demonstrated that Line Spectral Frequencies (LSF’s) are the most effective features for noise classification [6]. They also evaluated the effectiveness of four pattern recognition algorithms, namely, quadratic Gaussian classifier, least square linear classifier, nearest neighbor classifier, and the decision tree classifier for the noise classification task. In [7], Ma, Milner and Smith used hidden Markov models (HMMs) with mel frequency cepstral coefficients (MFCC’s) to classify noise, and also presented human accuracy for the same classification task.

In this paper, we employ the SVM framework for noise classification owing to its ability to train with small amounts of data [8]. Furthermore, we use the competitive Neyman-Pearson (CNP) VAD as it has a compact set of tunable parameters, cost-efficient implementation, and good performance in comparison to standard VAD algorithms [1]. By combining the CNP VAD and SVM based noise classifier, we form the environmentally aware VAD.

This study was supported in part by RADC under grant A40104 and Intervoice (Richardson, TX)
aware (EA) CNP VAD. Finally, we comprehensively test the EA-CNP VAD on the noisy speech data containing various acoustical backgrounds at different SNRs. The simulation results show appreciable improvement for the EA-CNP VAD over the CNP and AMR-1.2 VADs, especially for low SNR conditions.

2. Review of the CNP VAD

In the work presented here, we utilize the Competitive Neyman-Pearson (CNP) VAD proposed recently in [1] which uses five different parameters for its operation: noisy speech covariance matrix \((K_r)\), noise covariance matrix \((K_n)\), noisy speech update parameter \((\alpha)\), noise update parameter \((\beta)\), and test threshold \((\tau)\). Out of these parameters, \(K_r\) and \(K_n\) are updated through the functioning of the algorithm and \(\alpha\), \(\beta\) and \(\tau\) are fixed apriori. The mathematical form of the statistical likelihood ratio test (LRT) is reviewed briefly below:

Using the binary hypothesis model for the VAD problem, i.e.,

\[ H_0 : \text{pause}, \quad H_1 : \text{speech}, \quad (1) \]

and assuming that the mel filter bank energies (MFBEs) of speech and noise are statistically independent Gaussian random variables, we obtain the likelihood ratio test (LRT) as \([1]\):

\[
\frac{F \times (K_n^{-1} - K_r^{-1}) \times F^T}{\frac{d}{2}} \geq \frac{|F|}{|K_r|}, \quad (2)
\]

where \(|.|\) the determinant, \(F\) is the vector of MFBEs, and \(\bar{z}\) is the average prior SNR \([9]\), \(d\) is the normalized distance term \([10]\) and \(S(.)\) is the well known sigmoid function \([1]\).

During operation, the covariance matrices \(K_r\) and \(K_n\) are updated using a convex combination,

\[
K_r[j] = \alpha K_r[j - 1] + (1 - \alpha)(F[j] \times F^T[j]),
\]

and

\[
K_n[j] = \beta K_n[j - 1] + (1 - \beta)(F[j] \times F^T[j]).
\]

The test statistic (TS) in (2) can be shown to be an estimator of the speech energy in the speech subspace of the noisy speech signal \([1]\). Consequently, the performance of the detector directly depends upon the quality of the noise and speech covariances. Owing to the nature of the task, the constraints on the accuracy of the speech and noise covariances is not extremely demanding but a reasonable approximation should suffice for good operation.

The CNP approach towards building a statistical test was earlier introduced to be a natural extension of the Neyman-Pearson (NP) approach \([1]\). The CNP test can be viewed as a set of NP tests where the actual NP test used to test the speech/pause hypothesis depends upon an observed parameter carrying prior information (e.g., SNR). While the NP approach is guaranteed to be the most powerful test for a given probability of false-alarm \((P_f)\), its was earlier argued that a single choice of \(P_f\) is sub-optimal in the case of VAD where the signal and noise are non-stationary \([1]\). In fact, it was suggested that a \(P_f\) which varies directly with the SNR of the observed signal is more appropriate. For example, a large value of SNR strongly implies the presence of speech and demands the test be biased towards detecting speech and vice-versa. Herein, it is possible to bias the test using a variation in \(P_f\) which manifests in the form of a variable threshold for the statistical test. The need for a variable \(P_f\)

Figure 1: Illustrating the environmentally aware CNP (EA-CNP) VAD scheme.

(3) It is useful to note that the prior SNR \((\bar{z})\) is computed from \(K_r\) and \(K_n\). Moreover, it is straightforward to observe that \(\bar{Q}\) is a function of the noise type and SNR of the noisy speech signal, and therefore the functioning of the CNP VAD is dependent on the quality of the \(K_r\) and \(K_n\) estimates.

3. System Development

In classical settings, the initial estimates of the covariance matrices are generally set to random values which slowly adapt to their true values as the VAD operation stabilizes. The most critical period of adaptation are the initial few frames which are almost always assumed to be noise only. It is easy to see that it is very hard if not impossible to build a reasonable estimate of the noise covariance from such short data. Consequently, in poor SNR conditions the performance of the VAD deteriorates rapidly as the covariance estimates fail to converge to the true values. On the other hand, a reasonable initial estimate for the parameter set \(Q\) would ensure good performance.

We propose the EA-CNP scheme in order to achieve ideal initialization as shown in Fig. 1. Using the commonly encountered noise types, we build good quality estimates of the noise covariances offline. Simultaneously, a SVM based noise classifier is also trained using the available noise data. In operation, the initial 200ms of data is always assumed to be noise only, and that data is used to determine the noise type using the previously trained noise classifier. Upon ascertaining the noise type, the parameters \(K_n\), \(\beta\) and \(\tau\) are also fixed and suitably initialized following which the VAD operation resumes. As before, the speech and noise covariances are updated during speech and noise durations as and when discovered by the EA-CNP VAD.

4. Results and Discussion

The proposed VAD scheme is evaluated on a set of 21 speech samples consisting of 13 male and 8 female speakers which were drawn from the SWITCHBOARD database. Each speech sample was truncated to a duration of 1 minute, and it was ensured that a healthy balance of speech and pause duration is maintained within each file. Using this dataset and an additive noise model,
we created noisy speech samples of -10, -5, 0, 5 and 10dB SNR, using four types of noises: volvo, babble, f16 cockpit and m109 noise from the NOISEX database. A frame length of 20ms is chosen for the proposed VAD scheme.

Furthermore, the entire NOISEX database is employed in order to train the SVM based noise classifier, i.e., the following 15 noise types: babble (bab), bucaneer-1 (buc1), bucaneer-2 (buc2), destroyer-engine (dest-eng), destroyer-operations (destops), F-16 cockpit (f16), factory-1 (fac1), factory-2 (fac2), high-frequency channel (hfch), leopard military vehicle (leop), M-109 tank (m109), machine gun (mgm), pink, volvo (car) and white noise. The noise covariances for each noise type was also estimated offline and stored for later use.

4.1. Noise Classification using SVM

We employ an SVM framework for noise classification since it is well suited for training with limited data [8]. For an efficient fusion with the CNP VAD, we have chosen the MFBEs as the feature vector for the noise classification task. We used 30s of data per noise type for training the SVM with a frame size and overlap of 30ms and 15ms, respectively. In order to test the accuracy of the classifier we also used 17s of noise data per noise type for testing. Each noise frame was allocated to a noise type by the classifier, and the noise classification confusion matrix hence obtained is shown in Table 4.1. It is seen from the confusion matrix that the classification accuracy of most noise types is extremely high. The prominent confusions are shown to be m-109 tank, babble and factory-2 noises which are confused with leopard military vehicle, factory-2, and factory-1 noises, respectively.

In the proposed EA-CNP VAD system, using 200ms of initial noisy speech data we get as many as seven frames to determine the correct background noise type. Once each frame has been suitably classified and an appropriate noise type is assigned, we choose the maximally occurring noise type among the seven frames as the background noise type.

4.2. Choosing the Initial Parameter Set Q

While the appropriate value of $K_n$ is chosen after noise classification, the values of $\beta$ and $\tau$ also need to be determined to fix the operating point $Q$. In order to find the best value of $\beta$ and $\tau$ with varying SNR for a given noise type, we ran simulations of our dataset with various combinations of $\beta$ and $\tau$ across different noise types and SNR conditions. We chose the ideal operating point $Q$ as that combination of parameter values which yielded the best results in term of overall detection. The optimal values of the threshold and noise update rate parameter with varying SNR for different noise types is shown in Table 2. The simulation results are quite interesting as they suggest that the values of $\beta$ and $\tau$ are virtually SNR-independent given a good initial estimate of $K_n$. These results are very significant as a SNR independent approach reduces dependency on SNR estimation, which has proven to be a difficult task. Using the proposed procedure, the VAD can be guaranteed to operate very near the optimal operating point $Q$ thus ensuring best performance. It is also useful to mention that $\alpha$ and $K_T$ are also assigned constant values for the operation of EA-CNP VAD across all noises and SNRs.

4.3. VAD Performance

Figure 2 compares the performance of the proposed EA-CNP VAD scheme against the CNP, AMR 1 and AMR 2 VADs. The superiority of ideal parameter tuning becomes immediately apparent by comparing CNP and EA-CNP VAD performances. It is also seen that the biggest gains are achieved in low SNRs (0dB and below), notably are the improvements of 10-15% at -10dB SNR, m109 and f16 cockpit noises. In non-stationary noise environments like babble, it is observed that the gain in performance at high SNR is minimal which is expected as these noise conditions are time-varying and a good initialization condition alone does not guarantee superior performance. On the other hand, for relatively stationary noise environments like car noise, the performance improvements are consistent across SNR. Finally, it is also interesting to note that the performance of the EA-CNP VAD flattens out below 0dB SNR across all noise types, unlike the other VAD algorithms which rapidly deteriorate in sub-zero SNR conditions.

4.4. Robustness of Noise Classification

While the noise classification task is highly accurate, it has shown two prominent confusions (i.e., factory-2 with babble and m-109 with leopard vehicle). To study the impact of wrong noise classification on the performance of the EA-CNP VAD algorithm, we ran our simulation by swapping babble and m-109 models by factory-2 and leopard noise models, respectively. The impact of swapping the noise models is illustrated in Table 3 where the swapped model results are compared to the use of actual noise models. We have used overall detection (D), speech detection (S), and pause detection (P) to compare the results [1]. As observed, the swapped models outperform the actual models in most SNR conditions. In the case of fac-2 and bab noise mod-

Table 2: EA-CNP VAD operating point (Q) for noise types

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>$\beta$</th>
<th>$\tau$</th>
<th>$K_n$</th>
<th>$K_T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>bab</td>
<td>0.1</td>
<td>13</td>
<td>0.6</td>
<td>pre-trained</td>
</tr>
<tr>
<td>car</td>
<td>0.1</td>
<td>9</td>
<td>0.6</td>
<td>pre-trained</td>
</tr>
<tr>
<td>F-16 cockpit</td>
<td>0.1</td>
<td>9</td>
<td>0.6</td>
<td>pre-trained</td>
</tr>
<tr>
<td>tank</td>
<td>0.1</td>
<td>9</td>
<td>0.6</td>
<td>pre-trained</td>
</tr>
</tbody>
</table>
els, the overall detection (D) has improved due to improvement in speech detection rate (S). On the other hand, improvements in pause detection rate (P) helps leop model to perform better than m109 model. This is believed to be due to two reasons, first it is likely that the substitution of babble noise model by factory2 noise model achieves greater separability of the speech and noise probability distributions in the detection space. This is not very surprising as the long-term statistics (such as covariance) of speech and babble noise is expected to exhibit strong similarities and cause large overlaps in the detection space. Second, the relative stationarity of the confusing noise types has an impact on obtaining quality estimates of covariances. An analysis of the temporal-spectral structure of the noises showed that the leop and f16 noises have a more stationary structure than m109 and bab noises and provides better long term estimates.

5. Conclusion

A VAD which is aware of its environmental conditions is developed and shown to be superior in performance to the standard AMR VADs in noisy environments. We have identified that the use of initial frames of noisy signal may be insufficient in building usable noise models but they are sufficient to search within a broad class of noise conditions for the closest matching noise type. By building prior knowledge about commonly encountered noise types offline and subsequently detecting the acoustical background of a noisy speech signal online, we are able to ensure optimal initialization and operation of the CNP VAD. Further, we have also displayed the robustness of our SVM based noise classifier and illustrated the excellent tuning capabilities of the CNP VAD scheme, which ensures good performance in wide range of SNRs and different noise types. Our novel approach integrates environmental knowledge within the VAD scheme, offering attractive new directions for improved speech/language technology in adverse conditions.

6. References