Noise Tracking for Speech Systems In Adverse Environments

Nitish Krishnamurthy, John H. L. Hansen
Center for Robust Speech Systems, Department of Electrical Engineering, University of Texas at Dallas, Richardson, TX

Abstract

In the design of speech systems, the primary focus is the speech oriented task with the secondary emphasis on sustaining performance under varying operating conditions. Variation in environmental conditions is one of the most important factors that impact speech system performance. In this study, we propose a framework for noise tracking. The proposed noise tracking algorithm is compared with Martin’s [1] and Cohen’s [2] estimation scheme’s for speech enhancement in non-stationary noise conditions. The noise tracking scheme is evaluated over a corpus of three noise types including Babble (BAB), Large Crowd (LCR), and Machine Gun (MGN). The noise modeling scheme for tracking results in a measureable level of improvement for all the noise types (e.g., a 13.7% average relative improvement in Itakura-Saito(IS) measure over 9 noise conditions). This framework is therefore useful for speech applications requiring effective performance for non-stationary environments.

Index Terms: speech enhancement, noise tracking

1. Introduction

In most communication systems, we observe a single channel of speech which is typically corrupted with noise. For most speech enhancement applications, we assume the first few frames of the received signal to be noise and obtain a power density estimate from these initial frames to use in the speech enhancement process. This method works well for stationary noise (i.e., the statistical properties of noise do not change with time), but fails for non-stationary noise signals. The purpose of noise tracking is to estimate the noise in those parts of the input signal where the speech “corrupts” the noise signal. The best possible estimate of the degraded noise is needed to achieve an effective enhancement solution. Let y denote the received speech signal, n and x denote the noise and speech components of the signal respectively. Under additive noise assumption it can be written that,

\[ y = x + n \]  

(1)

if we assume further that the noise and speech are statistically uncorrelated and orthogonal it can be seen that

\[ R_{yy}(\tau) = R_{xx}(\tau) + R_{nn}(\tau) \]  

(2)

where, \( R_{yy}, R_{xx}, \) and \( R_{nn} \) are the received signal, speech and noise autocorrelation’s respectively. If this is a function of time the above equation can be written as,

\[ R_{yy}(\tau, t) = R_{xx}(\tau, t) + R_{nn}(\tau, t) \]  

(3)

This work was funded by RADC under contract A4104, and by Univ.of Texas at Dallas under projet EMMITT.

Another approach to this problem utilizes the fact that the power of the degraded speech is always greater than the power of the noise only part of the signal. Since speech is essentially an intermittent signal where, during voice communications the speaker will pause for short time periods, the noise can be tracked for short durations by tracking the minimum over a window of time.

\[ S_i(k) = \frac{S_{end}(k) - S_{begin}(k)}{N} + S_{begin}(k) \]  

(4)

where the minima of noise is tracked over a window length of \( 2L + 1 \) about the target frame. This approach for noise tracking was first proposed by Martin in [1]. Cohen et al. in [2, 3, 4] proposed approaches where the power spectral densities were weighted using speech presence probabilities before they were used to decide the minimum across the time frames. There has been much work towards obtaining an accurate estimate of the smoothing terms for the recursive estimation of noise and estimation of the signal presence probabilities, these have been noted in [2]. Loizou et al. [5] proposed advancements over the MCRA scheme that adapts faster to changing noise levels. The above cited methods essentially work on the premise that noise changes slowly compared to the change in the rate of speech over a window length,

\[ \frac{dR_{nn}(\tau, t)}{dt} \geq \frac{dR_{nn}(\tau, t)}{dt} \]  

(5)

where for an appropriate time window we can assume,

\[ \frac{dR_{nn}(\tau, t)}{dt} = 0 \]  

(6)

This assumption works well for slowly changing environments, but often fails for environments which change at a rate that is either comparable (babble) or greater (machine-gun) than the time rate of speech. Here, a new approach to the noise tracking problem is proposed. The basis of this approach is that the noise and the noise speech interaction in the environment could be statistically characterized over a period of time prior to noise tracking. With this, the pre-gathered information about the environment can be used for tracking. The large contiguous blocks of noise on either side of the speech utterance are used to build
noise models over time. Similar approaches were used by Sameti et al. in [7] and by Varga and Moore in [8] for speech recognition. The proposed approach here essentially is using a preobserved noise frame from the noise reservoir parts of the signal as a noise estimate. Using these noise only parts allows us to construct a degraded speech reservoir with an available clean speech side-corpus. When a noisy speech frame is observed, the closest matching degraded frame from the database is searched and the preobserved noise that was used to degrade this frame is used as the noise estimate. This works because in most environments even though the environment itself is non-stationary the underlying physical processes remain the same. In babble noise, this essentially implies the assumption that the speakers in the background do not change. Similarly for impulse period noise types like a jackhammer or machine gun noise, the signature of the device does not change with time. In section 2 we describe the algorithm and the clustering method developed for search speedup. Section 4 compares the performance in a speech enhancement framework, with comparisons to both Rainer’s algorithm and Cohen’s algorithm, finally, section 6 describes improvements and future work.

2. Noise Tracking Description

In the proposed approach, previously available information from the environment, or noise available from a reservoir surrounding the speech utterance is used to create noise statistical models. This noise reservoir is used to statistically model speech degraded with additive noise for the particular environment. The closest frame to the degraded speech at the current frame (which we call the target frame) is used to find the closest match of the training data to the test data. Next, the noise used to degrade this training frame is employed as a noise estimate for the current test frame. Define the noisy data as,

\[ y[n] = s[n] + d[n], \]

where \( y[n] \) is a frame of the received signal and \( s[n], d[n] \) are the speech and noise signal respectively. The power density spectrum is calculated using the assumption of zero mean noise that is independent of speech,

\[ |S_y(\omega)|^2 = |S_s(\omega)|^2 + |S_d(\omega)|^2. \]

Here, let \( \hat{d}[n] \) be an estimate of \( d[n] \) such that we minimize,

\[ \arg \min_{\hat{d}[n]} ||y - \hat{y}|^2 \]

where,

\[ \hat{y}[n] = \hat{s}[n] + \hat{d}[n] \]

and where \( \hat{d} \) and \( \hat{y} \) are extracted features from the test and the target frame. Extracted features are used in order to reduce the dimensionality of the data. Furthermore, in order to increase the efficiency of the process, the data is clustered into predetermined groups and after assigning the current test frame to a cluster a closest matching frame within the cluster is determined.

In the proposed setup described in Fig 1, 19 dimensional MFCC’s are used as feature vectors, since no direct one to one mapping exists from the MFCC to the signal, the MFCC’s are tagged along with the noise belonging to the frame and the noisy frame itself. The algorithm process is described below.

Step 1 Extract the noise only parts from the noisy speech signal.

Step 2 Use this noise data to degrade a secondary clean speech data corpus. Save this degraded data corpus.

Step 3 For each window frame, extract the feature vector and keep the noise signal used to degrade the secondary speech frame.

Step 4 Cluster all the extracted features from the secondary degraded speech into 128 clusters/mixtures using a GMM (Gaussian Mixture Model) [6].

Step 5 Extract the feature vectors from the input noisy speech.

Step 6 For each feature vector, find the most likely GMM mixture component and within this mixture component, find the nearest degraded MFCC vector to this particular feature vector (target degraded speech), using the Euclidean distance.

Step 7 Employ the noise that was used to degrade the target degraded speech frame as the noise estimate. Here a GMM was used over other clustering setups because for other applications, we can directly use the second order statistics provided by the GMM structure.

3. Implementation

As noted earlier, the noise reservoir extracted from the input audio stream is used to degrade a secondary clean speech corpus. The only separation in the model is based on the gender of the speaker. The combination of degraded speech, clean speech, and noise used to degrade the clean speech are frame indexed. To characterize speech, 19 dimensional Mel Frequency Cepstral Coefficients (MFCC’s) are extracted from the degraded speech frames to construct a single GMM with 128 mixtures. Each training MFCC frame is associated with the 128 GMM model-based on their scores across each of the Gaussian components. A frame is therefore associated with the maximum scoring Gaussian component. When a test degraded speech frame arrives, its MFCC vector is extracted and then scored against each of the 128 Gaussian components. The test vector is assigned to the maximum scoring Gaussian component. All the training MFCC vectors assigned to such a Gaussian component are compared to the test vectors using a Euclidean distance and the MFCC vector with the minimum distance is chosen. In this scheme, we have previously indexed/tagged the clean noise frame (training), the speech used to degrade the clean frame (training), and the degraded frame to this MFCC vector. We call this new frame the candidate frame. The noise frame used to degrade the candidate frame is employed as a noise estimate for the test frame. This operation is performed for the entire utterance resulting in a frame by frame noise estimate sequence.

4. Analysis

To evaluate the proposed noise tracking algorithm, a single test file was degraded from the 192 TIMIT core test sentence set. Only male speakers were used for constructing the above models. Babble noise at an SNR of 5 dB was used to degrade the speech file. To ensure that the test and train noise sequences are not the same, different noise observations were used, where a sample from a large crowd “booming” was employed. This environment was chosen because of its non-stationary nature and potential impact on speech enhancement algorithms. The degraded

1In principle only a single noise pdf is selected from the noise GMM. In practice however, it is possible that more than one noise is present for a given frame, or that a combination of perhaps two noise types, better models the noise in this frame. For this study, we concentrate on selection of a single noise pdf.
speech frames were clustered into a 128 mixture GMM. A noise tracking algorithm was used to estimate the noise distortion in the speech degraded sections on a per frame basis. This noise estimate is used as the true noise to enhance the speech using the Log-MMSE \cite{9} algorithm. While advancements to Log-MMSE have been developed (e.g., GMMSE-AMT-ERB \cite{10}), we selected the traditional Log-MMSE scheme to emphasize the noise tracking problem on a traditional, well accepted method. Fig. 4 shows the resulting waveforms and spectrograms of degraded and enhanced speech corrupted with babble noise. For purposes of comparison the Martin noise tracking \cite{1} methods was employed (Fig. 4d) along with the new noise tracking method (Fig. 4e,f). It can be observed that the speech has been enhanced under extremely noisy conditions. After the enhancement process, some level of music-like artifacts are present in the background, which are believed to be residual formants of the corrupting babble noise which persist after log-MMSE enhancement. Our focus here is not to formulate a better enhancement algorithm but to formulate better means of tracking noise across time. The speech portions of the original signal have been preserved. There are some artifacts in the beginning silence section of the utterance, which are believed to be due to non-matching noise frames in the test and train section.

5. Speech Enhancement Using Noise Tracking

Having illustrated the performance of the proposed noise tracking algorithm for a single sentence we now turn to a more comprehensive evaluation over a larger corpus to illustrate that the method scales up to general speech applications. The noise tracking algorithm was evaluated next for three noise conditions LCR (large crowd noise), BAB (Babble noise), and MGN (Machine Gun Noise). These four noise types have different levels of stationarity. BAB and MGN are non-stationary noise types where LCR is more stationary. These were used to degrade TIMIT sentences at SNR levels of -5 dB, 0 dB, and 5 dB. For these noise types, different noise samples were used for training and test phases. A set of 192 sentences were randomly chosen from the TIMIT corpus that were different from those used for training. A total of 6912 sentences were used to obtain the results (192 sentences X 3 SNR’s X 3 noise types X 4 algo’s). The Itakura-Saito (IS) distance measure was used to assess objective speech quality performance. As seen in table 1, the proposed noise tracking scheme either measurably outperforms other schemes (in 7 of 9 conditions) or produces comparable enhancement output for most cases. The relative improvements are calculated using,

$$R_i = \frac{IS_{degraded} - IS_{enhanced}}{IS_{degraded}} \times 100 \quad (12)$$

Table 1 shows the IS values of the degraded speech enhanced using noise estimated with established schemes and the speech enhanced using the proposed estimation scheme. As seen from these evaluations the quality of enhancement depends heavily on the stationarity of the noise. An average 13.71% improvement in IS measure is obtained using the new tracking algorithm. The stationarity of the noise signal can also be used to decide how many noise updates to use per frame when noise only frames are available. The noise estimates can be improved by incorporating information about the current state of noise.

Table 1: Comparison of enhancement performance in different environments. (a) Original degraded quality, (b) quality of enhanced speech using, (b) Cohen’s \cite{2}, (c) Martin’s \cite{1}, and (d) proposed noise tracking schemes.

<table>
<thead>
<tr>
<th>Noise Type</th>
<th>(a)</th>
<th>(b)</th>
<th>(c)</th>
<th>(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BAB</td>
<td>-5 dB</td>
<td>4.13</td>
<td>3.94</td>
<td>3.89</td>
</tr>
<tr>
<td></td>
<td>0 dB</td>
<td>3.46</td>
<td>3.27</td>
<td>3.28</td>
</tr>
<tr>
<td></td>
<td>5 dB</td>
<td>2.71</td>
<td>2.55</td>
<td>2.70</td>
</tr>
<tr>
<td>MGN</td>
<td>-5 dB</td>
<td>3.71</td>
<td>4.13</td>
<td>4.96</td>
</tr>
<tr>
<td></td>
<td>0 dB</td>
<td>3.22</td>
<td>3.60</td>
<td>4.45</td>
</tr>
<tr>
<td></td>
<td>5 dB</td>
<td>2.08</td>
<td>6.84</td>
<td>4.66</td>
</tr>
<tr>
<td>LCR</td>
<td>(a)</td>
<td>(b)</td>
<td>(c)</td>
<td>(d)</td>
</tr>
<tr>
<td></td>
<td>-5 dB</td>
<td>4.69</td>
<td>4.55</td>
<td>4.99</td>
</tr>
<tr>
<td></td>
<td>0 dB</td>
<td>4.01</td>
<td>4.87</td>
<td>3.97</td>
</tr>
<tr>
<td></td>
<td>5 dB</td>
<td>2.83</td>
<td>2.87</td>
<td>3.60</td>
</tr>
<tr>
<td>avg rel improv</td>
<td>-16.9%</td>
<td>-14.06%</td>
<td>13.71%</td>
<td></td>
</tr>
</tbody>
</table>

6. Conclusion

In this study a novel method for noise tracking in non-stationary environments was proposed. The method builds a degraded speech GMM using a secondary clean speech corpus, and matches the input degraded speech frames to the best pdf in the GMM, with the noise only data drawn from the tagged pdf’s in the GMM. Results from an extensive evaluation over 6912 sentences show that the proposed method outperforms Martin’s as well as Cohen’s noise tracking methods \cite{1,2}. For nonsta-
tionary noise types including babble (BAB) and machine-gun (MGN), improvement in IS objective speech quality was between 1-2. This framework is especially suited for highly non stationary environments where the rate of change of environment is higher than or comparable to the rate of change of the speech signal. One of the shortcomings of the nearest-neighbor frame selection approach is that each noise frame chosen is independent of the adjacent noise frames, so a loss in continuity in adjacent frames is experienced. It is possible to achieve better performance for the proposed system by using trajectory-based modelling/smoothing of the environment. Better approaches to finding the closest matching frame could also be employed. Advances such as MCRA and the proposed noise tracking scheme can be integrated into the general framework. This integration would enable performance over a larger set of noise classes.

7. Acknowledgment
The authors would like to thank Dr. Philip Loizou and Yang Lu from CRSS-UTD for providing us with the current noise tracking schemes to compare our algorithm.

8. References