Robust Automatic Speech Recognition using an Optimal Spectral Amplitude Estimator Algorithm in Low-SNR Car Environments

Zili Li, Hesham Tolba & Douglas O'Shaughnessy

INRS-EMT, Université du Québec
800 de la Gauchetière Ouest, Québec, H5A 1K6, Canada
{zili, tolba, dougo}@inrs-emt.uquebec.ca

Abstract

This paper addresses the problem of noise robustness of automatic speech recognition (ASR) systems in noisy car environments using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator (MMSE-STSA). This was accomplished by the integration of an adaptive time varying Noise Shaping Filter (NSF) with the MMSE-STSA algorithm in order to improve the speech enhancement performance by “whitening” the noisy speech signals. Experiments were conducted using a noisy version of speech signals extracted from the TIMIT database. The proposed NSF-based STSA algorithm is used as a processor of an ASR system in order to evaluate its robustness in severe interfering car noise environments. The HTK Hidden Markov Model Toolkit was used throughout our experiments. Results show that the proposed approach, when included in the front-end of an HTK-based ASR system, outperforms that of the conventional recognition process in severe interfering car noise environments for a wide range of SNRs down to -12 dB using a noisy version of the TIMIT database.

1. Introduction

One of the major challenges of the speech recognition problem is to make the system robust to background noise [1, 2, 3]. The performance of speech recognition systems dramatically decreases when they are trained (in noise-free) and used in different (noisy) environments. A recognizer can provide good performance even in very noisy background conditions if the exact same (or approximate) testing condition is used to provide the training material from which the reference patterns of the vocabulary are obtained.

Two main approaches to the problem of achieving robust speech recognition in noise can be defined: compensation during the data preprocessing stage; or compensation during the recognition stage. The first approach is classified into two classes. One suppresses the noise component in the speech signal before it is compared with the existing reference patterns in the recognizer. Well-known procedures of this type include Spectral Subtraction or Wiener filtering to remove an estimate of noise from noisy speech observation parameters [4]. The other one is focused on the development of distance measures that are robust to noise contamination. In this case, there will be no need to create noisy patterns or to process the signal before recognition. The second approach adapts the clean speech models to noise.

In this paper, the first approach of robustness of ASR systems in noisy car environments is adopted. The speech enhancement approach that is used to pre-process the speech is based on a modified version of the MMSE-STSA approach that was proposed in [5]. The MMSE-STSA algorithm has been shown asymptotically near optimal for signals corrupted by additive white noise [5]. This motivated us to integrate an adaptive time varying Noise Shaping Filter (NSF) in the MMSE-STSA algorithm that is used to enhance the noisy speech before the recognition process. The proposed NSF-based MMSE-STSA algorithm is used in such an environment to improve the speech recognition performance by “whitening” the noisy speech signals.

The outline of this paper is as follows. In section 2, an overview of the MMSE-STSA speech enhancement algorithm is given. Then in section 3, we describe our proposed approach that is based on the modification of the classical MMSE-STSA algorithm to deal with narrow-band noise signals. Then, we proceed in section 4 with the description of the database, the platform used in our experiments and the evaluation of the proposed recognizer in a noisy car environment and the comparison of such a recognizer to the baseline recognizer in order to evaluate its performance. Finally, in section 5 we conclude and discuss our present and future work.

2. Optimal Spectral Amplitude Estimator for Speech Enhancement

Ephraim and Malah presented in [5] an algorithm for enhancing speech degraded by uncorrelated additive noise when the noisy speech alone is available. Their basic approach is to optimally estimate the two components of the short-time Fourier transform (STFT) separately, rather than estimating the STFT itself. The estimation of both the short-time spectral amplitude (STSA) and the corresponding phase of the speech signal was based on a Minimum Mean Square Error (MMSE) criterion and an assumed statistical model. Then the optimal MMSE STSA estimator was combined with the optimal estimator of the complex exponential of the phase which does not affect the STSA estimation. The latter constrained complex exponential estimator was found to be the complex exponential of the noisy phase. Ephraim and Malah derived an STSA estimator which minimizes the mean-square error of the log-spectra under the assumption that the Fourier expansion coefficients of the original clean signal and the noise may be modeled as independent, zero-mean, Gaussian random signals [6].

The basic building blocks for such a speech enhancement system are shown in Figure 1. The inputs to the system are 16-kHz sampled noisy speech degraded by uncorrelated additive noise and its corresponding reference noise. Each frame of input noisy signal or reference noise is spectrally decomposed by means of a Fast Fourier Transform (FFT) after a Hanning window. The square values of spectral amplitudes of reference noise from different frames are averaged to produce $\lambda_k(k)$ as shown in Figure 1. The input noisy
speech undergoes a similar process to produce $R^k_a$, as shown in Figure 1, except that its phase $\theta_k$ is also derived and reserved for the construction of the estimated signal at the end. The $a$ posteriori SNR $\gamma_k$ is defined as the ratio of $\lambda_t(k)$ to $R^k_a$, while the $a$ priori SNR $\xi_k$ is estimated in a slightly more complex recursive process. For the first frame of noisy speech, an initial estimated value is given by

$$\xi^1_k = \alpha + (1 - \alpha)(\gamma^1_k - 1),$$  \hspace{1cm} (1)

where $\alpha$ is a constant of value 0.98. For the other frames $\{m, \ 1 < m \leq M\}$ where $M$ is the total number of frames,

$$\xi^m_k = \begin{cases} \alpha \xi^{m-1}_k + (1 - \alpha)(\gamma^m_k - 1) & \text{if } \gamma^m_k > 1 \\ \alpha \xi^{m-1}_k & \text{if } \gamma^m_k \leq 1. \end{cases}$$  \hspace{1cm} (2)

As shown in Figure 1, the desired STSA estimator $\hat{A}_k$ is calculated based on the weighting function, $H(\xi_k, \gamma_k)$ which depends on the $a$ priori and $a$ posteriori SNR values, $\xi_k$ and $\gamma_k$, respectively. The weighting function $H(\xi_k, \gamma_k)$ of the MMSE-STSA algorithm was also referred to as the gain of the estimator and was shown to increase when the SNR values decrease. $\hat{A}_k$ of each frame is calculated and combined with the complex exponential of the noisy phase $\theta_k$. The resulting DFT samples $\hat{X}(k)$ in each analysis frame are then transformed back into the time domain to obtain $\hat{x}(k)$, which are used to synthesize the enhanced speech signal using the overlap-and-add method. For more details see [5].

### 3. NSF-based Speech Enhancement

The MMSE-STSA algorithm introduced in section 2 has been shown to be asymptotically near optimal for the signals corrupted by additive white noises. However, for some narrow band noise or noise with evident and stable spectral peaks, an adaptive time varying filter can be used additionally to suppress the frequencies where the noise energy is high. This can help to improve the speech enhancement performance of the STSA by “whitening” the noisy speech signals, especially for some signals corrupted by narrow-band noise, e.g., car noise. The time varying Noise Shaping Filter (NSF) is designed for this purpose. The design of NSF has taken into account that the spectral characteristics of the noise change remarkably slower than those of the clean speech, as well as the bandwidth broadening effect of the noisy speech. After filtered by the NSF, amplitudes of the noisy speech at the frequencies where the reference noise has spectral peaks are de-emphasized slightly.

The above-mentioned NSF is designed via LPC analysis. As one of the most powerful speech processing techniques, LPC analysis can provide a reasonable estimation of the spectrum. LPC analysis and bandwidth expansion play important roles in the construction of the NSF filter. To design such a filter, the LPC analysis and bandwidth expansion are applied to the reference noise itself and not to the noisy speech. Moreover, average values of the LPC coefficients of several adjacent analysis frames are used instead of instantaneous values from only one frame, since it is a reasonable assumption that the additive noise changes more slowly than speech. This re-
inforces the consistent noise components, especially harmonics, at the expense of random noise components. Both the averaging operation and bandwidth expansion smear the spectrum and ignore abrupt changes, if any. The transfer function of the NSF is of the following form:

$$H(z) = \frac{A(z | \rho)}{A(z | \rho)} = \sum_{i=0}^{p} \tilde{a}_i \rho_i^2 z^{-i} \sum_{i=0}^{p} a_i \rho_i^2 z^{-i}$$

where $A(z)$ is related to the transfer function of the LPC synthesis filter as follows:

$$H(z) = \frac{1}{A(z)} = \sum_{i=0}^{p} a_i z^{-i}.$$ 

$p$ is the LPC order, and $\{\tilde{a}_i\}$ are the mean values of predictor coefficients from different analysis frames of the reference noise. The new coefficients $\{\tilde{a}_i\}$ in Eq. 3 are average values from the adjacent analysis frames of the reference noise, rather than the instantaneous predictor coefficients from a single analysis frame of noisy speech. The parameters $\rho_1$ and $\rho_2$ are bandwidth expansion factors that can expand formant bandwidths by moving poles or zeros radially inward towards the center of the unit circle. The bandwidth expansion factors $\rho_1$ and $\rho_2$ and the LPC analysis order $p$ have the following values: $\rho_1 = 0.95, \quad \rho_2 = 1, \quad p = 10$.

The implementation of the above-mentioned NSF algorithm is illustrated in the block diagram of Figure 2. The input to the system is the reference noise. Each analysis frame consists of 256 samples of reference noise using a Hanning window, and overlaps the previous analysis frame by 128 samples. In practical applications, the reference noise can be collected with a second microphone working simultaneously or recorded with the same microphone during short pauses of noisy speech. The core process is based on the LPC analysis, which generates the LPC coefficients $\{a_i\}$ of each frame of the reference noise. The values of $\{a_i\}$ from different reference noise are averaged to produce $\tilde{a}_i$ of Eq. 3. The mean LPC coefficients $\tilde{a}_i$ are then scaled by multiplying the powers of expansion factors $\rho_1$ and $\rho_2$, so as to expand the bandwidth of the peaks and the valleys, respectively. The scaled mean LPC coefficients are considered as the forward and backward coefficients of the desired NSF.

### 4. Experiments & Results

#### 4.1. Database

In the following experiments the TIMIT database, described in [7], was used. The TIMIT corpus contains broadband recordings of a total of 6300 sentences, 10 sentences spoken by each of 630 speakers from 8 major dialect regions of the United States, each reading 10 phonetically rich sentences. To simulate a noisy environment, car noise was added artificially to the clean speech. To study the effect of such noise on the recognition accuracy of the ASR system that we proposed, the reference templates for all tests were taken from clean speech. On the other hand, the $dr_1$ subset of the TIMIT database was chosen from the available database to evaluate the recognition system.

#### 4.2. Recognition Platform

In order to recognize the continuous speech data that has been enhanced as mentioned above, the HTK toolkit described in [8] has been used throughout all experiments. This toolkit is used to build an HMM-based speech recognition system. The HTK toolkit can be used for isolated or continuous whole-word-based speech recognition. The toolkit was designed to support continuous-density HMMs with any numbers of state and mixture components. It also implements a general parameter-tying mechanism that allows the creation of complex model topologies to suit a variety of speech recognition applications. For more details see [8].

#### 4.3. Tests & Results

In all our experiments, 12 MFCCs were calculated on a 30-msec Hamming window advanced by 10 msec each frame. Then, an FFT is performed to calculate a magnitude spectrum for the frame, which is averaged into 20 triangular bins arranged at equal Mel-frequency intervals. Finally, a cosine transform is applied to such data to calculate the 12 MFCCs. Moreover, the normalized log energy is also found, which is added to the 12 MFCCs to form a 13-dimensional (static) vector. This static vector is then expanded to produce a 39-dimensional vector (including 13 static coefficients, 13 delta coefficients and 13 acceleration coefficients) upon which the hidden Markov models (HMMs), that model the speech subword units, were trained. The baseline system used for the recognition task uses a triphone Gaussian mixture HMM system.

Applying the overall proposed recognizer to the noisy version of the TIMIT database under different SNRs, which vary between almost -12 and 20 dB, and performing some experiments proved that the recognition accuracy has increased significantly. In order to evaluate the performance of our proposed ASR system, we compared the performance of the NSF-based HTK recognizer to the baseline HTK recognition system. Table 1 shows a comparison of the percent word correctness rate ($\%C_{\text{word}}$), recognition accuracy ($\%\text{Acc}_{\text{\%}}$), and the degradations in the recognition performance represented by the deletion ($\%\text{Del}_{\text{\%}}$), substitution ($\%\text{Sub}_{\text{\%}}$) and insertion ($\%\text{Ins}_{\text{\%}}$) percentage errors of the MMSE-STSA-based HTK ASR system to the baseline HTK using single mixture triphones and the $dr_1$ subset of the TIMIT database when contaminated by additive car noise for different values of SNR.

Figure 3 illustrates the word recognition correctness rates obtained...
in these ASR tests and Table 1 gives some other detailed results. In Fig. 3, the dashed line at the top denotes the word recognition correctness rate (95.54%) of the clean speech. This can be considered as a baseline compared with that of the noisy speech and the enhanced speech. The lowest dashed line denotes the recognition correctness rates of the noisy speech (speech corrupted by the car noise). It decreases rapidly as the SNR level decreases and shows that ASR performance is sensitive to additive noise. The two solid lines give the word recognition correctness rates of the enhanced speech by both the MMSE-STSA and the NSF-based MMSE-STSA approaches.

As shown in Figure 3, the ASR results can be improved evidently after the corrupted speech is enhanced by the use of the proposed algorithm. The improvements of the ASR word correctness rates are up to 50%. Very good results were obtained by using the NSF-based MMSE-STSA preprocessor algorithm in the front-end, especially at low SNR. For example, for speech corrupted by car noise with a SNR value of -12 dB, using only the MMSE-STSA algorithm, the ASR correctness rate is improved by about 17%, whereas using the NSF-based MMSE-STSA algorithm, the ASR rate is improved by about 28%.

Table 1: Comparison of the percent word recognition performance (%CWrd), recognition accuracy (%AccWrd), deletion (%δDel), substitution (%δSub) and insertion (%δIns) percentage errors of the MMSE-STSA, NSF-based MMSE-STSA HTK ASR systems to the baseline HTK using the Dr1 subset of the TIMIT database when contaminated by additive car noise for different values of SNR.

<table>
<thead>
<tr>
<th></th>
<th>-12 dB</th>
<th>-8 dB</th>
<th>-4 dB</th>
<th>0 dB</th>
<th>4 dB</th>
<th>8 dB</th>
<th>12 dB</th>
<th>16 dB</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Noisy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%CWrd</td>
<td>5.21</td>
<td>10.83</td>
<td>25.31</td>
<td>42.08</td>
<td>60.73</td>
<td>73.02</td>
<td>80.21</td>
<td>88.02</td>
</tr>
<tr>
<td>%AccWrd</td>
<td>0.00</td>
<td>-4.90</td>
<td>-6.77</td>
<td>7.71</td>
<td>31.46</td>
<td>51.25</td>
<td>65.42</td>
<td>78.23</td>
</tr>
<tr>
<td>%δDel</td>
<td>37.50</td>
<td>17.92</td>
<td>5.63</td>
<td>3.54</td>
<td>3.02</td>
<td>1.98</td>
<td>1.25</td>
<td>1.15</td>
</tr>
<tr>
<td>%δSub</td>
<td>57.29</td>
<td>71.25</td>
<td>69.06</td>
<td>54.37</td>
<td>36.25</td>
<td>25.00</td>
<td>18.54</td>
<td>10.85</td>
</tr>
<tr>
<td>%δIns</td>
<td>5.21</td>
<td>15.73</td>
<td>32.08</td>
<td>34.37</td>
<td>29.27</td>
<td>21.77</td>
<td>14.79</td>
<td>9.79</td>
</tr>
<tr>
<td><strong>MMSE-STSA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%CWrd</td>
<td>21.56</td>
<td>44.38</td>
<td>61.04</td>
<td>71.25</td>
<td>77.81</td>
<td>85.31</td>
<td>87.08</td>
<td>90.83</td>
</tr>
<tr>
<td>%AccWrd</td>
<td>-6.77</td>
<td>18.65</td>
<td>38.75</td>
<td>54.79</td>
<td>66.04</td>
<td>77.08</td>
<td>80.94</td>
<td>87.19</td>
</tr>
<tr>
<td>%δDel</td>
<td>6.56</td>
<td>3.44</td>
<td>2.60</td>
<td>2.71</td>
<td>2.29</td>
<td>1.88</td>
<td>1.15</td>
<td>1.04</td>
</tr>
<tr>
<td>%δSub</td>
<td>71.87</td>
<td>52.19</td>
<td>36.35</td>
<td>26.04</td>
<td>28.04</td>
<td>19.90</td>
<td>12.81</td>
<td>11.77</td>
</tr>
<tr>
<td>%δIns</td>
<td>28.33</td>
<td>25.73</td>
<td>22.29</td>
<td>16.46</td>
<td>11.77</td>
<td>8.23</td>
<td>6.15</td>
<td>3.65</td>
</tr>
<tr>
<td><strong>NSF-MMSE-STSA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%CWrd</td>
<td>32.50</td>
<td>51.15</td>
<td>67.08</td>
<td>74.58</td>
<td>78.85</td>
<td>83.33</td>
<td>86.88</td>
<td>89.90</td>
</tr>
<tr>
<td>%AccWrd</td>
<td>6.25</td>
<td>31.46</td>
<td>52.92</td>
<td>64.58</td>
<td>71.35</td>
<td>77.40</td>
<td>82.19</td>
<td>86.35</td>
</tr>
<tr>
<td>%δDel</td>
<td>6.15</td>
<td>5.73</td>
<td>4.06</td>
<td>2.92</td>
<td>2.40</td>
<td>2.08</td>
<td>1.67</td>
<td>1.25</td>
</tr>
<tr>
<td>%δSub</td>
<td>61.35</td>
<td>43.12</td>
<td>28.85</td>
<td>22.50</td>
<td>18.75</td>
<td>14.58</td>
<td>11.46</td>
<td>8.85</td>
</tr>
<tr>
<td>%δIns</td>
<td>26.25</td>
<td>19.69</td>
<td>14.17</td>
<td>10.00</td>
<td>7.50</td>
<td>5.94</td>
<td>4.69</td>
<td>3.54</td>
</tr>
</tbody>
</table>

6. References


5. Conclusion

The integration of the proposed NSF in the MMSE-STSA algorithm [5, 6] as the front-end of an HMM-based recognizer leads to an improvement in ASR performance in low-SNR additive car noise environments. It is clear from Table 1 that the inclusion of the NSF-based MMSE-STSA algorithm in the front-end of our ASR system in noisy car environments reduces the word error rate for a wide range of SNR values down to -12 dB. This system is therefore a promising candidate for further automatic speech recognition studies and for practical applications, e.g., noise-robust speech recognition in mobile environments.