Maximum a posteriori estimation of noise from non-acoustic reference signals in very low signal-to-noise ratio environments

Ben Milner

School of Computing Sciences, University of East Anglia, UK

Abstract

This paper examines whether non-acoustic noise reference signals can provide accurate estimates of noise at very low signal-to-noise ratios (SNRs) where conventional estimation methods are less effective. The environment chosen for the investigation is Formula 1 motor racing where SNRs are as low as -15dB and the non-acoustic reference signals are engine speed, road speed and throttle measurements. Noise is found to relate closely to these reference signals and a maximum a posteriori method (MAP) is proposed to estimate airflow and tyre noise from these parameters. Objective tests show MAP estimation to be more accurate than a range of conventional noise estimation methods. Subjective listening tests then compare speech enhancement using the proposed MAP estimation to conventional methods with the former found to give significantly higher speech quality. 

Index Terms: speech enhancement, noise estimation, MAP

1. Introduction

The aim of this work is to use non-acoustic noise reference signals for noise estimation in very low signal-to-noise ratio (SNR) environments, and subsequently use this for speech enhancement. Requirements for speech enhancement in very low SNRs arise in many environments and for this work the task of improving driver to pit-crew communication in Formula 1 motor racing is considered. This is very noisy and highly nonstationary with SNRs commonly found as low as -15dB, with the primary noise sources being the car’s engine and tyres and from airflow passing over the car. A number of sensors are located on the car and measure road speed, engine speed and throttle opening. It is the aim of this work to utilise the non-acoustic data from these sensors to estimate the car noise and subsequently apply it to speech enhancement.

Many techniques have been proposed for noise estimation. Simple methods use voice activity detectors to identify non-speech regions from where noise estimates can be updated. These are susceptible to noise and at the very low SNRs encountered in this work are unsuitable. Minimum statistics methods of noise estimation do not rely on non-speech identification and instead track minimum amplitude levels in spectral bins which are assumed to be representative of the noise [1]. SNR-dependent noise estimation utilises SNRs in frequency bins to determine whether noise estimates should be updated [2]. An alternative criteria is to compute the probability of speech being absent in a spectral bin and update the noise estimate in proportion to the probability of speech being absent [3, 4]. These methods have been shown to operate effectively at positive SNRs, but have not been analysed in SNRs as low as the motor racing environment. As part of this work, analysis will be made of these methods at very low SNRs.

An investigation of the racing car in Section 2 establishes three primary noise sources to be present – engine noise, tyre noise and airflow noise. Previous work has studied the use of non-acoustic reference data for the cancellation of engine noise using an adaptive filter system [5, 6]. This work focuses on removing the remaining airflow and tyre noise, and applies the adaptive engine noise cancellation in a preprocessing stage. This gives a partially enhanced speech signal which forms the input to the airflow and tyre noise removal system. Section 3 proposes a maximum a posteriori method of airflow and tyre noise estimation using the non-acoustic reference data. Objective tests are presented in Section 4 that compare the proposed noise estimation with conventional methods of noise estimation. Subjective tests then compare the quality of speech enhanced using the MAP noise estimates to conventional methods.

2. Analysis of racing car environment

This section examines the noise sources in the racing car and investigates their relationship to the data parameters. Analysis into road car noise revealed the main sources to be the engine, tyres and airflow [7]. This is true also for the racing car, but the more powerful engine and very fast road speeds make these noise sources considerably louder. The racing driver’s head, and hence microphone, is also external to the car which further increases noise levels. The means that SNRs can be as low as -15dB in the racing car, compared to 20dB to 5dB in a road car.

2.1. Analysis of data parameters

Synchronised with the audio signal are sensor measurements taken from the car at a rate of 100Hz. These can be considered as a sequence of data vectors, \( p_i \), at each time instant \( i \),

\[ p_i = [r_i, v_i, o_i] \]  

where \( r_i \) is the engine speed, \( v_i \) the road speed and \( o_i \) the throttle opening. Figure 1 shows the spectrogram of a 13 second audio segment with the car accelerating from 100kmph to 300kmph and making six gear changes. Two dominant noises can be identified – narrow bands of energy from engine noise harmonics and lowpass noise from the tyres and airflow. Below the spectrogram are shown data parameters. These indicate a clear relation between engine speed and the engine noise harmonics, and that the low frequency airflow and tyre noise increases in proportion to the car’s speed.

2.2. Tyre and airflow noise characteristics

Studies of road car noise in a wind tunnel found the frequency characteristics of tyre noise to be lowpass [7]. As road speed increased the energy of the tyre noise increased with the spectral envelope shape largely unaffected. Similar observations
Figure 1: Spectrogram of a 13 second audio segment from the racing car, accelerating from 100kmph to 300kmph, with corresponding engine speed, road speed and throttle parameters.

were also made for airflow noise. By considering combined airflow and tyre noise from the racing car, similar observations can be made. Figure 2 shows the power spectrum of racing car noise taken at 155kmph with a fundamental engine frequency of 295Hz and fully open throttle. This represents a doubling of the car’s road speed whilst keeping engine speed the same and highlights spectral differences corresponding to airflow and tyre noise. This is confirmed by the similar harmonic structure of the two spectra, but a significant increase in low frequency energy at faster road speeds as a result of the greater airflow and tyre noise. Similar observations from other pairs of power spectra confirmed the increases in low frequency energy arising from greater airflow and tyre noise. This is confirmed by making joint feature vectors, $z_i$, that comprise a noise power spectrum, $d_i$, and a vector of data parameters, $q_i$.

$$z_i = [d_i, q_i]$$

(2)

Figure 2: Power spectrum of racing car noise in fast, 310kmph, (dashed line) and slow, 155kmph, (solid line) conditions.

The noise power spectrum is computed from 20ms frames of airflow and tyre noise, taken from the output of the engine noise removal system. Each frame is Hann windowed and a power spectrum computed to give

$$d_i = [\|D_i(0)\|^2, \ldots, \|D_i(k)\|^2, \ldots, \|D_i(K-1)\|^2]$$

(3)

$\|D_i(k)\|^2$ is the amplitude of the $k$th power spectral bin of the $i$th frame. Vector, $q_i$, is not yet defined strictly and can take different coefficients from data vector, $p_i$, including the addition of temporal derivatives. This is investigated in Section 4.1.

Expectation-maximisation (EM) is used to create a GMM, $\Psi$, from a set of training data vectors and models the joint density of noise power spectrum and data vector as

$$\Psi(z_i) = \sum_{c=1}^{C} \kappa_c \psi_c(z_i) = \sum_{c=1}^{C} \kappa_c N(z_i; \mu_c, \Sigma_c)$$

(4)

The GMM comprises $C$ Gaussian probability density functions that localise the joint density of the noise power spectrum and data vector, where $\mu_c$ and $\Sigma_c$ represent the mean and covariance of the joint vector within the $c$th Gaussian component

$$\mu_c = [\mu_c^d; \mu_c^q]$$

and

$$\Sigma_c = \begin{bmatrix} \Sigma_c^{dd} & \Sigma_c^{dq} \\ \Sigma_c^{qd} & \Sigma_c^{qq} \end{bmatrix}$$

(5)

The mean vector, $\mu_c$, comprises mean noise power spectrum, $\mu_c^d$, and mean data vector, $\mu_c^q$. The covariance matrix, $\Sigma_c$, comprises four components – covariance matrices of the noise power spectrum, $\Sigma_c^{dd}$, and data vector, $\Sigma_c^{qq}$, and cross-covariance matrices of noise power spectrum and data vector, $\Sigma_c^{dq}$ and $\Sigma_c^{qd}$. The prior probability, $\kappa_c$, reflects the proportion of training data vectors allocated to the $c$th component.

3.2. MAP estimation of noise

A MAP estimate of the noise power spectrum, $\hat{d}_c$, can be made from the data vector, $q_i$, and joint density, $\Psi$. For the $c$th component in the joint density, $\psi_c$, the noise estimate, $\hat{d}_c$, is

$$\hat{d}_c = \arg \max_{d_c} \{ p(d_c|q_i, \psi_c) \}$$

(6)

These estimates can be combined by the posterior probability, $h_c(q_i)$, of the $i$th data vector belonging to the $c$th component.
to give a weighted estimate of the power spectrum, \( \hat{d}_i \),
\[
\hat{d}_i = \sum_{c=1}^{C} h_c(q_i) \left\{ \mu_d^c + \Sigma_d^c \left( \Sigma_q^c \right)^{-1} (q_i - \mu_q^c) \right\} \quad (7)
\]
The posterior probability, \( h_c(q_i) \), of the \( i \)th data vector, \( q_i \), belonging to the \( c \)th component is computed
\[
h_c(q_i) = \frac{\kappa_c}{C} \frac{p(q_i|\psi_q^c)}{\sum_{c=1}^{C} \kappa_c p(q_i|\psi_q^c)} \quad (8)
\]
where \( \psi_q^c \) is the marginalised distribution of the data component for the \( c \)th component in the GMM.

The MAP noise estimate allows the \( a \) posteriori and \( a \) priori SNRs to be computed. The \( a \) posteriori SNR, \( \gamma_i(k) \), is given
\[
\gamma_i(k) = \frac{|X_i(k)|^2}{|D_i(k)|^2} \quad (9)
\]
\(|X_i(k)|^2\) is the power spectrum amplitude of the \( k \)th frequency bin of the \( i \)th frame of noisy speech, taken from the output of the engine noise removal system that is applied as a preprocessing stage to remove engine noise from the audio received from the car [5, 6]. \(|D_i(k)|^2\) is the MAP airflow and tyre noise power spectral estimate from equations (3) and (7). The \( a \) priori SNR, \( \xi_i(k) \), is defined as the ratio of the clean speech power spectrum to the noise power spectrum. As the clean speech power spectrum is unavailable a decision directed approach is used [8]. In this case the \( a \) priori SNR is computed recursively
\[
\xi_i(k) = \xi_i(k-1) \frac{|\tilde{S}_{i-1}(k)|^2}{|D_{i-1}(k)|^2} + (1 - \zeta) \max[\gamma_i(k) - 1, 0] \quad (10)
\]
\(|\tilde{S}_{i-1}(k)|^2\) is the clean speech power spectrum estimate of the previous frame, taken from the airflow and tyre noise removal stage described in Section 3.3. Weight, \( \zeta \), was set to 0.98.

3.3. Noise removal

Many methods of noise removal have been proposed for speech enhancement and can be loosely categorised into spectral subtraction, Wiener filtering, subspace and statistical methods. Several studies, [9], have compared these methods and have been in general agreement that the statistical method of log MMSE, [10], gives best performance, both objectively and subjectively. The aim of this paper has been to develop a novel method of noise estimation, rather than developing a new algorithm for noise removal. As such, airflow and tyre noise removal will use the log MMSE method which uses the \( a \) priori and \( a \) posteriori SNRs to give an estimate of the clean speech magnitude spectrum, \( |\tilde{S}_i(k)| \), from the noisy magnitude spectrum \( |X_i(k)| \). The noisy magnitude spectra are extracted from 20ms Hann windows, with a 10ms overlap, taken from the output of the engine noise removal system. The resulting clean speech magnitude spectra are combined with the noisy phase and transformed back to the time-domain and combined using the overlap-and-add method.

4. Experimental results

This section examines the effectiveness of the MAP airflow and tyre noise estimation. Objective tests are presented first that compare the proposed method to several conventional methods of noise estimation. Secondly, subjective listening tests are presented that compare the quality of speech enhanced using the proposed method of noise estimation against conventional methods. The experiments use data that was obtained during a racing car practice session. Audio was sampled from a single microphone in the driver’s helmet at a rate of 8kHz. The driver spoke only for short periods of time, saying brief sentences, which were not suitable for the tests. Instead, speech from the WSJCAM0 database was added to the noise to enable adjustment of SNRs and to allow for more suitable speech to be used in the tests.

4.1. Noise estimation accuracy

Noise estimation accuracy of the proposed MAP method is now compared to a range of conventional noise estimation methods under two different driving conditions. The first condition represents fast road speeds, high acceleration and multiple gear changes and has SNRs around -15dB. The second condition has higher SNRs of about +5dB and is characterised by lower roads speeds and less acceleration. Noise segments from the two conditions were extracted from the audio data and mixed with speech to give SNRs of either -15dB or +5dB. Each segment comprised 1 second of noise, followed by noisy speech (typically 3 seconds in duration), followed by 1 second of noise. These were pre-processed by the engine noise removal system before being applied to airflow and tyre noise estimation. Thirty such segments were created and these provided about 15,000 test frames. For each frame of audio the square error between the true and estimated airflow and tyre noise power spectrum, \(|D_i(k)|^2\) and \(|\tilde{D}_i(k)|^2\), was computed. This was averaged across all frequency bins, \( K \), and frames, \( N \), to give a mean square error measure, \( E_{MSE} \)
\[
E_{MSE} = \frac{1}{NK} \sum_{i=0}^{N-1} \sum_{k=0}^{K-1} (|D_i(k)|^2 - |\tilde{D}_i(k)|^2)^2 \quad (11)
\]
Mean square noise estimation errors are shown in Table 1 for MAP estimation with four different data vector configurations and also for four conventional noise estimation methods which serve as a comparison. These are minimum statistics [1], SNR-dependent updating of the noise estimate [2] and two minima controlled recursive averaging (MCRA) methods that utilise the probability of speech being absent – MCRA2 [4] and improved MRCA (IMCRA) [3].

<table>
<thead>
<tr>
<th>Method</th>
<th>SNR = -15dB</th>
<th>SNR = +5dB</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP((\nu))</td>
<td>0.210</td>
<td>0.105</td>
</tr>
<tr>
<td>MAP((\nu, r))</td>
<td>0.208</td>
<td>0.101</td>
</tr>
<tr>
<td>MAP((\nu, r, o))</td>
<td>0.207</td>
<td>0.100</td>
</tr>
<tr>
<td>MAP((\nu, r, o + \Delta))</td>
<td>0.194</td>
<td>0.094</td>
</tr>
<tr>
<td>Minimum statistics</td>
<td>0.384</td>
<td>0.249</td>
</tr>
<tr>
<td>SNR-dependent</td>
<td>0.311</td>
<td>0.233</td>
</tr>
<tr>
<td>MCRA2</td>
<td>0.371</td>
<td>0.247</td>
</tr>
<tr>
<td>IMCRA</td>
<td>0.315</td>
<td>0.159</td>
</tr>
</tbody>
</table>

Table 1: Mean square estimation error, \( E_{MSE} \), for MAP and conventional noise estimation methods at low and high SNRs
Table 2: CMOS test results at an SNR of -15dB.

<table>
<thead>
<tr>
<th>Test</th>
<th>Method 1</th>
<th>Method 2</th>
<th>CMOS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>NNC</td>
<td>ENG</td>
<td>0.64</td>
</tr>
<tr>
<td>2</td>
<td>NNC</td>
<td>ENG→AT(MAP)</td>
<td>1.43</td>
</tr>
<tr>
<td>3</td>
<td>NNC</td>
<td>CONV</td>
<td>0.66</td>
</tr>
<tr>
<td>4</td>
<td>ENG→AT(MAP)</td>
<td>CONV</td>
<td>-0.86</td>
</tr>
<tr>
<td>5</td>
<td>ENG→AT(MAP)</td>
<td>ENG→AT(IMCRA)</td>
<td>-0.65</td>
</tr>
</tbody>
</table>

0.210 in comparison to 0.315 which was achieved by MRCA2, the best performing of the four conventional estimation methods. Including engine speed and throttle opening in the MAP estimation, MAP(v, r, o) and MAP(v, r, o), gives a slight reduction in error which is attributed to residual engine noise being present in the audio used to train the joint density used in MAP estimation. Including temporal derivatives, MAP(v, r, o + \Delta t), allows the dynamics of the data parameters to be modelled and gives a reduction in error to 0.194. Results at +5dB show a similar trend across the different methods with lowest errors again given by MAP(v, r, o + \Delta t).

4.2. Speech quality

This section uses listening tests to compare the quality of enhanced speech using either the proposed MAP method of noise estimation or conventional methods. Most subjective tests employ mean opinion scores to measure speech quality but the very low SNRs encountered in the racing car make analysis difficult as most scores will be at the lowest level. Instead, a comparative MOS (CMOS) is used that allows listeners to compare the quality of one signal to another on a scale of -3 to +3. The magnitude of the scale ranges from zero which indicates no difference between signals, to +/-3 which indicates that one signal is much higher quality than the other. The tests were carried out in accordance with ITU recommendations [11] and listeners were played pairs of samples that correspond to two versions of the same audio but processed by different enhancement methods.

Five comparisons were made as shown in Table 2. NNC refers to no noise compensation and is the original audio from the car. ENG is the result of engine noise removal [5, 6], which serves as a preprocessor for the proposed airflow and tyre noise estimation methods. AT(MAP) and AT(IMCRA) represent log MMSE speech enhancement using either MAP or IMCRA noise estimation. Finally CONV is the result of applying IMCRA noise estimation to the original noisy signal from the car to estimate both engine noise and airflow and tyre noise and removing it with log MMSE enhancement. This provides a benchmark result of what may be expected from a conventional noise estimation/speech enhancement technique applied to the noisy audio from the racing car. Arrows indicate the output of one method is fed into the input of another. Each comparison of two methods uses eight different audio segments which gives a total of 40 audio samples in each listening test. These are played in a random order. Twenty listeners took part in the tests which were carried out in a sound-proof room using head-phones. The tests considered the low SNR condition only, with the audio adjusted to be around -15dB.

Tests 1, 2 and 3 compare different enhancement methods to the original noisy audio. Engine noise removal gives an improvement in speech quality of 0.64 over the original audio. Applying MAP airflow and tyre noise removal to the output of the engine noise removal system increases the comparative quality significantly to 1.43. Applying the conventional method of speech enhancement to the original noisy audio gives a quality increase of 0.66 which is significantly lower than achieved by the proposed method. In a direct comparison, test 4 shows audio from the MAP method to be of significantly higher quality than the conventional method by a score of 0.86. Test 5 compares airflow and tyre removal using either the proposed MAP method or the IMCRA method, with both methods taking as input the audio from engine noise removal. MAP estimation gives higher quality speech over IMCRA by a score of 0.65. The improved subjective quality of the MAP airflow and tyre noise estimation over IMCRA is supported by the objective comparison made in Section 4.1. For each of the comparative tests the statistical significance was measured and all were found to be significant at the 95% confidence level.

5. Conclusion

This work has shown that accurate estimates of noise can be obtained from non-acoustic noise reference signals within a MAP framework. Objective tests have shown that the proposed method obtains more accurate noise estimates than a range of conventional methods. Subjective tests have then compared the speech quality resulting from enhancement using MAP noise estimation to conventional methods. This has shown that enhancement using the MAP noise estimation gives significantly higher speech quality than conventional methods. A further advantage of using nonacoustic noise reference signals is that noise estimation continues when speech is present which is not the case for conventional noise estimation methods.

6. References