Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 55 (2013) 879–892

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Enhancing speech at very low signal-to-noise ratios using non-acoustic reference signals

Ben Milner *

School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK

Received 15 August 2012; received in revised form 19 April 2013; accepted 22 April 2013
Available online 17 May 2013

Abstract

An investigation is made into whether non-acoustic noise reference signals can be used for noise estimation, and subsequently speech enhancement, in very low signal-to-noise ratio (SNR) environments where conventional noise estimation methods may be less effective. The environment selected is Formula 1 motor racing where SNRs fall frequently to −15 dB. Analysis reveals three primary noise sources (engine, airflow and tyre) which are found to relate to data parameters measured by the car's onboard computer, namely engine speed, road speed and throttle opening. This leads to the proposal of a two stage noise reduction system that uses first engine speed to cancel engine noise within an adaptive filtering framework. Secondly, a maximum a posteriori (MAP) framework is developed to estimate airflow and tyre noise from data parameters which is subsequently removed. Objective measurements comparing noise estimation with conventional methods show the proposed method to be substantially more accurate. Subjective quality tests using comparative mean opinion score listening tests found that the proposed method achieves +1.43 compared to +0.66 for a conventional method. In subjective intelligibility tests, 81.8% of words were recognised correctly using the proposed method in comparison to 76.7% with no noise compensation and 66.0% for the conventional method.

Keywords: Noise estimation; Speech enhancement; Speech quality; Speech intelligibility

1. Introduction

The aim of this work is to examine whether non-acoustic noise reference signals can be used in speech enhancement at very low signal-to-noise ratios (SNRs), where conventional acoustic noise reference signals may be less effective. Such a need for speech enhancement at very low SNRs arises in many environments and in this work is applied to improving driver to pit-crew communication in Formula 1 motor racing. Although the domain of noise reduction within this motor racing environment is quite specific, the analysis and noise estimation techniques proposed could be applied to other high noise environments.

The motor racing environment is extremely noisy with SNRs of −15 dB being common. In many instances the speech signal is barely detectable, with SNRs too low for the driver's speech to be intelligible. The main sources of noise that contaminate the driver's speech come from the car's engine and tyres and from air flowing past the car. These sources are similar to those found when analysing noise from conventional road cars, although the more powerful engine and faster road speeds of the racing car makes these noises much louder than in road cars (Milner, 2011; Puder et al., 2003). The SNRs experienced in road cars varies according to driving conditions but typically does not fall below +5 dB.

In a typical single seat open cockpit racing car, speech from the driver is collected from a single microphone that is positioned inside the helmet immediately in front of the driver's mouth. The driver sits directly in the airflow

* Tel.: +44 1603 593339; fax: +44 1603 593345.
E-mail address: b.milner@uea.ac.uk

0167-6393/$ - see front matter © 2013 Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.specom.2013.04.004

passing over the car and close to the open tyres and engine intake. This, combined with a lack of acoustic shielding, contributes to the very high noise levels encountered. An on-board computer measures the outputs of various sensors on the car which includes the engine speed, road speed and throttle opening. These data stream parameters, together with the audio, are transmitted to the pit-crew.

Many different methods of speech enhancement have been proposed and these operate typically as a two-stage process of first noise estimation and secondly noise reduction (Vaseghi et al., 2000; Hu et al., 2006; Loizou, 2007). Many techniques for noise estimation have been proposed and these include voice activity detectors (VADs), minimum statistics tracking and recursive averaging approaches that update noise estimates according to criteria based on local SNRs or the probability of speech being present (Tucker, 1992; Martin, 2001; Hirsch and Ehrlicher, 1995; Cohen et al., 2003). A review of a range of noise estimation methods (Taghia et al., 2011) showed many to operate effectively, although the SNRs tested did not go as low as those encountered in the motor racing environment. A similarly wide range of noise reduction methods has also been proposed and these can be categorised broadly into spectral subtraction, Wiener filtering, statistical and subspace methods (Loizou, 2007). Techniques from all of these categories have been combined with various noise estimation methods and applied to speech enhancement in car noise although not at SNRs as low as the racing car environment.

The aim of this work is to examine whether more accurate noise estimates, and hence improved speech enhancement, can be made from non-acoustic reference signals taken from the data parameters obtained from the car's on-board computer. These parameters provide instantaneous measurements of the engine speed, road speed and throttle opening which exhibit a relation to the noises being generated by the racing car. Therefore, it is proposed to derive the noise estimates for speech enhancement from the data stream parameters directly without using the input audio signal.

The work begins in Section 2 with an analysis into the noise characteristics of the racing car environment which reveals three dominant noises – engine noise, airflow noise and tyre noise. For the purposes of enhancement the noise removal problem is separated into engine noise removal and airflow and tyre noise removal. As will be discussed in Section 2, engine noise is related closely to engine speed while airflow and tyre noise is related to road speed. Section 3 develops an adaptive engine noise removal method that uses only the non-acoustic engine speed parameter as a reference signal. Section 4 proposes a novel method for airflow and tyre noise reduction using a maximum a posteriori (MAP) noise estimation method that is based solely on the data parameters. Experimental results are presented in Section 5 which compare the accuracy of MAP noise estimation with conventional noise estimation methods. Subjective listening tests then compare the proposed speech enhancement method with conventional methods by measuring both the quality and intelligibility of the enhanced speech. Finally, spectrogram analysis is presented that shows the result of applying enhancement to speech recorded from the racing car driver.

2. Analysis of noise

This section analyses the characteristics of the noises present in the racing car environment and examines their relationship to the data stream parameters produced by the car's on-board computer. This is motivated by the need to develop models of the contaminating noises and data stream parameters that can be used for noise estimation. A description of the data used is given in Section 5.

The characteristics of noise produced from a racing car are considerably different from those of a normal road car. As discussed in Puder et al. (2003) the main noise sources in a road car are from the engine, tyres and airflow. While this is also true for the racing car, the high performance engine and very fast road speeds make these noise sources considerably more powerful than in a road car. There is also more variability in the noise sources found in the racing car. Engine speeds are more likely to reach their maximum and road speeds cover a wide range in comparison to road cars. The racing driver's helmet (and hence microphone) is positioned externally to the car which further increases noise levels. The effect of this is that SNRs can be as low as -15 dB in the racing car in comparison to 20 dB to 5 dB in a road car. However, in a racing car the noise sources are generally restricted to engine, airflow and tyres and do not include other noise sources such as windscreen wipers, indicators, fan and radio that the road car has. Therefore, a simplified model of the noisy speech, $x(n)$, that considers the dominant noises received from the racing car, is

$$x(n) = s(n) + d_E(n) + d_A(n) + d_T(n) \quad (1)$$

where $s(n)$ is the original speech sample at time n, $d_E(n)$ is the engine noise, $d_A(n)$ is the airflow noise and $d_T(n)$ is the tyre noise. These three noises can be considered independent as they come from different sources, namely from the engine, airflow passing over the car, and tyre contact with the road surface. However, as further analysis will show, the airflow and tyre noise can be considered jointly as they are both primarily related to road speed. The audio system used was designed specifically for motor racing applications and avoided non-linear effects such as clipping, making Eq. (1) valid across the conditions tested.

2.1. Trace data stream

For race monitoring purposes a data stream of vehicle parameters is measured and transmitted to the pit-crew along with the audio signal received from the driver's microphone. Three parameters relate to the contaminating noise, namely engine speed (the number of engine rotations

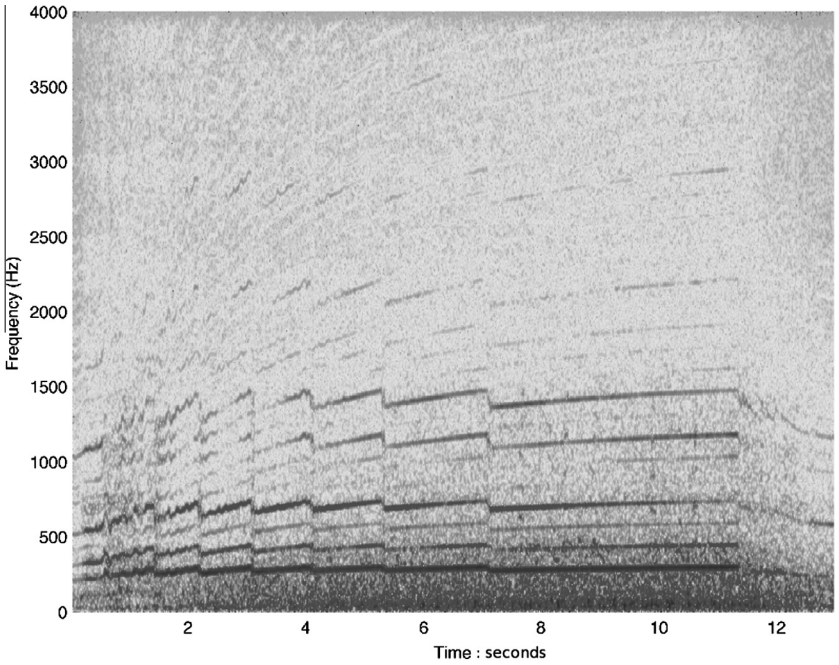

Fig. 1. Spectrogram of a 13 s audio segment from the racing car whilst accelerating from 100 kmph to 300 kmph and making six gear changes. Speech can be observed between 7.5 s and 10.5 s.

per second measured in revolutions per second (rps)), road speed (measured in kmph) and throttle opening (expressed as a percentage of being fully open). These parameters are sampled at a rate of 100 Hz and can be considered as a sequence of data stream vectors, $\boldsymbol{p}_i$, at each time instant i

$$\boldsymbol{p_i} = [r_i, v_i, o_i] \tag{2}$$

where r_i is the engine speed, v_i is the road speed and o_i is the throttle opening.

The relationship between the acoustic noise produced by the car and the data stream parameters is illustrated in Fig. 1. This shows the spectrogram of a 13 second audio segment produced when the car is accelerating from 100 kmph to 300 kmph and making six gear changes (at

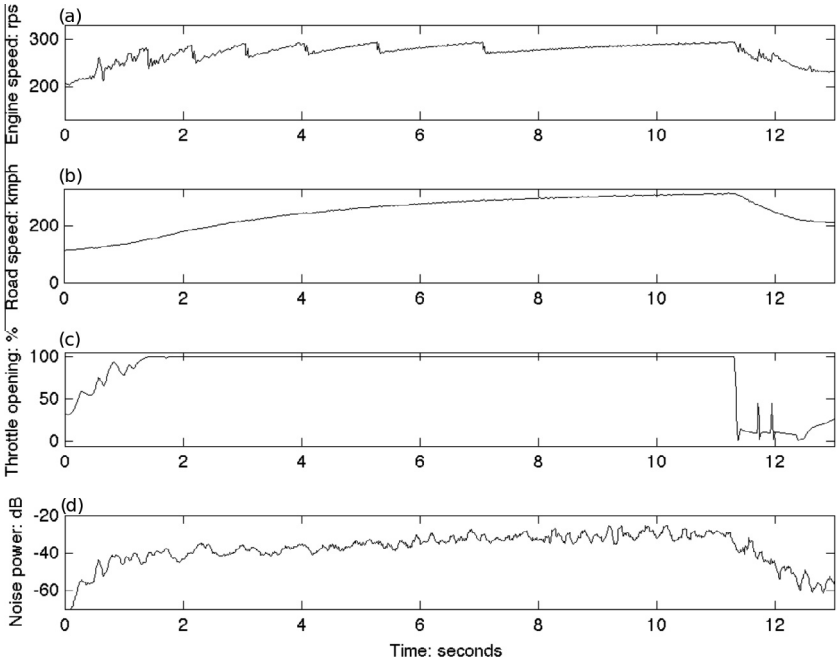

Fig. 2. Plots (a), (b) and (c) show data stream parameters of engine speed, road speed and throttle opening, corresponding to the 13 s audio signal shown in Fig. 1. Plot (d) shows the resulting noise power.

times 1.4 s, 2.2 s, 3.0 s. 4.1 s, 5.3 s and 7.1 s). Fig. 2(a)–(c) show the corresponding engine speed, road speed and throttle opening. The spectrogram shows two dominant noises to be present – narrow bands of high energy that correspond to harmonics and subharmonics of the engine noise and low frequency noise from airflow and tyres. Comparing the spectrogram to the engine speed shows a clear relation between the frequency of the engine noise harmonics and the engine speed. This is particularly noticeable as the engine speed increases between gear changes and falls abruptly during gear changes. Noise produced by the airflow passing over the car and from the tyres is low frequency in nature and increases in energy as the speed of the car increases. The effect of the throttle opening parameter on noise is not so obvious as it tends to be open fully during acceleration (0.8–10.3 s) and almost closed during deceleration (10.3–13.0 s). This characteristic was observed throughout the data. Fig. 2(d) shows the corresponding noise power of the audio which can be seen to increase in proportion to road speed with smaller fluctuations relating to the rise and fall of engine speed within each gear change.

Based on these observations of a relationship between data stream parameters and noise, the next two subsections formulate models of the engine noise and the tyre and airflow noise from these parameters.

2.2. Engine noise

Figs. 1 and 2 have shown a clear relation to exist between engine speed and the harmonic structure of the engine noise. This relationship is examined further in Fig. 3 which shows the power spectrum of a short-time frame of audio that was extracted with an engine speed of 286 rps. Harmonics can be observed integer multiples of 286 Hz (as indicated by circles) and also at half harmonic frequencies (as indicated by crosses). This is consistent with observations made of road car engine noise (Puder et al., 2000) with the precise relationship between engine speed and the harmonic structure of engine noise determined by several factors including the number of cylinders in the engine, their arrangement and the stroke pattern (Hillier, 2004).

Examining Figs. 1 and 2(c) show that the energies of the harmonics are also related to the throttle opening. The throttle governs the quantity of air drawn into the engine and consequently the volume of exhaust released after combustion, both of which contribute to engine noise. Fig. 2(c) shows that at time 11.2 s the throttle changes from fully open to almost closed and at this point Fig. 1 shows the power of the harmonic noise to reduce abruptly, even though engine speed remains high.

These observations suggest that engine noise, $d_E(n)$, can be modelled as a function, g_E, of engine speed and throttle opening,

$$d_E(n) \approx g_E(r_i, o_i) \tag{3}$$

In a power spectral representation of this function, $|D_E(f)|^2$, the engine noise can be modelled as a series of impulses, spaced at half integer harmonics of the engine speed, r_i

$$|D_E(f)|^2 \approx \sum_{m=1}^{M} h_m(o_i)\delta\left(f - \frac{mr_i}{2}\right) \tag{4}$$

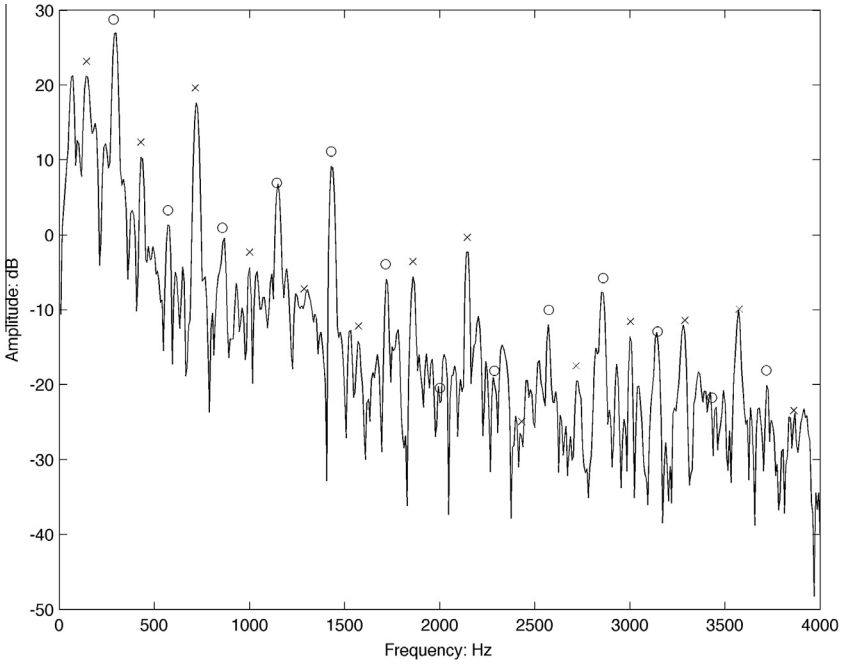

Fig. 3. Power spectrum of a short-time frame of audio extracted at an engine speed of 286 rps showing harmonic and half-harmonic structure to be present. Circles are positioned at harmonic frequencies and crosses at half-harmonic frequencies.

where $h_m(o_i)$ is the amplitude of the mth engine noise harmonic or half harmonic and is a function of the throttle opening o_i. M is the number of harmonics and half-harmonics present in the spectrum.

2.3. Tyre and airflow noise

The analysis in Section 2.1 has shown that airflow noise and tyre noise exhibit lowpass characteristics that increase in energy as road speed increases. From the available data it is not possible to examine tyre noise and airflow noise as individual components as the audio was collected during real driving conditions. Studies of a road car in a wind tunnel (Puder et al., 2000) found the frequency characteristics of tyre noise to be lowpass. As road speed increased the energy of the tyre noise increased although the spectral envelope shape was largely unaffected. Similar observations were made for airflow noise, although while still lowpass in character it had wider bandwidth than the tyre noise. By considering combined tyre and airflow noise from the racing car, similar observations can be identified in comparison to the road car. To illustrate this Fig. 4 shows the short-time power spectrum of racing car noise taken at a speed of 155 kmph with an engine speed of 295 rps and fully open throttle. Shown in the same figure is the power spectrum with the car travelling at 310 kmph but with the same engine speed of 295 rps and the throttle fully open. This represents a doubling of the car's road speed whilst keeping engine speed the same. A frequency scale from 0 Hz to 2 kHz is used in the figure to highlight the region where most differences occur.

Selecting two audio frames with equal engine speeds and throttle openings, but significantly different road speeds, highlights spectral differences that correspond to airflow and tyre noises. This is confirmed by the similar location of the harmonics in the two spectra and their similar amplitudes. However, the low frequency energy (up to about 1 kHz) in the power spectrum from the faster road speed is significantly greater than that of the slower road speed. Similar observations from other pairs of power spectra confirmed the increases in low frequency energy arising from the greater airflow and tyre noise which is predominantly low frequency in nature.

These observations suggest that the combined tyre and airflow noise, $d_{TA}(n)$, can be modelled as a function, g_{TA}, of the road speed, v_i

$$d_{TA}(n) \approx g_{TA}(v_i) \tag{5}$$

2.4. Proposed noise reduction system

The breakdown of car noise into a harmonically structured engine component and a lowpass airflow and tyre component together with their relation to data stream parameters leads to the proposal of a two-stage noise reduction system, which is shown in Fig. 5. The first stage uses the engine speed parameter to estimate and subsequently reduce engine noise to generate a partially enhanced speech signal, $s_E(n)$. The second stage then estimates and removes airflow and tyre noise to give the final enhanced speech signal, $s_A(n)$. The next two sections describe the two noise reduction systems.

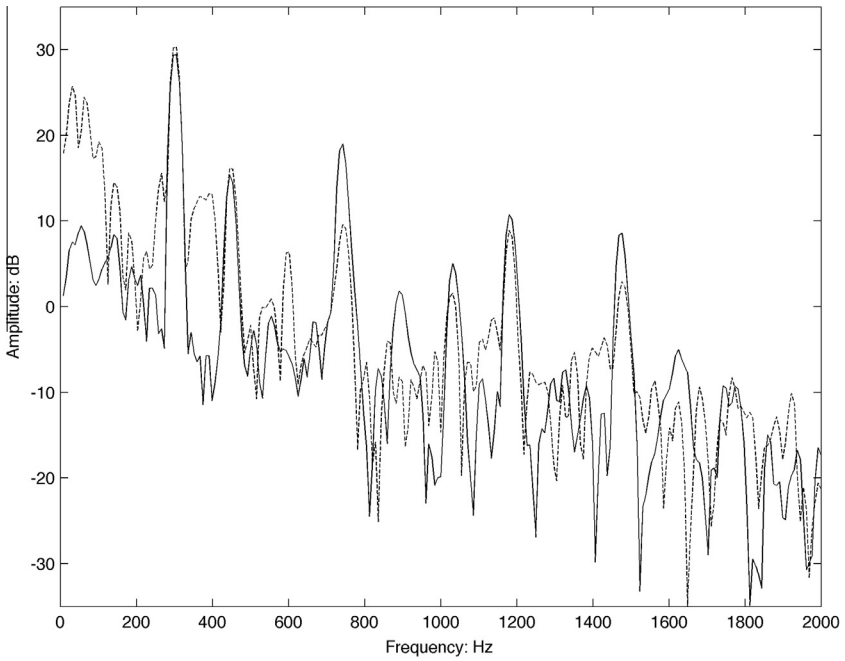

Fig. 4. Power spectrum of racing car noise under two different driving conditions. The solid line shows a road speed of 155 kmph and engine speed of 295 rps while the dashed line shows a road speed of 310 kmph and engine speed of 295 rps.

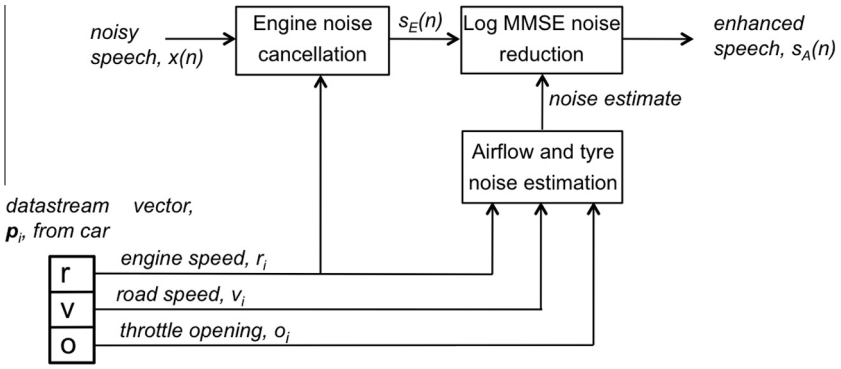

Fig. 5. Block diagram of the proposed two-stage noise reduction system that is based on first engine noise cancellation and second on airflow and tyre noise reduction.

3. Engine noise reduction

This section describes the first stage of noise removal which is to remove engine noise from the audio signal. The analysis made in Section 2.2 has shown that engine noise comprises a series of harmonics that are related to the engine speed. When the frequency of the noise is stationary, optimal methods such as Wiener filtering provide good solutions by creating sharp notch filters to remove the sinusoidal components. However, when the harmonic noise is non-stationary it is necessary to track these changes and adapt the frequency of the notches accordingly. Such adaptive filters typically use an acoustic reference signal to provide harmonic information that enables the filter to position correctly the notches to remove the noise. In this work no such acoustic reference signal is available as only a single microphone is used. Instead it is proposed to replace the acoustic reference signal with a non-acoustic reference signal computed from the engine speed from the data stream. This is synchronised with the audio and as shown in Section 2.2 relates closely to the fundamental frequency of the engine noise.

Within the framework of the adaptive filter the non-acoustic engine speed reference is used to synthesise an estimate of the engine noise, $\widehat{d_E}(n)$, as a set of M sinusoids with their frequencies, $f_m(n)$, amplitudes, $a_m(n)$, and phases, $\theta_m(n)$ matching the harmonics of the engine noise. The frequency of each sinusoid is computed from the engine speed parameter while syncronisation of the amplitudes and phases to match the engine noise is achieved by the adaptive filter. This estimate of the engine noise is subtracted from the noisy input signal, $x(n)$, to produce an estimate of the engine noise-free output signal, $s_E(n)$,

$$\begin{aligned} s_E(n) &= x(n) - \widehat{d_E}(n) \\ &= x(n) - \sum_{m=1}^{M} a_m(n) sin(2\pi f_m(n) n + \theta_m(n)) \end{aligned} \tag{6}$$

Implementation of the adaptive filter to adjust the amplitude and phase of each sinusoid is achieved more easily using the trigonometric identity $Csin(x+\theta) = Asin(x) + Bcos(x)$ to reformulate the filter in terms of a pair of sine and cosine reference signals for each sinusoid to give

$$s_E(n) = x(n) - \sum_{m=1}^{M} \left(u_m^c(n) w_m^c(n) + u_m^s(n) w_m^s(n) \right) \tag{7}$$

where $u_m^c(n)$ and $u_m^s(n)$ are the pair of cosine and sine reference signals with associated filter weights $w_m^c(n)$ and $w_m^s(n)$. The task of the adaptive filter is to estimate the filter weights. Fig. 6 illustrates the proposed engine noise removal system and the next two subsections describe the generation of the reference signals and then calculation of the filter weights.

3.1. Generation of non-acoustic reference signal

The analysis made in Section 2.2 identified engine noise harmonics at integer and half-integer multiples of the engine speed. The adaptive filter must therefore generate notches at harmonic and half-harmonics frequencies. This requires a pair of reference signals, $u_m^c(n)$ and $u_m^s(n)$, and associated filter weights, $w_m^c(n)$ and $w_m^s(n)$, for each notch m. Together, these synthesise a sinusoid corresponding to either a harmonic or half-harmonic which tracks the amplitude and phase of the engine noise component at that frequency. Considering a reference signal pair, $u_m^c(n)$ and $u_m^s(n)$, these take the form of a pair of sinusoids at the instantaneous frequency of the $\frac{m}{2}th$ harmonic with a $\frac{\pi}{2}$ phase shift between them,

$$u_m^c(n) = \cos(\phi_m(n)) \quad \text{and} \quad u_m^s(n) = \sin(\phi_m(n)) \tag{8}$$

The phase, $\phi_m(n)$, of the mth sinusoid pair is derived from the non-acoustic reference signal taken from the engine speed parameter, r_i. First the engine speed measurement is upsampled from 100 Hz to the sampling frequency of the audio, f_s, to give $r^{up}(n)$. The normalised fundamental angular frequency of the engine noise, $\omega_0(n)$, is then computed,

$$\omega_0(n) = \frac{2\pi r^{up}(n)}{f_s} \tag{9}$$

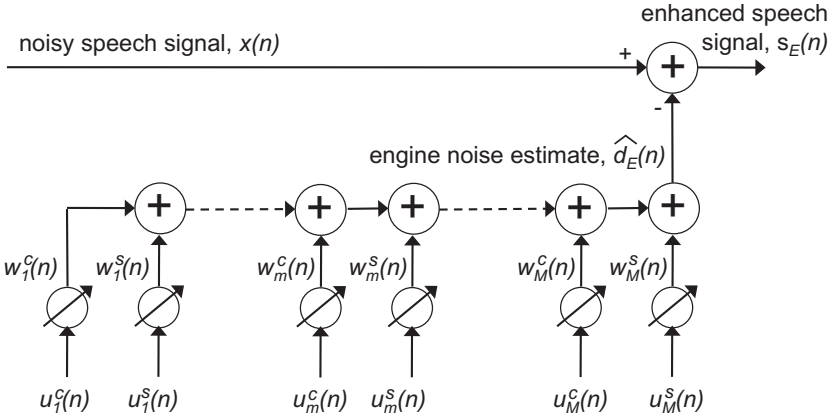

Fig. 6. Adaptive engine noise removal system showing synthesis of engine noise from pairs of sinusoidal inputs and associated filter weights.

The phase, $\phi_m(n)$, for the mth reference signal pair, $u^c_m(n)$ and $u^s_m(n)$, is computed by measuring the angle that has been moved around the unit circle up to time instant n and scaling by half of the reference signal index, i.e. $\frac{m}{2}$,

$$\phi_m(n) = \sum_{i=0}^{n} \frac{m}{2}\omega_0(i) \tag{10}$$

This can be represented recursively as,

$$\phi_m(n) = \phi_m(n-1) + \frac{m}{2}\omega_0(n) \tag{11}$$

Thus, from the engine speed parameter a series of sine and cosine pairs are created that, with appropriate filter weightings, track the amplitude and phase of the engine noise harmonics and half-harmonics. The next subsection discusses computation of these filter weights.

3.2. Updating of filter weights

The filter weights, $w^c_m(n)$ and $w^s_m(n)$, are updated using the least mean squares (LMS) algorithm (Widrow et al., 1985), which provides simple update equations for the weight pairs associated with each harmonic and half-harmonic,

$$w^c_m(n+1) = w^c_m(n) + \mu^c_m s_E(n) u^c_m(n) \tag{12}$$

$$w^s_m(n+1) = w^s_m(n) + \mu^s_m s_E(n) u^s_m(n) \tag{13}$$

where μ^c_m and μ^s_m are the step sizes for weights $w^c_m(n)$ and $w^s_m(n)$ of the adaptive filter. The choice of step size is crucial in producing an effective adaptive filter. A step size too large will give rapid convergence of the filter but a high mean square error. Conversely, a step size too small will give a low mean square error but the convergence will be slow. For the application of modelling engine noise both fast convergence (during rapid engine noise changes in gear changes) and low mean square error (during stable periods between gear changes) are required. Such performance can be obtained using a variable step size that is adjusted according to parameters associated with the filter such as the instantaneous error energy or error gradient (Kwong and Johnston, 1992; Aboulnsar and Mayyas, 1997). In this work the step size is updated recursively according to the output error of the adaptive filter (Kwong and Johnston, 1992), to give variable step sizes, $\mu^c_m(n)$ and $\mu^s_m(n)$, that replace static μ^c_m and μ^s_m in (12) and (13),

$$\mu^c_m(n+1) = \alpha\mu^c_m(n) + \gamma s_E(n)^2 \tag{14}$$

$$\mu^s_m(n+1) = \alpha\mu^c_m(n) + \gamma s_E(n)^2 \tag{15}$$

where $0 \leqslant \alpha \leqslant 1$ and $\gamma \geqslant 0$, with suitable values found to be $\alpha = 0.9$ and $\gamma = 0.01$. In situations where speech is present on the audio, and particularly when SNRs are high, the increased speech power present on signal $s_E(n)$ can lead to an overestimation of the step sizes. To limit the dynamic range of step size values, maximum and minimum limits are imposed where $\mu_{min} = 0.00001$ and $\mu_{max} = 0.01$.

The maximum number of harmonics and half-harmonics that can be cancelled by the filter is equal to the number of filter weight pairs, M. This depends on the maximum engine speed, r_{max}, and sampling frequency, f_s, and is computed,

$$M = 2\frac{\left(\frac{f_s}{2}\right)}{r_{max}} = \frac{f_s}{r_{max}} \tag{16}$$

At an engine speed of r_{max} the highest engine noise harmonic will be at the Nyquist frequency and will be cancelled by the Mth filter weight pair in the adaptive filter. At lower engine speeds, higher order harmonics will be present in the sampled audio but these will not have an associated notch filter for their cancellation. However, it has been observed that at these lower engine speeds, the power of higher frequency harmonics is considerably less than the lower frequency harmonics, making their cancellation perceptually unimportant. In this work r_{max} was 300 rps leading to an adaptive filter with $M = 52$ notches.

4. Airflow and tyre noise reduction

This section describes the proposed airflow and tyre noise reduction stage which is applied after engine noise

removal. As with most noise reduction systems the proposed method operates in two stages, first estimating the airflow and tyre noise and secondly removing it from the noisy speech.

Many methods of noise estimation have been proposed and these typically attempt to identify non-speech regions from where to update noise statistics. For example, a common method of noise estimation is to use a VAD to identify non-speech regions from where a noise estimate can be made (Tucker, 1992). VADs are effective at moderate SNRs but become increasingly less reliable as SNRs reduce. Furthermore, noise estimation stops during speech periods which is the time it is needed most accurately. This is particularly important when the noise is non-stationary, such as in the racing car environment. Preliminary investigations of VAD performance in the racing car environment found their output to be highly erroneous and unsuitable for noise estimation. Minimum statistics methods of noise estimation do not require speech/non-speech classification and instead track minimum power levels in spectral bins across a window comprising a small number of frames with the assumption that this is representative of the noise (Martin, 2001). In practice the amplitudes in the spectral bins fluctuate rapidly which can lead to underestimation of the noise. Estimation is improved through recursive temporal averaging of the noise estimates and estimation of a bias to compensate for underestimation. A further set of noise estimation methods also utilise recursive temporal averaging and only update noise estimates when appropriate to do so. One such criteria is to use the SNR of a frequency bin and update the noise estimate either when the SNR is below a threshold or proportionally according to the inverse of the SNR (Hirsch and Ehrlicher, 1995). An alternative criteria is to compute the probability of speech being absent in a spectral bin of a frame of audio. The noise estimate is then updated in proportion to the probability of speech being absent (Cohen et al., 2003; Rangachari and Loizou, 2006).

A recent study into the accuracy of noise estimation methods showed many to be effective across a range of noise conditions and considered SNRs down to -5 dB (Taghia et al., 2011). In the racing car environment SNRs fall below this which raises questions about the effectiveness of conventional noise estimation methods. To further investigate noise estimation an alternative method is proposed whereby a non-acoustic reference signal forms the basis of noise estimation within a MAP framework. The analysis of airflow and tyre noise in Section 2.3 suggested that tyre and airflow noise can be modelled as a function g_{TA} of the road speed, v_i. However, in the proposed two-stage noise removal system residual engine noise may still be present after engine noise removal, meaning that the remaining noise comprises airflow and tyre noise and possibly residual engine noise. Therefore, noise estimation in this section will, in addition to the road speed parameter, also consider engine speed and throttle parameters.

4.1. Noise estimation

The proposed noise estimation method utilises a model of the joint density of the noise power spectrum and data parameters, created from a set of training data. In an ideal situation, for airflow and tyre noise estimation, the available training data would contain only airflow and tyre noise. However, such data is not available from the audio recordings and would require rolling road and wind tunnel facilities and be unlikely to match the true noise conditions encountered during driving. Instead, a practical approach is taken whereby the joint density is trained on noise data from the output of the engine noise removal system in Section 3. In addition to airflow and tyre noise this data will also contain residual engine noise. This is also likely to be present during testing and so is useful for matching training and testing conditions.

4.1.1. Training of the noise model

The joint density of the noise power spectrum and data stream parameters is modelled using a Gaussian mixture model (GMM). Training begins by creating joint feature vectors, $\boldsymbol{z}_i$, that comprise a noise power spectrum, $\boldsymbol{d}_i$, and a vector of data stream parameters, $\boldsymbol{q}_i$

$$\boldsymbol{z}_i = [\boldsymbol{d}_i, \boldsymbol{q}_i] \tag{17}$$

The noise power spectrum is computed from 20ms frames of audio taken from the output of the engine noise removal system, $s_E(n)$, with frames having a 10 ms overlap. Each frame is Hann windowed, a Fourier transform taken and a power spectrum computed to give,

$$\boldsymbol{d}_i = [|D_i(0)|^2, \ldots, |D_i(k)|^2, \ldots, |D_i(K-1)|^2] \tag{18}$$

where $|D_i(k)|^2$ is the amplitude of the kth power spectral bin of the ith audio frame, extracted from noise-only periods following engine noise removal. $K = 128$ power spectral bins were used in this work.

At this stage the data vector, $\boldsymbol{q}_i$, is not defined strictly and can take different forms. For example, $\boldsymbol{q}_i$ could simply equal v_i and so noise estimation would rely solely on road speed. Alternatively, $\boldsymbol{q}_i$ could be set to $\boldsymbol{p}_i$ and include all data stream parameters. Other variations include $\boldsymbol{q}_i$ incorporating temporal derivatives of data stream parameters. Details on the choice of $\boldsymbol{q}_i$ are investigated in Section 5.1.

From a training data set of joint feature vectors a GMM, Ψ, is trained using expectation-maximisation (EM) to model the joint density of noise power spectrum and data stream vector (Therrien et al., 1992). Initial cluster positions for EM were obtained using the Linde–Buzo–Gray (LBG) algorithm (Linde et al., 1980). EM clustering was terminated when either no change occurred in successive iterations or when the number of iterations exceeded 120. The GMM is defined as

$$\Psi(\boldsymbol{z}_i) = \sum_{c=1}^{C} \kappa_c \psi_c(\boldsymbol{z}_i) = \sum_{c=1}^{C} \kappa_c N(\boldsymbol{z}_i, \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c) \tag{19}$$

Within the GMM, a set of C Gaussian probability density functions (PDFs) localises the joint density of the noise power spectrum and data stream vector, where $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ represent the mean and full covariance of the joint vector within the cth Gaussian distribution

$$\boldsymbol{\mu}_c = \begin{bmatrix} \boldsymbol{\mu}_c^d \\ \boldsymbol{\mu}_c^q \end{bmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_c = \begin{bmatrix} \boldsymbol{\Sigma}_c^{dd} & \boldsymbol{\Sigma}_c^{dq} \\ \boldsymbol{\Sigma}_c^{qd} & \boldsymbol{\Sigma}_c^{qq} \end{bmatrix} \tag{20}$$

The mean vector, $\boldsymbol{\mu}_c$, comprises a K-dimensional mean noise power spectrum vector, $\boldsymbol{\mu}_c^d$, and an L-dimensional mean data vector, $\boldsymbol{\mu}_c^q$. The covariance matrix, $\boldsymbol{\Sigma}_c$, comprises four components: a KxK covariance matrix of the noise power spectrum, $\boldsymbol{\Sigma}_c^{dd}$, an LxL covariance matrix of the data vector, $\boldsymbol{\Sigma}_c^{qq}$, and KxL and LxK cross-covariance matrices of the noise power spectrum and data vector, $\boldsymbol{\Sigma}_c^{qd}$ and $\boldsymbol{\Sigma}_c^{dq}$. A prior probability, κ_c, reflects the proportion of training data vectors allocated to the cth cluster.

4.1.2. MAP estimation of noise

A MAP estimate of the noise power spectrum, $\hat{\boldsymbol{d}}_i$, can be made from the data vector, $\boldsymbol{q}_i$, and the joint density, Ψ. For the cth cluster in the joint density, ψ_c, a noise estimate, $\hat{\boldsymbol{d}}_i^c$, can be made

$$\hat{\boldsymbol{d}}_i^c = \arg\max_{\boldsymbol{d}_i} \{p(\boldsymbol{d}_i|\boldsymbol{q}_i, \psi_c)\} \tag{21}$$

Noise estimates from each cluster of the joint density can be combined by weighting according to the posterior probability, $h_c(\boldsymbol{q}_i)$, of the ith data vector belonging to the cth cluster. With Gaussian probability density functions the MAP estimate is equivalent to the minimum mean square error (MMSE) estimate which gives a weighted estimate of the power spectrum, $\hat{\boldsymbol{d}}_i$ (Afify et al., 2009; Vaseghi et al., 2006)

$$\hat{\boldsymbol{d}}_i = \sum_{c=1}^{C} h_c(\boldsymbol{q}_i) \left\{ \boldsymbol{\mu}_c^d + \boldsymbol{\Sigma}_c^{dq} \left(\boldsymbol{\Sigma}_c^{qq}\right)^{-1} \left(\boldsymbol{q}_i - \boldsymbol{\mu}_c^q\right) \right\} \tag{22}$$

The posterior probability, $h_c(\boldsymbol{q}_i)$, of the ith data stream vector belonging to the cth cluster is computed as

$$h_c(\boldsymbol{q}_i) = \frac{\kappa_c p(\boldsymbol{q}_i|\psi_c^q)}{\sum_{c=1}^{C} \kappa_c p(\boldsymbol{q}_i|\psi_c^q)} \tag{23}$$

where ψ_c^q is the marginalised distribution of the data stream component for the cth cluster in the GMM.

The MAP noise estimate enables the *a posteriori* and *a priori* SNRs to be computed. The *a posteriori* SNR, $\gamma_i(k)$, is computed

$$\gamma_i(k) = \frac{|S_i^E(k)|^2}{|\widehat{D_i}(k)|^2} \tag{24}$$

where $|S_i^E(k)|^2$ is the power spectrum amplitude of the kth frequency bin of the ith frame of noisy speech from the output of the engine noise removal system and $|\widehat{D_i}(k)|^2$ is the noise power spectral estimate from (22). 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 to compute the *a priori* SNR is used (Ephraim and Malah, 1984). In this case the *a priori* SNR is computed recursively

$$\xi_i(k) = \zeta \frac{|S_{i-1}^A(k)|^2}{|\widehat{D_{i-1}}(k)|^2} + (1-\zeta)\max[\gamma_i(k) - 1, 0] \tag{25}$$

$|S_{i-1}^A(k)|^2$ is the clean speech power spectrum estimate of the previous frame, estimated from the noise removal stage (described in Section 4.2). ζ is a weighting factor which for this work is set at $\zeta = 0.98$.

4.2. Noise removal

Many methods of noise removal have been proposed for speech enhancement that use either an estimate of the contaminating noise power spectrum or the *a priori* and *a posteriori* SNRs. In this section the aim has been to propose a novel method of airflow and tyre noise estimation using non-acoustic reference signals rather than develop a new algorithm for noise removal. As such the airflow and tyre noise removal system uses an existing noise removal technique that takes as its input the SNR estimates produced from the noise estimation method.

Noise removal methods can be loosely categorised into spectral subtraction, Wiener filtering, subspace and statistical methods. Several studies (Loizou, 2007) have compared these methods and have been in general agreement that statistical methods give best performance (both subjectively and objectively) and in particular the log MMSE enhancement method (Ephraim and Malah, 1985). Based on these findings the airflow and tyre noise removal in this work uses log MMSE with the required *a priori* and *a posteriori* estimates of the SNR determined from (24) and (25). Using log MMSE the estimate of the clean speech spectral magnitudes, $|S_i^A(k)|$, is computed as

$$|S_i^A(k)| = \left[\frac{\xi_i(k)}{1+\xi_i(k)} \exp\left(\frac{1}{2} \int_{v_i(k)}^{\infty} \frac{e^{-t}}{t} dt \right) \right] |S_i^E(k)| \tag{26}$$

where $|S_i^E(k)|$ is the spectral magnitude of the noisy speech and $v_i(k)$ is defined

$$v_i(k) = \frac{\xi_i(k)\gamma_i(k)}{1+\xi_i(k)} \tag{27}$$

For implementation, frames of audio are extracted using 20 ms duration Hann windows with 10 ms overlap. This is identical to the training stage of the MAP noise estimator. Each frame of audio is transformed into its magnitude and phase spectral representations using a discrete Fourier transform (DFT). Clean speech magnitude spectra, $|S_i^A(k)|$, are estimated using (26) and combined with the noisy phase before being transformed back to the time-domain using an inverse DFT. Each time-domain frame is then overlapped with its neighbouring frames by 50% and the sample ampli-

tudes added to generate the final enhanced speech signal, $s_A(n)$.

5. Experimental results

This section analyses the effectiveness of the engine noise and tyre and airflow noise reduction systems. Objective measures are first used to examine the effectiveness of airflow and tyre noise estimation. Secondly, the results of subjective listening tests, measuring both speech quality and speech intelligibility, are presented that compare the proposed methods of engine noise removal and airflow and tyre noise removal with conventional speech enhancement techniques that use acoustic reference signals.

The data used in these experiments was collected during a car testing session at the Valencia racing circuit in Spain. The audio was collected from a single microphone located in the driver's helmet and was PCM encoded using 16 bits per sample and sampled at a rate of 8 kHz. Examination of the audio revealed it to be free from coding distortions such as clipping or transmission errors. During the recordings the driver spoke only occasionally and then for only for short periods of time, typically repeating brief sentences about the car's performance or driving conditions. The limited quantity and restricted content of these sentences made them unsuitable for objective and subjective testing and instead speech was added artificially to the noise. This enabled manipulation of SNRs for testing and allowed for more suitable speech to be used in the listening tests. However, artificially adding speech removes any Lombard effects that may have been present in the original speech as a result of the high noise environment. Examination of the driver's speech did reveal the Lombard effect to be present at times. In informal tests, the effect of speech enhancement on this noisy Lombard speech was found to be no different to enhancement of non-Lombard noisy speech. This indicates that experimental results obtained when artificially adding speech give a good indication to performance when the Lombard effect is present. Section 5.3 examines further the effect of enhancement using spectrogram analysis, specifically when applied to real speech recordings made by the driver.

5.1. Airflow and tyre noise estimation

This section examines the accuracy of MAP airflow and tyre noise estimation and compares performance to a range of conventional estimation methods. The joint density of noise spectrum and data vector was trained on 70 seconds of audio data that represented a range of driving conditions in terms of engine speed, road speed and throttle opening. The audio data was taken from the output of the engine noise removal system. A set of preliminary experiments determined that best performance was obtained using $C = 8$ modes in the GMM.

Section 2 has shown that driving conditions are highly variable and non-stationary. Rather than measuring noise estimation accuracy globally across all conditions, two different situations have been considered. The first represents periods of fast road speeds, high acceleration and multiple gear changes and results in a non-stationary audio signal. Typical SNRs in these regions are around −15 dB. The second condition is characterised by lower road speeds and deceleration and has SNRs in the region of +5 dB. Analysis of driving conditions revealed that the car is almost solely in one or other of these two conditions. Noise segments representing the two driving conditions were extracted from the output of the engine noise removal system and mixed with speech (from the WSJCAM0 database (Fransen et al., 1994)) to give SNRs of either −15 dB or +5 dB, according to the driving condition. After adding speech, each audio segment was adjusted to comprise 1 second of noise followed by the noisy speech (about 3 seconds in duration) followed by 1 second of noise, giving a total duration of about 5 seconds.

For evaluation a set of 30 audio segments were created and together these comprised over 15,000 test frames. For each frame the mean square error between the true noise power spectrum, $|D_i(k)|^2$ and the estimated noise power spectrum, $|\widehat{D_i}(k)|^2$ was measured. This was averaged across all frequency bins, K, and frames, N, to give a mean square estimation error measurement, E_{MSE}

$$E_{MSE} = \frac{1}{NK}\sum_{i=0}^{N-1}\sum_{k=0}^{K-1}\left(|D_i(k)|^2 - |\widehat{D_i}(k)|^2\right)^2 \tag{28}$$

Table 1 shows the MAP mean square estimation error using four different data vectors. MAP(v) uses only the road speed in the joint density, MAP(v, r) includes engine speed and MAP(v, r, o) includes throttle opening. MAP($v, r, o + \Delta$) uses all three parameters and augments them with their velocity temporal derivatives. Temporal derivatives were computed using a window of 5 frames as this was found to give best performance. For comparison Table 1 also shows the performance of four conventional noise estimation methods on the same task. These were discussed in Section 4 and include minimum statistics (Martin, 2001), SNR-dependent updating of the noise estimate (Hirsch and Ehrlicher, 1995) and two minima controlled recursive averaging (MCRA) methods that utilise the prob-

Table 1
Mean square estimation error, E_{MSE}, for MAP and conventional noise estimation methods at low and high SNRs.

Method	Low SNR (−15 dB)	High SNR (+5 dB)
Minimum statistics	0.584	0.249
SNR-dependent	0.511	0.233
MCRA2	0.371	0.247
IMCRA	0.315	0.159
MAP(v)	0.210	0.105
MAP(v, r)	0.208	0.101
MAP(v, r, o)	0.207	0.100
MAP($v, r, o + \Delta$)	0.194	0.094

ability of speech being absent – MRCA2 (Rangachari and Loizou, 2006) and improved MRCA (IMCRA) (Cohen et al., 2003).

The results show that MAP estimation using only the road speed parameter, MAP(v), gives substantially lower estimation errors than the conventional noise estimation methods at both low and higher SNRs. For example at an SNR of −15 dB the MAP(v) method has a mean square error of 0.210 in comparison to 0.315 which was achieved by MRCA2. This was the best performing of the four conventional noise estimation methods. Including engine speed and throttle position into the MAP estimation, MAP(v, r, o), reduced errors slightly down to 0.207 at −15 dB. This relatively small reduction in error is attributed to engine speed and throttle having low correlation to the airflow and tyre noise. The correlation that does exist is attributed to residual engine noise being present in the audio signal used to train and test the GMM used in the MAP estimation. Augmenting the data parameters by their temporal derivatives, MAP($v, r, o + \Delta$), enables the model to capture the dynamics of the data parameters and gives a further reduction in estimation error to 0.194. Results at an SNR of +5 dB show a similar trend in the accuracy of noise estimation, with lowest errors also being given by the MAP($v, r, o + \Delta$) system, which again outperformed the conventional methods of noise estimation.

These results confirm the analysis made in Section 2.3 that the airflow and tyre noise can be modelled effectively by the road speed. The slight reduction in estimation error when including the engine speed parameter may be a result of the noise model now being able to model the residual engine noise remaining after engine noise removal. Throttle position makes almost no difference in error which is attributed to the fact that the throttle is almost bimodal in that it is either fully open or almost closed.

5.2. Subjective tests

The previous section has shown objectively that the proposed use of non-acoustic reference signals provides accurate estimates of airflow and tyre noise. To examine the effect on speech enhancement it is necessary to undertake subjective measurements of the enhanced speech signal. This section presents the results of subjective quality and intelligibility tests for a variety of different speech enhancement configurations.

5.2.1. Subjective quality

A common subjective measure of speech quality is the mean opinion score where listeners rate quality using an absolute scale from 1 to 5. However, the very low SNRs encountered during racing conditions are likely to cause most listeners to rate the audio close to its lowest level of quality making analysis and comparison difficult. Instead, a comparative mean opinion score (CMOS) is employed where listeners evaluate the quality of one signal in comparison to another. The CMOS tests were conducted in accordance with the ITU guidelines (ITU-T, 1996; ITU-T, 2003) where listeners are played pairs of samples that correspond to two versions of the same segment of audio, processed by two different methods. Listeners rate the second sample in comparison to the first using a scale of −3 to +3 as shown in Table 2. The tests were carried out in a soundproof room with listeners wearing headphones.

Five different comparisons were made and these are shown in Table 3. NNC refers to no noise compensation and represents the original unprocessed noisy speech. ENG is the result of the engine noise removal of Section 3. A&T is the result of the log MMSE-based airflow and tyre noise removal using either the MAP estimation method of Section 4, to give A&T(MAP), or the IMCRA method to give A&T(IMCRA) which serves as a comparative noise removal method using conventional noise estimation. Finally ENG + A&T(IMCRA) is the result of applying IMCRA noise estimation to the original noisy signal to estimate both engine noise and airflow and tyre noise and using this within the log MMSE enhancement method. This provides a benchmark result of what may be expected from a conventional noise estimation technique applied to the noisy audio signal from the racing car. Arrows indicate that the output of one method is fed into the input of another. For example, method 2 in test 2, ENG → A&T(MAP), is the result of engine noise removal followed by airflow and tyre noise removal using MAP noise estimation.

For each comparison of two methods, eight different audio segments were played to listeners, with the ordering of the two methods balanced. This gives a total of 40 audio files in each test which were played in a random order. Twenty listeners took part in the tests and these were carried out in a sound-proof room using headphones. As with the tests in Section 5.1, sentences from WSJCAM0 were

Table 2
Comparative Mean Opinion Scores (CMOS): quality of the second sample compared to the quality of the first sample is:

Score	Description
+3	Much better
+2	Better
+1	Slightly better
0	About the same
−1	Slightly worse
−2	Worse
−3	Much worse

Table 3
Comparative mean opinion score (CMOS) test results at an SNR of −15 dB.

Test	Method 1	Method 2	CMOS
1	NNC	ENG	0.64
2	NNC	ENG → A&T(MAP)	1.43
3	NNC	ENG + A&T(IMCRA)	0.66
4	ENG → A&T(MAP)	ENG + A&T(IMCRA)	−0.86
5	ENG → A&T(MAP)	ENG → A&T(IMCRA)	−0.65

added to the noise. The tests concentrated on low SNR conditions which were adjusted to be around −15 dB.

Considering first tests 1, 2 and 3 which compare no noise compensation to three enhancement methods, engine noise removal on its own gives a comparative improvement of 0.64 over NNC. Applying airflow and tyre noise removal to this increases the comparative quality of the enhanced audio to 1.43. In comparison, applying IMCRA-based speech enhancement to the original noisy audio results in a comparative quality improvement of 0.66 in comparison to NNC, which is substantially lower than that attained by the proposed non-acoustic noise estimation method, ENG → A&T(MAP). Test 4 confirms this result by comparing the two noise compensation methods directly where the proposed method outperforms the IMCRA method by a score of 0.86. Test 5 examines the effectiveness of the airflow and tyre noise estimation and removal using either the proposed MAP method or the IMCRA method. The output from the engine noise removal system is applied to log MMSE noise reduction using either MAP noise estimation, ENG → A&T(MAP), or IMCRA noise estimation, ENG → A&T(IMCRA). Comparative tests show MAP estimation to give higher quality speech than IMCRA by a score of 0.65. This improved subjective quality of MAP airflow and tyre noise estimation over IMCRA is supported by the objective comparison made in Section 5.1. For each of the comparative tests the statistical significance was measured and all were found to be within the 95% confidence interval.

5.2.2. Subjective intelligibility

The experiments in this section examine whether the intelligibility of the enhanced speech has changed. Many studies into speech enhancement have reported gains in speech quality but studies very rarely report improvements in speech intelligibility (Loizou, 2007; Loizou and Kim, 2011). In most cases intelligibility actually falls as a result of the enhancement method. Given the very low SNRs of the racing car environment this forms an interesting domain in which to examine intelligibility.

Five different speech enhancement configurations were evaluated for intelligibility and these are shown in Table 4 and comprise: (1) no noise compensation, (2) engine noise removal only, (3) engine noise removal followed by airflow and tyre noise removal using MAP noise estimation, (4) engine noise removal followed by airflow and tyre noise removal using IMCRA noise estimation and (5) combined engine noise and airflow and tyre noise removal using MAP noise estimation. Method 3 corresponds to the proposed method and method 5 to a conventional method of speech enhancement. Each listening test comprised 50 digit strings that were played to the subject. Each string was made up from 4 digits taken from the Aurora 1 database (Pearce and Hirsch, 2000). Digits were chosen as the evaluation data in the intelligibility tests as there is no context with the digits and it was relatively straightforward for listeners to remember and record the digits that they heard in each string. For each of the five enhancement configurations, 10 sets of digit strings were played to listeners in a random order. Each digit test string comprised 1 second of noise followed by the noisy digit string and then followed by another 1 second of noise with SNRs adjusted to be around −15 dB. Twenty listeners participated in the intelligibility tests and the tests were carried out in a soundproof room with listeners wearing headphones.

Table 4
Intelligibility test results at an SNR of −15 dB.

Test	Method	Digit accuracy (%)
1	No noise compensation	76.7
2	Engine noise removal only	76.0
3	Engine noise removal *followed by* airflow and tyre noise removal using MAP noise estimation (proposed method)	81.8
4	Engine noise removal *followed by* airflow and tyre noise removal using IMCRA noise estimation	71.4
5	*Combined* engine noise and airflow and tyre noise removal using IMCRA noise estimation (conventional method)	66.0

For each of the five configurations, the percentage of digits that were recognised correctly was computed across all 20 listeners with the resulting digit accuracies shown in Table 4. With no noise compensation (method 1) a digit accuracy of 76.7% was attained and this forms the baseline intelligibility. Applying engine noise removal (method 2) had negligible effect on intelligibility. Applying the MAP-based airflow and tyre noise removal to the output of the engine noise removal system (method 3) increased digit accuracy to 81.8% which was a statistically significant improvement at the 95% confidence level over the unprocessed audio (Gillick et al., 1989). Conversely, applying IMCRA-based airflow and tyre noise removal to the output of the engine noise removal system (method 4) reduced accuracy to 71.4%. Finally, applying conventional noise estimation and removal to the original noisy audio (method 5) gave lowest intelligibility of 66.0%. Statistical analysis revealed the difference in intelligibility between this result and the proposed method to be statistically significant at the 95% confidence level.

5.3. Spectrogram analysis of enhanced speech

The aim of this section is to examine the effect of speech enhancement on real speech recorded from the driver and to use spectrograms to show the effect of both the proposed enhancement method and a conventional method on the noisy speech.

Fig. 7 shows three spectrograms of a 4.5 seconds speech utterance with the driver saying "*one-two-three, one-two-three, one-two-three, one-two-three*". Fig. 7(a) shows the original noisy speech spoken by the driver with the car accelerating from 297 kmph to 312 kmph with an engine speed that rises from 16,700 rpm to 17,600 rpm. The harmonics of the

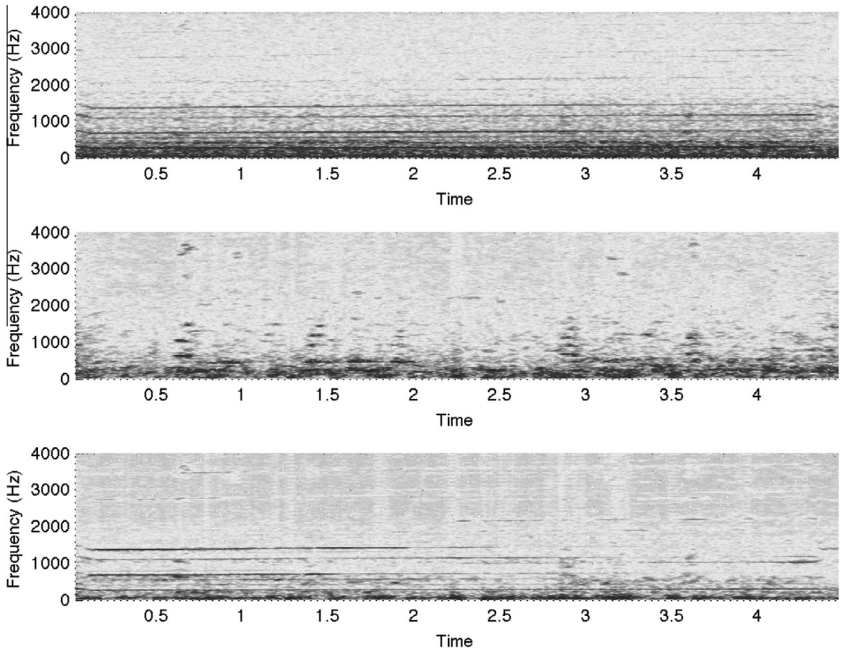

Fig. 7. Spectrograms showing: (a) original driver's speech recorded in racing car noise, (b) speech enhanced using the proposed method and (c) conventional method of speech enhancement using log MMSE with IMCRA to provide a noise estimate.

engine noise are visible clearly as is the lowpass tyre and airflow noise. The result of applying the proposed method of speech enhancement is shown in Fig. 7(b). The engine noise harmonics are largely removed as is the majority of low frequency tyre and airflow noise, the result of which shows the speech signal to be substantially more prominent. Fig. 7(c) shows the result of applying conventional IMCRA noise estimation to the noisy speech followed by log MMSE. This has removed some of the noise but has left much harmonic noise from the engine and some of the low frequency airflow and tyre noise. This has also been at the expense of removing much of the speech signal when compared to Fig. 7(b)

Informal listening tests on this utterance and other similar real recordings made from the driver confirms the effectiveness of the proposed method on real speech recordings.

6. Conclusion

This work has developed an effective method of noise estimation and speech enhancement that uses non-acoustic noise reference signals. This has been shown to operate effectively at very low SNRs and to give significant improvements in both speech quality and speech intelligibility over a conventional speech enhancement method. Objective comparisons of the proposed MAP method of airflow and tyre noise estimation with conventional noise estimation methods found the proposed method to be more accurate at both very low SNRs (−15 dB) and at higher SNRs (+5 dB). These tests were carried out under similar driving conditions for training and testing the noise model. It is likely that if testing conditions changed, for example using a different car or moving from dry to wet conditions, then noise estimation accuracy would reduce. Performance in these situations could be improved by training the noise model on a range of driving conditions across different cars, or creating multiple noise models and then selecting the most appropriate according to the probability of the noise estimate from each model. For engine noise removal, given that this uses only the engine speed parameter, it is likely that this would be effective across different cars and driving conditions.

A range of comparative listening tests on the enhanced speech found that engine noise reduction and airflow and tyre noise reduction both contributed to significant improvements in speech quality. Comparison against a conventional speech enhancement method using IMCRA noise estimation and log MMSE noise reduction found the proposed method to provide significantly higher quality speech. Listening tests measuring the intelligibility of the enhanced speech found that the proposed method gave significantly higher intelligibility over both the original noisy audio and the conventional speech enhancement method. The same tests also found the conventional enhancement method to reduce significantly the intelligibility below that of the original noisy audio, which is consistent with other research (Loizou, 2007). Although the proposed method has been shown effective at compensating against engine, airflow and tyre noise, transient noise events in motor racing can also occur, such as when other cars either pass by or are overtaken. In this situation no adjustment of the noise estimate takes place. Analysis of noise levels made from other cars found them to be substantially lower in power than the combined engine, airflow and tyre noise, making their effect perceptually small. A further advantage

of the proposed method of using non-acoustic noise reference signals is that noise estimation can continue when speech is present. This is not the case for many conventional noise estimation methods and contributes further to the accuracy of the proposed method.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.specom.2013.04.004.

References

Aboulnsar, T., Mayyas, K., 1997. A robust variable step-size LMS-type algorithm: analysis and simulations. IEEE Transactions on Signal Processing 45 (3), 631–639.

Afify, M., Cui, X., Gao, Y., 2009. Stereo-based stochastic mapping for robust speech recognition. IEEE Transactions on Audio, Speech and Language Processing 17 (7), 1325–1334.

Cohen, I., 2003. Noise spectrum estimation in adverse environments: improved minima controlled recursive averaging. IEEE Transactions on Speech Audio Processing 11 (5), 466–475.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 32 (6), 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing 33 (2), 443–445.

Fransen, J., Pye, D., Robinson, T., Woodland, P., Young, S., 1994. WSJCAM0 corpus and recording description, Cambridge University Engineering Department, Cambridge, UK, Tech. Rep. CUED/F-INFENG/TR.192.

Gillick, L., Cox, S., 1989. Some statistical issues in the comparison of speech recognition algorithms. In: ICASSP, pp. 532–535.

Hillier, V., 2004. Fundamentals of Motor Vehicle Technology. Nelson Thornes.

Hirsch, H., Ehrlicher, C., 1995. Noise estimation techniques for robust speech recognition. In: ICASSP, USA, pp. 153–156.

Puder, H., Steffens, F., 2000. Improved noise reduction for hands-free car phones utilising information on vehicle and engine speeds. In: EUSIPCO, pp. 1851–1854.

Hu, Y., Loizou, P., 2006. Subjective comparison of speech enhancement algorithms. In: ICASSP, France, pp. 153–156.

ITU-T, P.800: Methods for subjective determination of transmission quality. ITU-T recommendation, 1996.

ITU-T, P.835: Subjective test methodology for evaluating speech communication systems that include noise suppression algorithms. ITU-T recommendation, 2003.

Kwong, R., Johnston, E., 1992. A variable step size LMS algorithm. IEEE Transactions on Signal Processing 40 (7), 1633–1642.

Linde, Y., Buzo, A., Gray, R., 1980. An algorithm for vector quantisation design. IEEE Transactions on Communications 28 (1), 94–95.

Loizou, P., 2007. Speech Enhancement: Theory and Practice. CRC Press Inc..

Loizou, P., Kim, G., 2011. Reasons why current speech-enhancement algorithms do not improve speech intelligibility and suggested solutions. IEEE Transactions on Audio, Speech and Language Processing 19 (1), 47–56.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Transactions on Speech Audio Processing 9 (5), 504–512.

Milner, B., 2011. Maximum a posteriori estimation of noise from non-acoustic reference signals in very low signal-to-noise ratio environments. In: Interspeech, Italy.

Pearce, D., Hirsch, H.-G., 2000. The AURORA experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ICSLP, vol. 4, Beijing, China, pp. 29–32.

Puder, H., 2003. Speech enhancement for hands-free car phones by adaptive compensation of harmonic engine noise components. In: Eurospeech, Switzerland, pp. 1397–1400.

Rangachari, S., Loizou, P., 2006. A noise estimation algorithm for highly nonstationary environments. Speech Communication 48 (2), 22–231.

Taghia, J., Taghia, J., Mohammadiha, N., Sang, J., Bouse, V., Martin, R., 2011. An evaluation of noise power spectral density estimation algorithms in adverse acoustic environments. In: ICASSP.

Therrien, C., 1992. Discrete Random Signals and Statistical Signal Processing. Prentice-Hall, ISBN 0-13-217985-7.

Tucker, R., 1992. Voice activity detection using a periodicity measure. IEE Proceedings of Communications, Speech and Vision I 139 (4), 377–380.

Vaseghi, S., 2006. Advanced Digital Signal Processing and Noise Reduction, third ed. Wiley.

Vaseghi, S., Chen, A., McCourt, P., 2000. State-based subband LP Wiener filters for speech enhancement in car environments. In: ICASSP, Turkey, pp. 213–216.

Widrow, B., Stearns, S., 1985. Adaptive Signal Processing. Prentice-Hall.