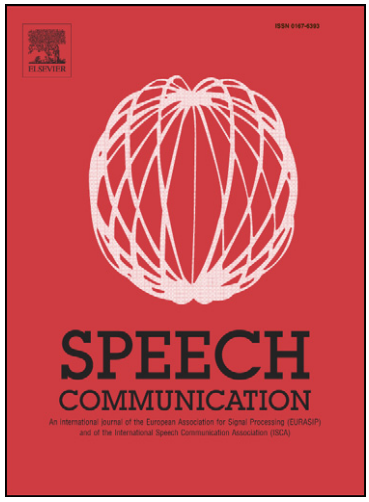
Combined Speech Enhancement and Auditory Modelling for Robust Distributed Speech Recognition

Ronan Flynn, Edward G. Jones



Combined Speech Enhancement and Auditory Modelling for Robust Distributed Speech Recognition

Ronan Flynn, Edward Jones

PII: S0167-6393(08)00077-0
DOI: 10.1016/j.specom.2008.05.004
Reference: SPECOM 1719

To appear in: *Speech Communication*

Received Date: 3 July 2007
Revised Date: 6 May 2008
Accepted Date: 13 May 2008

Please cite this article as: Flynn, R., Jones, E., Combined Speech Enhancement and Auditory Modelling for Robust Distributed Speech Recognition, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.05.004

Combined Speech Enhancement and Auditory Modelling for Robust Distributed Speech Recognition

Ronan Flynn[1] *and Edward Jones*[2]

[1]Department of Electronic Engineering, Athlone Institute of Technology, Ireland
[2]Department of Electronic Engineering, National University of Ireland, Galway, Ireland
rflynn@ait.ie, edward.jones@nuigalway.ie

Abstract

The performance of Automatic Speech Recognition (ASR) systems in the presence of noise is an area that has attracted a lot of research interest. Additive noise from interfering noise sources, and convolutional noise arising from transmission channel characteristics both contribute to a degradation of performance in ASR systems. This paper addresses the problem of robustness of speech recognition systems in the first of these conditions, namely additive noise. In particular, the paper examines the use of the auditory model of Li *et al.* (2000) as a front-end for a HMM-based speech recognition system. The choice of this particular auditory model is motivated by the results of a previous study by Flynn and Jones (2006) in which this auditory model was found to exhibit superior performance for the task of robust speech recognition using the Aurora 2 database (Hirsch and Pearce, 2000). In the speech recognition system described here, the input speech is pre-processed using an algorithm for speech enhancement. A number of different methods for the enhancement of speech, combined with the auditory front-end of Li *et al.*, are evaluated for the purpose of robust connected digit recognition. The ETSI basic (ETSI ES 201 208, 2003) and advanced (ETSI ES 202 050, 2007) front-ends proposed for DSR are used as a baseline for comparison. In addition to their effects on speech recognition performance, the speech enhancement algorithms are also assessed using perceptual speech quality tests, in order to examine if a correlation exists between perceived speech quality and recognition performance. Results indicate that the combination of speech enhancement pre-processing and the auditory-model front-end provides an improvement in recognition performance in noisy conditions over the ETSI front-ends.

Keywords: Speech enhancement; Auditory front-end; Robust speech recognition.

1. Introduction

The front-end processor in Automatic Speech Recognition (ASR) systems converts the incoming speech signal into a format that is later used in a classification stage. The front-end extracts a feature from the speech signal that ideally should be independent of the speaker (for speaker independent recognition tasks) and background noise, and distortion introduced by the transmission channel. It is well known that the presence of noise severely degrades the performance of speech recognition systems, and much research has been devoted to the development of techniques to alleviate this effect. One aspect of the robustness of an ASR system is its ability to maintain its recognition accuracy under conditions that are different from the original training conditions. One common approach to improving system performance in noise is to use front-ends that produce robust features. Some of these approaches involve modifications of well established techniques, such as cepstral mean subtraction. Other approaches involve the use of auditory-based front-ends in order to improve robustness (e.g. Ghitza, 1988; Seneff, 1988; Dau *et al.*, 1996).

Another method that has been proposed to improve the robustness of ASR systems is to enhance the speech signal before feature extraction. Enhancement of noisy speech signals is designed to improve the perception of the speech by human listeners or to improve the processing of the speech by ASR systems. It may also have benefits in enhancing robustness in ASR systems. Speech enhancement can be particularly useful in cases where a significant mismatch exists between training and testing conditions, such as where a recognition system is trained with clean speech and then used in noisy conditions. Inclusion of speech enhancement can help to reduce the mismatch.

The enhancement of noisy speech can be described as an estimation problem in which the original clean signal is estimated from a degraded version of the signal. A significant amount of research has been carried out on speech enhancement, and a number of approaches have been well documented in the literature. Ephraim and Cohen (2006) present a survey of a number of approaches to speech enhancement from a single microphone. Many enhancement techniques are based on the concept of noise spectral estimation coupled with spectral subtraction. The advantage of these methods is

a reduction in noise and an improvement in the signal-to-noise ratio. A disadvantage is the introduction of speech distortion and a residual noise called 'musical noise'.

Two measures that can be used to perceptually evaluate speech are its *quality* and its *intelligibility*. Speech quality is a subjective measure and is dependent on the individual preferences of listeners. It is a measure of how comfortable a listener is when listening to the speech under evaluation. The intelligibility of the speech can be regarded as an objective measure, and is calculated based on the number or percentage of words that can be recognised by listeners. The intelligibility and the quality of speech are not correlated and it is well known that improving one of the measures can have a detrimental effect on the other one. Speech enhancement algorithms give a trade-off between noise reduction and signal distortion. A reduction in noise can lead to an improvement in the subjective quality of the speech but a decrease in the measured speech intelligibility (Ephraim and Cohen, 2006). The quality and the intelligibility of speech can be evaluated using listening tests. There are however a number of mathematically based tools available that facilitate the evaluation of speech quality and speech intelligibility without the need for listeners. Speech enhancement can have a negative impact on subjective speech intelligibility if the spectral cues and the gross temporal envelope cues in the speech are not adequately preserved by the enhancement algorithm. For example, Hu and Loizou (2007) found that single-microphone speech enhancement algorithms do not improve subjective intelligibility in normal-hearing listeners and that with certain enhancement algorithms the intelligibility was impaired.

When using speech enhancement in an ASR system, the speech is enhanced before feature extraction and recognition processing. The advantage of this is that there is no impact on the computational complexity of the feature extraction or the recognition processes as the enhancement is independent of both. However, every speech enhancement process will introduce some form of signal distortion and it is important that the impact of this distortion on the recognition process is minimised.

Kleinschmidt *et al.* (2001) combined the model of auditory perception (PEMO) described by Tchorz and Kollmeier (1999), with the noise reduction algorithm proposed by Ephraim and Malah (1984). This noise reduction algorithm is well known, and has been found to exhibit good performance. Kleinschmidt *et al* (2001) compared the performance of this combination with the performance of a front-end based on the standard Mel Frequency Cepstral Coefficient (MFCC) framework, for the task of recognition of an isolated German digit database, and found that the combination of speech enhancement and auditory model resulted in better performance.

This paper extends this paradigm by examining the performance of the auditory model proposed by Li *et al.* (2000), in combination with a number of different speech enhancement algorithms. Many computational auditory models have been proposed for use in speech recognition systems, often with excellent results, particularly in the presence of noise. In this work, the auditory model of Li *et al.* (2000) is used. The choice of this auditory front-end is motivated by previous work carried out by Flynn and Jones (2006) where a number of auditory front-ends were investigated in a comparative study of robust speech recognition with the widely-used Aurora 2 database (Hirsch and Pearce, 2000). In that study, there was no pre-processing or enhancement of the speech utterances. The front-ends investigated were Perceptual Linear Prediction (PLP) proposed by Hermansky (1990), the PEMO algorithm proposed by Tchorz and Kollmeier (1999), and the front-end processor proposed by Li *et al.* (2000). For the task of connected digit recognition using the Aurora 2 database, the front-end proposed by Li *et al.* gave the best overall recognition results of all the auditory models examined, and with an overall reduction in recognition error compared to the ETSI basic front-end (ETSI ES 201 208, 2003) which was used as a baseline for comparison. The ETSI front-ends have been proposed for use in a Distributed Speech Recognition (DSR) paradigm, wherein the front-end would typically be implemented in a mobile handset, while the recognition engine would be implemented on a centrally-located server. The proposed system could also be implemented as part of a DSR framework, therefore, consideration needs to be given to the computational complexity associated with embedded implementation of any front-end algorithm, in particular focusing on the additional computational cost associated with the extra processing for speech enhancement.

In this paper, both the ETSI basic front-end (ETSI ES 201 208, 2003) and the ETSI advanced front-end (ETSI ES 202 050, 2007), in combination with the different speech enhancement methods, are used for comparison with Li *et al.* (2000). The recognition problem examined in the present paper is also connected digit recognition using the Aurora 2 database. The motivation behind the creation of the Aurora database was to provide a framework that allowed comparison of different ASR systems in noisy conditions, thus providing a good basis for comparison between researchers. The noisy conditions include subway, airport, restaurant, train station, street, exhibition hall, car and babble noise. The classifier used for the recognition experiments is the HMM recogniser (implemented using HTK) specified for use with the Aurora database. Of particular interest in this paper is

the condition where a mismatch exists between training and test conditions, so the emphasis here is on training using clean speech and testing using noisy speech. Performance analysis is based on speech recognition performance, and perceptual speech quality as measured using the Perceptual Evaluation of Speech Quality (PESQ) algorithm (ITU-T Rec. P.862, 2001).

The layout of the paper is as follows. Section 2 of this paper gives an overview of the auditory models used in the feature extraction for connected digit recognition. The speech enhancement techniques considered are described in Section 3. The Aurora 2 database is discussed in Section 4 and this is followed by a description of the recognizer used in Section 5. The algorithm used for the evaluation of speech quality is described in Section 6. Connected digit recognition results using the Aurora 2 database are presented in Section 7. The computational complexity of the speech enhancement algorithms used is discussed in Section 8. The results are presented in Section 9 with conclusions and suggestions for further work outlined in Section 10.

2. Front-End Processors

This section briefly describes the front-end processors that were examined in this work. The first., Mel Frequency Cepstral Coefficients, is implemented according to ETSI guidelines for DSR. Two versions, the ETSI basic and advanced front-ends (ETSI ES 201 208, 2003 and ETSI ES 202 050, 2007), are examined. The second front-end processor is based on the auditory model of Li *et al*. (2000). In this case, two versions are also examined.

2.1 Mel Frequency Cepstral Coefficients Front-End

Feature extraction based on mel-frequency cepstral coefficients (MFCCs) has been well documented (Davis and Mermelstein, 1980). While the details of implementation for MFCCs are well-known, a brief description is included here for completeness, as well as to provide a framework for discussion of the specific variations and parameter values used for this research. The speech signal first undergoes pre-emphasis in order to compensate for the unequal sensitivity of human hearing across frequency. Following pre-emphasis, a short-term power spectrum is obtained by applying an FFT to a frame of Hamming windowed speech. Critical band analysis is carried out using a bank of overlapping, triangular shaped, bandpass filters, whose centre frequencies are equally spaced on the mel scale. The FFT magnitude coefficients are grouped into the appropriate critical bands and then weighted by the triangular filters. The energies in each band are summed, creating a filter bank vector of spectral energies in the mel scale. The size of this vector of spectral energies is equal to the number of triangular filters used. A non-linearity in the form of a logarithm is applied to the energy vector. The final step is the application of a discrete cosine transform (DCT) to generate the MFCCs.

The ETSI basic front-end (ETSI ES 201 208, 2003) and the ETSI advanced front-end (ES 202 050, 2007) both use MFCCs with the following parameters. Speech, sampled at 8 kHz, is blocked into frames of 200 samples with an overlap of 60%. A logarithmic frame energy measure is calculated for each frame before any processing takes place. In the case of the basic front-end, pre-emphasis is carried out using a filter coefficient equal to 0.97 while the advanced front-end uses a value of 0.9. A Hamming window is used in both the ETSI basic and advanced front-ends prior to taking an FFT. In the ETSI advanced front-end a power spectrum estimate is used before performing the filter-bank integration. This results in higher noise robustness when compared with using a magnitude spectrum estimate (Macho *et al.*, 2002) as used in the ETSI basic front-end. The basic front-end generates a feature vector consisting of 13 coefficients made up of the frame log-energy measure and cepstral coefficients C_1 to C_{12}. The feature vector produced by the advanced front-end contains the cepstral coefficients C_1 to C_{12} along with a weighted combination of cepstral coefficient C_0 and the frame log-energy measure. In the recognition experiments, velocity and acceleration coefficients are appended to the 13 static features above, to give a total of 39 elements in each feature vector.

2.2 The Auditory Model of Li *et al.*

The auditory feature extraction algorithm proposed by Li *et al.* (2000) is based on an analysis of the human auditory system. The functions of the outer ear, middle ear, cochlea, hair cells and nerve system are modelled from an information and signal processing point of view. The steps involved in the feature extraction are shown in Figure 1.

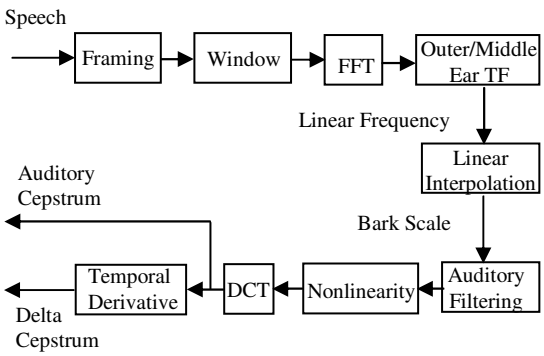


Figure 1: Feature extraction proposed by Li *et al*. (2000)

Speech is sampled at 8 kHz and blocked into frames of 240 samples. The frame overlap is 66.7% and a Hamming window is used prior to taking an FFT. An outer/middle ear transfer function (see Figure 2) that models pressure gain in the outer and middle ears is applied to the spectrum magnitude. The spectrum is then subjected to a nonlinear frequency transformation to convert it to the Bark scale.

After conversion of the spectrum to the Bark scale, the transfer function output is processed in the frequency domain by an auditory filter that is derived from psychophysical measurements of the frequency response of the cochlea. The auditory filters used are symmetric on the Bark frequency scale and an example is shown in Figure 3.

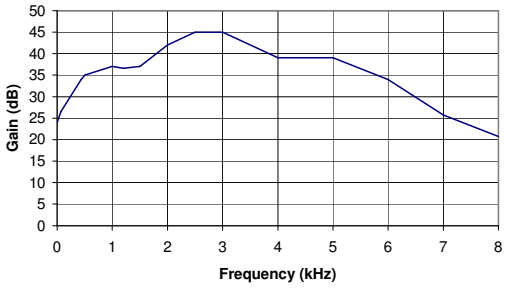


Figure 2: Outer/middle ear transfer function (Li *et al*., 2000).

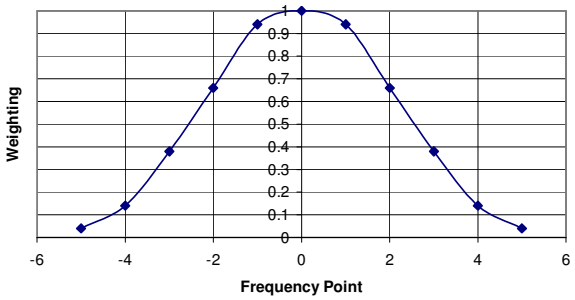


Figure 3: Frequency response of an 11-point auditory based filter (Li *et al*., 2001).

Like the MFCCs, a nonlinear function in the form of a logarithm, followed by a DCT, is applied to the filter outputs to generate the cepstral coefficients. The recognition experiments use vectors that include energy and 12 cepstral coefficients along with delta and acceleration coefficients. Again, this results in vectors with an overall dimension equal to 39. In their work, Li *et al.* (2000) found this model to be superior to LPCC and MFCC features, on a connected digit recognition task using two CDMA wireless speech databases.

For the purpose of this paper, two versions of the Li *et al.* front-end are used. The first generates a feature vector consisting of 13 coefficients made up of the frame log-energy measure and the cepstral coefficients C_1 to C_{12}. The second version generates a feature vector that contains the cepstral coefficients C_1 to C_{12} along with a weighted combination of cepstral coefficient C_0 and the frame log-energy measure. In this paper, the two versions will be referred to as Li *et al*. (I) and (II). The specifics of the two versions allow for a closer comparison with the ETSI basic front-end (ETSI ES 201 208, 2003) and the ETSI advanced front-end (ES 202 050, 2007), respectively. This is discussed further in Section 7.1 below.

2.3 Comparison of MFCC and Li *et al.* front-ends

It is interesting to compare the structure of the two front-end processors considered (MFCC and the auditory model of Li *et al*), since a number of structural similarities exist. In previous work, Milner (2002) presented a comparative analysis of the processing stages involved in using MFCCs and a different auditory model (Perceptual Linear Prediction) for speech recognition, and noted that there were several commonalities, and some differences, between the algorithms. A similar comparison is carried out here for MFCCs and the front-end of Li *et al*. (2000). Processing stages in each of the front-ends that are similar are linked by dashed lines in Figure 4.

The first processing step in the MFCC front-end is a pre-emphasis of the input speech signal in the time domain. This stage is required in order to flatten the spectral tilt of speech signals reflected in the transfer function of the outer ear. Pre-emphasis is achieved using a first-order high pass filter. The equivalent processing stage in Li et *al.* is the outer/middle ear TF block. Li *et al*. scale the frequency spectrum using an outer/middle ear transfer function (Figure 2) that is based on psychoacoustic measurements. In the spectral analysis, the Bark scale is used by Li et *al*. The Bark scale is designed to represent the tonotopic outputs of the critical band filters along the basilar membrane in the ear. In each critical band, the contributions from the short-term power spectrum of the speech signal are summed. Filters, derived from psychophysical measurements of the frequency response of the cochlea (Figure 3), are equally spaced on the Bark scale. In MFCC analysis, triangular shaped critical band filters are equally spaced on the Mel frequency scale.

The front-end of Li *et al.* (2000) is designed to more accurately model the human auditory process from the outer ear, through the middle ear, to the inner ear. The use of the Bark scale showed a slight advantage over other auditory scales in recognition experiments (Li *et al.*, 2001). Spectral data are equally spaced in the Bark

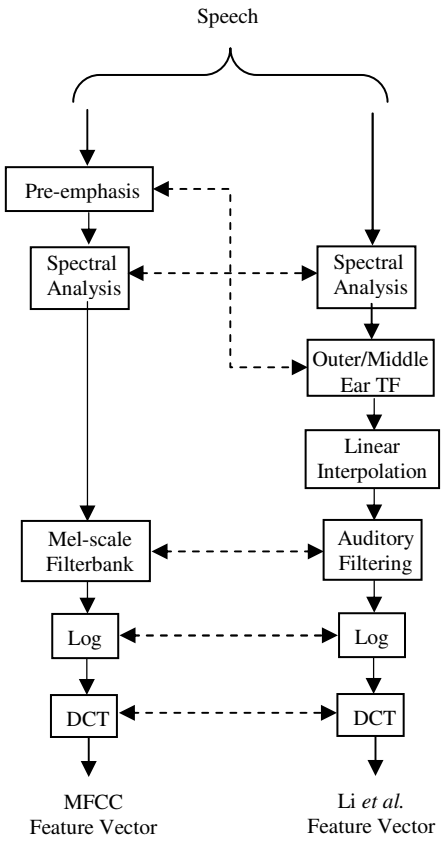


Figure 4: Comparison of front-end processors.

scale through linear interpolation before auditory filtering. The resolution of the filtered output on the spectrum is preserved compared to the lower resolution obtained in the MFCC front-end. Li *et al.* (2001) also found that speech recognition performance was partially improved by the outer/middle ear transfer function (Figure 2). However, they found that the most significant factor in the improvement of speech recognition results was the new set of auditory filters.

It is well documented that the human auditory perception process is believed to carry out spectral analysis through the use of a bank of bandpass auditory filters. Psychoacoustic research carried out by Patterson, Moore and Glasberg *et al.* (Moore, 2003) showed that the auditory filter can be approximated using a symmetric Gaussian curve, not unlike the shape of the filters used in the auditory model of Li *et al*. The triangular filters used in the MFCC front-end are a simplistic approximation to the Gaussian shape of the human auditory filters and are adopted for computational efficiency. Mak *et al.* (2004) have taken the front-end of Li *et al.* (2000) and studied the effect of the shape of the auditory filters in speech recognition. Results show that the auditory-based features of Li *et al.* following discriminative training of the auditory filters are more robust than MFCCs in mismatched testing environments.

3. Speech Enhancement

In this section, the various speech enhancement algorithms that were examined are briefly described. The algorithms range from well-established algorithms like that of Ephraim and Malah (1984), to very recently-proposed ones like that of Rangachari and Loizou (2006). Furthermore, the algorithms cover a range of paradigms, including spectral subtraction-based algorithms using the FFT for spectral analysis, as well as methods based on auditory filterbanks.

3.1 Ephraim & Malah

Ephraim and Malah (1984) present a minimum mean-square error short-time spectral amplitude (MMSE STSA) estimator. The estimator is derived based on modelling speech and noise spectral components as statistically independent Gaussian random variables. The enhanced speech is constructed using the MMSE STSA estimator combined with the original phase of the noisy signal. Analysis is carried out in the frequency domain and the signal spectrum is estimated using an FFT.

In a noisy signal $x(t)$, the MMSE amplitude estimator of the k^{th} spectral component is given by

$$\hat{A}_k = G_k R_k \tag{1}$$

where R_k is the amplitude of the k^{th} spectral component in $x(t)$ and G_k is given by

$$G_k = \frac{\sqrt{\pi}}{2} . \frac{\sqrt{v_k}}{SNR_{post\_k}} . M\left[v_k\right] . \tag{2}$$

In equation (2), v_k is calculated as

$$v_k = \left(\frac{SNR_{prio\_k}}{1 + SNR_{prio\_k}} \right) SNR_{post\_k} \tag{3}$$

where SNR_{prio_k} and SNR_{post_k} are the *a priori* and *a posteriori* signal-to-noise ratios respectively. The function $M[\ ]$ in equation (2) is evaluated as follows:

$$M\left[\theta\right] = \exp\left(\frac{-\theta}{2} \right) \left[(1+\theta) I_0\left(\frac{\theta}{2} \right) + \theta I_1\left(\frac{\theta}{2} \right) \right] . \tag{4}$$

I_0 and I_1 in equation (4) represent the modified Bessel functions of zero and first order, respectively.

The *a priori* SNR for the k^{th} spectral component in the n^{th} analysis frame is determined by

$$SNR_{prio\_k}(n) = \alpha\left(\frac{\hat{A}_k^2(n-1)}{\lambda_k(n-1)}\right) + (1-\alpha)P[SNR_{post\_k}(n)-1] \quad (5)$$

where $0 \leq \alpha < 1$, λ_k is the variance of the k^{th} spectral component of the noise and $P[\ ]$ is a half-wave rectification operator which is defined by

$$P[x] = \begin{cases} x & if\ x \geq 0 \\ 0 & otherwise \end{cases} . \quad (6)$$

The *a posteriori* signal-to-noise ratio, SNR_{post_k}, is determined using $(R_k)^2$, the amplitude-squared of the k^{th} spectral component, and the current estimate of the noise power.

3.2 Westerlund *et al.*

Westerlund *et al.* (2005) present a speech enhancement technique in which the input signal is first divided into a number of sub-bands. The signal in each sub-band is individually multiplied by a gain factor in the time domain based on an estimate of the short term SNR in each sub-band at every time instant. High SNR values indicate the presence of speech and the sub-band signal is amplified. Low SNR values indicate the presence of noise only and the sub-band signal remains unchanged.

Westerlund *et al.* (2005) consider a discrete time speech signal, $s(n)$, corrupted by a noise signal, $w(n)$, that results in a noise corrupted speech signal $x(n)$, where

$$x(n) = s(n) + w(n) . \quad (7)$$

After filtering $x(n)$ by a bank of K bandpass filters, $x(n)$ can be written as

$$x(n) = \sum_{k=0}^{K-1} x_k(n) = \sum_{k=0}^{K-1} s_k(n) + w_k(n) \quad (8)$$

where $x_k(n)$ is the sub-band noisy speech signal. Westerlund *et al.* (2005) calculate a gain function, $G_k(n)$, for each sub-band and this function weights the input signal sub-bands based on the ratio of $s_k(n)$ to $w_k(n)$. The enhanced signal is given by

$$y(n) = \sum_{k=0}^{K-1} G_k(n) x_k(n) \quad (9)$$

In each sub-band, the short term exponential magnitude average, $A_{x,k}(n)$ is based on $|x_k(n)|$; and an estimate of the noise floor level, $\underline{A}_{x,k}(n)$, are calculated according to equations (10) and (11) respectively.

$$A_{x,k}(n) = (1-\alpha_k)A_{x,k}(n-1) + \alpha_k |x_k(n)| \quad (10)$$

$$\underline{A}_{x,k}(n) = \begin{cases} (1+\beta_k) \times \underline{A}_{x,k}(n-1) \\ \qquad if\ A_{x,k}(n) > \underline{A}_{x,k}(n-1) \\ A_{x,k}(n) \quad otherwise \end{cases} \quad (11)$$

In equation (10), α_k is a positive constant that controls how sensitive the response is to rapid changes in input signal amplitude in sub-band k. The parameter β_k, in equation (11) is a constant that controls how fast the noise floor level estimate in sub-band k adapts to changes in the noise environment. The gain function in equation (9) is then calculated as

$$G_k(n) = \left(\frac{A_{x,k}(n)}{\underline{A}_{x,k}(n)}\right)^{p_k}, \ p_k \geq 0, \ \underline{A}_{x,k}(n) > 0 , \quad (12)$$

where p_k controls the gain exponent individually applied to each of the sub-band signals. To prevent excessively large values, the gain function is limited according to

$$G_k(n) = \begin{cases} G_k(n) & if\ G_k(n) \leq L_k \\ L_k & otherwise \end{cases} , \quad (13)$$

where L_k is a positive constant.

Westerlund *et al.* (2005) claim that their algorithm performs well in different noise environments with minimal parameter adjustment, and that the algorithm computational complexity is low.

3.3 Martin

Martin (1994) presented an algorithm for the enhancement of noisy speech signals by means of spectral subtraction, in particular through a method for estimation of the noise power. Martin's noise estimation method is based firstly on the independence of speech and noise and secondly on the observation that speech energy in an utterance falls to a value close to or equal to zero for brief periods. Such periods of low speech energy occur between words or syllables in an utterance and during speech pauses. The energy of the speech during these periods reflects the noise power level. Martin's minimum statistics noise estimation method tracks the short term power spectral density estimate of the noisy speech signal in each frequency bin separately. The minimum power within a defined window is used to estimate the noise floor level. The minimum tracking method requires a bias compensation since the minimum

power spectral density of the noisy signal is smaller than the average value.

In (Martin, 2001), Martin further develops the noise estimation algorithm by using a time and frequency dependent smoothing parameter when calculating the smoothed power spectral density. A method to calculate an appropriate time and frequency dependent bias compensation is also described in (Martin, 2001) as part of the algorithm. The smoothed power spectral density of the noisy signal is

$$P(\lambda,k) = \alpha(\lambda,k)P(\lambda-1,k) + (1-\alpha(\lambda,k))\,|Y(\lambda,k)|^2 \quad (14)$$

where λ and k are the time and frequency indices respectively, $Y(\lambda,k)$ is the DFT of the windowed noisy signal and $\alpha(\lambda,k)$ is a dynamic smoothing parameter:

$$\alpha(\lambda,k) = \frac{\alpha_{\max}\alpha_c(\lambda)}{1+(P(\lambda-1,k)/\hat{\sigma}_N^2(\lambda-1,k)-1)^2} \quad ,(15)$$

where α_{max} is a constant close to unity, α_c is a time dependent correction factor and the noise power spectral density is $\hat{\sigma}_N^2$.

The additional bias factor for compensating the minimum of the noisy signal power spectral density is derived by Martin (2001) as

$$B_{\min}(\lambda,k) \approx 1+(D-1)\frac{2}{\tilde{Q}_{eq}(\lambda,k)} \quad , \quad (16)$$

where D is the window length over which the minimum is found and $\tilde{Q}_{eq}(\lambda,k)$ is called "equivalent degrees of freedom". The unbiased noise estimate is

$$\hat{\sigma}_N^2(\lambda,k) = B_{\min}(\lambda,k)P_{\min}(\lambda,k) \quad . \quad (17)$$

The algorithm requires that the minimum of D power spectral density estimates $P(\lambda, k)$ be found. To improve the speed of the noise tracking, the window of D samples is divided into U sub-windows of V samples each. The maximum delay in responding to a rising noise power is $D+V$.

3.4 Rangachari and Loizou

Rangachari and Loizou (2006) recently proposed an algorithm for the estimation of noise in highly non-stationary environments. The smoothed power spectrum of the noisy speech signal is given by

$$P(\lambda,k) = \eta P(\lambda-1,k) + (1-\eta)\,|Y(\lambda,k)|^2 \quad (18)$$

where is the frame index, k the frequency index, a smoothing constant and $|Y(,k)|^2$ is the short-time power spectrum of the noisy speech. The local minimum of the noisy speech power spectrum, $P_{min}(,k)$, is found using equation (19) in which and are experimentally determined constants.

$$P_{\min}(\lambda,k) = \begin{cases} \gamma P_{\min}(\lambda-1,k) + \dfrac{1-\gamma}{1-\beta}(P(\lambda,k) \\ -\beta P(\lambda-1,k)) \\ \qquad if\ P_{\min}(\lambda-1,k) < P(\lambda,k) \\ P(\lambda,k) \quad otherwise \end{cases} \quad (19)$$

The decision as to whether or not speech is present is based on a comparison of an experimentally determined frequency dependent threshold, (k), with the ratio of the noisy speech power spectrum to its local minimum:

$$\begin{aligned} &if \quad \frac{P(\lambda,k)}{P_{\min}(\lambda,k)} > \delta(k) \\ &\quad I(\lambda,k) = 1 \quad speech\ present \\ &else \\ &\quad I(\lambda,k) = 0 \quad speech\ absent \\ &end \end{aligned} \quad (20)$$

Using equation (20), the speech-presence probability is updated as follows:

$$p(\lambda,k) = \alpha_p p(\lambda-1,k) + (1-\alpha_p)I(\lambda,k) \quad (21)$$

where $_p$ is a smoothing constant. A time-frequency dependent smoothing factor is determined using equation (22) in which $_d$ is a constant.

$$\alpha_s(\lambda,k) = \alpha_d + (1-\alpha_d)\,p(\lambda,k) \quad (22)$$

The noise power spectrum estimate, $D(,k)$, is then updated as

$$D(\lambda,k) = \alpha_s(\lambda,k)D(\lambda-1,k) + (1-\alpha_s(\lambda,k))\,|Y(\lambda,k)|^2 \quad . \quad (23)$$

The estimated clean speech spectrum is evaluated as

$$C(\lambda,k) = \max\left\{|Y(\lambda,k)|^2 - D(\lambda,k), vD(\lambda,k)\right\} \quad (24)$$

where v is a small positive constant.

Rangachari and Loizou (2006) combined the noise-estimation algorithm with a Wiener-type speech-enhancement algorithm that has the following spectral gain function with over subtraction factor μ_k:

$$G(\lambda,k) = \frac{C(\lambda,k)}{C(\lambda,k)+\mu_k D(\lambda,k)} \quad (25)$$

3.5 Agarwal and Cheng

A technique for the removal of noise from degraded speech using two filtering stages was proposed by Agarwal and Cheng (1999). The first filtering stage coarsely reduces the noise and whitens any residual noise while the second stage attempts to remove the residual noise. Filtering is based on the Wiener filter concept and filter optimisation is carried out in the mel-frequency domain. The algorithm, described as a two-stage mel-warped Wiener filter noise reduction scheme, is a major component of the ETSI advanced front-end standard for DSR (ETSI ES 202 050, 2007).

The implementation of the noise reduction in the ETSI advanced front-end (see Figure 5) is summarised by Macho *et al* (2002). Speech, sampled at 8 kHz, is divided into frames of 25 ms duration with a 60% overlap. A 256-point FFT is applied to the Hanning windowed speech to obtain the signal spectrum estimate.

The FFT spectrum length is reduced to 65 through an averaging process and the power spectral density (PSD) mean is calculated. A voice activity detector for noise estimation (VADNest) makes a speech/non-speech decision for the current frame based on the frame log energy and a long-term estimate of non-speech log energy. Frames labelled as non-speech are used for updating the noise estimation. The Wiener filter magnitude response is estimated based on the current frame spectrum and the decision output from the VADNest block. The magnitude response is smoothed and transformed to the mel-frequency scale using 23 triangular, equally spaced, mel-warped frequency windows. The impulse response of the Wiener filter is obtained by using a mel-warped inverse discrete cosine transform. Convolving the noisy input speech signal with the Wiener filter impulse response produces the enhanced speech signal.

The output of the first stage in Figure 5 is fed directly to the input of the second stage. The second filter stage is very similar to the first, the differences being no VADNest block and an additional gain factorisation block. The gain factorisation block carries out a dynamic, SNR-dependent noise reduction that is aggressively applied to purely noisy frames. Less aggressive noise reduction is applied in the second stage to frames that contain speech and noise.

The full detail of the two-stage mel-warped Wiener filter noise reduction algorithm can be found in (recognition (ETSI ES 202 050, 2007).

4. The Aurora 2 Database

The Aurora 2 database was designed to evaluate the performance of speech recognition algorithms in noisy conditions (Hirsch and Pearce, 2000). The speech database is derived from utterances of isolated digits and

connected digit sequences spoken by US-American adults in the TIDigits database. The speech in the TIDigits database is sampled at 20 kHz and is down-sampled to 8 kHz in the Aurora database. Some additional filtering is applied to the down-sampled data in order to take into account the frequency characteristics of equipment used in telecommunications systems. The channel characteristics used are G.712 and MIRS. The down-sampled, filtered speech corresponds to "clean" data in the Aurora database.

The Aurora database also contains "noisy" data. This corresponds to clean data with noise artificially added at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB and –5 dB. The noise signals added are chosen to reflect environments in which telecommunication terminals are used. In total there are eight different noise types used: subway, babble, car, exhibition hall, restaurant, street, airport and train station.

The Aurora framework also includes a set of standard test conditions for evaluation of front-end processors. For the purpose of training the speech recogniser, two modes are defined. The first mode is training on clean data and the second mode is multi-condition training on noisy data. The same 8440 utterances, taken from the training part of the TIDigits, are used for both modes. For the multi-condition training, the clean speech signals are used, as well as speech with four different noise types (subway, babble, car and exhibition hall), added at SNRs of 20 dB, 15 dB, 10 dB and 5 dB. However, for the recognition experiments described in this paper, only training in clean conditions was used, as the inclusion of speech enhancement is intended to reduce the mismatch between training and testing conditions. The paradigm of training with clean speech only was also previously used by Kleinschmidt *et al.* (2001).

There are three different test sets defined for recognition testing, with the test utterances taken from the testing part of the TIDigits database. Test Set A (28028 utterances) employs the same four noises as used for the multi-condition training. Test Set B uses the same utterances as Test Set A but uses four different noise types (restaurant, street, airport and train station). In both Test Sets A and B, the frequency characteristic used in the filtering of the speech and noise is the same as that used in the training sets, namely G.712. The frequency characteristic of the filter used in Test Set C (14014 utterances) is MIRS, and is different from that used in the training sets. Subway and street noises are used in Test Set C. In all three test sets, noise is added at SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, –5 dB and the clean condition.

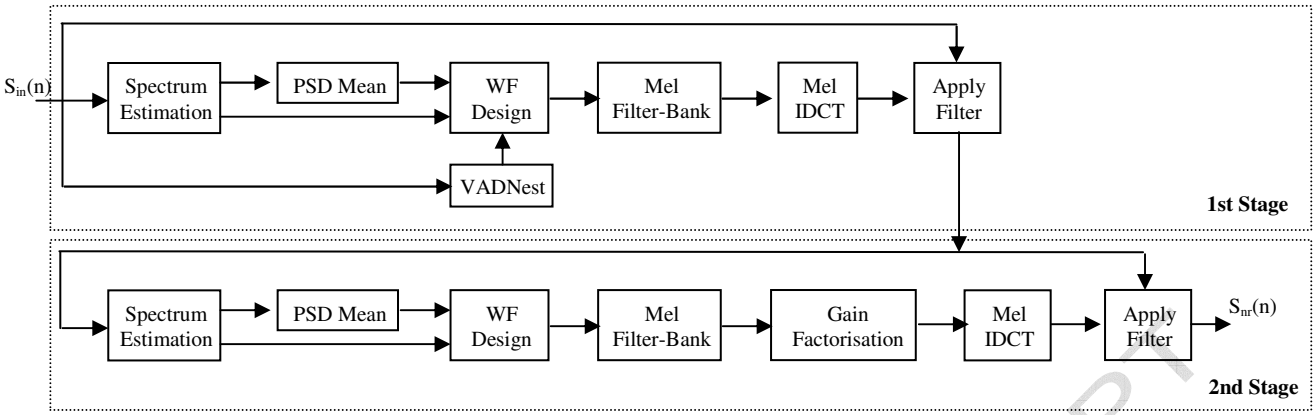


Figure 5: Block diagram of a two-stage mel-warped Wiener filter noise reduction scheme (Macho *et al.*, 2002)

5. The HTK Recogniser

For the experiments reported here, the HMM based recogniser architecture specified for use with the Aurora 2 database is used (Hirsch and Pearce, 2000), and the recogniser is implemented with the widely used HTK package. The use of a well-known specification provides a common framework with which to compare different front-ends and feature vectors for the purpose of connected digit recognition. There are eleven whole word HMMs each with 16 states and with each state having 3 Gaussian mixtures. Two pause models, "sil" and "sp", are defined. The "sil" model has 3 states and each state has 6 mixtures. The "sp" model has a single state. Script files provided with the Aurora 2 database for the purpose of training and testing a HTK based recogniser were used in the evaluation of the front-ends. The version of HTK used was HTK 3.3.

6. Perceptual Evaluation of Speech Quality

A further element in the evaluation of the speech enhancement algorithms was to estimate the improvement in perceptual quality of the enhanced speech produced by each algorithm. ITU-T Recommendation P.862 (2001) details an algorithm used for the Perceptual Evaluation of Speech Quality (PESQ). A reference C implementation of the algorithm is provided by ITU, and this was used for the purposes of this research. PESQ is an "intrusive" algorithm in that it compares a reference signal with a test signal (often a degraded version of the reference), generating an output Mean Opinion Score (MOS) that is a prediction of the perceived quality that would be assigned to the test signal by subjects in a subjective listening test. An overview of the basic philosophy used in PESQ is shown in Figure 6.

The reference signal and the test signal are converted into an internal representation based on a perceptual model. Time alignment is used to ensure that the two versions are synchronised in time. Differences in the two internal representations determine the audible difference between the two signals. The cognitive model computes two error parameters based on the differences between the signals and these parameters are combined to give an objective listening quality MOS, the range of which is between 0.5 and 4.5. For this work, the reference signal is the original clean utterance, while the test utterance is the utterance with additive noise, after application of a speech enhancement algorithm.

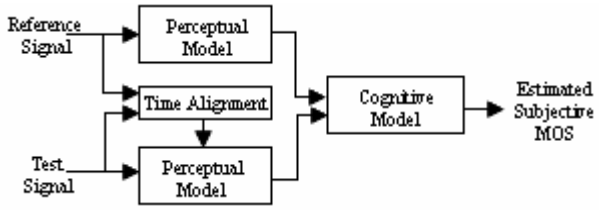


Figure 6: PESQ score (ITU-T Rec. P.862, 2001)

7. Results

This section presents results from the evaluation of the combination of the speech enhancement algorithms described in Section 3 and the auditory models described in Section 2.

7.1 Recognition Tests

The primary purpose of the paper is to examine the performance of speech enhancement algorithms in combination with the auditory model proposed by Li *et al.* (2000). As noted in Section 2.2, two versions of this algorithm were examined, Li *et al.* (I) and Li *et al.* (II). For comparison purposes, these two versions were

compared with baselines provided by the ETSI basic front-end (ETSI ES 201 208, 2003) and the ETSI advanced front-end (ETSI ES 202 050, 2007) respectively. In all cases training was carried out using clean data, so that the effect of the speech enhancement in removing mismatch could be examined. The speech enhancement algorithms were used on both the (clean) training speech as well as the (noisy) test speech. Feature vectors are extracted directly from the enhanced speech with no intermediate processing. The ETSI advanced front-end includes a SNR-dependent waveform processing block that is applied after noise reduction and before feature extraction. However, this work is looking primarily at the effect of speech enhancement or noise reduction alone on the connected digit recognition accuracy. Therefore, the waveform processing block in the ETSI advanced front-end was disabled. A detailed description of the waveform processing block can be found in (Macho and Cheng, 2001).

7.1.1 Li *et al.* (I) and the ETSI basic front-end

The two front-ends being compared in this case generate a feature vector consisting of 13 coefficients made up of the frame log-energy measure and the cepstral coefficients C_1 to C_{12}. There is no post-processing of the feature vectors carried out. The recognition results using the Aurora2 database for Li *et al.* (I), for each speech enhancement algorithm, are given in Table 1 and the corresponding results for the ETSI basic front-end are given in Table 2. The word accuracies are calculated according to Hirsch and Pearce (2000) which defines the performance measure for a test set as the word accuracy averaged over all noises and over all SNRs between 0 dB and 20dB. The overall word accuracy is calculated as the average over the three test sets A, B and C.

7.1.2 Li *et al.* (II) and the ETSI advanced front-end

The feature vector produced by the ETSI advanced front-end contains the cepstral coefficients C_1 to C_{12} along with a weighted combination of cepstral coefficient C_0 and the frame log-energy measure. In addition, the ETSI advanced front-end carries out post-processing in the cepstral domain in the form of blind equalisation as described by Mauuary (1998). To ensure a closer match with the ETSI advanced front-end, the weighted combination of C_0 and frame log-energy measure is included here in the feature vector generated by Li *et al.* (2000). The feature vectors produced by Li *et al.* undergo post-processing in the cepstral domain by means of cepstral mean subtraction (CMS). The recognition results are for Li *et al*. (II), for each speech enhancement algorithm, are detailed in Table 3 and the recognition results for the ETSI advanced front-end are detailed in Table 4. The word accuracies are again calculated as defined by Hirsch and Pearce (2000) which is the word accuracy averaged over all noises and over all SNRs between 0 dB and 20dB. The overall word accuracy is calculated as the average over the three test sets A, B and C.

	Absolute Word Accuracy %			
Enhancement	Set A	Set B	Set C	Overall
None	62.16	64.31	57.76	62.14
Ephraim & Malah	78.85	79.38	74.78	78.25
Westerlund *et al.*	75.87	76.32	70.45	74.97
Martin	72.47	71.96	70.21	71.81
Rangachari & Loizou	74.50	73.16	74.29	73.92
Agarwal & Cheng	86.33	84.87	81.86	84.85

Table 1: Recognition results - Li *et al.* (I)

	Absolute Word Accuracy %			
Enhancement	Set A	Set B	Set C	Overall
None	61.34	55.75	66.14	60.06
Ephraim & Malah	76.34	75.91	73.71	75.64
Westerlund *et al.*	76.04	72.54	72.36	73.90
Martin	67.98	67.57	68.24	67.87
Rangachari & Loizou	63.58	61.57	67.82	63.62
Agarwal & Cheng	84.39	82.75	78.72	82.60

Table 2: Recognition results - ETSI basic front-end

	Absolute Word Accuracy %			
Enhancement	Set A	Set B	Set C	Overall
None	67.34	69.18	63.44	67.30
Ephraim & Malah	80.36	81.03	79.34	80.42
Westerlund *et al.*	78.70	80.02	78.44	79.18
Martin	73.07	72.93	72.17	72.83
Rangachari & Loizou	76.08	76.16	75.94	76.08
Agarwal & Cheng	87.03	86.85	84.58	86.47

Table 3: Recognition results - Li *et al.* (II)

	Absolute Word Accuracy %			
Enhancement	Set A	Set B	Set C	Overall
None	65.92	65.48	70.07	66.57
Ephraim & Malah	77.92	77.61	78.64	77.94
Westerlund *et al.*	79.09	79.13	79.70	79.23
Martin	71.26	72.91	72.71	72.21
Rangachari & Loizou	73.77	73.35	78.85	74.62
Agarwal & Cheng	85.92	85.66	83.89	85.41

Table 4: Recognition results - ETSI advanced front-end

7.2 Perceptual Quality Tests

The PESQ (ITU-T Rec. P.862, 2001) algorithm was used to evaluate the perceptual quality of the enhanced speech. The results are detailed in Table 5 and are the averages of the PESQ mean opinion scores obtained for noisy test utterances with SNRs of 20 dB, 15 dB, 10 dB, 5 dB and 0 dB. To give some idea of the range of performance obtained across different SNRs, the PESQ scores obtained across the SNR range for Set A when enhanced using Martin's algorithm, varied from 3.2153 (SNR=20dB) to 1.8456 (SNR=0dB). This compares with PESQ scores of 2.7962 (SNR=20dB) and 1.6069 (SNR=0dB) when the speech is not enhanced.

	Average PESQ MOS score			
Enhancement	Set A	Set B	Set C	Overall
None	2.1903	2.2730	2.1433	2.2140
Ephraim & Malah	2.4255	2.4689	2.3563	2.4290
Westerlund *et al.*	2.3198	2.3885	2.2818	2.3397
Martin	2.5645	2.5658	2.4351	2.5391
Rangachari & Loizou	2.4275	2.4696	2.3632	2.4315
Agarwal & Cheng	2.4869	2.5203	2.3986	2.4826

Table 5: Perceptual quality evaluation score

8. Computational Complexity Analysis

It is of interest to compare the computational complexity of the enhancement algorithms evaluated, since such systems would need to be implemented on a mobile handset if used as part of a DSR framework. Table 6 presents an estimate of the number of multiplications and additions required for each enhancement algorithm, for a frame of speech of size N samples. The complexity estimate for the algorithm proposed by Westerlund *et al.* is taken directly from (Westerlund *et al,* 2005) where K is the number of filter sub-bands used. For the work presented here, the input signal was divided into 12 sub-bands. The analysis focuses on a straightforward implementation of the core signal processing blocks, and does not take into account additional overhead due to buffering of data, control loops etc., as these are assumed to be approximately the same for all the enhancement algorithms.

A number of assumptions have been made in the computational complexity analysis of the speech enhancement algorithms. In calculating an N-point FFT, radix-2 is assumed which requires $0.5N\log_2 N$ complex multiplications and $N\log_2 N$ complex additions. In equation (4) of Ephraim and Malah's (1984) algorithm, it is assumed that the Bessel functions are implemented by means of a look-up table. For transcendental functions, a power series approximation is used and this is limited to a third-order approximation as it is a reasonable trade-off between accuracy and computational requirements.

Both Ephraim and Malah (1984) and Rangachari and Loizou (2006) operate in the frequency domain and make use of the Fast Fourier Transform (FFT) and the inverse-FFT. Both algorithms show comparable complexity. Martin's (2001) enhancement algorithm also operates in the frequency domain but requires more multiplications and additions. The two-stage algorithm of Agarwal and Cheng (1999) requires almost twice as many multiplications and fifty percent more additions compared to that of Martin's algorithm for a frame with N samples of speech. For Westerlund *et al.* (2005), when K is equal to 12, the number of multiplications required is comparable to that of Agarwal and Cheng (1999) and the number of additions required is comparable to that of Martin (2001). In Agarwal and Cheng's (1999) algorithm, calculation of the time-domain impulse response of the Wiener filter and the application of this filter in the time-domain by means of a convolution operation take place in each of the two stages of the algorithm (see Figure 5). These two operations contribute to the large number of multiplication and addition operations in Table 6.

Enhancement	Multiplications	Additions
Ephraim & Malah	$4N\log_2 N+11N$	$4N\log_2 N+2.5N$
Westerlund *et al.*	$(1+8K)N$	$(1+4K)N$
Martin	$4N\log_2 N+23N+14$	$4N\log_2 N+15.5N+4$
Rangachari & Loizou	$4N\log_2 N+10.5N$	$4N\log_2 N+6N$
Agarwal & Cheng	$4N\log_2 N+55N+4054$	$4N\log_2 N+22N+3970$

Table 6: Computational analysis for frame size N

In the evaluation, the frame size and the frame overlap used were not the same for each of the speech enhancement algorithms. As a further comparison, the number of operations required by each speech enhancement algorithm to process 1 second of speech was estimated and these are presented in Table 7. This gives some idea of the MIPS capability required for implementation using an embedded microcontroller or DSP. From this, it can be clearly seen that, perhaps not surprisingly, the two-stage algorithm of Agarwal and Cheng (1999) is the most computationally demanding.

The overall processing complexity of the speech enhancement algorithms could be reduced by investigating the use of more efficient implementations of some of the sub-blocks in the individual algorithms. For example, a proposal is presented by Li *et al.* (2004) to improve the efficiency of the implementation of the speech enhancement algorithm proposed by Agarwal

and Cheng (1999) in the ETSI advanced front-end (ETSI ES 202 050, 2007). Their proposed structure for the Wiener filtering algorithm has a computational load that is one third of that of the original algorithm proposed by Agarwal and Cheng (1999).

Enhancement	Multiplications	Additions
Ephraim & Malah	1079000	866000
Westerlund *et al.*	776000	392000
Martin	860000	742000
Rangachari & Loizou	664000	593000
Agarwal & Cheng	2580000	1744000

Table 7: Computational analysis for 1 second of speech

9. Discussion

Table 8 provides an overall view of the relative performance of the different speech enhancement algorithms for each of the four front-end versions considered. The last column in Table 8 indicates relative placement from the perspective of perceptual quality as estimated by the PESQ algorithm.

Ignoring speech enhancement, comparing Tables 1 and 2, the performance of Li *et al*. (I) exceeds the baseline ETSI front-end (ETSI ES 201 108, 2003) by 2.08% overall. From Table 3 and Table 4, again without speech enhancement applied, there is a difference in recognition accuracy of 0.73% in favour of Li *et al.* (II) when compared with the ETSI advanced front-end (ETSI ES 202 050, 2007).

The other results in Tables 1 to 4 show that enhancement of the speech prior to feature extraction significantly improves the overall recognition performance. This improvement in recognition accuracy is observed for both the ETSI basic (ETSI ES 201 108, 2003) and advanced (ETSI ES 202 050, 2007) front-ends and the front-end proposed by Li *et al.* (2000). A comparison of Table 1 with Table 2 shows that Li *et al.* (I) outperforms the ETSI basic front-end for each of the speech enhancement techniques evaluated. Furthermore, from Tables 3 and 4, it is seen that Li *et al.* (II) again outperforms the ETSI advanced front-end for all speech enhancement methods except Westerlund *et al.* (2005), for which the overall recognition results are quite close.

For Li *et al.* (I), Li *et al.* (II), the ETSI basic front-end and the ETSI advanced front-end, the best overall recognition accuracy is obtained for speech enhancement using the algorithm proposed by Agarwal and Cheng (1999). The combination of auditory front-end and the two-stage, mel-warped, Wiener filter noise reduction scheme results in an overall recognition accuracy that is approximately 6% better overall compared with the next ranked front-end and speech enhancement combination. After Agarwal and Cheng (1999), the next best performance across the board is obtained using Ephraim and Malah (1984), and Westerlund *et al*. (2005). This suggests that the choice of speech enhancement algorithm for best speech recognition performance is somewhat independent of the choice of front-end (though clearly this would have to be validated by further testing with other front ends).

On the other hand, taking the PESQ scores into account, there is no obvious relationship between perceptual quality and recognition performance, though there is a slight tendency for algorithms that perform well in recognition (Westerlund *et al*., 2005; Ephraim and Malah, 1984) to perform less well from a quality perspective. (though Agarwal and Cheng, 1999, produces reasonably good quality speech as well as good speech recognition performance). At the same time, the overall recognition performance when using Martin's (1994, 2001) enhancement algorithm is poor compared to the other algorithms investigated. In contrast to its recognition performance, speech enhanced using Martin's (1994, 2001) algorithm obtained the highest perceptual quality score using PESQ.

The performance rankings in Table 8 support the notion that speech enhancement to obtain better speech quality and speech enhancement for the purpose of intelligibility (as reflected in speech recognition performance) should possibly be regarded as two separate tasks. Speech enhancement algorithms may need to be designed with one target in mind, speech quality or speech intelligibility, but not both. However, the algorithm of Agarwal and Cheng (1999) does seem to produce a good compromise in both applications.

10. Conclusions

This paper has examined the speech recognition performance of a number of speech enhancement algorithms combined with the auditory model front-end proposed by Li *et al.* (2000). A range of speech enhancement algorithms was considered, including both well established techniques, and more recently proposed methods. For comparison purposes, the recognition results obtained using the ETSI basic front-end (ETSI ES 201 108, 2003) and the ETSI advanced front-end (ETSI ES 202 050, 2007) were used as a baseline. The perceptual quality of speech produced by each enhancement algorithm was also estimated using the PESQ algorithm (ITU-T Rec. P.862, 2001).

Overall, speech enhancement was found to improve recognition performance, in particular by compensating for mismatch between training and testing conditions when the recogniser was trained using clean speech only. The results indicate that the auditory-based front-end processor of Li *et al.* (2000) gives improved recognition results when compared with the ETSI basic (ETSI ES 201 108, 2003) and advanced (ETSI ES 202

Rank	Li *et al.* (I)	ETSI basic front-end	Li *et al.* (II)	ETSI advanced FE	PESQ
1	Agarwal & Cheng	Agarwal & Cheng	Agarwal & Cheng	Agarwal & Cheng	Martin
2	Ephraim & Malah	Ephraim & Malah	Ephraim & Malah	Westerlund *et al.*	Agarwal & Cheng
3	Westerlund *et al.*	Westerlund *et al.*	Westerlund *et al.*	Ephraim & Malah	Rangachari & Loizou
4	Rangachari & Loizou	Martin	Rangachari & Loizou	Rangachari & Loizou	Ephraim & Malah
5	Martin	Rangachari & Loizou	Martin	Martin	Westerlund *et al.*

Table 8: Performance ranking of enhancement algorithms

050, 2007) front-ends. The best recognition performance is obtained with the combination of the speech enhancement method of the two-stage, mel-warped, Wiener filter noise reduction scheme proposed by Agarwal and Cheng (1999) and the auditory model front-end. However, analysis of both speech recognition performance, and perceptual quality scores, suggests that the most optimal speech enhancement performance for both applications may not be possible with a single algorithm.

Future work will include additional testing to further validate some of the results presented here. In addition, further analysis of the computational complexity of the front-end and speech enhancement combinations for embedded implementation will be carried out.

Furthermore, the evaluation of the auditory front-end and the speech enhancement algorithms in this paper used a HMM-based recogniser architecture, since this is well-known and has been standardised within the Aurora framework. However, it has been proposed that HMMs have a number of modelling inadequacies arising from assumptions that are made in order to simplify the speech recognition task (Tsontzos et al., 2007). Linear Dynamical Models (LDMs) have been proposed as a method of enhancing speech recognition performance in the presence of noise. Future work will investigate the combination of the auditory model of Li *et al.* (2000) and an LDM based classifier.

References

Agarwal, A., Cheng, Y.M., 1999. Two-stage mel-warped wiener filter for robust speech recognition. In: Proceedings of Automatic Speech Recognition and Understanding Workshop, Keystone, Colorado, USA, pp. 67-70.

Dau, T., Püschel, D., Kohlrausch, D., 1996. A quantitative model of the 'effective' signal processing in the auditory system: I. Model structure. J. Acoust. Soc. Amer., vol. 99, no. 6, pp. 3615-3622.

Davis, S., Mermelstein, P., 1980. Comparision of parameteric representations for monosyllable word recognition in continuous spoken sentences. IEEE Trans. on Acoustics, Speech and Signal Processing, vol. 28(4), pp. 357-366.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process., vol ASSP-32, no. 6, pp. 1109-1121.

Ephraim, Y., Cohen, I., 2006. Recent advancements in speech enhancement. In: The Electrical Engineering Handbook, CRC Press, 2006

ETSI ES 201 108 Ver. 1.1.3, 2003. Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Front-end feature extraction algorithm; Compression algorithms.

ETSI ES 202 050 Ver. 1.1.5, 2007. Speech Processing, Transmission and Quality Aspects (STQ); Distributed speech recognition; Advanced front-end feature extraction algorithm; Compression algorithms.

Flynn R., Jones, E., 2006. A comparative study of auditory-based front-ends for robust speech recognition using the Aurora 2 database. In: Proceedings of the IET Irish Signals and Systems Conference, Dublin, Ireland, pp. 111-116.

Ghitza, O., 1988. Temporal non-place information in the auditory-nerve firing patterns as a front-end for speech recognition in a noisy environment. Journal of Phonetics, vol 16, pp. 109-123.

Hermansky, H., 1990. Perceptual linear prediction (PLP) analysis of speech. J. Acoust. Soc. Amer., vol. 87, no. 4, pp. 1738-1752.

Hirsch, H.G., Pearce, D., 2000. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: Proc. ISCA ITRW ASR2000, Paris, France, pp. 181-188.

Hu, Y., Loizou, P., 2007. A comparative intelligibility study of single-microphone noise reduction algorithms. J. Acoust. Soc. Amer., vol. 122, no. 3, pp. 1777-1786.

ITU-T Rec. P.862, 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs.

Kleinschmidt, M., Tchorz, J., Kollmeier, B., 2001. Combining speech enhancement and auditory feature extraction for robust speech recognition. Speech Communication, 34, 75-91.

Li, Q., Soong, F.K., Siohan, O., 2000. A high-performance auditory feature for robust speech recognition. In: Proc. of 6th International Conference on Spoken Language Processing (ICSLP), pp. III 51-54.

Li, Q., Soong, F.K., Siohan, O., 2001. An auditory system-based feature for robust speech recognition. In: Proc. of Eurospeech, 2001. vol. 1, pp. 619-622.

Li, J., Liu, B., Wang, R., Dai, L., 2004. A complexity reduction of ETSI advanced front-end for DSR. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '04), vol. 1, pp. 61-64.

Macho, D., Cheng, Y.M., 2001. SNR-dependent waveform processing for robust speech recognition. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '01), pp. 305-308.

Macho, D. *et al.*, 2002. Evaluation of a noise-robust DSR front-end on Aurora databases. In: Proceedings of International Conference on Speech and Language Processing, Denver, Colorado, USA, pp. 17-20.

Mak, B., Tam, Y., Li, P., 2004. Discriminative auditory-based features for robust speech recognition. IEEE Trans. on Speech and Audio Processing, vol 12, no. 1, pp. 27-36.

Martin, R., 1994. Spectral subtraction based on minimum statistics. In: Proc. Eur. Signal Processing Conference, pp 1182-1185.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. on Speech and Audio Processing, vol 9, no. 5, pp. 504-512.

Mauuary, L., 1998. Blind Equalization in the cepstral domain for robust telephone based speech recognition. In: Proc. EUSPICO '98, Vol. 1, pp. 359-363.

Milner, B., 2002. A Comparison of front-end configurations for robust speech recognition. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '02), pp. I 797-800.

Moore, B., 2003. An Introduction to the Psychology of Hearing, 5th ed. Academic Press.

Rangachari, S., Loizou, P., 2006. A noise-estimation algorithm for highly non-stationary environments. Speech Communication, 48, 220-231.

Seneff, S., 1988. A joint synchrony/mean-rate model of auditory speech processing. Journal of Phonetics, vol 16, pp. 55-76.

Tchorz, J., Kollmeier, B., 1999. A model of auditory perception as front end for automatic speech recognition. J. Acoust. Soc. Amer., vol. 106, no. 4, pp. 2040-2050.

Tsontzos, G., Diakoloukas, V., Koniaris, C., Digalakis, V., 2007. Estimation of general identifiable linear dynamic models with an application in speech recognition. In: Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP '07), vol. IV, pp. 453-456.

Westerlund, N., Dahl, M., Claesson, I., 2005. Speech enhancement for personal communication using an adaptive gain equalizer. Speech Communication, 85, 1089-1101.