Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 53 (2011) 51–61

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Use of speech presence uncertainty with MMSE spectral energy estimation for robust automatic speech recognition

Anthony Stark, Kuldip Paliwal *

Signal Processing Laboratory, Griffith University, Nathan, QLD 4111, Australia

Received 21 April 2010; received in revised form 25 June 2010; accepted 2 August 2010

Abstract

In this paper, we investigate the use of the minimum mean square error (MMSE) spectral energy estimator for use in environment-robust automatic speech recognition (ASR). In the past, it has been common to use the MMSE log-spectral amplitude estimator for this task. However, this estimator was originally derived under subjective human listening criteria. Therefore its complex suppression rule may not be optimal for use in ASR. On the other hand, it can be shown that the MMSE spectral energy estimator is closely related to the MMSE Mel-frequency cepstral coefficient (MFCC) estimator. Despite this, the spectral energy estimator has tended to suffer from the problem of excessive residual noise. We examine the cause of this residual noise and show that the introduction of a heuristic based speech presence uncertainty (SPU) can significantly improve its performance as a front-end ASR enhancement regime. The proposed spectral energy SPU estimator is evaluated on the Aurora2, RM and OLLO2 speech recognition tasks and can be shown to significantly improve additive noise robustness over the more common spectral amplitude and log-spectral amplitude estimators.

Keywords: Robust speech recognition; MMSE estimation; Speech enhancement methods

1. Introduction

The development of robust automatic speech recognition (ASR) is an important goal. While performance for state of the art ASR is impressive during ideal conditions, its recognition accuracy tends to degrade rapidly in the presence of additive background noise. Since it is often impossible to eliminate all noise from the operating environment, the problem of ASR robustness has been receiving considerable attention. Several approaches have been proposed in the literature, most of which fall under two categories: front-end speech/feature enhancement and back-end model adaptation. Back-end adaptation seeks to modify the acoustic models of the recognizer to better match the noisy operating environment. Font-end enhancement on the other hand, seeks to remove the effects of noise prior to recognition, either from the speech signal or from the parameterized features directly.

In this paper, we are interested in methods that perform enhancement on the speech signal. Several methods falling into this category have been reported in the literature (Loizou, 2007). This includes spectral subtraction (Berouti et al., 1979), minimum mean square error (MMSE) estimation (Ephraim and Malah, 1985), Wiener filtering (linear MMSE) (Wiener, 1949), Kalman filtering (Paliwal and Basu, 1987) and subspace (Ephraim and Trees, 1995) methods. These algorithms are specifically designed to improve the subjective quality of an acoustic signal for human listeners. For example, the MMSE log-spectral amplitude (LSA) estimator is often favored because of its psychoacoustic considerations. While many of the aforementioned algorithms have been used for robust ASR (Lathoud et al., 2005; Ephraim and Trees, 1991; Gemello et al., 2006;

* Corresponding author. Tel.: +61 7 3875 6536; fax: +61 7 3875 5198.
E-mail addresses: a.stark@griffith.edu.au (A. Stark), k.paliwal@griffith.edu.au (K. Paliwal).
URL: http://maxwell.me.gu.edu.au/spl/ (K. Paliwal).

0167-6393/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2010.08.001

Hermus et al., 2007; Fujimoto and Ariki, 2000), there are clear differences between the objectives of robust ASR and speech enhancement.

For subjective human listening, it is often held that noise suppression is most effective when applied to the log-spectral domain. The LSA estimator was derived under such an assumption. However, the typical ASR system does not operate directly on the log-spectral domain. Instead, higher level features such as Mel-frequency cepstral coefficients (MFCCs) are used. As a result, the complicated suppression rule of the LSA estimator may not be fully justified for use in ASR-based speech enhancement.

In this paper, we examine a similar estimator for use in robust ASR; namely the spectral energy (SE) estimator. Specifically, we investigate its suitability for estimating clean speech MFCCs, from speech corrupted with additive noise. We show that the suppression rule of the SE estimator is closely related to the MMSE MFCC estimator. That is, an estimator that produces a cepstral estimate $\hat{\boldsymbol{c}}_x$ that minimizes the square error from the true, clean MFCC vector $\boldsymbol{c}_x$. Despite this, the SE estimator has several shortcomings that must be addressed before it can be used for robust ASR. Foremost among these problems is its tendency to under-suppress noise at low signal to noise ratios (SNRs). We identify two causes of this under-suppression: (1) an inherent positive bias when using the SE estimator to derive log-filterbank energies and (2) the tendency of the SE estimator to over-estimate the *a priori* SNR within a decision-directed framework (Ephraim and Malah, 1984). Later, we show that both of these issues may be corrected with the use of a heuristic based speech presence uncertainty (SPU). The proposed SE SPU estimator offers a number of advantages over the LSA estimator. First, its suppression rule is more efficiently implemented and second, it offers better recognition performance across a wide range of noise types and SNRs.

The rest of this paper is organized as follows. In Section 2, we cover the statistical framework used to derive the common short-time spectral amplitude estimators. In Section 3, we investigate the use of the SE estimator for deriving MFCC features. Firstly, we examine the optimality of the SE estimator in the context of MFCC estimation. Secondly, we highlight the considerations that must be taken into account for practical implementation of the SE estimator. In Section 4, we first describe the use of SPU within the spectral estimation framework. We then show how SPU may be used to overcome the limitations of the SE estimator. In Section 5, we present experimental ASR results for the RM (Price et al., 1988), OLLO2 (Wesker et al., 2005) and Aurora2 (Pearce and Hirsch, 2000) ASR tasks. Lastly in Section 6, we present concluding remarks.

2. Statistical framework for short-time spectral amplitude estimation

The discrete short-time Fourier transform (DSTFT) of corrupted speech signal $y(n)$ is given by

$$Y(m,k) = \sum_{n=-\infty}^{\infty} y(n)w(mS-n)\exp(-j2\pi kn/K), \tag{1}$$

where k denotes the kth discrete frequency of K uniformly spaced frequencies, $w(n)$ is an analysis window function, m is the short-time frame index and S is the analysis frame shift (in samples). In this paper, we consider an additive noise model. Here, the corrupted speech DSTFT may also be represented as[1]

$$Y(k) = X(k) + D(k), \tag{2}$$

where $X(k)$ and $D(k)$ are the DSTFT expansion coefficients for the kth discrete frequency bin of the clean speech signal and noise signals, respectively.

DSTFT expansion coefficients $X(k)$ and $D(k)$ are assumed to be independent complex zero-mean Gaussian variables, with expected power $\lambda_x(k) = E[|X(k)|^2]$ and $\lambda_d(k) = E[|D(k)|^2]$, where $E[\cdot]$ is the expectation operator. A detailed justification of this statistical assumption may be found in (Ephraim and Malah, 1984).

The general goal of speech enhancement is to derive an estimate of the clean speech given the observed noisy speech and a noise estimate. To accomplish this, it can be useful to split DSTFT coefficients $Y(k)$ and $X(k)$, into spectral amplitude and phase:

$$Y(k) = R(k)\exp(j\vartheta(k)), \tag{3}$$

$$X(k) = A(k)\exp(j\theta(k)), \tag{4}$$

where $R(k)$ and $\vartheta(k)$ are the amplitude and phase spectrums of the noisy speech, respectively, while $A(k)$ and $\theta(k)$ are the amplitude and phase spectrums of the clean speech, respectively.

For typical Fourier analysis-modification-synthesis (AMS) based speech enhancement (Ephraim and Malah, 1985; Ephraim and Malah, 1984; Boll, 1979), an estimate $\widehat{A}(k)$ is obtained from the noisy signal and combined with the noisy spectral phase $\vartheta(k)$ to produce the estimated clean speech spectrum $\widehat{X}(k)$

$$\widehat{X}(k) = \widehat{A}(k)\exp(j\vartheta(k)) = Y(k)\cdot G(k), \tag{5}$$

where

$$G(k) = \frac{\widehat{A}(k)}{R(k)} \tag{6}$$

is the spectral amplitude gain of the speech enhancement system and $\widehat{A}(k)$ is the estimated clean DSTFT coefficient amplitude. Using the DSTFT estimate $\widehat{X}(k)$, enhanced time-domain speech may then be synthesized with an inverse discrete Fourier transform (IDFT) and overlap-add-synthesis (Crochiere, 1980). A block diagram of the typical AMS-based speech enhancement framework is given in Fig. 1.

Several spectral amplitude estimators have been suggested in the literature. The spectral amplitude gain func-

[1] For notational convenience, we have dropped the frame index m and dependence on this subscript is implicitly assumed unless stated otherwise.

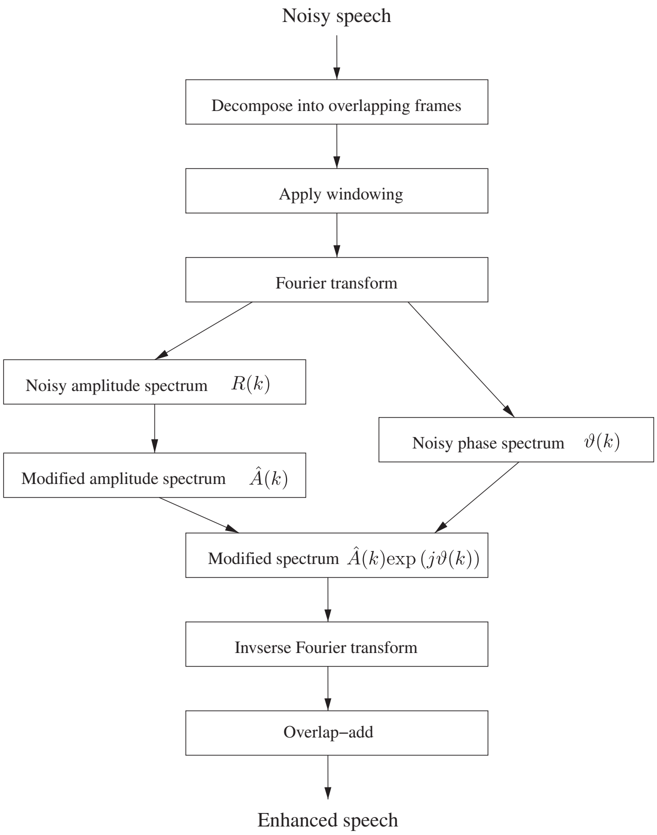
Noisy speech
Decompose into overlapping frames
Apply windowing
Fourier transform
Noisy amplitude spectrum $R(k)$
Noisy phase spectrum $\vartheta(k)$
Modified amplitude spectrum $\hat{A}(k)$
Modified spectrum $\hat{A}(k)\exp(j\vartheta(k))$
Invserse Fourier transform
Overlap–add
Enhanced speech

Fig. 1. Block diagram of the short-time analysis-modification-synthesis speech enhancement framework.

tions for the spectral Wiener (SW) (Loizou, 2007), MMSE spectral amplitude (SA) (Ephraim and Malah, 1984), MMSE log-spectral amplitude (LSA) (Ephraim and Malah, 1985) and MMSE spectral energy (SE) filters are given below:

$$G_{SW}(k) = \frac{\xi(k)}{1+\xi(k)}, \tag{7}$$

$$G_{SA}(k) = \frac{\sqrt{\pi v(k)}}{2\gamma(k)} \exp\left(\frac{-v(k)}{2}\right) \cdot \left[(1+v(k)) I_0\left(\frac{v(k)}{2}\right) + v(k) I_1\left(\frac{v(k)}{2}\right)\right], \tag{8}$$

$$G_{LSA}(k) = \frac{\xi(k)}{1+\xi(k)} \exp\left(\frac{1}{2}\int_{v(k)}^{\infty} \frac{\exp(-t)}{t} dt\right), \tag{9}$$

$$G_{SE}(k) = \frac{\xi(k)}{1+\xi(k)} \sqrt{1+\frac{1}{v(k)}}, \tag{10}$$

where

$$v(k) = \frac{\xi(k)}{1+\xi(k)}\gamma(k), \tag{11}$$

$$\xi(k) = \frac{\lambda_x(k)}{\lambda_d(k)}, \tag{12}$$

$$\gamma(k) = \frac{[R(k)]^2}{\lambda_d(k)}. \tag{13}$$

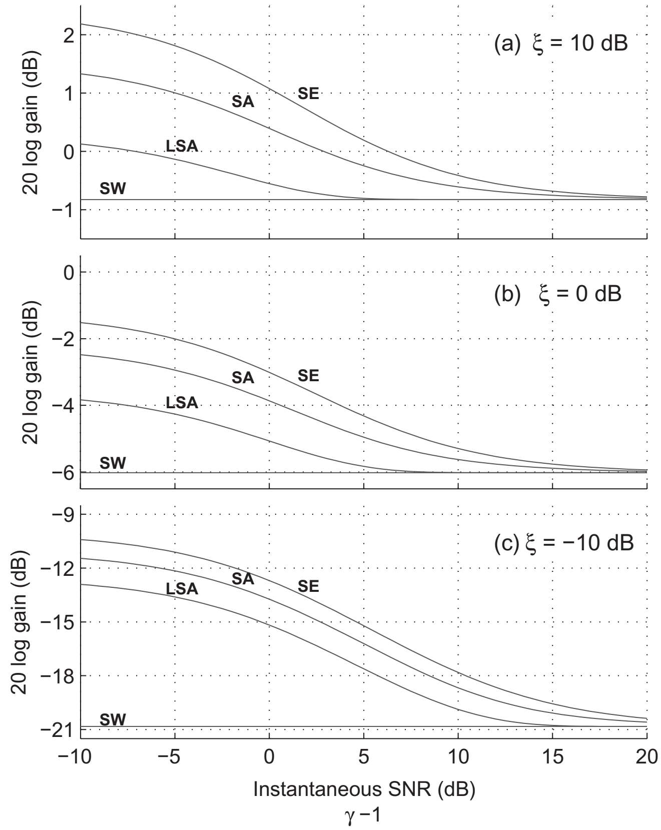

Fig. 2. Common spectral amplitude gain functions. Gain functions shown are: spectral energy (SE), spectral amplitude (SA), log-spectral amplitude (LSA) and spectral Wiener (SW) estimators. Subplots: (a) *A priori* SNR = 10 dB, (b) *a priori* SNR = 0 dB, (c) *a priori* SNR = −10 dB.

The parameters ξ and γ are interpreted as the *a priori* signal to noise ratio (SNR) and *a posteriori* SNR, respectively. $I_0(\cdot)$ and $I_1(\cdot)$ are given as the zeroth and first order modified Bessel functions, respectively. Fig. 2 shows the spectral amplitude gain functions for each estimator over several SNR values. Interestingly, all gains become equivalent to the spectral Wiener gain at high SNRs; i.e., when $v(k) \gg 1$.

3. Use of the SE estimator for ASR

As stated earlier, our goal in this paper is to investigate the SE estimator for use in ASR-based speech enhancement. One immediate justification for this is the simple gain rule (10), which requires less computation than both the SA and LSA estimators. A second reason for investigating the spectral energy estimator, is that it is closely related to the log-filterbank energy estimator – the intermediate stage of the popular MFCC feature set (Huang et al., 2001).

Despite these reasons, use of the SE estimator is not common in the ASR field. This is largely due to its poor noise suppression in low SNR environments. Consequently, this problem must be understood and compensated if the SE estimator is to be used in ASR. Two causes of noise under-suppression are identified in the remainder of this section.

3.1. Sub-optimality of the SE estimator for generating MFCCs

Since MFCCs are currently the dominant speech parameterization, a suitable goal for ASR-centric speech enhancement is optimal estimation of the MFCC vector. In this sub-section we determine the relationship between the SE estimator and the optimal MMSE MFCC estimator. The optimal MMSE MFCC estimator is given as

$$\hat{\boldsymbol{c}}_x = E[\boldsymbol{c}_x|\boldsymbol{Y}], \tag{14}$$

where $\hat{\boldsymbol{c}}_x \in \mathbb{R}^{Q\times 1}$ is the MFCC estimate that minimizes the square error from the true, clean MFCC vector $\boldsymbol{c}_x$ and $\boldsymbol{Y} = [Y(0), Y(1), \ldots, Y(K-1)]^T$ is a spectral frame of noisy speech. The MFCC vector is related to the log-filterbank energy vector via the discrete cosine transform (DCT). Since the DCT is a unitary operator, the total squared error in both the MFCC and log-filterbank domains is equivalent. This allows us to recast our problem into MMSE estimation of log-filterbank energies

$$\hat{\boldsymbol{c}}_x = \boldsymbol{C}\cdot E[\boldsymbol{L}_x|\boldsymbol{Y}], \tag{15}$$

where $\boldsymbol{L}_x$ is the clean speech log-filterbank energy and $\boldsymbol{C} \in \mathbb{R}^{Q\times Q}$ is the DCT matrix. Under normal circumstances, higher order cepstral coefficients are truncated from the DCT matrix (Huang et al., 2001). Strictly speaking, (15) is not the MMSE estimate for the truncated MFCC vector. However it remains a good approximation since truncated coefficients themselves tend to have very small energies. Using this approximation, we may now directly determine the suitability of the SE estimator for ASR. The SE criterion for estimating clean spectral amplitudes is given by

$$\widehat{A}(k) = \sqrt{E[[A(k)]^2|\boldsymbol{Y}]}. \tag{16}$$

Assuming filterbank energies are accumulated off spectral energies (and not amplitudes), log-filterbank energies will be given by

$$\begin{aligned}\widehat{L}_x^{SE}(q) &= \log\left(\sum_k h(q,k)E[[A(k)]^2\middle|\boldsymbol{Y}]\right)\\ &= \log E\left[\left(\sum_k h(q,k)[A(k)]^2\right)\middle|\boldsymbol{Y}\right],\end{aligned} \tag{17}$$

where $\widehat{L}_x^{SE}(q)$ is the SE estimate of the qth log-filterbank energy and $h(q,k)$ is the filterbank gain of the qth filterbank and kth frequency bin. By Jensen's inequality we can show that these estimates will be positively biased, i.e.:

$$E[\log(B(q)|\boldsymbol{Y})] \leqslant \log E[B(q)|\boldsymbol{Y}], \tag{18}$$

where filterbank energy $B(q) = \sum_k h(q,k)[A(k)]^2$. The first term of (18) is the ideal (MMSE) log-filterbank energy estimate, and the second term is the estimate produced by the SE estimator. Both terms are quite similar, differing by only the position of the logarithm. The positive bias arises from the fact that this logarithm is a concave function. If a convex operator was used instead, the inequality would be reversed. If a linear operator was used, then the inequality would become an equality. This is important, because over small dynamic ranges, the logarithm is approximately linear. Fig. 3 shows this. Here, both filterbank variables B_1 and B_2 have the same mean, but different variances. The variable B_1 exists on a fairly small dynamic range. Over this range the logarithm is approximately linear. The variable B_2 exists on a much larger dynamic range. Over this range the logarithm is highly concave.

This suggests that the SE estimator would perform quite well at higher SNRs – conditions where there is little uncertainty/variance in the estimation of filterbanks $B(q)$. In such a case, the concavity of the logarithm would play a minor role, meaning the SE estimator would (effectively) produce unbiased log-filterbank energies. Conversely, at lower SNRs we would expect the positive bias to become worse. Large amounts of noise energy will introduce large variance into the estimation of $B(q)$ – making the logarithm a highly non-linear, concave operation.

3.2. Considerations for estimation of a priori SNR

A more practical consideration of the SE estimator involves estimation of the *a priori* SNR parameter ξ. While the estimation of *a posteriori* SNR γ is relatively straightforward, several considerations must be taken into account when estimating ξ. The typical method for estimating ξ is the recursive, decision-directed approach presented in (Ephraim and Malah, 1984). This approach assumes the *a priori* SNR to be a slowly evolving parameter. Here the *a priori* SNR for the mth analysis frame and kth frequency bin $\xi(m,k)$, is estimated as the weighted sum of two terms

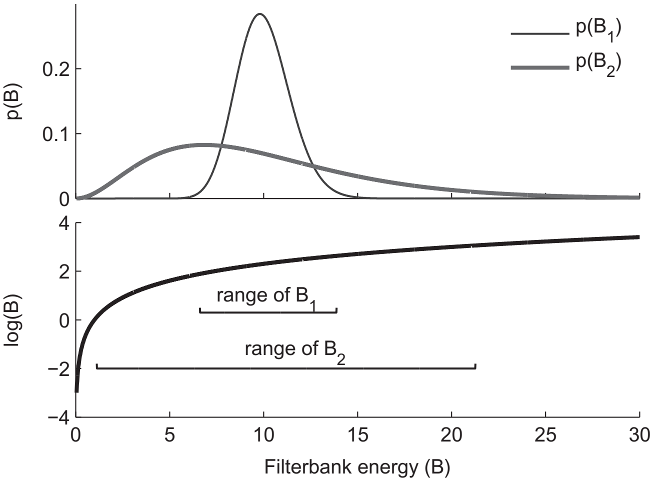

Fig. 3. Effect of the logarithm on a filterbank variable. Top figure shows probability density functions for two filterbank variables B_1 and B_2. Variable B_1 exists on a small dynamic range, so the logarithm is a predominantly linear effect. Variable B_2 has much larger dynamic range, over which the logarithm is highly concave.

$$\xi(m,k) = \alpha\frac{[G(m-1,k)R(m-1,k)]^2}{\lambda_d(m-1,k)} + (1-\alpha)P[\gamma(m,k)-1], \tag{19}$$

where mixing constant $\alpha \approx 0.98$ and

$$P[x] = \begin{cases} x, & \text{if } x \geqslant 0, \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

The term $P[\gamma(m,k)-1]$ in (19) can be interpreted as the instantaneous SNR, and is derived as the maximum likelihood estimate of ξ. The proceeding term in (19) is an estimate of ξ derived from the previous, enhanced frame. The inclusion of spectral gain $G(m-1,k)$ makes this term highly dependent on the enhancement regime. This is of particular concern for the SE estimator, as it has a relatively mild spectral gain (w.r.t the SA and LSA estimators, see Fig. 2). Because of this, residual noise often makes its way into the *a priori* SNR estimate. This leads to over-estimation of ξ, which itself leads to less suppression and more residual noise in subsequent frames.

4. Use of speech presence uncertainty to improve the spectral energy estimator

In the previous section, we have highlighted two causes of noise under-suppression in the SE estimator:

1. The inherent positive bias of the SE estimator to derive log-filterbank energies.
2. The tendency to over-estimate the *a priori* SNR ξ within the decision-directed framework.

Combined, these problems degrade ASR performance substantially in low SNR environments. To address both of these problems, we investigate the use of speech presence uncertainty (SPU) (McAulay and Malpass, 1980).

4.1. Overview of speech presence uncertainty within the spectral estimation framework

SPU does not assume speech to be present at all times and at all frequencies. Instead, speech presence is represented as a probabilistic variable. A two-state speech (absent/present) hypothesis can be incorporated into the conditional probability density function (PDF) $p(A(k)|Y(k))$ as follows:

$$\begin{aligned} p(A(k)|Y(k)) = {} & p(H_0(k)|Y(k)) \cdot p(A(k)|Y(k),H_0(k)) \\ & + p(H_1(k)|Y(k)) \cdot p(A(k)|Y(k),H_1(k)), \end{aligned} \tag{21}$$

where $H_0(k)$ and $H_1(k)$ represent the hypotheses of speech absence and presence, respectively for the k'th frequency bin. Given an *a posteriori* probability of speech presence $\varphi(k) \triangleq p(H_1(k)|Y(k))$ for the kth frequency bin, a SPU modified conditional PDF $p(A(k)|Y(k))$ can be given as follows:

$$\begin{aligned} p(A(k)|Y(k)) = {} & \varphi(k) \cdot p(A(k)|Y(k),H_1(k)) \\ & + (1-\varphi(k)) \cdot \delta(A(k)), \end{aligned} \tag{22}$$

where $\delta(\cdot)$ is the Dirac delta function. Here we have assumed that under signal absence hypothesis $H_0(k)$, the clean spectral amplitude $A(k)$ is most surely zero. The form and derivation of $p(A(k)|Y(k),H_1(k))$ may be found in (Ephraim and Malah, 1984). The SE SPU estimate for the clean spectral amplitude $\widehat{A}(k)$ is now given by

$$\widehat{A}(k) = \sqrt{\varphi(k)E[[A(k)]^2|Y(k),H_1(k)]}, \tag{23}$$

or, as a spectral amplitude gain

$$G_{SE-SPU}(k) = \frac{\sqrt{\varphi(k)v(k)[1+v(k)]}}{\gamma(k)}. \tag{24}$$

The *a posteriori* speech presence probability $\varphi(k)$ is given as (Ephraim and Malah, 1984)

$$\varphi(k) = \frac{\Lambda(k)}{1+\Lambda(k)}, \tag{25}$$

where $\Lambda(k)$ is the generalized speech presence ratio

$$\Lambda(k) = \frac{p(H_1(k))}{p(H_0(k))} \cdot \frac{p(Y(k)|H_0(k))}{p(Y(k)|H_1(k))} = \frac{1-q(k)}{q(k)} \cdot \frac{\exp(v(k))}{1+\xi_k}, \tag{26}$$

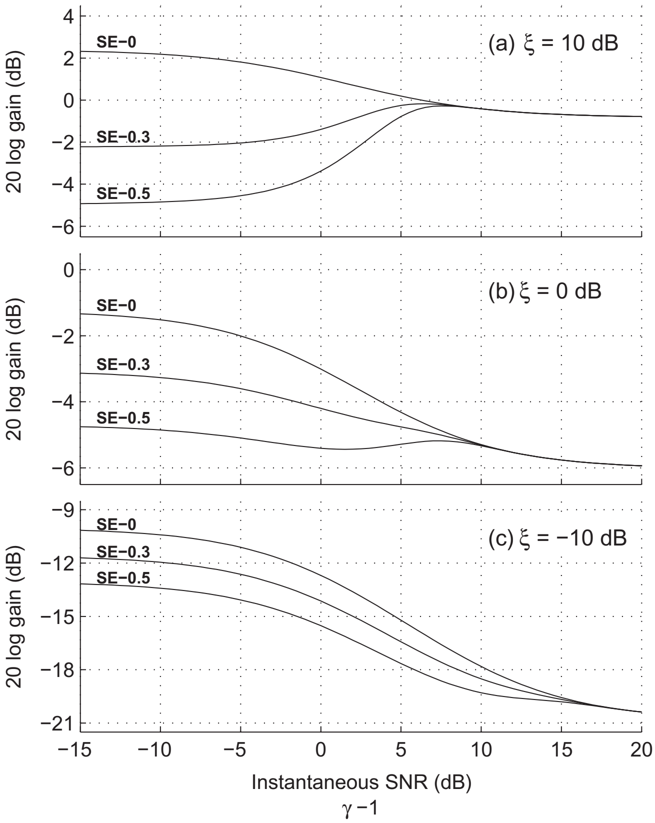

Fig. 4. Effect of SPU on the spectral energy estimator spectral amplitude gain. Three settings are tested: 0%, 30% and 50% *a priori* likelihood of speech absence. Subplots: (a) *a priori* SNR = 10 dB, (b) *a priori* SNR = 0 dB, (c) *a priori* SNR = −10 dB.

where $q(k) \triangleq p(H_0(k))$ is given as the *a priori* speech absence probability and is regarded as a tuneable parameter. Fig. 4 shows the SE spectral amplitude gain for $q(k) = 0$, $q(k) = 0.3$ and $q(k) = 0.5$. When $q(k) = 0$, the SE SPU gain simplifies to the standard SE estimator, while for $q(k) = 0.5$, the SE estimator is transformed into a very aggressive noise suppressor. It can be seen that this added aggressiveness manifests in regions where *a posteriori* SNR γ is small. However, increasing the sensitivity of spectral gain to the parameter γ can sometimes be undesirable. This is because γ has much higher estimation variance than ξ (Cappe, 1994). For these reasons, it can be useful to limit the maximum aggressiveness of the SPU.

4.2. Application of speech presence uncertainty to improve the spectral energy estimator

The net result of SPU is an additional mechanism to suppress noise-like spectral energy. From Fig. 4 we can see that increasing the value of $q(k)$ will increase the aggressiveness of the estimator. Doing so would mitigate the bias problem of the SE estimator at lower SNRs. However, this would come at the cost of speech degradation in cleaner conditions where the bias is negligible. In order to make the choice of suppression more flexible, we use a simple noise-driven heuristic to modify the *a priori* speech absence probability $q(k)$

$$q(k) = U\left[\left(1 + \kappa\sqrt{\frac{E[[A(k)]^2]}{\lambda_d(k)}}\right)^{-1}\right], \tag{27}$$

where

$$U[x] = \begin{cases} x, & \text{if } x \leqslant q_{\max}, \\ q_{\max}, & \text{otherwise.} \end{cases} \tag{28}$$

The term $E[[A(k)]^2]$ is an ensemble spectral energy average generated from a clean speech corpus[2] and κ, $q_{\max} > 0$ are tuneable parameters. The heuristic presented in (27) consists of several components: an ensemble clean speech spectral energy average, a noise power/energy estimate, and scaling parameters κ and $q_{\max}$. The ensemble energy average is introduced to bring frequency dependence to the heuristic. Effectively, its introduction increases the probability of speech presence within *speech-like* frequency regions – typically between 500 and 1000 Hz (see Fig. 5). The noise power estimate $\lambda_d(k)$ is the main component of the heuristic and is incorporated to directly address the SE estimator bias problem. As the noise power $\lambda_d(k)$ rises, the aggressiveness of the SPU is increased to compensate. This relationship is controlled further by the variables κ and $q_{\max}$.

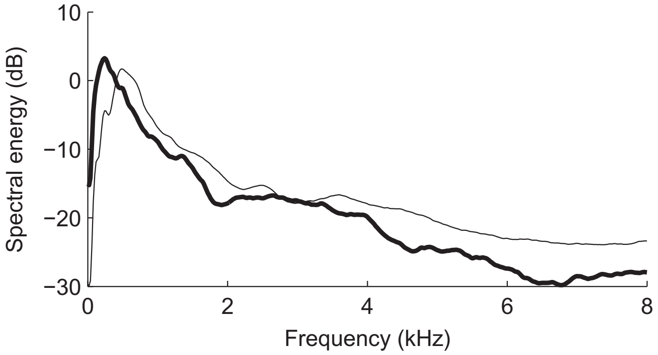

Fig. 5. Ensemble average clean speech energy (in log-spectral energy domain). $E[[A(k)]^2]$ was generated off the vowel–consonant–vowel OLLO2 training dataset (dark line), and the RM training set (thin line).

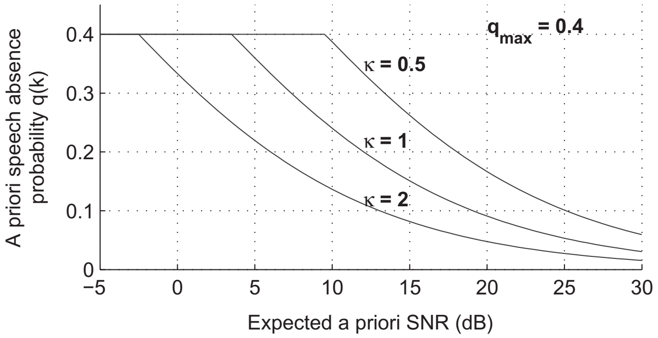

Fig. 6. Effect of parameters κ and $q_{\max}$ on the determination of *a priori* speech absence probability $q(k)$. The expected *a priori* SNR axis is given by $E[[A(k)]^2]/\lambda_d(k)$, where $E[[A(k)]^2]$ is an ensemble spectral energy average generated from a clean speech corpus (see Fig. 5). Lower values of κ increases the aggressiveness of the SPU, assigning a higher probability of speech absence at a given SNR.

Fig. 6 shows the effect of κ and $q_{\max}$ for determining the *a priori* speech absence probability $q(k)$. In low SNR conditions, the value of $q(k)$ is large. This increases the aggressiveness of the SPU in order to compensate the bias of the SE estimator and prevent over-estimation of ξ. At higher SNRs, the value of $q(k)$ decreases, eventually switching off SPU (at $q(k) = 0$) in clean conditions. This reflects the belief that in high SNR conditions there is negligible bias to compensate. The parameter κ controls the rate at which the SPU is scaled. Small values of κ increase overall SPU aggressiveness and vice-versa. The parameter $q_{\max}$ sets the maximum allowed SPU strength.

In may be noted that in (27), we have used a spectral amplitude ratio (between the ensemble speech average and noise) rather than a spectral power/energy ratio. Our choice for this was motivated empirically, rather than mathematically. In our experiments, the amplitude ratio appeared to give a good balance between noise reduction and speech degradation across a wide range of SNRs.

5. Experimental results

5.1. Enhancement system description

For our experiments, we decompose speech utterances into overlapping frames. Each analysis frame is 25 ms in

[2] All speech utterances used for experimentation are scaled/normalized, such that the maximum analysis frame energy (of a given utterance) is one.

length, and overlaps the previous analysis frame by 15 ms. Each analysis frame has a Hamming window applied before being enhanced with a given regime. Enhanced frames are then synthesized into coherent utterance with the overlap-add method (Crochiere, 1980). To derive the noise estimate $\lambda_d(m,k)$, we use a simple voice activity detector (VAD). An initial noise estimate is generated from the first 125 ms of each speech stimulus, and recursively updated. The recursive update is given as follows:

$$\lambda_d(m,k) = \eta\lambda_d(m-1,k) + (1-\eta)[R(k)]^2, \tag{29}$$

where $\eta = 0.98$ in the case that a noise-only frame has been detected and $\eta = 1$ otherwise. The *a posteriori* SNR can be then be calculated via (13). To calculate the *a priori* SNR ξ, we use the decision-directed approach covered in Section 3.2. For the proposed SE SPU estimator, the scaling factor κ must be empirically determined. To do this, we degraded several training utterances from each dataset with additive white Gaussian noise at 5 dB. The degraded sentences were enhanced with the proposed SE SPU estimator over several values of κ. The optimal value of κ was selected to minimize the word error rate (WER) using a clean speech HMM recognizer. For the following experiments, we set the heuristic parameter $q_{\max}$ to 0.4.

5.2. Automatic speech recognition system description

To test ASR performance, we use a standard MFCC feature vector in conjunction with the HTK recognition framework (Young et al., 2000). Speech utterances are first decomposed into 25 ms long frames, each shifted by 10 ms. Frames then have a Hamming window applied before MFCCs are calculated. We accumulate 26 log-filterbank energies, and retain the first 12 cepstral coefficients (not including the zeroth coefficient). For all front-end estimators, we replace the zeroth cepstral coefficient with a log-frame energy measure. For the parallel model combination method (PMC), the zeroth coefficient is retained. Given the static feature vector, we append delta and acceleration coefficients to give a 39 dimensional feature vector. For each of the front-end estimators presented, we produce MFCCs from their respective time-domain enhanced signals. For the PMC method, noisy speech MFCCs are fed into an adapted HMM back-end recognizer. An overview of the ASR system used in our experiments is given in Table 1. Training is provided by clean, unaltered utterances. We give results for the parallel model combination (PMC) back-end model adaptation method (Gales, 1995), spectral Wiener (SW), spectral subtraction (SS), MMSE log-spectral amplitude (LSA) (Ephraim and Malah, 1985), MMSE spectral amplitude (SA) (Ephraim and Malah, 1984) and MMSE spectral energy (SE) estimator. For the SE estimator, we also give results for a static $q(k) = 0.3$ SPU (SE-0.3), a data-driven SPU (SE-DD) proposed in (Malah et al., 1999) and the proposed heuristic-driven SPU (SE-prop). We conduct experiments over three speech databases, each of which covers a different ASR topology.

- *Resource management (Price et al., 1988). Continuous triphone-based recognition, medium vocabulary with structured language model.*
- *OLLO2 (Wesker et al., 2005). Single token, monophone-based recognition, medium vocabulary with no language model.*
- *Aurora2 digits (Pearce and Hirsch, 2000). Continuous word-based recognition, small vocabulary with no language model.*

5.3. Resource management word recognition

A speaker independent section of the DARPA resource management (RM) database is used for medium vocabu-

Table 1
Overview of the ASR parameters used for experimental analysis.

Parameter	Speech database		
	RM	OLLO	Aurora2
Acoustic model	3-state triphone HMMs, eight Gaussian mixtures per state	3-state monophone HMMs, 32 Gaussian mixtures per state	16-state word HMMs, three Gaussian mixtures per state
Sampling freq. (kHz)	16	16	8
Frame length (ms)	25	25	25
Frame shift (ms)	10	10	10
Analysis window	Hamming	Hamming	Hamming
No. Mel filterbanks	26	26	26
Filterbank freq. range (kHz)	0–8	0–8	0–4
Cepstral coefficients	1 through 12	1 through 12	1 through 12
Cepstral lifter factor	22	22	22
Cepstral mean subtraction	Yes	No	Yes
Appended features	Log-frame energy, Δ's, Δ^2's	Log-frame energy, Δ's, Δ^2's	Log-frame energy, Δ's, Δ^2's
Δ window (frames)	±2	±2	±3
Δ^2 window (frames)	±2	±2	±5
Feature dimension	39	39	39

lary recognition (Price et al., 1988). The database was recorded in clean conditions (sample rate of 16 kHz) and has a vocabulary of approximately 1000 words. For training, there are 3990 sentences spoken by 109 speakers. For testing, we use the February'89 test set which has 300 sentences spoken by 10 different speakers. White, Volvo and babble noises are artificially added at several SNRs. For recognition, we train triphone-level HMMs, having three states with eight Gaussian mixtures each. Cepstral mean subtraction (CMS) is applied as a standard post-processor. For the proposed SE SPU estimator, a SPU scaling factor of $\kappa = 0.5$ was used. ASR word error rate (WER) scores are given in Table 2. The standard SE estimator has good performance in light noise conditions. However it begins to struggle from 20 dB SNR onward. A static SPU ($q(k) = 0.3$) worked quite well for this dataset, giving it better overall performance than both the SA and LSA estimator. For the proposed estimator, SPU aggressiveness is increased further at low SNRs leading to an overall improvement over the static SE-SPU estimator.

Table 2
RM ASR word error rates.

Treatment	SNR (dB)				
	Clean	30	20	10	AVG
White noise					
None	4.30	5.48	11.89	47.13	17.20
PMC	4.73	5.20	9.78	34.53	13.56
SS	5.32	6.80	12.48	47.32	17.98
SA	4.26	5.04	7.43	27.02	10.94
LSA	4.42	5.36	7.55	23.39	10.18
SW	9.86	9.78	22.02	61.17	25.71
SE	4.18	5.04	9.11	37.86	14.05
SE-0.3	4.50	5.01	7.00	23.03	9.89
SE-DD	4.26	4.93	7.35	25.58	10.53
SE-prop	4.07	5.04	7.16	22.21	9.62
Babble noise					
None	4.30	4.73	8.21	38.87	14.03
PMC	4.73	4.93	7.43	21.24	9.58
SS	5.32	5.24	8.21	30.86	12.41
SA	4.26	4.89	7.78	28.90	11.46
LSA	4.42	5.08	8.56	32.58	12.66
SW	9.86	13.96	28.55	73.52	31.47
SE	4.18	4.73	8.06	31.17	12.04
SE-0.3	4.50	4.93	7.74	30.00	11.79
SE-DD	4.26	5.04	7.78	30.47	11.89
SE-prop	4.07	4.89	7.67	27.06	10.92
Volvo noise					
None	4.30	4.03	5.01	7.86	5.30
PMC	4.73	5.12	5.12	5.24	5.05
SS	5.32	4.85	5.44	7.67	5.82
SA	4.26	4.42	4.34	4.73	4.44
LSA	4.42	4.61	4.46	4.93	4.61
SW	9.86	8.21	7.12	8.64	8.46
SE	4.18	4.18	4.22	6.22	4.70
SE-0.3	4.50	4.69	4.42	4.93	4.64
SE-DD	4.26	4.54	4.26	5.04	4.53
SE-prop	4.07	4.26	4.07	4.50	4.23
Average					
None	4.30	4.75	8.37	31.29	12.18
PMC	4.73	5.08	7.44	20.34	9.40
SS	5.32	5.63	8.71	28.62	12.07
SA	4.26	4.78	6.52	20.22	8.95
LSA	4.42	5.02	6.86	20.30	9.15
SW	9.86	10.65	19.23	47.78	21.88
SE	4.18	4.65	7.13	25.08	10.26
SE-0.3	4.50	4.88	6.39	19.32	8.77
SE-DD	4.26	4.84	6.46	20.36	8.98
SE-prop	4.07	4.73	6.30	17.92	8.26

5.4. OLLO2 logatome recognition

In this section we present results for a subset of the Oldenburg logatome database (OLLO) (Wesker et al., 2005). OLLO2 is unique from the RM and Aurora2 databases in several areas. Firstly, it is not a continuous recognition task, requiring only a single logatome per speech file to be recognized. Secondly, there is very little context available for recognition. Logatomes are nonsense words, and consist of every possible vowel–consonant pairing, making the acoustic model relatively sensitive to noise.

We use the vowel–consonant–vowel stimulus for recognition – abba, adda, egge etc, for a total of 70 logatomes. Each sound file has a single spoken logatome digitized at 16 kHz and is recorded under one of several conditions (slow, fast, loud, quiet, questioning and normal). The testing and training datasets are matched in terms of regional dialect and recording conditions and consist of roughly 27,000 utterances each. We degrade the testing stimulus with (stationary) additive white and F16 noise at various SNRs. We train monophone HMMs, with three states per phoneme and 32 Gaussian mixtures per state. Because of the short-duration of the utterances, we do not apply cepstral mean subtraction. For the proposed SE SPU estimator, we set $\kappa = 0.5$. Word error rates are determined by treating the entire logatome as a word. ASR WER scores are given in Table 3.

Again, it can be seen that the standalone SE estimator performs quite well in clean and light noise environments. However, performance rapidly degrades as the SNR falls below 20 dB. The introduction of standard SPU ($q(k) = 0.3$) did make the SE estimator considerably more robust, but like the SA and LSA estimators this appeared to involve a trade-off between performance at higher and lower SNRs. For example, increasing $q(k)$ beyond 0.3 would improve noisy condition performance but would introduce speech degradation in cleaner conditions. A bigger improvement can be seen for the heuristic SE SPU. Here the aggressiveness of the SPU is scaled back at higher SNRs, limiting the amount of speech distortion. Overall, this gave the proposed SE SPU estimator superior performance compared with the other estimators.

5.5. Aurora2 digit recognition

Aurora2 is a speaker independent database for connected digit recognition (Pearce and Hirsch, 2000). Unlike the RM database, Aurora2 lacks a language model, though

Table 3
OLLO2 ASR word error rates.

Treatment	SNR (dB)				
	Clean	30	20	10	AVG
White noise					
None	16.95	35.87	79.12	96.78	57.18
PMC	16.99	20.85	34.21	57.77	32.46
SS	17.26	19.06	23.25	38.63	34.31
SA	17.19	19.74	32.84	74.73	36.13
LSA	17.84	19.30	26.42	58.57	30.53
SW	19.82	23.36	34.00	49.89	31.77
SE	17.03	23.56	56.47	89.07	46.53
SE-0.3	17.55	19.15	28.75	67.89	33.34
SE-DD	17.10	21.30	50.21	85.56	43.54
SE-prop	17.01	19.42	24.24	41.39	25.52
F16 noise					
None	16.95	19.96	29.37	65.25	32.88
PMC	16.99	18.13	23.94	40.37	24.86
SS	17.26	19.06	23.25	38.63	34.31
SA	17.19	18.03	21.79	33.86	22.72
LSA	17.84	18.12	21.89	31.80	22.42
SW	19.82	21.94	30.19	43.69	28.91
SE	17.03	18.63	24.35	42.33	25.59
SE-0.3	17.55	18.11	21.25	31.69	22.15
SE-DD	17.10	18.39	23.10	37.36	23.99
SE-prop	17.01	18.16	21.13	30.67	21.74
Average					
None	16.95	27.92	54.25	81.02	45.04
PMC	16.99	19.49	29.08	49.07	28.66
SS	17.26	19.06	23.25	38.63	34.31
SA	17.19	18.89	27.32	54.30	29.42
LSA	17.84	18.71	24.16	45.19	26.48
SW	19.82	22.65	32.10	46.79	30.34
SE	17.03	21.10	40.41	65.70	36.06
SE-0.3	17.55	18.63	25.00	49.79	27.74
SE-DD	17.10	19.85	36.66	61.46	33.77
SE-prop	17.01	18.79	22.69	36.03	23.63

Table 4
Aurora2A ASR word error rates.

Treatment	SNR (dB)					
	20	15	10	5	0	AVG
Subway noise						
None	2.95	5.59	12.34	30.49	70.56	24.39
PMC	1.90	4.45	10.78	26.62	60.88	20.93
SS	10.75	19.83	32.61	53.48	74.73	38.28
SA	3.07	4.54	9.12	17.68	39.70	14.82
LSA	3.90	5.68	11.30	20.94	42.95	16.95
SW	31.56	38.75	47.80	59.66	78.85	51.32
SE	2.98	4.94	9.58	19.62	45.87	16.60
SE-0.3	3.13	5.04	10.04	18.91	41.05	15.63
SE-DD	3.07	4.91	9.27	19.07	41.94	15.65
SE-prop	3.10	4.57	8.69	17.32	40.59	14.85
Babble noise						
None	2.03	4.63	16.51	53.02	90.84	33.41
PMC	2.27	3.99	10.13	30.96	67.62	22.99
SS	23.31	32.59	46.67	63.69	81.74	49.60
SA	3.42	7.56	17.23	38.33	68.80	27.07
LSA	8.86	15.57	26.57	45.77	70.95	33.54
SW	39.60	46.58	58.43	74.61	87.30	61.30
SE	1.93	4.05	11.12	31.32	67.96	23.28
SE-0.3	6.05	11.88	22.76	43.11	70.28	30.82
SE-DD	2.93	6.92	16.63	37.82	69.56	26.77
SE-prop	2.06	4.44	13.51	35.43	68.35	24.76
Car noise						
None	2.09	4.38	11.39	34.39	81.12	26.67
PMC	2.09	3.82	12.02	37.46	75.51	26.18
SS	7.34	14.35	28.27	50.22	72.17	34.47
SA	1.58	2.80	5.31	14.08	34.30	11.61
LSA	1.70	3.40	6.53	16.10	38.98	13.34
SW	25.14	32.09	41.87	60.93	83.54	48.71
SE	1.70	3.04	6.53	17.54	46.64	15.09
SE-0.3	1.37	2.86	5.25	14.02	34.33	11.57
SE-DD	1.64	2.92	6.05	14.79	36.74	12.43
SE-prop	1.76	2.95	6.35	14.43	35.07	12.11
Exhibition noise						
None	3.73	6.97	15.21	38.88	82.14	29.39
PMC	2.41	4.57	12.53	31.22	64.79	23.10
SS	9.69	19.32	35.27	55.14	77.54	39.39
SA	3.70	6.17	13.45	28.51	51.96	20.76
LSA	6.76	9.97	18.82	35.51	57.76	25.76
SW	40.39	48.63	61.28	75.87	86.27	62.49
SE	3.15	5.49	11.88	25.67	54.27	20.09
SE-0.3	4.75	7.56	15.49	31.78	54.92	22.90
SE-DD	3.58	6.33	13.45	29.44	54.12	21.38
SE-prop	3.15	5.58	11.79	26.75	53.69	20.19
Set A averages						
None	2.70	5.39	13.86	39.20	81.17	28.47
PMC	2.17	4.21	11.37	31.57	67.20	23.30
SS	12.77	21.52	35.71	55.63	76.55	40.44
SA	2.94	5.27	11.28	24.65	48.69	18.57
LSA	5.31	8.66	15.81	29.58	52.66	22.40
SW	34.17	41.51	52.34	67.77	83.99	55.96
SE	2.44	4.38	9.78	23.54	53.69	18.77
SE-0.3	3.83	6.84	13.39	26.96	50.15	20.23
SE-DD	2.80	5.27	11.35	25.28	50.59	19.06
SE-prop	2.52	4.39	10.09	23.48	49.43	17.98

its acoustic models are relatively sparse. Spoken digits in the database consist of zero through nine as well as ‘oh’, giving a vocabulary size of 11. Testing and training utterances were downsampled to 8 kHz and filtered with G712 characteristics. Finally, utterances have had noise artificially added at several SNRs. Word-level HMMs are built, each with 16 states and three Gaussian mixtures per state. CMS was applied as a standard post-processor. For the proposed SE SPU estimator we used a scaling factor of $\kappa = 5$. ASR WER scores are given in Tables 4 and 5 for recognition tasks A and B, respectively.

The Aurora2 recognition task was particularly sensitive to speech distortion. This was immediately apparent when determining the SPU scaling parameter κ. For the Aurora2 task, κ was found to be much higher than both the RM and OLLO2 tasks (5 versus 0.5). As a result, the SPU addition played a minor role for the Aurora2 task. Here, the SE and SE SPU estimators performed similarly, only showing significant divergence at the 0 dB noise level. The use of a white noise development set to train the heuristic SPU worked well for most noise environments. Notable exceptions were the restaurant, airport and babble Aurora2 noise tasks, where the heuristic SPU (trained on white noise) was too aggressive. However this may also reflect the limitations of using VAD to track non-stationary noises.

Table 5
Aurora2B ASR word error rates.

Treatment	SNR (dB)					
	20	15	10	5	0	AVG
Restaurant noise						
None	2.33	4.70	16.18	45.96	78.97	29.63
PMC	1.93	3.25	7.74	24.65	60.06	19.53
SS	22.51	32.08	45.47	62.27	79.61	48.39
SA	5.13	11.30	23.64	42.03	67.58	29.94
LSA	11.45	18.02	31.01	47.80	71.45	35.95
SW	37.80	46.85	57.48	72.89	86.83	60.37
SE	2.64	5.74	13.94	35.55	65.61	24.70
SE-0.3	8.66	15.17	28.03	45.41	69.42	33.34
SE-DD	4.33	9.27	21.86	41.57	67.73	28.95
SE-prop	2.86	6.17	16.21	38.62	66.84	26.14
Street noise						
None	2.60	5.11	13.45	37.64	75.09	26.78
PMC	2.39	4.96	13.36	38.94	72.22	26.37
SS	12.97	24.06	36.55	57.35	76.42	41.47
SA	3.33	5.20	9.95	22.61	47.85	17.79
LSA	4.41	7.65	13.48	28.96	54.38	21.78
SW	32.83	40.08	51.57	67.68	85.07	55.45
SE	2.36	4.17	8.71	23.67	53.75	18.53
SE-0.3	3.75	5.89	11.34	24.97	49.79	19.15
SE-DD	3.48	5.32	10.10	24.67	50.85	18.88
SE-prop	2.54	4.72	8.80	23.07	48.70	17.57
Airport noise						
None	1.79	4.09	11.72	37.16	73.90	25.73
PMC	1.55	3.07	7.31	26.16	60.45	19.71
SS	15.96	24.49	38.20	57.23	74.62	42.10
SA	2.54	5.91	13.09	29.76	53.80	21.02
LSA	5.13	9.54	18.70	35.67	58.75	25.56
SW	33.97	39.43	50.91	68.83	84.40	55.51
SE	1.76	3.34	9.54	24.72	54.91	18.85
SE-0.3	4.00	7.72	16.28	34.09	56.49	23.72
SE-DD	2.51	5.64	13.36	31.20	55.95	21.73
SE-prop	1.76	3.67	10.53	27.02	53.12	19.22
Train noise						
None	1.54	4.32	11.66	34.74	80.31	26.51
PMC	1.36	3.21	9.94	33.42	69.85	23.56
SS	9.81	17.59	30.85	51.47	72.88	36.52
SA	2.93	4.75	8.05	20.98	40.76	15.49
LSA	4.17	6.39	11.08	24.07	44.99	18.14
SW	28.36	34.37	45.97	64.79	84.94	51.69
SE	1.82	3.89	7.28	20.92	49.40	16.66
SE-0.3	3.67	5.37	9.04	21.88	41.56	16.30
SE-DD	2.62	4.47	8.08	21.51	43.51	16.04
SE-prop	2.07	4.04	7.31	20.12	41.93	15.09
Set B averages						
None	2.07	4.56	13.25	38.88	77.07	27.16
PMC	1.81	3.62	9.59	30.79	65.65	22.29
SS	15.31	24.56	37.77	57.08	75.88	42.12
SA	3.48	6.79	13.68	28.85	52.50	21.06
LSA	6.29	10.40	18.57	34.13	57.39	25.36
SW	33.24	40.18	51.48	68.55	85.31	55.76
SE	2.15	4.29	9.87	26.22	55.92	19.69
SE-0.3	5.02	8.54	16.17	31.59	54.32	23.13
SE-DD	3.24	6.17	13.35	29.74	54.51	21.40
SE-prop	2.31	4.65	10.71	27.21	52.65	19.51

6. Conclusion

In this paper we have investigated the use of the spectral energy estimator for use in robust ASR. Traditionally, the spectral energy estimator has suffered from the problem of residual noise. In order to improve the SE estimator for use in robust ASR, we identified the causes of the residual noise. These problems were then addressed with a simple, heuristic based SPU. Experimental results show a significant improvement in robustness, over both the baseline results and the more common log-spectral amplitude estimator. The improvement gained by the heuristic SPU is especially evident at lower SNRs, where the standalone SE estimator typically struggles.

References

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: IEEE Internat. Conf. on Acoust. Speech Signal Process., ICASSP, pp. 208–211.

Boll, S., 1979. Suppression of acoustic noise in speech using spectral subtraction. IEEE Trans. Acoust. Speech Signal Process. ASSP-27, 113–120.

Cappe, O., 1994. Elimination of the musical noise phenomenon with the Ephraim and Malah noise suppressor. IEEE Trans. Speech Audio Process. 2, 345–349.

Crochiere, R., 1980. A weighted overlap-add method of short-time Fourier analysis/synthesis. IEEE Trans. Acoust. Speech Signal Process. ASSP-28, 99–102.

Ephraim, Y., Malah, D., 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 32, 1109–1121.

Ephraim, Y., Malah, D., 1985. Speech enhancement using a minimum mean-square error log-spectral amplitude estimator. IEEE Trans. Acoust. Speech Signal Process. 33, 443–445.

Ephraim, Y., Trees, H.V., 1991. Constrained iterative speech enhancement with application to speech recognition. IEEE Trans. Signal Process. 39, 795–805.

Ephraim, Y., Trees, H.V., 1995. A signal subspace approach for speech enhancement. IEEE Trans. Speech Audio Process. 3, 251–266.

Fujimoto, M., Ariki, Y., 2000. Noisy speech recognition using noise reduction method based on Kalman filter. In: IEEE Internat. Conf. on Acoust. Speech Signal Process., Vol. 3, pp. 1727–1730.

Gales, L.F.J., 1995. Model-Based Techniques For Robust Speech Recognition. Ph.D. Thesis, University of Cambridge, UK.

Gemello, R., Mana, F., Mori, R., 2006. Automatic speech recognition with a modified Ephraim–Malah rule. IEEE Signal Process. Lett. 13, 56–59.

Hermus, K., Wambacq, P., Hamme, H.V., 2007. A review of signal subspace speech enhancement and its application to noise robust speech recognition. EURASIP J. Appl. Signal Process. 2007, 195–197.

Huang, X., Acero, A., Hon, H.W., 2001. Spoken Language Processing. Prentice Hall.

Lathoud, G., Magimai-Doss, M., Mesot, B., Bourlard, H., 2005. Unsupervised spectral subtraction for noise-robust ASR. In: Proc. 2005 IEEE ASRU Workshop, pp. 343–348.

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

Malah, D., Cox, R.V., Accardi, A.J., 1999. Tracking speech-presence uncertainty to improve speech enhancement in non-stationary noise environments. In: IEEE Internat. Conf. on Acoust. Speech Signal Process., pp. 789–792.

McAulay, R., Malpass, M., 1980. Speech enhancement using a soft-decision noise suppression filter. IEEE Trans. Acoust. Speech Signal Process. 28, 137–145.

Paliwal, K., Basu, A., 1987. A speech enhancement method based on Kalman filtering. In: IEEE Internat. Conf. on Acoust. Speech Signal Process., ICASSP, pp. 297–300.

Pearce, D., Hirsch, H.G., 2000. The aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In: ISCA ITRW ASR, pp. 29–32.

Price, P., Fisher, W., Bernstein, J., Pallett, D., 1988. The DARPA 1000-word resource management database for continuous speech recognition. In: IEEE Internat. Conf. on Acoust. Speech Signal Process., pp. 651–654.

Wesker, T., Meyer, B., Wagener, K., Anemüller, J., Mertins, A., Kollmeier, B., 2005. Oldenburg logatome speech corpus (OLLO) for speech recognition. In: Proc. Interspeech, pp. 1273–1276.

Wiener, N., 1949. The Extrapolation, Interpolation, and Smoothing of Stationary Time Series with Engineering Applications. Wiley, New York.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2000. The HTK Book Version 3.0. Cambridge University Press.