Available online at www.sciencedirect.com
SCIENCE DIRECT
Speech Communication 48 (2006) 96–109
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Speech enhancement in nonstationary noise environments using noise properties

Kotta Manohar, Preeti Rao *

Department of Electrical Engineering, Indian Institute of Technology, Powai, Bombay 400 076, India

Received 5 September 2004; received in revised form 17 August 2005; accepted 19 August 2005

Abstract

Traditional short-time spectral attenuation (STSA) speech enhancement algorithms are ineffective in the presence of highly nonstationary noise due to difficulties in the accurate estimation of the local noise spectrum. With a view to improve the speech quality in the presence of random noise bursts, characteristic of many environmental sounds, a simple post-processing scheme is proposed that can be applied to the output of an STSA speech enhancement algorithm. The post-processing algorithm is based on using spectral properties of the noise in order to detect noisy time–frequency regions which are then attenuated using a SNR-based rule. A suitable suppression rule is developed that is applied to the detected noisy regions so as to achieve significant reduction of noise with minimal speech distortion. The post-processing method is evaluated in the context of two well-known STSA speech enhancement algorithms and experimental results demonstrating improved speech quality are presented for a data set of real noise samples.

Keywords: Speech enhancement; Nonstationary noise; Spectral subtraction

1. Introduction

The problem of enhancing speech degraded by additive background noise, when only a single channel is available, remains challenging due to the limitations of existing methods in difficult though realistic noise conditions. Single-channel speech enhancement algorithms are generally based on short-time spectral attenuation (STSA). Widely researched and applied examples of STSA speech enhancement are spectral subtraction as originally proposed by Berouti et al. (1979) and the Ephraim–Malah MMSE short-time spectral amplitude estimator (Ephraim and Malah, 1984). These methods can be viewed in terms of applying a spectral gain to each frequency bin in a short-time frame of the noisy speech signal. Since the spectral components are assumed to be statistically independent, the gain is adjusted individually as a function of the relative local SNR at each frequency. With low SNR regions attenuated relative to high SNR regions, an improvement in the overall SNR is achieved. A good estimate of the instantaneous noise spectrum is crucial in the estimation of the local SNR, without which quality would degrade

* Corresponding author. Tel.: +91 2225720083.
E-mail addresses: manohar@ee.iitb.ac.in (K. Manohar), prao@ee.iitb.ac.in (P. Rao).

0167-6393/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2005.08.002

due to the presence of either high residual noise or high speech distortion in the enhanced speech.

A common method of noise estimation involves the use of a voice activity detector (VAD) to detect pauses in speech. The noise estimate is then obtained by a recursively smoothened adaptation of noise during the detected pauses. In stationary background noise, such an estimator is generally reliable. However nonstationary noises, with noise spectrum levels changing in time, cannot be tracked adequately by a recursive noise estimation method that adapts only during detected speech pauses. This is especially true of environmental noises such as factory or battlefield noise that are characterized by large, irregular random bursts embedded in a relatively stationary background. Even if the VAD is reliable (which is unlikely at low SNRs and in nonstationary noise), changes in the noise spectrum occurring during active speech cannot influence the noise estimate in a timely manner. Due to the difficulty in tracking highly nonstationary noise spectra, STSA-based algorithms are effective only in suppressing the stationary noise component generally leaving noise bursts unattenuated in the enhanced speech.

In this paper, we focus on the extreme but common form of nonstationary noise, one with randomly occurring high-energy noise bursts embedded in a stationary background. In the next section, we discuss the shortcomings of available methods in dealing with such noise. We propose next a method which exploits known differences in the spectro-temporal properties of noise and speech to selectively attenuate noisy time–frequency regions remaining in STSA-enhanced signals. This post-processing method is evaluated in the context of two well-known STSA speech enhancement algorithms and experimental results on speech quality are presented for a data set of real noise samples.

2. Suppressing nonstationary noise

Realizing the limitation of traditional STSA speech enhancement methods in the presence of nonstationary noise, research efforts have been directed over the past decade to devise new solutions. The proposed solutions generally fall into two categories, namely, improvements to the noise estimator and modifications of the suppression rule. While the former class of methods essentially targets the limitation of VAD-based noise estimation, the suppression rule modifications are based on some prior knowledge of the speech signal. In this section we provide a brief overview of some of these methods with a discussion of their shortcomings.

As discussed in Section 1, accurate estimation of the instantaneous noise spectrum would make for an effective STSA speech enhancement algorithm in any background noise condition. Hence much research has been focused on improving the noise estimation block. Recognizing the limitations of VAD accuracy in low SNR and varying background noise conditions, a number of methods for noise spectrum estimation without explicit speech pause detection have been proposed (Hirsch and Ehrlicher, 1995; Stahl et al., 2000; Martin, 2001). These methods are based on tracking some statistic (e.g. minimum or median) of past power spectral values for each frequency bin over several frames. However the buffer length necessary to bridge peaks of speech activity makes it difficult to follow any rapid variations in noise spectrum. Fig. 1 illustrates the limitation of this approach for speech in nonstationary noise characterized by short-duration noise bursts. The true instantaneous noise power in a frequency bin near 800 Hz is compared with the noise power estimated using a VAD-based estimator and using the QBNE (quantile based noise estimation, Stahl et al., 2000) method. In the QBNE method a buffer of 0.64 s duration was employed with quantile value 0.5. The test file has a continuous stretch of speech frames as indicated by the corresponding high/low pulse in the plot. Factory noise is nonstationary in nature having stationary noise background with occasional random bursts to which the sudden peaks in the instantaneous noise power spectra of Fig. 1 can be attributed; frames where the 'random bursts' of noise have been detected manually are indicated by a high/low pulse in the plot. We observe that the VAD estimator tracks the noise burst level only when speech is absent. The QBNE estimator, on the other hand, responds to the noise burst only approximately and with a delay. That such direct estimation methods for noise fail in conditions such as factory noise has also been noted by Ris and Dupont (2001).

A different approach to carry out the adaptation of noise during both speech absence and presence is via a speech absence probability based on an estimate of SNR (Malah et al., 1999). However any sudden increase in the background noise level is not easily distinguished from speech and results in high estimated SNR making the method relatively less effective in highly nonstationary noise. Also

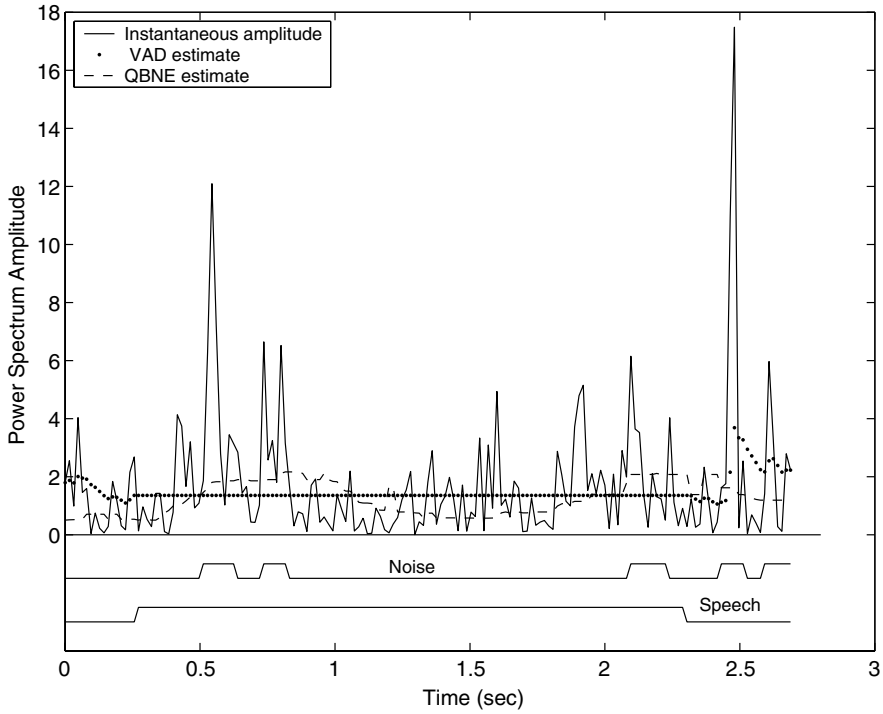

Fig. 1. Instantaneous and estimated noise power spectrum values for a frequency bin at 800 Hz in the power spectrum of factory noise-corrupted speech.

based on the concept of speech absence probability is work of Cohen (2003) wherein the noise estimate is obtained by averaging past spectral power values, using a time–varying frequency-dependent smoothing parameter that is controlled by the signal presence probability. The speech presence probability is determined by observing the minima of a smoothed periodogram. However, the only nonstationary noise for which performance results are reported is white Gaussian noise with level increasing at 2 dB/s which is a poor approximation to the fast modulations of the bursty noises under consideration here. Besides, as mentioned earlier, no direct estimation methods can track highly nonstationary noises accurately even if the noise estimate is updated in every frame. To compensate for inaccuracies in the estimation of background noise, Malah et al. (1999) suggest a multiplicative modifier to the STSA enhancement gain function based on the a priori probability of speech absence in each spectral component of the noisy speech. However, the a priori probability is a function of estimated SNR and hence is not accurate enough when the noise is highly nonstationary and of high energy.

Acknowledging the fact that in spite of applying speech enhancement methods such as spectral subtraction, certain spectro-temporal regions will remain noisy, Cooke et al. (2001) propose missing data methods for robust ASR. A two-stage approach is followed in which spectral subtraction is employed to suppress the stationary noise component and then the recognition processor is conditioned on the estimated reliability of spectro-temporal regions of the signal as determined by various speech spectrum cues. The authors have commented upon the difficulty of detecting unreliable regions when the nonstationary noise component is intermittent and impulsive as is the case with factory noise. More importantly, the missing data approach cannot be easily extended to enhancement in noise for general speech perception. A similar concept applicable to speech enhancement is the use of statistical models of clean speech (Attias et al., 2001; Yao and Lee, 2004) or trained codebooks where a priori information in the form of spectral envelope shapes is stored for both speech and noise (Srinivasan et al., 2003). A joint or iterative optimization over assumed speech and noise models is carried out for each frame of noisy speech to determine the noise estimate which is then used in a noise reduction algorithm. Apart from the computational complexity of such methods, their

performance would be expected to depend critically on a good match between training and actual usage conditions.

The work presented in this paper is targeted towards a robust algorithm for the suppression of random noise bursts with minimal speech distortion. It is motivated by the possibility of using available knowledge to distinguish between speech and noise in order to identify, and further attenuate, unreliable spectro-temporal regions in signals enhanced by traditional STSA based on a simple noise estimator. To achieve improved speech quality using this approach requires solutions to two problems: (1) determining reliable cues for identifying noisy spectro-temporal regions, and (2) finding a suitable suppression rule applicable to the detected noisy regions so as to achieve significant reduction of noise with minimal speech distortion.

3. Proposed post-processing algorithm

As mentioned earlier, we focus on a common form of nonstationary noise characterized by high-energy randomly occurring noise bursts in a stationary background. Such noise bursts typically lead to the partial or complete corruption of the short-term spectrum of speech at their instants of occurrence. A traditional STSA algorithm based on a simple noise estimator, such as discussed in the previous section, can effectively suppress only the stationary background noise leaving mainly the speech together with residual noise bursts in the enhanced signal. The proposed post-processing algorithm involves identifying regions in the spectrogram of the STSA-enhanced speech that are dominated by the residual noise. These regions are selectively attenuated further with the goal to improve the overall quality of the enhanced speech. The post-processed, enhanced speech is then reconstructed from the resulting magnitude spectrum and the original noisy speech phase spectrum as in traditional STSA enhancement.

Although it would be ideal to have speech/noise classification available at the time–frequency bin level for the purpose of selective suppression of noise, finding reliable cues to achieve this is difficult. We instead attempt to identify time–frequency regions at the coarser level of broad frequency bands in each time frame that are dominated by noise similar in approach to the missing features methods for speech recognition (Seltzer et al., 2000; Cooke et al., 2001). The post-processing scheme thus comprises the following steps:

1. Divide the spectrum of each frame of the STSA-enhanced speech into several, possibly overlapping, frequency bands in view of the fact that the noise spectrum may be localized in frequency.
2. Carry out speech/noise classification to detect frequency bands that are dominated by residual noise. Possible features to effect this classification are discussed in the next section.
3. Using a suitable suppression rule, attenuate the spectral values in the identified noisy bands.

The suppression rule should ideally depend on the bin SNR in a manner as to apply more attenuation in low SNR regions. This would help to minimize speech distortion while achieving an overall improvement in the SNR. The definition of estimated SNR is crucial in the design of the suppression rule. Since the "noise" estimated by the STSA noise estimator is essentially the stationary background component, an increase in SNR would be linked to a local increase in the energy either due to speech or from a noise burst. Thus noise bursts too would give rise to high a posteriori SNR and therefore high spectral gains.

However if the identification of noisy frequency bands in Step 2 is reasonably reliable, a local SNR increase in an identified nonspeech bin would signal the onset of a noise burst. Given this, an appropriate definition for the estimated SNR is given by the "average a priori SNR" computed as in (1)

$$\xi(k) \approx (1-\eta)\underbrace{\frac{|S(k)|^2_{\text{est}}}{|\widehat{D}(k)|^2}}_{\approx \text{ current SNR}} + \eta \underbrace{\frac{|\widehat{S}(k)|^2_{\text{prev}}}{|\widehat{D}(k)|^2_{\text{prev}}}}_{\text{previous SNR}} \tag{1}$$

where $|S(k)|^2_{\text{est}} = \text{Max}(|Y(k)|^2 - |\widehat{D}(k)|^2, 0)$ and 'η' is a smoothing constant fixed at '0.99'. $Y(k)$ is the noisy speech spectrum and while $|\widehat{S}(k)|^2_{\text{prev}}$ refers to the final enhanced speech power spectrum of previous frame (obtained by STSA enhancement and the post-processing combined), $|\widehat{D}(k)|^2$ refers to the average noise power spectrum estimate as obtained from the noise estimator of the STSA. This decision-directed, averaging formulation of SNR is the same as that used in the gain function of the Ephraim and Malah (1984) STSA algorithm. However, there it is motivated by the ability to suppress random musical tones, an annoying artifact of spectral subtraction processing based on instantaneous estimated SNR. In the present problem, the advantage of using the "average a priori" form of SNR definition is that it serves as a good spectral cue

for crude classification of speech/nonspeech bins in an identified "noisy" band. Whenever a noise burst occurs during speech, the $\xi(k)$ value for a nonspeech bin (such as inter-harmonic bins) would be low. This would hold for the first several frames after the onset of the noise burst because of the recursive smoothing in its computation. Thus nonspeech bins would be attenuated more as compared to the speech bins in an identified noisy band leading to effective noise suppression with relatively low speech spectrum distortion. This aspect is especially useful in the case of noises where the noise bursts are of short duration.

A functional form of the suppression rule that was found to work well in practice is to vary the attenuation linearly with the estimated SNR in dB (similar to Berouti spectral subtraction (Berouti et al., 1979)). The *attenuation factor* $\lambda(k)$ is varied linearly with the estimated a priori SNR $\xi(k)$ in dB but restricted to the range of 0.05–0.9.

$$\lambda(k) = \begin{cases} 0.05 & \xi(k) < SNR\_low \\ f_0 + s \times \xi(k) & SNR\_low \leqslant \xi(k) \leqslant SNR\_high \\ 0.9 & \xi(k) > SNR\_high \end{cases} \tag{2}$$

In (2) f_0 is the value at 0 dB SNR and 's' is the slope of the line (whose values depend on the parameters '*SNR_low*' and '*SNR_high*'). The suppression rate can be controlled by varying the parameters '*SNR_low*' and '*SNR_high*'. After obtaining the attenuation factors, we recalculate the speech estimate as in (3) of an ith 'noisy band' limiting the values to a spectral floor; β is the spectral floor gain parameter.

$$\begin{aligned} |\widehat{S}_i(k)|^2_{\text{final}} &= \lambda(k) * |\widehat{S}_i(k)|^2_{\text{STSA}}, \\ &\quad \text{if } |\widehat{S}_i(k)|^2_{\text{final}} > \beta|\widehat{D}_i(k)|^2; \\ &= \beta|\widehat{D}_i(k)|^2 \quad \text{otherwise} \end{aligned} \tag{3}$$

We note here that since overlapping of bands is allowed, a particular frequency bin may fall in more than one band. In this case, identification as "noisy" in at least one band suffices to label the frequency bin as falling in a noisy band.

4. Features for detection of noise-dominated regions

As discussed in Section 3, it is required to identify time–frequency bins in the STSA-enhanced speech that are dominated by residual noise. The available estimate of the stationary background noise spectrum can be used to locate regions of energy level higher than that of the background. The higher energy of these regions may be due to either (1) speech or (2) a high-energy nonstationary noise component. Since it is not possible to distinguish the two possibilities based on instantaneous energy alone, we turn to alternative features for speech–noise discrimination.

Voice activity detectors often employ time-domain features based on indications of voicing such as zero-crossing count or autocorrelation peak level (Itoh and Mizushima, 1997). However these cannot be applied to detect noise in localized frequency regions. Differences in spectral characteristics could be more useful, with residual noise in the enhanced speech showing nearly flat spectra over extended frequency regions. On the other hand, at low SNRs only the resolved harmonics in the noisy speech spectrum which are not masked by noise can be reconstructed on applying STSA enhancement; unvoiced speech and unvoiced high-frequency bands which have low energy will be suppressed and hence the enhanced speech in the absence of residual noise is mostly harmonic in nature. The relatively flat spectral structure of noise-dominated regions can be captured by the measures, discussed below, that quantify "spectral flatness".

4.1. Spectral flatness based classifiers

Based on the assumption that the STSA-enhanced speech contains primarily harmonic speech and frequency-localized noise bursts, we investigate various implementations of spectral flatness measurement for the detection of noise-dominated regions at the frequency band level. Let $X[k]$ denote the magnitude spectrum values computed via a DFT. The ith frequency band comprises L frequency bins with bin index k in the range $[b_i, e_i]$. For instance, with a 256-point DFT at sampling frequency of 8 kHz, the 0–1 kHz band will be bounded by the bin indices: $b_i = 0$ and $e_i = 31$. We note here that the frequency bands must be wide enough to provide the needed averaging in the estimation but may also take into account the noise frequency-domain structure. The measures investigated are:

(1) *SFM* (*spectral flatness measure*): It is defined as the ratio of the geometric mean to the arithmetic mean of the magnitude spectrum values as given in Eq. (4) (Johnston, 1988).

$$\mathrm{SFM}_i = \frac{\left(\prod_{k=b_i}^{e_i} X[k]\right)^{1/L}}{\frac{1}{L}\sum_{k=b_i}^{e_i} X[k]} \tag{4}$$

The SFM lies in the range $[0,1]$ taking low values for harmonic regions representing speech, and high values for noise-dominated regions which have a relatively flat spectrum.

(2) *Energy-normalized variance*: The harmonic structure or deviation from flatness of the spectrum in any chosen frequency band is reflected in the energy-normalized variance of the spectral values computed as

$$n\_var_i = \frac{\sum_{k=b_i}^{e_i} (X[k] - \overline{X}_i)^2}{\sum_{k=b_i}^{e_i} (X[k])^2} \tag{5}$$

where $\overline{X}_i$ is the mean of the magnitude spectrum values ($X[k]$) within the band i. Eq. (5) is expected to take high values for harmonic regions representing speech, and low values for noise-dominated regions, with the range of values spanning $[0,1]$.

(3) *Entropy*: A related measure is "entropy" as used in the VAD of Renevey and Drygajlo (2001) on the assumption that the signal spectrum is more organized during speech segments than during noise segments. It can be redefined for a frequency band as

$$H_i = -\frac{1}{\log(L)} \sum_{k=b_i}^{e_i} P(|X(k)|^2) \log(P(|X(k)|^2)) \tag{6}$$

where $P(|X(k)|^2) = \frac{|X(k)|^2}{\sum_{k=b_i}^{e_i} |X(k)|^2}$ is the "*probability*" of the frequency bin 'k'. H takes maximum value of '1' when the signal is a white noise, and minimum value of '0' when it is a pure tone (sinusoid). Hence, the entropy based method is well suited for speech detection in white or quasi-white noises.

4.2. Experimental comparison of classifiers

A comparative evaluation of the different classifiers can be achieved by experimental observations in a typical application situation, i.e. by comparing the receiver operating characteristics (ROC), or the hit rate versus false-alarm rate plots, for noisy band detection in a typical noise condition. By changing the threshold criterion in the algorithm's decision rules, false-alarm rates can be decreased at the cost of a decrease in the hit rates. A better classifier would be characterized by a lower false-alarm rate for a given hit rate. While 'hit-rate' for a particular band refers to the fraction of all the actual 'noisy bands' (i.e. noise-burst time–frequency regions) that are correctly detected as 'noisy band', 'false-alarm rate' refers to the number of times 'speech bands with no noise burst' are erroneously detected as 'noisy band', as a fraction of the total speech bands. The actual noisy bands are determined from the pure noise files.

Berouti spectral subtraction (BSS) (Berouti et al., 1979) is used as front-end STSA algorithm to obtain the enhanced speech estimate as given in Eq. (7)

$$\begin{aligned} |\widehat{S}(k)|^2_{\mathrm{STSA}} &= |Y(k)|^2 - \alpha|\widehat{D}(k)|^2, \\ &\quad \text{if } |\widehat{S}(k)|^2_{\mathrm{STSA}} > \beta|\widehat{D}(k)|^2; \\ &= \beta|\widehat{D}(k)|^2 \quad \text{otherwise} \end{aligned} \tag{7}$$

where α is the oversubtraction factor which is a linear function of segmental a posteriori SNR. The parameter α_0 (value of α at 0 dB SNR) is set to equal to 5 and spectral floor gain parameter β in Eqs. (3), (7) is chosen as 0.01. The noise estimate is updated during the silence frames by using an averaging rule. For these experiments, an ideal VAD is assumed (i.e. frames are identified as speech or silence manually).

Fig. 2 shows the ROC curves for SFM, energy-normalized variance and entropy based features in the case of an 8 kHz sampled speech test file comprising of eight concatenated sentences from the TIMIT database (Fisher et al., 1986) with factory noise added to achieve an SNR of 0 dB. Next the noisy band detection algorithms based on each feature are applied to the enhanced speech by dividing the spectrum into four nonoverlapping frequency bands each of 1 kHz width. The hit rate and false-alarm rate are calculated for each band separately and the exercise is repeated by varying the thresholds to obtain the ROC plots for the three features. The steepness or slope of the ROC curves determines the suitability of the feature in terms of providing an adequate level of discrimination between speech and noise. We note that all the three features examined have fairly steep ROCs and thus can be considered to exhibit discriminability suitable for the speech/noise classification at the band level. However since the ROC curve of SFM falls below the ROC curves of *entropy* and *n_var* (energy-normalized variance) in a large portion of the operating region, *n_var* and *entropy* can be considered to provide slightly better discriminability than the *SFM*. Further, energy-normalized variance exhibits slightly better ROC in bands 2 and 3. Hence in our post-processing scheme we chose

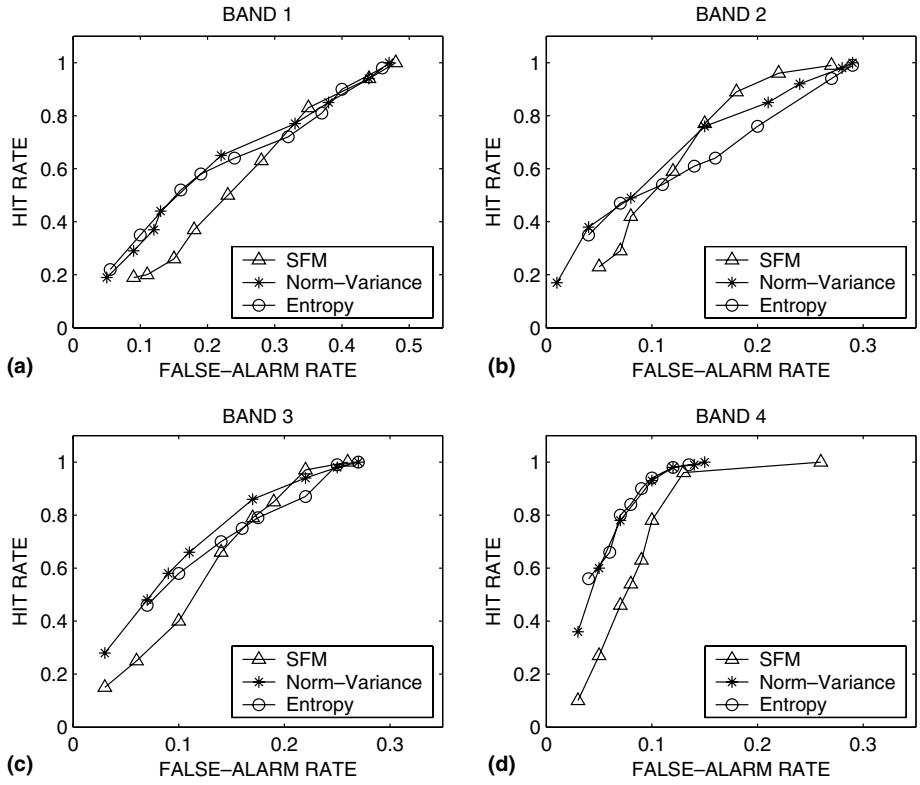

Fig. 2. ROC plots of the energy-normalized variance, SFM and entropy in the detection of noisy regions for factory noise-corrupted speech at 0 dB SNR.

energy-normalized variance as the feature for speech/noise classification based on applying a threshold to n_var_i in (5) to identify bands below the threshold as noise dominated. We observe from Fig. 2 that the ROC curve for band 1 (note the difference in the x-axis scaling for band 1) is worse (i.e. low hit rate for a given false-alarm rate) than that for the other bands for all features; this is explained by the fact that speech energy occupies largely the frequency range of 0–1 kHz with typically strong harmonic structure. Therefore, for the residual noise to influence the classification feature, the noise level must be particularly high.

We note here that the proposed noise detection method is based on the assumption that enhanced speech in the absence of residual noise is chiefly harmonic. At high SNRs however, unvoiced speech or unvoiced high-frequency energy may also be restored by STSA enhancement invalidating the harmonicity assumption. To avoid the needless suppression of speech in such circumstances, an automatic post-processing switch off/on feature may be incorporated. It is found adequate to observe an estimate of the frame-level a posteriori SNR (based on the noisy speech energy and average noise energy as determined by the STSA noise estimator) over a duration of 1 s, and switch the post-processing off for the following 1 s if the SNR is above a threshold (10 dB) for more than 25% of the time.

5. Experimental evaluation

The performance of the post-processing algorithm is evaluated for three real environmental noises viz. factory noise, machine gun noise and train interior noise; all the three noises are highly fluctuating, characterized by random energetic bursts. Noise-corrupted speech at selected SNRs is generated by adding speech and noise digitally. The global SNR is computed according to ITU P.56 standard (ITU-T Recommendation, 1993) based on the rms "active speech level" and rms noise level of the test sentence. The active speech level and noise level are estimated by leaving out silence and idle segments but including grammatical/

structural pauses (i.e. those within 300 ms), and therefore corresponds typically to a much lower value in terms of segmental SNR. Two standard STSA algorithms, the Berouti spectral subtraction (BSS) (Berouti et al., 1979) and the multiplicatively modified log spectral amplitude estimator (MM-LSA) (Malah et al., 1999) are chosen as the front-end STSA algorithms. Spectral subtraction is a simple and robust method that has been extensively applied in practice. The MM-LSA evolved from the MMSE estimator of Ephraim and Malah (1984) which was motivated by the desire to eliminate the musical noise artifact, characteristic of spectral subtraction enhanced speech. This relatively new class of STSA algorithms do successfully eliminate musical noise but are generally considered to be difficult to tune in terms of the noise reduction–speech distortion tradeoff (Accardi and Cox, 1999). For background noise estimation as required by the STSA algorithm, a recently proposed VAD-based method is used (Marzinzik and Kollmeier, 2002) which achieves a low false-alarm rate even in low SNR conditions. The parameters of BSS algorithm and the MM-LSA algorithm are selected as shown in Table 1 (notations follow the referenced publications). These values are chosen to achieve a good trade-off between noise reduction and speech distortion and serve to largely suppress the stationary background noise. In all experiments, a 32 ms Hamming window with 50% overlap is applied to 8 kHz sampled speech. The spectrum is computed using a 256-point DFT. The post-processing algorithm is applied to the time–frequency bin levels of the STSA-enhanced signal. The parameters of the post-processor are adjusted considering the gross properties of the noise as explained below.

5.1. Noise properties and post-processing parameter settings

Three examples of environmental noises with strongly fluctuating characteristics are chosen for the experimental validation of the post-processing algorithm. Table 2 lists some properties of the three

Table 1
Parameters selected for the STSA algorithms used in the evaluation

Berouti spectral subtraction (BSS) Berouti et al. (1979)	
$\alpha_0 = 5$	Value of oversubtraction factor at 0 dB SNR
$\beta = 0.01$	Spectral gain floor
Multiplicatively modified LSA (MM-LSA) Malah et al. (1999)	
$\alpha = 0.92$	Weighting factor for the a priori SNR estimation
$\eta_{\min} = -25$ dB	Lower limit for the a priori SNR
$\gamma_{\text{TH}} = 0.8$	Threshold for hypothesis testing
$\alpha_q = 0.95$	Smoothing parameter for q estimation

Table 2
Properties and post-processing algorithm parameters for the test noises

Noise	Properties	Selected post-processor parameters			
		Band width	Variance threshold	SNR_low	SNR_high
Factory	Frequency-localized bursts of 50–200 ms duration with local noise power 3–5 dB higher than that of the stationary background	1 kHz	0.775 for bands below 2 kHz and 0.682 for the rest	−5 dB for bands below 2 kHz and 0 dB for other bands	10 dB for all bands
Machine gun	Bursts of 30–40 ms duration with low frequency localization	1 kHz	0.775 for band 1 and 0.65 for the other bands	−5 dB for bands below 2 kHz and 0 dB for other bands	10 dB for all bands
Train	Broadband noise bursts of 50–150 ms duration with local noise power 6–10 dB higher than that of the stationary noise	2 kHz	0.775 for first two bands and 0.682 for the last band	0 dB for all bands	12 dB for all bands

noise samples. Factory noise from the SPIB database (Signal Processing Information Base, 2004) contains randomly occurring events such as hammer blows embedded in a more homogenous background noise. Machine gun noise, also from the same database, is a series of gunshots recorded in a quiet environment. To make it more realistic, a white background noise, also from the SPIB database, has been added to the machine gun noise. The third noise sample considered for testing is "train noise" (obtained from Essential Indian Sound Effects, CD-ROM, 1999). It is sound recorded in the interior of an Indian electric train with windows open (i.e. the noise arises from the moving mechanical parts of the train). Spectrograms are shown for samples of 5 s duration for the three noises in Fig. 3. We observe that while factory noise contains frequency-localized noise bursts, train noise has broadband bursts of varying intensities with respect to a relatively stationary background. The machinegun noise is characterized by short-duration noise bursts with strength decreasing rapidly in the frequency region beyond about 1 kHz. We next provide guidelines on how the post-processing parameters can be tuned to specific noise properties for situations where the noise characteristics are known.

Table 2 depicts the selected noise specific post-processing algorithm parameters for each of the noises investigated. We note that the parameters chosen differ only slightly across the noise types and are influenced by the gross properties as listed in Table 2 and the ROC plots of Fig. 2 besides informal listening tests. The frequency bandwidth for the variance-based noise detection in (5) is selected to provide a high-frequency resolution for noisy region detection while keeping in mind the need for adequate averaging for a reliable estimate. For instance, the frequency band should span at least two harmonics of high-pitched voiced speech. In general, a bandwidth of 1 kHz provides a good compromise between the required averaging and noise burst frequency localization. The bandwidth is set to 1 kHz for the factory and machine gun noises since the noise bursts are localized in frequency. On the other hand in the train noise, noise bursts have energy distributed over the entire spectrum. Hence a higher bandwidth of 2 kHz has been chosen for better averaging in (5). In all cases, the total number of frequency bands searched

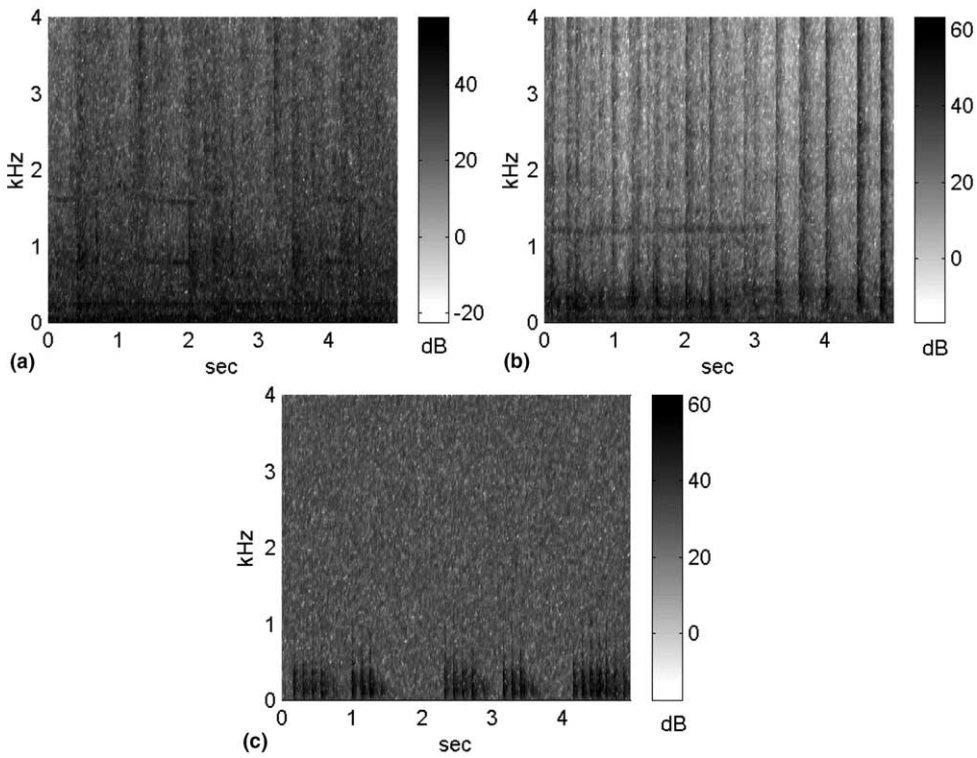

Fig. 3. Spectrograms of segments of (a) factory, (b) train and (c) machinegun noise (spectrogram parameters: Hamming window of 32 ms, 50% overlap, DFT size = 256).

corresponds to 50% overlapping of the bands with the specified bandwidths.

The choice of decision threshold on the variance in (5) for the detection of noise-dominated bands should be based on the desired hit rate or tolerable false-alarm rate. A low false-alarm rate helps to minimize speech distortion. Informal listening showed that a false-alarm rate in the region of 0.2 achieved good noise reduction with low speech distortion. Accordingly, the thresholds for the different bands were chosen from a study of ROC plots for each noise (as demonstrated for factory noise in Fig. 2). The slightly lower threshold (implying a bias toward voicing) chosen for the high-frequency bands is justified by the typically reduced harmonicity of speech in these bands compared to that in the lower frequency bands.

Coming to the *attenuation factor* calculation in (2), the parameters *SNR_low* and *SNR_high* determine the amount of attenuation as a function of the estimated a priori SNR. As such, the a priori SNR is expected to take on higher values for noise bursts with levels increasingly above the stationary noise floor. Hence the slightly higher parameter values for the train noise case, leading to a more aggressive suppression rule relative to that for the other noises. In the case of factory and machine gun noise, it was necessary to increase the suppression in the high-frequency region due to the reduced speech energy which leads to the reduced masking of residual noise in this region.

In order to verify that the performance gains from post-processing are not greatly dependent on the tuning of the post-processor parameters to the specific noise, the objective performances with train and machine gun noises are computed again with the parameter settings selected for factory noise in Table 2.

5.2. Measuring speech quality improvement

The speech quality obtained after post-processing, relative to that before, is expected to reflect the amount of random burst noise suppression achieved by the algorithm. Apart from noise reduction, naturalness and intelligibility of speech output are important attributes of the performance of any speech enhancement system. Since achieving a high degree of noise suppression is often accompanied by speech signal distortion, it is important to evaluate both quality and intelligibility. In fact, for nearly all methods of speech enhancement, significant gains in noise reduction are accompanied by a decrease in speech intelligibility (Lim and Oppenheim, 1979). Formal subjective listening tests are the best indicators of achieved overall quality. A–B comparison tests of sentences processed by competing processing methods can be used to obtain comparative quality rankings. The chief attributes tested here are the naturalness or overall quality of the processed speech. Speech intelligibility, on the other hand, is not easily quantified by general sentence level testing. Rhyming words (Quackenbush et al., 1988) and semantically unpredictable sentences have been used in the past in the subjective testing of speech intelligibility.

The subjective test employed in this work for overall quality ranking is A–B comparison involving four listeners and eight distinct sentences from the TIMIT database (Fisher et al., 1986), each from a different speaker (four male and four female). Each sentence pair presented for listening comparison comprises of the processed versions of a single sentence, before and after post-processing. To avoid bias, the order A and B are interchanged and randomized across sentences and listeners. Thus for each test condition, a total of 64 subjective judgements is obtained across listeners and test sentences. Speech intelligibility is tested by the SUS (semantically unpredictable sentences) test, originally proposed for evaluating synthetic speech (Benoit et al., 1996). Thirty SU sentences, six of each of five syntax structures, were generated and played in random order to each of four listeners who were asked to write down the sentences they hear. To avoid listener familiarity with a specific noise sample, segments of the noise file to be added to the sentences were chosen randomly from a larger noise sample and digitally added to the clean speech. Given the large number of conditions that need to be evaluated in this work, formal subjective listening tests were limited to a subset with the remaining conditions evaluated only via objective distance measurements.

There are a large number of objective measures that quantify the degradation in quality of processed speech with respect to a reference speech sample (Quackenbush et al., 1988). However, not all objective measures may be appropriate for specific kinds of distortion. For instance, the popular segmental SNR measure is useful only when the distorted speech has the same phase alignment as the reference speech as occurs in waveform coding. For the purpose of speech enhancement, Marzinzik

(2000) has investigated various objective measures to predict sound quality in noise reduction algorithms and found LAR distance and PSQM to be most correlated with subjective judgements of quality degradation. The PSQM, originally designed for the assessment of narrowband speech codecs, incorporates sophisticated auditory and cognitive models. The ITU standard PESQ (ITU-T Recommendation P.862, 2001) is an advanced version of the PSQM which predicts subjective MOS for a variety of speech distortions in transmission systems. On the other hand, a simple aural measure that is known to show good correlation to subjective data is the weighted spectral slope (WSS) distance based on critical band filtering and comparison of the slopes in each frequency band (Quackenbush et al., 1988; Hansen and Pellom, 1998). Our own informal listening tests and comparisons with the predictions of several objective measures revealed that the WSS distance measure is an accurate predictor of subjective quality differences across noise reduction algorithms. Based on the above considerations, we use PESQ and WSS in the experiments to measure quality gains, if any, achieved due to post-processing.

6. Results and discussion

The A–B comparison test was carried out to compare enhanced speech obtained using Berouti spectral subtraction (BSS) alone over that obtained after added post-processing (BSS + PP). Table 3 shows the results obtained in percentage of the number of times a configuration was preferred across sentences and listeners for each of the three noises at 0 dB and 3 dB SNR. We observe that there is a clear listener preference for the post-processed speech over that before post-processing. Further, the SUS speech intelligibility test was carried out for the case of factory noise-corrupted speech. The score results are computed by considering the overall percentage of correct words for the whole corpus of 144 words for each of the four listeners. The percentage word intelligibility scores averaged across the listeners are 60.7, 51.7 and 50.6 at 3 dB SNR for the three configurations of noisy, BSS and BSS + PP respectively. Although there is the expected decrease in intelligibility in going from noisy to noise-reduced speech, we note that there is no significant change in intelligibility scores due to the post-processing. In a few specific cases, even a slight improvement in intelligibility was noticed in the case of stop consonants closely following a noise burst. The forward temporal masking effect from the burst was significantly diminished due to its attenuation by the post-processor.

Table 3
A–B comparison test results as % preferred for BSS and BSS + PP on various noisy speech samples

Preferred configuration	Factory		Machine gun		Train	
	0 dB	3 dB	0 dB	3 dB	0 dB	3 dB
BSS	14.1	15.6	12.5	14.1	14.1	12.5
BSS + PP	75.0	71.9	78.1	70.3	79.6	78.1
Neutral	10.9	12.5	9.4	15.6	6.3	9.4

That the perceptual effects of noise bursts are reduced by post-processing is also evident by an examination of the spectrograms of the enhanced speech, an example of which appear in Fig. 4. The narrowband spectrogram (analysis Hamming window length = 32 ms) representation of a 2.5 s sample of clean speech is shown together with that of the same sample corrupted with 0 dB factory noise. Also shown are the spectrograms of the BSS-enhanced noisy speech and that obtained after further post-processing. We note the nearly complete removal of the stationary component of the background noise in the BSS-processed speech of Fig. 2(c). However the noise bursts are clearly visible particularly in the low energy speech regions; also evident is the loss of some low-amplitude speech harmonics. Fig. 2(d) indicates that post-processing leads to a noticeable reduction in the strengths of the noise bursts with respect to the speech regions.

Tables 4–6 list the objective quality scores of noisy and enhanced speech (with respect to the corresponding clean speech file) for a speech file (comprising of the concatenation of eight sentences: four male and four female from the TIMIT database) corrupted by the three noises at various SNRs. The different configurations tested include post-processing applied to each of the two STSA algorithms as discussed in Section 5, namely, BSS and the MM-LSA. The objective results are based on the WSS distance and the PESQ MOS measure. We observe that the WSS distance indicates a consistent decrease (implying an improvement in quality) with post-processing from that obtained with STSA enhancement alone. The anomaly in estimated WSS distance for noisy speech (i.e. the WSS score for noisy speech is better than that of enhanced speech) is due to the simplicity of the objective

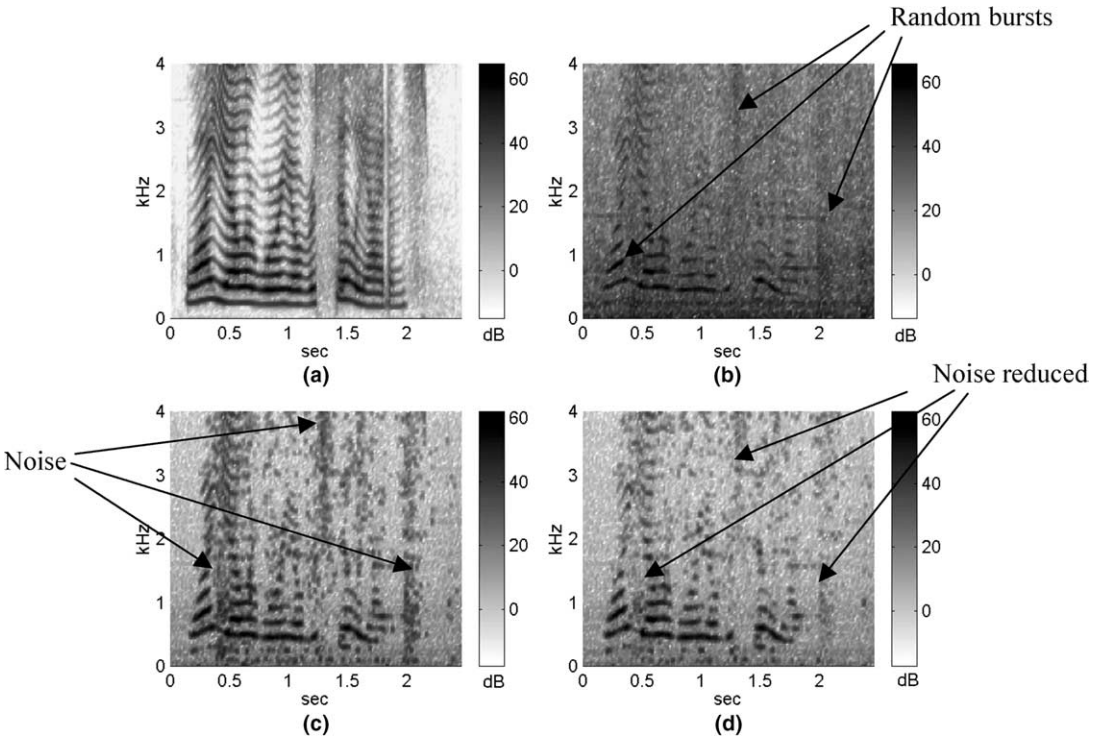

Fig. 4. Narrowband spectrograms of (a) clean, (b) noisy, (c) BSS-enhanced speech and (d) after post-processing, for a speech segment in factory noise (spectrogram parameters: Hamming window of 32 ms, 50% overlap, DFT size = 256).

Table 4
WSS distance and PESQ MOS scores for noisy and processed speech in factory noise

SNR (dB)	Noisy		BSS		BSS + PP		MM-LSA		MM-LSA + PP	
	WSS	MOS	WSS	MOS	WSS	MOS	WSS	MOS	WSS	MOS
−5	76.4	1.31	90.6	1.64	83.0	1.71	96.3	1.68	86.3	1.73
−3	71.6	1.33	86.5	1.76	77.9	1.78	91.4	1.90	82.2	1.87
0	66.6	1.65	77.7	2.01	70.5	2.01	82.9	2.09	76.8	2.10
3	60.5	1.84	72.5	2.20	66.8	2.20	76.4	2.29	73.4	2.29
5	54.7	2.01	65.6	2.37	63.2	2.30	68.7	2.42	67.8	2.43

Table 5
WSS distance and PESQ MOS scores for noisy and processed speech in machine gun noise

SNR (dB)	Noisy		BSS		BSS + PP		MM-LSA		MM-LSA + PP	
	WSS	MOS	WSS	MOS	WSS	MOS	WSS	MOS	WSS	MOS
−5	63.3	1.21	77.4	1.28	70.8	1.55	78.0	1.41	65.5	1.61
−3	59.3	1.35	72.9	1.49	67.0	1.73	73.2	1.63	62.1	1.79
0	52.9	1.59	67.1	1.75	62.2	1.90	66.7	1.90	57.4	1.97
3	46.7	1.85	60.6	2.01	56.6	2.14	59.9	2.16	52.0	2.20
5	42.6	2.01	56.6	2.14	52.8	2.26	55.5	2.31	48.6	2.32

measure and its consequent inability to predict the subjectively perceived quality of distinctly different classes of degradation. The PESQ MOS on the other hand is consistent with the subjectively perceived trend of an improvement in speech quality with STSA enhancement over that of noisy speech, and a further improvement in quality with post-processing. The improvement with post-processing

Table 6
WSS distance and PESQ MOS scores for noisy and processed speech in train noise

SNR (dB)	Noisy		BSS		BSS + PP		MM-LSA		MM-LSA + PP	
	WSS	MOS	WSS	MOS	WSS	MOS	WSS	MOS	WSS	MOS
−5	69.3	1.34	81.1	1.17	76.8	1.54	77.6	1.28	71.1	1.45
−3	64.9	1.48	77.0	1.31	72.8	1.70	73.5	1.49	67.1	1.70
0	58.0	1.73	69.9	1.59	66.6	1.84	66.7	1.77	61.1	1.92
3	50.9	1.93	63.2	1.88	60.2	2.06	59.8	2.03	55.5	2.14
5	46.3	2.07	58.9	2.05	56.2	2.16	55.4	2.19	51.1	2.26

is particularly significant in the case of train noise and machine gun noise corruptions. Further, both the objective measures indicate that post-processing has a greater influence at the lower SNRs relative to that at higher SNRs.

Table 7 lists the objective quality scores obtained after post-processing on machine gun noise and train noise using the post-processor parameter settings corresponding to factory noise in Table 2. Comparing the scores in Table 7 with the corresponding columns of Tables 5 and 6, we note that the performance gains due to post-processing do not change significantly with the change in the algorithm parameters. The objective quality improvement over speech before post-processing continues to be clearly evident.

It may be mentioned that while our present study employed an available VAD (Marzinzik and Kollmeier, 2002) to identify the silence frames for the noise estimation in the STSA algorithm, the relative improvement obtained due to post-processing remains unaltered when an ideal VAD (i.e. frames are identified as speech or silence manually) is used.

Table 7
Objective scores obtained using the factory noise settings of Table 2 for the post-processing of speech in (Panel A) machine gun noise, and (Panel B) train noise

SNR (dB)	BSS + PP		MM-LSA + PP	
	WSS	MOS	WSS	MOS
Panel A				
−5	69.1	1.52	72.8	1.54
−3	64.7	1.70	68.4	1.73
0	60.5	1.89	61.6	1.94
3	55.3	2.12	55.5	2.17
5	52.0	2.24	51.2	2.31
Panel B				
−5	77.0	1.49	64.9	1.53
−3	72.6	1.68	61.4	1.72
0	66.2	1.86	57.0	1.92
3	60.3	2.06	52.0	2.15
5	56.0	2.17	49.0	2.26

7. Conclusion

Traditional STSA speech enhancement algorithms perform inadequately in application to speech corrupted by highly nonstationary noise. While stationary background noise components are effectively suppressed, random noise fluctuations characteristic of burst noises remain unattenuated in STSA-enhanced speech output. A frequency-domain post-processing algorithm to follow STSA speech enhancement has been proposed with a view to improve speech quality in the presence of random noise bursts. The post-processing is based on the detection of noise-dominated time–frequency regions in the STSA-enhanced speech followed by selective bin-dependent attenuation based on a measure of SNR. The usual statistically optimal speech–noise classifiers being signal energy-based cannot distinguish speech from high-energy noise bursts. Hence the detection of noise-dominated regions is based on exploiting the difference in the spectral flatness of noise spectra from that of harmonic speech. With limited added complexity, the post-processing algorithm is effective in significantly reducing the perceived effects of the noise bursts at low SNRs without further speech distortion. While the onsets of noise bursts are greatly attenuated, bursts of long duration are not suppressed completely due to the difficulties in the reliable classification of bins as speech or noise dominated within an identified noise burst band. More detailed acoustic models for the noise, over the present simple spectrum flatness cue, would be expected to improve the noise suppression capability of the post-processing algorithm further.

References

Accardi, A.J., Cox, R.V., 1999. A modular approach to speech enhancement with an application to speech coding. In: Proc. ICASSP-99.

Attias, H., Platt, J.C., Acero, A., Deng, L., 2001. Speech denoising and dereverbertaion using probabilistic models.

In: Advances in Neural Information Processing Systems, Vol. 13, pp. 758–764.

Benoit, C., Grice, M., Hazan, V., 1996. The SUS test: a method for the assessment of text-to-speech synthesis intelligibility using semantically unpredictable sentences. Speech Comm. 18, 381–392.

Berouti, M., Schwartz, R., Makhoul, J., 1979. Enhancement of speech corrupted by acoustic noise. In: Proc. ICASSP-79, pp. 208–211.

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

Cooke, M., Green, P., Josifovski, L., Vizinho, A., 2001. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Comm. 34, 267–285.

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

Essential Indian Sound Effects, CD-ROM, BMG Crescendo (India) Ltd., 1999.

Fisher, W.M., Doddington, G.R., Goudie-Marshall, K.M., 1986. The DARPA speech recognition research database: specification and status. In: Proc. DARPA Speech Recognition Workshop, 1986.

Hansen, J.H.L., Pellom, B.L., 1998. An effective quality evaluation protocol for speech enhancement algorithms. In: Proc. ICSLP-1998.

Hirsch, H.G., Ehrlicher, C., 1995. Noise estimation methods for robust speech recognition. In: Proc. ICASSP-1995, pp. 153–156.

Itoh, K., Mizushima, M., 1997. Environmental noise reduction based on speech/non-speech identification for hearing aids. In: Proc. ICASSP-1997.

Johnston, J.D., 1988. Transform coding of audio signals using perceptual noise criteria. IEEE J. Selected Areas Comm. 6 (2), 314–323.

Lim, J.S., Oppenheim, A.V., 1979. Enhancement and bandwidth compression of noisy speech. Proc. IEEE 67, 1586–1604.

Malah, D., Cox, R.V., Accardi, A.J., 1999. Tracking speech-presence uncertainty to improve speech enhancement in nonstationary noise environments. In: Proc. ICASSP-99, pp. 789–792.

Martin, R., 2001. Noise power spectral density estimation based on optimal smoothing and minimum statistics. IEEE Trans. Speech Audio Process. 9, 504–512.

Marzinzik, 2000. Noise reduction schemes for digital hearing aids and their use for the hearing impaired. Ph.D. dissertation, University of Oldenburg.

Marzinzik, M., Kollmeier, B., 2002. Speech pause detection for noise spectrum estimation by tracking power envelope dynamics. IEEE Trans. Speech Audio Process. 10, 109–118.

Objective measurement of active speech level. ITU-T Recommendation P.56, March 1993.

Perceptual Evaluation of Speech Quality (PESQ), an objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. ITU-T Recommendation P.862, February 2001.

Quackenbush, S.R., Barnwell, T.P., Clements, M.A., 1988. Objective Measures of Speech Quality. Prentice-Hall, NJ.

Renevey, P., Drygajlo, A., 2001. Entropy based voice activity detection in very noisy conditions. In: Proc. Eurospeech-2001, pp. 1887–1890.

Ris, C., Dupont, S., 2001. Assessing local noise level estimation methods: application to noise robust ASR. Speech Comm. 34, 141–158.

Seltzer, M.L., Raj, B., Stern, R.M., 2000. Classifier-based mask estimation for missing feature methods of robust speech recognition. In: Proc. ICSLP-2000.

Signal Processing Information Base, "Noise data". Available from: <http://spib.rice.edu/spib/select_noise.html>, February 2004.

Srinivasan, S., Samuelsson, J., Kleijn, W.B., 2003. Speech enhancement using a priori information. In: Proc. Eurospeech-2003.

Stahl, V., Fischer, A., Bippus, R., 2000. Quantile based noise estimation for spectral subtraction and wiener filtering. In: Proc. ICASSP-2000, pp.1875–1878.

Yao, K., Lee, T.-W., 2004. Speech enhancement by perceptual filter with sequential noise parameter estimation. In: Proc. ICASSP-2004, pp. 693–696.