Estimating time-varying noise introduced by CVSD for speech enhancement

Z. Goh
K.-C. Tan
B.T.G. Tan

Indexing terms: Speech enhancement, Delta modulation, CVSD noise

Abstract: The authors examine the possibility of improving the quality of speech obtained with continuously variable slope delta-modulation (CVSD), a speech coding method that has been quite widely employed in mobile radios. CVSD is well-known for its robustness and implementation simplicity, but the quality of the resultant speech (commonly at 16 kbps) is often not very satisfactory. Through their study, the authors obtain two crucial findings on the characteristics of the disturbance/noise introduced by CVSD. Based on their findings, they develop a method for estimating the short-time power spectra of CVSD noise, and this enables them to apply existing speech enhancement methods to effectively suppress CVSD noise. Performance assessments based on objective measures such as signal-to-noise ratio (SNR), segmental SNR, and COSH distortion measure, and informal subjective listening tests have all indicated significant improvements in speech quality.

1 Introduction

There is a great demand for the use of mobile radios in recent years in both military and civilian contexts. To achieve mobility, it is almost essential for such radios to transmit signals which carry information on speech, image, data, etc., via wireless means. Unfortunately, the allowable transmission bandwidth is limited in many practical scenarios, and this prohibits more extensive use of mobile radios. One common approach to tackling the problem resulting from bandwidth limitation is to reduce the bit rate required for signal transmission. In this connection, considerable research attention has been given to the reduction of the bit rate for transmitting speech signals during the process of speech coding.

IEE Proceedings online no. 19981746

Paper first received 3rd June and in revised form 28th October 1997

Z. Goh and K.-C. Tan are with the Centre for Signal Processing, Level B4, S2-B4-08, School of Electrical and Electronic Engineering, Nanyang Technological University, Nanyang Avenue, Singapore 639798, Republic of Singapore

B.T.G. Tan is with the Faculty of Science, National University of Singapore, Singapore 119260, Republic of Singapore

Although a major objective of speech coding is to achieve bit-rate reduction (while maintaining reasonably high speech quality and intelligibility), there are other crucial considerations in designing speech coding algorithms. In particular, the robustness against transmission channel errors and implementation simplicity are often of concern. Consequently, continuously variable slope delta-modulation (CVSD) [1, 2], which is well-known for robustness and implementation simplicity, remains attractive despite the fact that the quality of the resultant speech is not completely satisfactory (though the speech intelligibility is often very acceptable), and that modern methods, for example those proposed in [3, 4], are able to achieve higher bit-rate reduction than CVSD.

The main objective of this work is to investigate methods for improving the quality of speech obtained with CVSD (commonly operating at 16 kbps). As a matter of fact, the 'noise' CVSD introduces, which arises from quantisation, is of considerably large bandwidth and quite annoying. Therefore, some existing speech enhancement methods capable of suppressing wide-band noise, such as those proposed in [5, 6], appear to be applicable. These speech enhancement methods require reasonably good estimates of the short-time power spectra of the noise, which are often obtained from the short-time power spectra of those frames containing only noise. Such estimates would be good if the noise characteristics remain quite stationary over time. Unfortunately, the characteristics of CVSD quantisation noise vary as rapidly as those of the speech itself, and this prohibits direct application of the existing enhancement methods.

We begin with a detailed analysis of the characteristics of CVSD quantisation noise. Note that although there are studies of quantisation noise introduced by general delta-modulation coding schemes [1, 7, 8], we have not come across one specifically devoted to CVSD. Also, these studies adopt a speech model which is statistically stationary. However, speech is stationary only during a somewhat short period, and thus it is not easy to develop a general model which incorporates the time varying nature of speech. In addition, it is not straightforward to characterise the effect due to a variety of factors such as speaker, language, conversation content, etc.. Therefore, we carry out the analysis through statistical means, using some 'representative' input speech from the TIMIT database [9].

Through our analysis, we quantify the dependence of CVSD quantisation noise on the uncoded speech. In addition, we work out a crucial relationship between

the quantisation noise and the CVSD decoded speech. Based on these two findings, we develop a method for estimating the short-time power spectra of CVSD noise, and this enables us to apply existing speech enhancement methods [5, 6] to effectively suppress CVSD noise. One encouraging observation is that our method works reasonably well even for speech which is different from that used for our analysis (to be elaborated in Section 5). Note that the main ideas of this work have been presented in [10].

2 Discussion on speech enhancement

We shall present a brief description of two relevant existing speech enhancement methods, namely spectral subtraction [5] and the Ephraim–Malah method [6]. Spectral subtraction is attractive because of its simplicity and (computational) efficiency. However, it often introduces a specific disturbance, commonly known as 'musical noise'. The Ephraim–Malah method, on the other hand, does not usually introduce such disturbance and will be of good use in some demanding applications. But the trade-off is high complexity and computational overheads.

The two enhancement methods are based on the following model:

$$y[n] = s[n] + w[n] \tag{1}$$

where $y[n]$, $s[n]$ and $w[n]$ denote discrete-time samples of noisy speech, clean speech and noise respectively. A sampled short-time Fourier transform (SSTFT) on eqn. 1 leads to

$$Y_r[k] = S_r[k] + W_r[k] \tag{2}$$

where $Y_r[k]$, $S_r[k]$ and $W_r[k]$ denote respectively the SSTFTs of $y[n]$, $s[n]$ and $w[n]$ for the rth frame and the index k denotes the kth frequency component. To carry out spectral subtraction, we first obtain $|\hat{S}_r[k]|$s, the magnitudes of the SSTFT of the enhanced speech, with the following equation:

$$|\hat{S}_r[k]| = \begin{cases} (|Y_r[k]|^2 - \beta E(|W_r[k]|^2))^{1/2} \\ \qquad \text{if } |Y_r[k]|^2 > \beta E(|W_r[k]|^2) \\ 0 \qquad \text{otherwise} \end{cases} \tag{3}$$

where $E(|W_r[k]|^2)$ denotes the statistical mean of $|W_r[k]|^2$, the short-time power spectrum of noise, and β is a constant which controls the amount of noise suppression. Note that larger β gives more noise suppression and reduction of musical noise, but speech intelligibility will be compromised. In conducting the experiments to be reported in Section 5, we shall use $\beta = 3.7$, which often yields a reasonable compromise between noise suppression and intelligibility preservation. For the Ephraim–Malah method, the spectral magnitudes $|\hat{S}_r[k]|$s are obtained in a more sophisticated way (please refer to [6] for details).

Next, the spectral phases $\arg(\hat{S}_r[k])$s of the enhanced speech are taken to be those of the noisy speech. With $|\hat{S}_r[k]|$ and $\arg(\hat{S}_r[k])$, the enhanced speech can be obtained via inverse fast Fourier transform and the standard overlap-add processing [5, 6].

It is clear that both enhancement methods involve the use of $E(|W_r[k]|^2)$, the statistical mean of the short-time power spectrum of the noise, which is usually estimated by $\hat{E}(|W_r[k]|^2)$, the average of the short-time power spectra for those frames of noisy speech containing only the noise. In general, it can be expected that if $|W_r[k]|^2$, the actual short-time power spectrum of the noise for that frame (i.e., frame r), is available and is used as a substitute for $E(|W_r[k]|^2)$, the enhancement result should be better. In cases where the noise is stationary, $\hat{E}(|W_r[k]|^2)$ is a reasonably good approximation of $|W_r[k]|^2$, and the enhancement result would be good. However, when the noise is nonstationary, $\hat{E}(|W_r[k]|^2)$ will differ significantly from $|W_r[k]|^2$, and the enhancement result may not be satisfactory. Here, we are concerned with the noise introduced by CVSD which is nonstationary in nature, and thus $\hat{E}(|W_r[k]|^2)$ will not be a good estimate of $|W_r[k]|^2$. Consequently, we shall attempt to obtain a better estimate in Section 4.

3 An analysis of CVSD noise

In this Section, we shall study the characteristics of the noise introduced by CVSD, in the hope of obtaining a good estimate of the short-time power spectra of the noise.

3.1 Speech-dependent nature of CVSD quantisation noise

It can be deduced by listening that CVSD quantisation noise (the difference between CVSD decoded speech and the original speech) often contains some speech components. A simple way to model this is:

$$q[n] = c \cdot s[n] + d[n] \tag{4}$$

where $q[n]$ denotes CVSD quantisation noise, $c \cdot s[n]$ denotes the 'speech' component with $s[n]$ being the original (uncoded) speech and c a non-zero constant, and $d[n]$ denotes the 'noise' component. For convenience, we shall call $d[n]$ the 'uncorrelated' noise.

To make use of the above model (i.e., eqn. 4) to estimate the power spectra of CVSD noise, the value of c has to be first determined. In this connection, the approach we adopt is to minimise over c a measure of correlation between $d[n]$ (= $q[n] - c \cdot s[n]$), the 'uncorrelated' noise, and $s[n]$, the speech itself, for a particular class of speech. The value of c giving rise to the minimum correlation measure will be chosen and will be used in the estimation of CVSD noise when the class of speech is encountered. We choose the measure to be $K(c)$, the correlation between the normalised short-time power spectra of $d[n]$ and $s[n]$:

$$K(c) = \frac{\sum_{r,k}(X[r,k] - \overline{X[r,k]})(Y[r,k] - \overline{Y[r,k]})}{\left\{\sum_{r,k}(X[r,k] - \overline{X[r,k]})^2 \sum_{r,k}(Y[r,k] - \overline{Y[r,k]})^2\right\}^{\frac{1}{2}}} \tag{5}$$

where

$$X[r,k] = \frac{|D_r[k]|^2}{\frac{1}{N}\sum_{l=1}^{N}|D_r[l]|^2}$$

$$Y[r,k] = \frac{|S_r[k]|^2}{\frac{1}{N}\sum_{l=1}^{N}|S_r[l]|^2} \tag{6}$$

$\overline{X[r,k]}$ and $\overline{Y[r,k]}$ denote the means of $X[r, k]$ and $Y[r,k]$ respectively, and $|D_r[k]|^2$ and $|S_r[k]|^2$ denote the short-time power spectra of $d[n]$ (= $q[n] - c \cdot s[n]$) and $s[n]$ respectively. The rationale behind choosing the above short-time power spectrum correlation measure is that we find it more appropriate to assess the close-

ness of two signals based on their short-time power spectra than their time-domain waveforms.

The minimisation of $K(c)$ is done by statistical means in the following manner. For each value of c, we compute $K(c)$ using 50 speech sentences (which contain 8563 short-time spectra) taken from the TIMIT database [9]. To achieve as much diversity in the short-time spectra as possible, we choose sentences spoken by 25 male and 25 female speakers which are phonetically quite distinct. We then take the representative value of c to be one giving rise to the smallest $K(c)$ — assigning such a value to c for the model specified by eqn. 4 will yield a $d[n]$ that is statistically most uncorrelated with the speech signal itself.

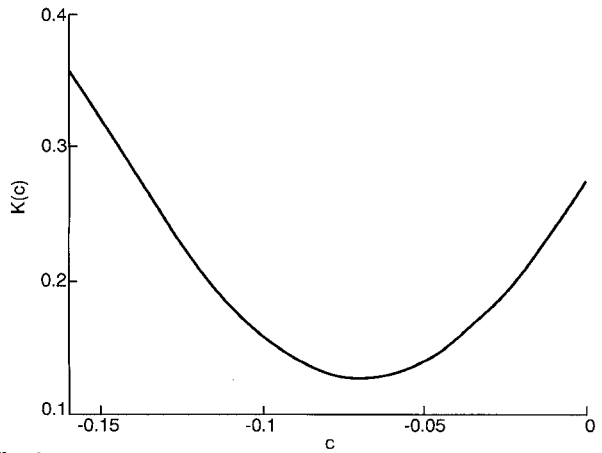

Fig. 1 *The correlation measure $K(c)$ against c*

Fig. 1 shows the graph of $K(c)$ versus c. At $c = -0.07$, $K(c)$ attains its minimum amounting to 0.13. (In comparison, $K(c)$ attains a much higher value of 0.28 at $c = 0$, the case where $d[n]$ is exactly $q[n]$.) Consequently, we adopt the following model hereafter:

$$q[n] = -0.07 \cdot s[n] + d[n] \tag{7}$$

Basically, the result we obtain here is that $q[n]$, the CVSD quantisation noise, in fact contains some speech components, and this is in agreement with our listening assessment. Therefore, one should be concerned with suppression of $d[n]$, the 'uncorrelated' noise, rather than $q[n]$ as a whole, since suppression of $q[n]$ (which has speech components) will affect the desired speech signal.

In essence, it is the 'uncorrelated' noise, rather than the CVSD quantisation noise as a whole, that one should suppress. Thus one should focus on estimating the short-time power spectra of the 'uncorrelated' noise. In this connection, our approach is to first establish a relationship between the 'uncorrelated' noise and CVSD decoded speech (to be discussed in the next subsection), and then make use of this relationship to estimate the short-time power spectra of the 'uncorrelated' noise (to be discussed in Section 4).

Remark: It is interesting to note that in cases where the 'correlated' part of CVSD quantisation noise (i.e., $c \cdot s[n]$) is of concern, it could be removed by amplitude scaling.

3.2 Dependence of the 'uncorrelated' noise on CVSD decoded speech

To analyse the dependence of the 'uncorrelated' noise on CVSD decoded speech, we use the same 50 speech sentences mentioned in Section 3.1. Indeed, we first generate CVSD decoded speech from the original speech for all the 50 speech sentences. We then compute $q[n]$, the CVSD quantisation noise, as the difference between the CVSD decoded speech and $s[n]$, the original speech. Subsequently, we compute $d[n]$, the 'uncorrelated' noise, using the formulation $q[n] + 0.07\ s[n]$, for all the 50 speech sentences. Finally, we segment the CVSD decoded speech and the 'uncorrelated' noise into 8563 overlapping frames, each of which is 32 msec long (overlap by 28 msec). These 8563 frames of the 'uncorrelated' noise and 8563 frames of the CVSD decoded speech are then used in our subsequent analysis.

Our preliminary analysis reveals that there is in fact a close relationship between the 'uncorrelated' noise and CVSD decoded speech. Indeed, it is apparent in Fig. 2, which shows a scattered diagram of the 8563 short-time energies of the 'uncorrelated' noise versus that of the CVSD decoded speech, that there could be a linear relationship between them. However, the spread of the data is somewhat large and this prevents one from exploiting the result directly. This then motivated us to search for another 'energy' measure which yields a closer relationship, and in turn leads to a more accurate estimate of the short-time power spectra of the 'uncorrelated' noise.

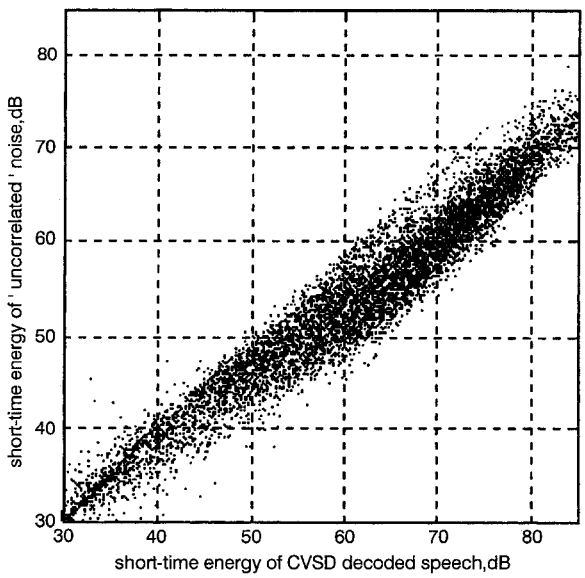

Fig. 2 *Scattered diagram of short-time energy of the 'uncorrelated' noise against that of CVSD decoded speech for a total of 8563 speech frames*
Each dot represents a point (EN^s, EN^n), where EN^s and EN^n denote the short-time energy of CVSD decoded speech and that of the 'uncorrelated' noise respectively for a particular frame

Indeed, we have found that the mean of the short-time log power spectrum (we shall refer to as *short-time CEP1* hereafter) is a good candidate. Let us first examine Fig. 3, which shows a scattered diagram of the 8563 short-time CEP1s of the 'uncorrelated' noise versus that of the CVSD decoded speech. There is clearly a close relationship. We then perform regression analysis on the data and obtain the following cubic polynomial which best fits the data in the least-square-error sense (refer to [11] for details of regression analysis):

$$y = (5.58 \times 10^{-5})x^3 + (5.51 \times 10^{-3})x^2 + 1.02x + 1.66 \tag{8}$$

where x and y denote the short-time CEP1 of CVSD decoded speech and that of the 'uncorrelated' noise respectively. Note that such regression analysis will be

carried out once and for all for a particular class of speech signals.

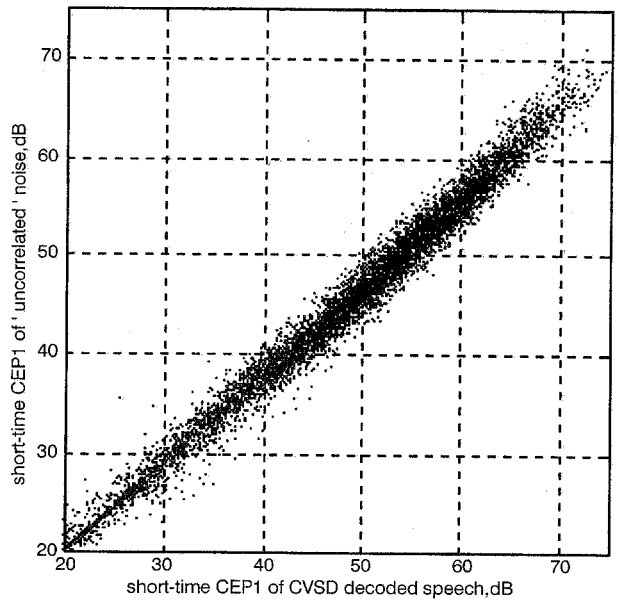

Fig. 3 *Scattered diagram of short-time CEP1 of 'uncorrelated' noise against that of CVSD decoded speech for a total of 8563 speech frames*
Each dot represents a point ($CEP1^s$, $CEP1^n$), where $CEP1^s$, $CEP1^n$, denote the short-time CEP1 of CVSD decoded speech and that of the 'uncorrelated' noise respectively for a particular frame

Although eqn. 8, relative to all other possible cubic polynomials, best fits the data in the least-square sense, one should also be concerned with how well it fits the data in an absolute sense (i.e., how 'good' the goodness-of-fit is in statistical terms). In this connection, we propose using two measurements. The first is the coefficient-of-determination [11]. It gives a value of 0.98 (a value of 1 means perfect fit), indicating that the regression model (eqn. 8) fits the data very well. The other measurement is the standard error, which is the standard deviation of the estimation error [11]. It gives a value of 1.36dB, and this implies that 95% of the estimated values are at most only 2.67dB (1.96 times of the standard error) away from the actual values.

Our analysis here indicates that there is indeed a close relationship between the 'uncorrelated' noise and CVSD decoded speech. Moreover, the relationship established, i.e., eqn. 8, allows one to obtain a reasonably good estimate of the short-time CEP1 of the 'uncorrelated' noise from that of the CVSD decoded speech. This in turn allows one to obtain a reasonably good estimate of the power spectra of the 'uncorrelated' noise from CVSD decoded speech, which is the main topic of Section 4.

3.3 Remarks

Note that although the analysis has been carried out for CVSD, it could similarly be carried out for other waveform coders such as constant factor delta-modulation (CFDM), hybrid companding delta-modulation (HCDM), adaptive differential pulse code modulation (ADPCM), etc.. Like CVSD, CFDM and HCDM belong to the class of delta-modulation (DM) methods. Thus the analysis/results presented here should be relevant. As for ADPCM, the quantisation noise it produces is usually not as annoying as those of DM methods, and so the pay-offs of carrying out such analysis on ADPCM might not be as rewarding.

For vocoders and hybrid coders such as linear predictive coder (LPC), code excited LPC (CELP), multiband excitation vocoder (MBE), mixed excitation LPC (MELP), etc., one can also consider using the approach we presented. However, the residual noise/distortions generated by vocoders/hybrid coders are usually more complex than those generated by waveform coders (in particular, they can be both additive and convolutional). Consequently, a more comprehensive analysis on such noise will be necessary.

4 Method for estimating CVSD noise and the complete enhancement procedure

By exploiting the specific characteristics of the noise introduced by CVSD as discussed in Section 3, we propose a method for estimating the short-time power spectra of the 'uncorrelated' noise. First, we establish that the short-time power spectra of the 'uncorrelated' noise can be sensibly split into two components, one of which is of high variation and varies as rapidly as the characteristics of speech while the other is of low variation. Next, we suggest a method for estimating these two components separately. After obtaining the estimates of these two components, we combine them and then apply the existing speech enhancement methods discussed in Section 2.

4.1 Splitting the noise power spectrum into high-variation and low-variation components

To facilitate our analysis, we express $|D_r[k]|^2$, the short-time power spectra of the 'uncorrelated' noise, in logarithmic scale:

$$P_r[k] = 10\log_{10}|D_r[k]|^2 \tag{9}$$

At this juncture, it is worthwhile to recall that the subscript r denotes the rth frame and k the kth frequency component. Our study reveals that if one estimates $P_r[k]$ by the statistical mean $\bar{P}_r[k]$ given by

$$\bar{P}_r[k] = 10\log_{10} E(|D_r[k]|^2) \tag{10}$$

where the expectation operation is performed over frames, then the standard deviations of the estimation errors (which will be referred to as *standard error* hereafter) of such estimates for all frequency components are generally quite high (more details will be presented in Section 4.3). Therefore, we will have to take another approach. Indeed, we first express $P_r[k]$ as a sum of two components:

$$P_r[k] = A_r + B_r[k] \tag{11}$$

where A_r denotes the short-time CEP1 of the 'uncorrelated' noise which varies considerably over speech frames, and $B_r[k] = P_r[k] - A_r$ which is relatively more stationary. As a matter of fact, it is apparent in Fig. 3 that A_r, being the short-time CEP1 of the 'uncorrelated' noise, can vary as much as 50dB over different speech frames. On the other hand, some fairly extensive experiments indicate that the variation of $B_r[k]$ is only about 10dB for each frequency component. By splitting $P_r[k]$ as a sum of A_r and $B_r[k]$, it facilitates a relatively more accurate way of estimating $P_r[k]$.

4.2 Estimation of high-variation and low-variation components

Before we discuss the estimation procedure, we shall highlight the fact that A_r (for each and every frame), the high variation component, needs to be estimated for every enhancement process, while $B_r[k]$, the low

variation component, needs to be estimated only once and for all, during the 'training' process.

We estimate each A_r based on eqn. 8. Indeed, for each frame r, we first compute the short-time CEP1 of the CVSD decoded speech. We then substitute it into eqn. 8 and obtain $\hat{A}_r$, the estimated short-time CEP1 of the 'uncorrelated' noise for the same frame. We would like to reiterate that for each speech frame, we have to carry out such an estimation once to obtain the $\hat{A}_r$.

For $B_r[k]$, we propose estimating it by $E(B_r[k])$, its statistical mean, and do it once and for all during the 'training' process. Indeed, using the short-time power spectra of the 'uncorrelated' noise in 8563 different speech frames obtained from the same 50 speech sentences mentioned in Section 3.1, one can compute an estimate of $E(B_r[k])$ by averaging over the 8563 speech frames for each k (i.e., for each frequency component). More precisely, we first generate CVSD decoded speech from the original speech for all the 50 speech sentences. We then compute $q[n]$, the CVSD quantisation noise, which is the difference between the CVSD decoded speech and $s[n]$, the original speech. Next, we compute $d[n]$, the 'uncorrelated' noise, from $q[n]$ and $s[n]$ according to eqn. 7. Subsequently, for each frame r, we compute $|D_r[k]|^2$, the short-time power spectrum of $d[n]$, and then compute $P_r[k]$ according to eqn. 9. Next, we compute the short-time CEP1 of $d[n]$, which is the *exact* A_r for the current frame (note that during such training process, we do not use eqn. 8 to estimate A_r since $d[n]$ is available and thus A_r can be precisely computed). We then compute $B_r[k]$ by $P_r[k] - A_r$ for each r and k. Finally, we will have a total of 8563 $B_r[k]$s (since there are 8563 frames altogether) for each k and use them to obtain an estimate of the statistical mean $E(B_r[k])$. The values of the estimated $E(B_r[k])$s, for all k, is shown in Fig. 4.

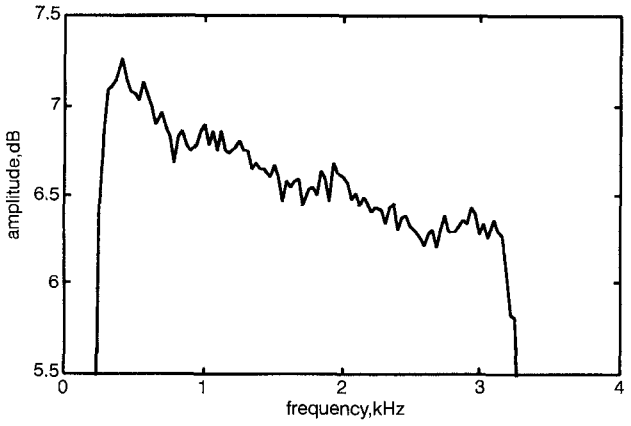

Fig. 4 *Graph of $E(B_r[k])$ (amplitude) versus k (frequency)*

4.3 The proposed estimator for $P_r[k]$ and its performance

In Section 4.1, we have mentioned that $\bar{P}_r[k]$ as given by eqn. 10 will result in a poor estimate for $P_r[k]$. Here, we propose a better estimator $\hat{P}_r[k]$ for $P_r[k]$ as follows, using the estimates of the high-variation and low-variation components discussed in the previous subsection:

$$\hat{P}_r[k] = \hat{A}_r + E(B_r[k]) \tag{12}$$

Now we shall assess how good our estimator is. We use the same 50 speech sentences (which contain 8563 speech frames) mentioned in Section 3.1 as our analysis data. For each of the 8563 speech frames, we first compute the exact value of $P_r[k]$, and then compute its estimate $\hat{P}_r[k]$ according to our proposed formulation given by eqn. 12. For comparison, we also compute $\bar{P}_r[k]$, the straight forward estimator, according to eqn. 10. Subsequently, we compute the standard error of estimating $P_r[k]$ by $\hat{P}_r[k]$, and also that by $\bar{P}_r[k]$. It turns out that our proposed estimator $\hat{P}_r[k]$ gives a significantly smaller standard error than $\bar{P}_r[k]$ for every k. In absolute terms $\bar{P}_r[k]$, gives an average standard error of 18.9dB which is more than three times that of our proposed estimator $\hat{P}_r[k]$, which is only 5.5dB.

Our objective is to obtain a good estimator for $|D_r[k]|^2$, the short-time power spectrum of the 'uncorrelated' noise. With $\hat{P}_r[k]$, which is a reasonably good estimator for $P_r[k]$, one can obtain $|\hat{D}_r[k]|^2$, the estimate for $|D_r[k]|^2$, directly using the relationship given by eqn. 9 as follows:

$$|\hat{D}_r[k]|^2 = 10^{\hat{P}_r[k]/10} \tag{13}$$

which should give us a satisfactory result for the estimation.

4.4 The complete enhancement procedure

Now we shall present the complete enhancement procedure. First, CVSD decoded speech (the noisy speech of concern) is buffered into overlapping frames, each of which is 32 msec long (overlap by 28 msec). Each frame is then multiplied by a Hamming window and transformed to the frequency domain via a fast Fourier transform (FFT). The FFT magnitude is used to compute the short-time CEP1 of the CVSD decoded speech, which is then used to compute $\hat{A}_r$, the estimate for the short-time CEP1 of the 'uncorrelated' noise via eqn. 8. Next, the estimated short-time CEP1, together with the estimated $E(\mathbf{B}_r[k])$ obtained during training (see Section 4.2), will be used to compute $\hat{P}_r[k]$ via eqn. 12, and then $|\hat{D}_r[k]|^2$ via eqn. 13. Subsequently, enhancement has to be carried out according to the spectral subtraction formulation (see eqn. 3) or the Ephraim–Malah formulation [6], with $E(|W_r[k]|^2)$ being replaced by $|\hat{D}_r[k]|^2$ to compute $|\hat{S}_r[k]|$, the SSTFT magnitude of the enhanced speech. The $|\hat{S}_r[k]|$ so obtained is then combined with the spectral phase of CVSD decoded speech to obtain the resultant spectral value followed by an inverse FFT. Finally, the time domain signals are overlap-added to obtain the enhanced speech.

5 Performance assessments

We shall now assess the performance of our method. We use 20 phonetically balanced speech sentences, of which 10 are produced by male speakers and 10 by female, taken from the TIMIT database [9]. It should be noted that these speakers and sentences are entirely different from those (25 male and 25 female sentences) used in Section 3 and 4 for training. To measure performance, we rely on both objective measures, in particular signal-to-noise ratio (SNR), segmental SNR (SEGSNR), and COSH spectral distortion measure [12], and informal subjective listening tests. The measures SNR and SEGSNR yield the 'closeness' of the processed signal to the original signal in the time-domain, whereas COSH yields that in the frequency-domain. The COSH distortion measure, although less frequently used in the speech processing community, is employed here because it can be considered as a symmetric version of the well-established Itakura–Saito measure (see [12] for advantages of using symmteric

measures). Since COSH is a distortion measure, small values will indicate less distortion. For example, a speech signal corrupted by white gaussian noise at SNRs of 20dB and 30dB will yield COSH values of about 400 and 40 respectively. (For SNR and SEGSNR, the higher the values are, the better the quality.)

We compute SNR, SEGSNR and COSH with the use of all the 20 sentences mentioned above (which contain 1633 speech frames), for the following signals: CVSD decoded speech, 'enhanced' speech obtained with spectral subtraction using $E(|D_r[k]|^2)$, the straight-forward estimator, enhanced speech obtained with spectral subtraction using $|\hat{D}_r[k]|^2$, our proposed estimator given by eqn. 13, 'enhanced' speech obtained with the Ephraim–Malah method using the straight-forward estimator, and enhanced speech obtained with the Ephraim–Malah method using our proposed estimator. Note that the computation is based on a concatenation of all sentences. The results, as tabulated in Table 1, show that the enhanced speeches obtained with both enhancement methods using our proposed estimator yielded significant improvements over the noisy CVSD decoded speech in terms of all three objective measures: SNR improves by 2.8 dB, SEGSNR improves by at least 2.6 dB, and COSH improves by at least 50 points. On the other hand, the 'enhanced' speeches obtained with both enhancement methods using the straightforward estimator result in degradation in terms of SNR and COSH, and only slight improvement in terms of SEGSNR (at most 1dB). One main reason for the poor performance of the straight-forward estimator is that it requires the noise to be stationary, but here, the noise of concern (i.e. CVSD noise) is highly nonstationary.

Table 1: Various objective measurements

	SNR (dB)	SEGSNR (dB)	COSH*
CVSD decoded speech (noisy speech	11.7	6.5	110
'Enhanced' speech by spectral subtraction using $E(\|D_r[k]\|^2)$, the straight-forward estimator	11.4	7.3	130
Enhanced speech by spectral subtraction using $\|\hat{D}_r[k]\|^2$, our proposed estimator given by eqn. 13	14.5	9.1	60
'Enhanced' speech by Ephraim-Malah method using $E(\|D_r[k]\|^2)$, the straight-forward estimator	11.5	7.5	131
Enhanced speech by Ephraim-Malah method using $\|\hat{D}_r[k]\|^2$, our proposed estimator given by eqn. 13	14.5	9.3	30

*Note that for COSH, lower value implies better quality

It is interesting to note that although spectral subtraction and the Ephraim–Malah method (with the use of our proposed estimator) give similar performance in terms of SNR, their performances differ quite significantly in terms of COSH, with the Ephraim–Malah method giving better performance. As a matter of fact, SNR does not often give reliable indication of the speech quality. Therefore, many performance measures (including SEGSNR, Itakura–Saito measure, COSH) have been proposed to supplement SNR for performance assessment. In this case, the COSH measure indicates that the Ephraim–Malah method outperforms spectral subtraction and this is further confirmed by the informal subjective listening tests (see the next paragraph).

Finally, we conduct informal subjective listening tests as follows. There are two tests and for each test, 20 listeners (10 males and 10 females) are involved. The first test is to assess which of the two methods, namely Ephraim–Malah method and spectral subtraction, are better when being used in conjunction with our proposed estimator for the noise spectrum. We first estimate the noise spectrum and then obtain the enhanced speeches using the two methods, for the 20 speech sentences. Each pair of enhanced speeches are then randomly arranged. Subsequently, the listeners are asked to assess the 20 pairs of enhanced speeches and indicate their choices for each pair of speeches. The three choices are that he/she prefers the first one, prefers the second one, has no preference. The result is that 37% preferred the Ephraim–Malah method, 18% preferred spectral subtraction, and 45% indicated no preference. In response to our further questions, some listeners mentioned that they prefer the Ephraim–Malah method since slight 'musical noise' could be heard in the enhanced speech obtained with spectral subtraction but not Ephraim–Malah method. The second test is to examine whether the enhanced speech obtained using our proposed estimator (with the Ephraim–Malah method) is preferred to the raw (original) CVSD decoded speech. The setup of this test is the same as the first one. The result is that 84% preferred the enhanced speech, 6% preferred the raw speech, and 10% indicated no preference.

6 Conclusion and discussion

Using a statistical analysis, we have obtained crucial insights into the characteristics of the disturbance introduced by CVSD. With such insights, we have developed a method for estimating the short-time power spectra of CVSD noise, and this enables us to apply existing speech enhancement methods to effectively suppress CVSD noise.

Objective assessments based on SNR, SEGSNR and COSH, and informal subjective listening tests all indicated that our method is reasonably effective. One encouraging observation is that our method works reasonably well even for speeches that are different from those used for analysis and training.

The extra delay introduced by the proposed enhancement procedure is about two speech frames (i.e., 64 msec) and this will not cause significant disturbance in some applications, especially those not involving satellite communications. On computation, we would like to highlight that considerable processing is needed for obtaining the constant c mentioned in Section 3.1, the regression curve mentioned in Section 3.2, as well as the estimator for $E(B_r[k])$ mentioned in Section 4.2. Fortunately, this processing needs to be carried out only once, for a particular class of speech. As far as the online enhancement is concerned, the computational overhead is quite low. On applying our method in real-world scenario, we would like to mention that more research effort, in addition to that reported in this paper, is needed. For example, when implementing the method on a mobile system, other disturbances such as background noise, channel noise, electrical noise, etc.

will have to be taken into consideration. Some of these issues will be looked into in future work.

7 Acknowledgments

The authors are grateful to the anonymous reviewers for their useful and encouraging comments.

8 References

1 JAYANT, N.S., and NOLL, P.: 'Digital coding of waveforms' (Prentice-Hall, 1994), pp. 372–427
2 UN, C.K., and LEE, H.S.: 'A study of the comparative performance of adaptive delta modulation systems', *IEEE Trans. Commun.*, 1980, **28**, pp. 96–101
3 CAMPBELL, J.P., TREMAIN, T.E., and WELCH, V.C.: 'The proposed federal standard 1016 - 4800 bps voice coder: CELP', *Speech Technol. (USA)*, 1990, pp. 58–64
4 HARDWICK, J.C., and LIM, J.S.: 'A 4.8 kbps multi-band excitation speech coder'. Proceedings of the IEEE international conference on *Acoustics, speech and signal processing*, 1988, pp. 374–377
5 LIM, J.S., and OPPENHEIM, A.V.: 'Enhancement and bandwidth compression of noisy speech', *Proc. IEEE (USA)*, 1979, **67**, pp. 1586–1604
6 EPHRAIM, Y., and MALAH, D.: 'Speech enhancement using a minimum mean-square error short-time spectral amplitude estimator', *IEEE Trans. Acoust. Speech Signal Process.*, 1984, **32**, pp. 1109–1121
7 STEELE, R.: 'Delta modulation systems' (Pentech Press, London, 1975(, pp. 78–117
8 JAYANT, N.S.: 'A first-order Markov model for understanding delta modulation noise spectra', *IEEE Trans. Commun.*, 1978, pp. 1316–1318
9 National institute of standards and technology (NIST), DARPA TIMIT acoustics-phonetic continuous speech corpus, NIST Speech Disc 1-1.1, 1990
10 GOH, Z., TAN, K.C., and TAN, B.T.G.: 'A post-processing procedure for adaptive delta-modulation'. Proceedings of the IEEE Singapore international conference on *Communication systems*, 1996, pp. I:249–252
11 GUNST, R.F., and MASON, R.L.: 'Regression analysis and its application' (Marcel Dekker, Inc., 1980), pp. 52–84
12 RABINER, L., and JUANG, B.H.: 'Fundamentals of speech recognition' (Prentice Hall, 1993), pp. 154–194