Proc. IEEE Int. Conf. on Acoustics, Speech, Signal Processing (Atlanta, GA), pp. 761-764, May 1996

CLASSIFIED NONLINEAR PREDICTIVE VECTOR QUANTIZATION OF SPEECH SPECTRAL P ARAMETERS

James H.Y. Loo [†,1] *Wai-Yip Chan* [‡] *Peter Kabal* [†]

[†]Department of Electrical Engineering, McGill University, Montreal, Quebec, Canada H3A 2A7
[‡]Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616-3793

ABSTRACT

Nonlinear predictive split vector quantization (NPSVQ) and classified NPSVQ (CNPSVQ) are introduced to exploit the correlation among the speech spectral parameters from two adjacent analysis frames. By interleaving intraframe SVQ with forward predictive SVQ, error propagation is limited to at most one adjacent frame. At an overall bit rate of about 21 bits/frame, NPSVQ can provide similar coding quality as intraframe SVQ at 24 bits/frame. Voicing classification is used in CNPSVQ to obtain an additional average gain of 1 bit/frame for unvoiced frames. Therefore, an overall bit rate of 20 bits/frame is obtained for unvoiced frames. The particular form of nonlinear prediction we use incurs virtually no additional encoding computational complexity. We have verified our comparative performance results using subjective listening tests.

1. INTRODUCTION

In low-bit-rate speech coding, the short-term spectral envelope of the speech signal is often modeled by the magnitude frequency response of an all-pole synthesis filter. The filter coefficients are usually obtained by performing a linear prediction (LP) analysis of a frame of input speech signal. The filter coefficients are then quantized with sufficient accuracy to maintain speech intelligibility and quality. Numerous quantization schemes have been explored in pursuit of higher spectral coding efficiency. Grass and Kabal [1] explored using *vector-scalar quantization* at 20–30 bits/frame[2]. Paliwal and Atal [2] demonstrated that *transparent coding* quality can be achieved using *split vector quantization* (SVQ) at about 24 bits/frame. Paksoy *et al.* [3] obtained a bit rate of 21 bits/frame by employing rather elaborate VQ techniques. The above schemes all employ line spectral frequency (LSF) representation of the filter coefficients and are recent examples of *intraframe* coding.

Intraframe coding using the same quantizer for all frames ignores the non-stationary statistics and perceptual modality of the speech signal. *Multimodal* or *classified* coding has been used to improve performance wherein the coder changes its configuration in accordance with the *class* of the speech signal being processed. For different classes, the bit allocations among coder components may vary, and so may the number of bits generated per frame. A simple *voicing* classification strategy is to distinguish between a voiced (V) and an unvoiced (UV) frame of speech. Some speech coders already transmit such voicing information as part of their encoded data. For instance, as part of its multimodal coding strategy, the GSM half-rate standard speech coder [4] transmits two mode bits to indicate the strength of voicing for each frame.

Interframe coding can also be used to improve coding efficiency by exploiting the temporal redundancy of the LP spectral envelopes. Farvardin and Laroia [5] reported "strong" correlation between neighbouring frames (10 ms frame-shift interval) of LSF parameters. Unfortunately, prediction that is based on the recursive reconstructions of the decoder can suffer from the propagation of channel errors over numerous frames. Ohmuro *et al.* [6] proposed a *moving average* (MA) prediction scheme that can limit error propagation to a number of frames given by the order of the MA predictor. In a similar direction, de Marca [7] explored a scheme wherein the LSF parameters of every other frame are intraframe coded with SVQ; the LSF parameters of an intervening frame are linearly predicted from the quantized LSF parameters of the previous frame and the prediction residual vector is then coded with SVQ. Thus, if the bits of a quantized LSF vector contain errors, no more than two adjacent frames will be affected (actually, the adverse effect might propagate further through the memory of the synthesis filter). For *transparent coding* quality [2], de Marca reported an average bit rate of 27 bits/frame. In de Marca's scheme, the prediction is a function of a quantized LSF vector. In the scheme of Ohmuro *et al.*, the prediction is a function of the quantized prediction residuals of several frames. Thus, de Marca's scheme has the potential of furnishing a higher prediction gain; however, this gain is offset by the fact that only half of the frames are predictively coded. If longer error propagation can be tolerated, intraframe coding can be used less often while the intervening frames are all (recursively) interframe coded.

2. LINEAR PREDICTIVE SVQ

In principle, nonlinear prediction can outperform linear prediction. To gauge the performance gain from nonlinear prediction, we employ de Marca's non-recursive prediction framework [7]. The framework is depicted in Figure 1, where the box labeled "predictor" is instrumented as a linear scalar predictor in de Marca's scheme and as a nonlinear vector predictor in our scheme. We now describe the frame-

[0]This work was supported in part by the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche, the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Telecommunications Research, and the National Science Foundation.

[1]James Loo is currently with Bell-Northern Research, 16 Place du Commerce, Verdun, Quebec, Canada H3E 1H6.

[2]In this paper, unless otherwise specified, a frame is one of a sequence of consecutive 20 ms speech segments.

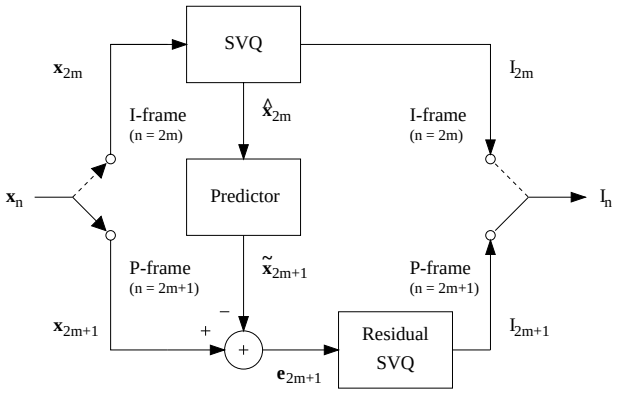

Figure 1. Predictive SVQ Encoder

work without specifying whether the predictor is linear or nonlinear. Let $\{\mathbf{x}_n\}$ be a sequence of 10-dimensional LSF vectors to be encoded. The LSF vectors are grouped together into contiguous pairs $(\mathbf{x}_{2m}, \mathbf{x}_{2m+1})$. The vector $\mathbf{x}_{2m}$ is encoded with intraframe SVQ and is said to be from an *intra-coded frame* (I-frame). We let $\hat{\mathbf{x}}_{2m}$ denote the quantized $\mathbf{x}_{2m}$ and $\mathbf{I}_{2m}$ denote the corresponding codevector indices that are transmitted to the decoder.

From $\hat{\mathbf{x}}_{2m}$, the predictor generates a prediction $\tilde{\mathbf{x}}_{2m+1}$ of the LSF vector $\mathbf{x}_{2m+1}$, which is said to be from a *predicted frame* (P-frame). The prediction error vector $\mathbf{e}_{2m+1} = \mathbf{x}_{2m+1} - \tilde{\mathbf{x}}_{2m+1}$ is then quantized to $\hat{\mathbf{e}}_{2m+1}$ using a different SVQ ("residual SVQ" in Figure 1); the transmitted codevector indices are $\mathbf{I}_{2m+1}$. The quantized LSF vector of the P-frame can be reconstructed by adding the quantized residual to the prediction:

$$\hat{\mathbf{x}}_{2m+1} = \tilde{\mathbf{x}}_{2m+1} + \hat{\mathbf{e}}_{2m+1}.$$

This encoding process is then repeated by alternately applying intraframe SVQ to $\mathbf{x}_{2m}$ and predictive SVQ to $\mathbf{x}_{2m+1}$ for $m = 1, 2, 3, \ldots$.

Given an LSF vector sequence $\{\mathbf{x}_n\}$, the M-th order *vector linear predictor* generates a prediction $\tilde{\mathbf{x}}_n$ of the current vector $\mathbf{x}_n$ based on M preceding reconstructed vectors $\hat{\mathbf{x}}_{n-i}$, as

$$\tilde{\mathbf{x}}_n = \sum_{i=1}^{M} P_i \hat{\mathbf{x}}_{n-i}$$

where $P_i, i = 1, \ldots, M$ are 10×10 prediction matrices. When all M prediction matrices are diagonal, vector linear prediction reduces to the special case of scalar linear prediction. As in de Marca's scheme [7], which we call *scalar linear predictive SVQ* (PSVQ), each LSF vector component in the P-frame is predicted only from the corresponding (quantized) LSF vector component in the preceding I-frame. When the prediction matrices are not diagonal, we have vector linear prediction; the predictive SVQ framework then specializes to *vector predictive SVQ* (VPSVQ). Vector linear prediction additionally exploits the intercomponent correlation, if any, between neighbouring frames. We have explored both scalar and vector linear prediction but the order M of the prediction has been restricted to one. By exploiting the interframe redundancy of LSF parameters, fewer bits are required to encode the prediction residual vector in the P-frames than those required to encode the LSF vector in the I-frames. However, if every frame must be allocated the same number of bits for spectral coding, an additional buffering delay of one frame would be incurred.

3. NONLINEAR PREDICTIVE SVQ

In our work, we have sought to determine whether additional performance gain can be garnered by using nonlinear instead of linear prediction, and whether nonlinear prediction combined with voicing classification offers performance gain over classified intraframe coding. When a nonlinear predictor is used for the "predictor" block in the structure of Figure 1, we say that the P-frame vector is encoded using *nonlinear predictive SVQ* (NPSVQ). Our nonlinear prediction scheme is based on applying Gersho's nonlinear interpolative VQ [8] to SVQ and multistage VQ structures [9]. We note that the minimum mean-square error (MMSE) estimate (prediction) $\tilde{\mathbf{Y}}$ of a random vector $\mathbf{Y}$ given another random vector (observation) $\mathbf{X}$ is the conditional expectation of $\mathbf{Y}$ given $\mathbf{X}$:

$$\tilde{\mathbf{Y}}(\mathbf{X}) = E[\mathbf{Y}|\mathbf{X}].$$

If the joint probability distribution of $\mathbf{X}$ and $\mathbf{Y}$ is not known, we can generally assume that the conditional expectation is a nonlinear function. If the observation $\mathbf{X}$ is quantized to a finite set of possible values $\{\hat{\mathbf{x}}^{(i)}\}$, there is also only a finite number of possible conditional expectation values $\{\tilde{\mathbf{y}}^{(i)}\}$, where

$$\tilde{\mathbf{y}}^{(i)} = E[\mathbf{Y}|\hat{\mathbf{x}}^{(i)}].$$

Thus, even without knowing the functional form of the MMSE estimator, a table of conditional expectation values can be found as part of the process of designing the quantizer for $\mathbf{X}$ [8]. Accordingly, associated with the i-th partition region of the VQ encoder is the decoder output $\hat{\mathbf{x}}^{(i)}$ as well as the MMSE estimate value $\tilde{\mathbf{y}}^{(i)}$. It is straightforward to extend the above VQ-based nonlinear prediction scheme to first-order nonlinear prediction of LSF vectors: replace the random vectors $\mathbf{X}$ and $\mathbf{Y}$ with the I-frame and P-frame LSF vectors respectively.

In NPSVQ, the nonlinear predictor is constructed as a codebook of conditional expectations, one for each distinct value of the quantized observation $\hat{\mathbf{x}}_{2m}$. However, there can be as many as 2^b distinct values for $\hat{\mathbf{x}}_{2m}$ where b is the number of bits of the I-frame SVQ (e.g. $b = 24$). We have chosen to use the same *product codebook* [9] structure for the predictor as for the I-frame SVQ. If $\mathbf{x}_{2m}$ is quantized to $\hat{\mathbf{x}}_{2m}$ using L-way SVQ (L-SVQ), then the prediction $\tilde{\mathbf{x}}_{2m+1}$ will also be split into L subvectors in exactly the same manner as splitting $\hat{\mathbf{x}}_{2m}$. Hence, for each distinct value of a subvector of $\hat{\mathbf{x}}_{2m}$, we assign one value (obtained during codebook training) to the corresponding subvector of the prediction $\tilde{\mathbf{x}}_{2m+1}$. In this work, for each L-SVQ configuration we explore, the I-frame, P-frame and prediction vectors are all split in identical fashions. Although NPSVQ requires more codebook storage than PSVQ, the computational complexity remains virtually unchanged.

4. CLASSIFIED NONLINEAR PREDICTIVE SVQ

We have explored using voicing classification to enhance our spectral coding schemes. Our voicing algorithm is based on the classifier used in the U.S. Federal Standard 1015 (LPC-10E) vocoder [10], where each speech frame is labeled as either voiced (V) or unvoiced (UV). As unvoiced speech does

not generally exhibit a distinct pattern of formants, fewer bits may be employed to encode the LSF vectors from UV frames than those from V frames. Hagen *et al.* [11] reported that, based on subjective listening tests, transparent coding quality can be achieved in a CELP coding context by using 9 bits for spectral coding in the unvoiced frames (corresponding to an average spectral distortion for unvoiced frames of 2.1 dB) and 24 bits in the voiced frames.

With classification, the intraframe SVQ for the I-frame becomes *classified SVQ* (CSVQ), wherein different sets of SVQ codebooks are used for the V and UV classes. Since there are four possible combinations of V/UV classifications for the I-frame and the P-frame jointly, there are four corresponding sets of nonlinear predictor and P-frame NPSVQ codebooks. We call this classification-enhanced scheme *classified nonlinear predictive SVQ* (CNPSVQ).

5. PERFORMANCE RESULTS

Our performance results are based on a *training set* and a separate *test set* [3] [12] of LSF vectors. A database of approximately 24.5 minutes of silence-removed speech, which has been lowpass filtered at 3.4 kHz and sampled at 8 kHz, is used to construct the training sequence. An additional 2.5 minutes of similarly filtered speech are used for the test set. Tenth order LPC analysis is performed using the modified covariance method with high frequency compensation. A 20 ms Hamming window is used for analysis over every consecutive 20 ms time interval. With non-overlapping analysis windows, the correlation between adjacent LSF vectors is kept to a minimum.

We have experimented with two different splitting configurations. In one configuration (2-SVQ), the I-frame vector is split into two subvectors of dimensions (4, 6). In the other configuration (3-SVQ), the I-frame vector is split into three subvectors of dimensions (3, 3, 4). With the training set, we first design the I-frame SVQ codebooks and their corresponding nonlinear predictor codebooks [8]. A set of residual training vectors is then obtained by subtracting the prediction vector for the n-th frame from the LSF training vector of the n-th frame, where n indexes all the vectors in the training set. The codebooks for the P-frame SVQ are then designed using the set of residual training vectors.

We first compare first-order linear and nonlinear predictive SVQ using the prediction gain data tabulated in Table 1. The prediction gain for the i-th LSF vector component, $PG^{(i)}$, is measured as the ratio in dB between the sample variance of component $x^{(i)}$ and that of its prediction residual:

$$PG^{(i)} = 10 \log_{10} \frac{\sigma^2_{x^{(i)}}}{E[(x_n^{(i)} - \tilde{x}_n^{(i)})^2]} \text{ dB}.$$

The prediction gain is measured as a function of the predictor type and splitting configuration. For both 2-SVQ or 3-SVQ, the I-frame LSF vector is quantized using 24 bits. Table 1 suggests that nonlinear prediction outperforms both scalar and vector linear prediction. The advantage of nonlinear over linear prediction is greater for 2-SVQ than 3-SVQ, and also greater for the lower order than the higher order LSFs. Moreover, the coarser quantization of the I-frame due to using 3-SVQ instead of 2-SVQ degrades the prediction gain.

Table 2 presents spectral distortion (SD) [2] measurements for 2-SVQ and 3-SVQ at bit rates of 21–24 bits/frame. These measurements are compared with the SD data for PSVQ and NPSVQ in Table 3. The bit allocation for the I-frame is kept constant at 24 bits while that for the P-frame is varied. Overall, our results indicate that nonlinear predictive SVQ provides a modest improvement over linear predictive SVQ. In terms of average SD performance, NPSVQ achieves a gain of 5–6 bits for the P-frames or an average gain of up to 3 bits for all frames. While the amounts of spectral outliers for NPSVQ at overall rates of 21–22 bits/frame are higher than that for SVQ at 24 bits/frame, they still fall below the amounts for SVQ at 21–22 bits/frame.

Predictor Type	Prediction Gain (dB)				
	LSF 1	LSF 2	LSF 3	LSF 4	LSF 5
2-PSVQ	4.35	3.94	3.59	4.22	5.99
2-VPSVQ	4.37	4.02	3.78	4.37	6.06
2-NPSVQ	4.83	4.55	4.31	4.76	6.38
3-PSVQ	4.19	3.88	3.56	4.14	5.85
3-VPSVQ	4.21	3.96	3.73	4.29	5.91
3-NPSVQ	4.31	4.10	3.78	4.29	5.94
Predictor Type	Prediction Gain (dB)				
	LSF 6	LSF 7	LSF 8	LSF 9	LSF 10
2-PSVQ	5.26	4.68	4.37	3.51	3.09
2-VPSVQ	5.42	4.75	4.45	3.57	3.15
2-NPSVQ	5.76	5.13	4.79	3.82	3.33
3-PSVQ	5.06	4.70	4.44	3.77	2.80
3-VPSVQ	5.26	4.78	4.52	3.83	2.91
3-NPSVQ	5.17	4.81	4.58	3.86	3.05

Table 1. Prediction gain for each of the 10 LSF vector components as a function of the predictor type and splitting configuration. The I-frame vector is encoded using 24 bits.

SD performance results for 3-CSVQ and 3-CNPSVQ are shown in Tables 4 and 5, respectively. With 3-CSVQ, we obtain the same SD performance using 22 bits for UV frames and 24 bits for V frames (corroborating the results of Hagen *et al.* [11]). Consequently, 3-CNPSVQ (Table 5) achieves an additional gain of 2 bits per UV I-frame. While NPSVQ at 21–22 bits/frame can provide similar coding quality as SVQ at 24 bits/frame, CNPSVQ can yield equivalent results at 20–21 bits/frame for unvoiced speech and at 21–22 bits/frame for voiced speech.

6. SUBJECTIVE LISTENING TESTS

In addition to obtaining SD measurements, we also performed listening tests on the reconstructed speech. Speech is reconstructed using a synthesis filter with quantized coefficients, and with the filter excited by the unquantized linear prediction residual signal. The tests were conducted with 12 listeners using 4 different test-set sentences from a male speaker and a female speaker. In each test, a listener would listen to the original sentence and two encoded versions of the sentence. The listener was then asked to choose which encoded version was more similar to the reference.

When asked to choose between 2-SVQ at 24 bits/frame and 2-NPSVQ at 21 bits/frame, the listeners chose 2-NPSVQ over 2-SVQ in 52% of the test cases. In a comparison of 3-SVQ (24 bits/frame) and 3-NPSVQ (21 bits/frame), 3-NPSVQ was preferred over 3-SVQ approximately 44% of the time. The listeners favoured 3-CNPSVQ, using 21 bits for UV frames and 22 bits for V frames, in 56% of the test trials over 3-CSVQ, using 22 bits for UV frames and 24 bits for V frames. However, in only 40% of the cases was 3-CNPSVQ at 20 bits for UV frames and 21 bits for V frames preferred over 3-CSVQ.

7. CONCLUSION

When NPSVQ is interleaved with intraframe SVQ, nonlinear prediction furnishes an average gain of 3 bits/frame relative to no prediction (24-bit intraframe SVQ for all frames). With NPSVQ, error propagation is limited to at most one adjacent frame. By classifying frames as U or UV and employing class-specific SVQ, the number of bits for the UV frames can be reduced by 2. When voicing classification is combined with NPSVQ, the coding gains for classification and nonlinear prediction are additive.

REFERENCES

[1] J. Grass and P. Kabal, "Methods of improving vector-scalar quantization of LPC coefficients," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, Toronto, pp. 657–660, May 1991.

[2] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *IEEE Trans. Speech and Audio Proc.*, pp. 3–14, January 1993.

[3] E. Paksoy, W.-Y. Chan and A. Gersho, "Vector quantization of speech LSF parameters with generalized product codes," *Proc. Int. Conf. Spoken Language Proc.*, Banff, Canada, pp. 33–36, October 1992.

[4] W.-Y. Chan, I. A. Gerson and T. Miki, "Half-rate standards," in *The Mobile Communications Handbook*, J. D. Gibson, ed., CRC Press, 1995.

[5] N. Farvardin and R. Laroia, "Efficient encoding of speech LSP parameters using the discrete cosine transform," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, Glasgow, pp. 168–171, May 1989.

[6] H. Ohmuro, T. Moriya, K. Mano and S. Miki, "Coding of LSP parameters using interframe moving average prediction and multi-stage vector quantization," *IEEE Workshop Speech Coding for Telecom.*, Sainte-Adèle, Canada, pp. 63–64, October 1993.

[7] J. R. B. de Marca, "An LSF quantizer for the North-American half-rate speech coder," *IEEE Trans. Vehic. Tech.*, pp. 413–419, August 1994.

[8] A. Gersho, "Optimal nonlinear interpolative vector quantization," *IEEE Trans. Comm.*, vol. COM-38, no. 9-10, pp. 1285–1287, September 1990.

[9] W.-Y. Chan and A. Gersho, "Generalized product code vector quantization: A family of efficient techniques for signal compression," *Digital Signal Processing*, pp. 95–126, April 1994.

[10] J. P. Campbell, Jr. and T. E. Tremain, "Voiced/unvoiced classification of speech and applications to the U.S. Government LPC-10E algorithm," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, Tokyo, pp. 473–476, April 1986.

[11] R. Hagen, E. Paksoy and A. Gersho, "Variable rate spectral quantization for phonetically classified CELP coding," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, Detroit, pp. 748–751, May 1995.

[12] W.-Y. Chan and D. Chemla, "Low-complexity encoding of speech LSF parameters using constrained-storage TSVQ," *Proc. Int. Conf. Acoust., Speech, Signal Proc.*, Adelaide, pp. I-521–I-524, April 1994.

SVQ Config.	Bits per Frame	Average SD (dB)	SD Outliers (%) 2-4 dB	SD Outliers (%) > 4 dB
2-SVQ	24	1.17	2.38	0.03
2-SVQ	23	1.24	2.65	0.03
2-SVQ	22	1.34	4.69	0.03
2-SVQ	21	1.40	6.01	0.04
3-SVQ	24	1.23	2.38	0.03
3-SVQ	23	1.30	2.87	0.03
3-SVQ	22	1.35	3.55	0.03
3-SVQ	21	1.46	8.18	0.03

Table 2. Spectral distortion performance of SVQ.

P-Frame Quantizer	(I,P) Bit Allocation	Average SD (dB)	SD Outliers (%) 2-4 dB	SD Outliers (%) > 4 dB
2-PSVQ	(24,24)	1.01	1.64	0.01
2-PSVQ	(24,19)	1.16	4.83	0.09
2-PSVQ	(24,18)	1.20	5.91	0.14
2-NPSVQ	(24,24)	1.01	1.69	0.03
2-NPSVQ	(24,19)	1.16	4.70	0.08
2-NPSVQ	(24,18)	1.20	5.60	0.12
P-Frame Quantizer	(I,P) Bit Allocation	Average SD (dB)	SD Outliers (%) 2-4 dB	SD Outliers (%) > 4 dB
3-PSVQ	(24,24)	1.08	2.29	0.03
3-PSVQ	(24,19)	1.21	4.83	0.09
3-PSVQ	(24,18)	1.25	5.91	0.14
3-NPSVQ	(24,24)	1.07	2.26	0.03
3-NPSVQ	(24,19)	1.20	4.70	0.08
3-NPSVQ	(24,18)	1.24	5.60	0.12

Table 3. Spectral distortion performance of linear and nonlinear predictive SVQ.

CSVQ Config.	UV Frames Bits	UV Frames SD (dB)	Average SD (dB)	SD Outliers (%) 2-4 dB	SD Outliers (%) >4 dB
3-CSVQ	24	1.18	1.25	2.24	0.02
3-CSVQ	23	1.24	1.27	2.41	0.02
3-CSVQ	22	1.29	1.29	2.66	0.02

Table 4. Spectral distortion performance of classified intraframe 3-SVQ. Only the results for the unvoiced (UV) frames are shown. The voiced (V) frames are encoded with 24 bits, yielding an average SD of 1.28 dB.

P-Frame Quantizer	P-frm. Bits UV	P-frm. Bits V	Average SD (dB)	SD Outliers (%) 2-4 dB	SD Outliers (%) > 4 dB
3-CNPSVQ	24	24	1.11	2.62	0.04
3-CNPSVQ	19	19	1.24	4.70	0.13
3-CNPSVQ	18	19	1.26	5.03	0.16
3-CNPSVQ	19	18	1.26	5.30	0.14
3-CNPSVQ	18	18	1.28	5.49	0.17

Table 5. Spectral distortion performance of classified nonlinear predictive 3-SVQ. Bit-allocations for the I-frame are kept constant at 22 bits for an unvoiced (UV) I-frame and 24 bits for a voiced (V) I-frame. Only the P-frame bit allocations are varied in the table.