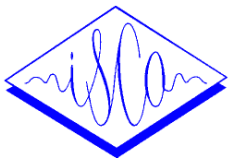

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

A TWO-STAGE CODING OF SPEECH LSP PARAMETERS BASED ON KLT TRANSFORM AND 2D-PREDICTION

Fu-Rong Jean and Hsiao-Chuan Wang

Department of Electrical Engineering, National Tsing Hua University,
Hsinchu, Taiwan, 30043, Republic of China

ABSTRACT

In this paper, a two-stage approach based on KLT transform and 2D prediction is proposed for encoding the LSP parameters. At the first stage, KLT transform is applied to the even part of LSP parameters which are mean-removed. Then DPCM coding is performed on the transformed parameters. At the second stage, the even part of LSP parameters is reconstructed, and a third-order two-dimensional prediction is used to estimate the odd part of LSP parameters. This approach considers both interframe and intraframe correlations simultaneously. The simulation on the database provided by different speakers shows that 1 dB^2 difference limen of spectral distortion can be achieved at 19 bits per frame by using the partitioned vector quantization (PVQ) to the residuals of LSP parameters. The residuals of LSP parameters are partitioned into even part and odd part according to the processing of stage sequence. The outlier frames which have spectral distortion greater than 2 dB are 3.63 %. A switched classifier is built to reduce the outlier frames down to about 0.27 % and make no frames with spectral distortion greater than 4 dB at an average bit-rate below 19 bits per frame.

I. INTRODUCTION

The transparent quantization of LPC information was proposed in [1], it means that the LPC quantization does not introduce any audible distortion in the coded speech if the following requirements are satisfied simultaneously: (1) the average spectral distortion is about 1 dB, (2) there is no outlier frames having spectral distortion larger than 4 dB, and (3) the number of outlier frames having spectral distortion in the range 2-4 dB is less than 2 %. *Line spectrum pair* (LSP) is a set of parameters derived from LPC parameters which was first introduced by Itakura [2]. In low bit-rate speech coders, the LSP parameters were found to be efficient in representing the short-time spectrum envelope information for their good quantization and interpolation characteristics.

This paper proposes a new LSP encoding method not only to reduce the bit-rate at the 1 dB difference limen but also to reduce the number of outlier frames. To solve the former problem, a two-stage approach based on KLT transform and 2D prediction is proposed for this purpose. To solve the latter problem, a switched classifier is built to reduce the outlier frames and make no frames with spectral distortion greater than 4 dB.

II. TWO-STAGE CODING (TSC) SCHEME

2.1 Distortion Measure

Given a p-th order LPC model, the parametric distortion defined by

$$PD_n^2 = \sum_{i=1}^{p} g_i^2(n)\left(\omega_i(n) - \hat{\omega}_i(n)\right)^2 \qquad (1)$$

is used as the design criterion in this study. Where $\omega_i(n)$ is the i-th parameter of the n-th frame LSP vector, and $\hat{\omega}_i(n)$ is its reconstructed value. The adaptive weighting factor $g_i^2(n)$ is the so-called *spectral sensitivity* with respect to $\omega_i(n)$ [3], which is defined as

$$g_i^2(n) = \frac{1}{\pi}\int_0^{\pi}\left|\frac{\partial \log S_n(\omega)}{\partial \omega_i(n)}\right|^2 d\omega . \qquad (2)$$

Furthermore, in this study we divide the parametric distortion into two parts for convenience, i.e.,

$$PD_n^2 = \sum_{i=1}^{p/2} g_{2i}^2(n)\left(\omega_{2i}(n) - \hat{\omega}_{2i}(n)\right)^2 + \sum_{i=1}^{p/2} g_{2i-1}^2(n)\left(\omega_{2i-1}(n) - \hat{\omega}_{2i-1}(n)\right)^2$$
$$= PD_{n,E}^2 + PD_{n,O}^2, \qquad (3)$$

where $PD_{n,E}^2$ is the portion of the spectral weighted quantization error produced at the first stage and $PD_{n,O}^2$ is the portion at the second stage.

2.2 First Stage of TSC Scheme

The LSP vector at the n-th frame of speech signals is denoted by

$$\boldsymbol{\omega}(n) = \left[\omega_1(n) \quad \omega_2(n) \quad \cdots \quad \omega_p(n)\right]^{\mathrm{T}}, \qquad (4)$$

where p is the order of the LPC model and is assumed to be an even integer. The even part of this vector is noted by

$$\boldsymbol{\omega}_{\mathrm{E}}(n) = \left[\omega_2(n) \quad \omega_4(n) \quad \cdots \quad \omega_p(n)\right]^{\mathrm{T}}. \qquad (5)$$

Let the mean vector of $\boldsymbol{\omega}_{\mathrm{E}}(n)$ is defined by

$$\bar{\boldsymbol{\omega}}_{\mathrm{E}} = \left[\bar{\omega}_2\, \bar{\omega}_4 \quad \cdots \quad \bar{\omega}_p\right]^{\mathrm{T}}, \qquad (6)$$

where $\bar{\omega}_i$ is the mean value of $\omega_i(n)$, i.e., $\bar{\omega}_i = E\left[\omega_i(n)\right]$.

The orthogonal KLT transform matrix, $\mathbf{A} = \left[a_{ij}\right]$, is performed on the mean removed even part of LSP vector,

$$\boldsymbol{\Omega}_E(n) = \left[\Omega_2(n) \quad \Omega_4(n) \quad \cdots \quad \Omega_p(n)\right]^{\mathrm{T}} = \mathbf{A}\left(\boldsymbol{\omega}_E(n) - \bar{\boldsymbol{\omega}}_E\right). \qquad (7)$$

For each $\Omega_i(n)$, an m_i-th order linear predictor is used to calculate its residual,

$$r_i(n) = \Omega_i(n) - \sum_{j=1}^{m_i} h_i(j)\hat{\Omega}_i(n-j), \quad for \quad i = 2,4,\cdots,p, \qquad (8)$$

where $h_i(j), j = 1,2,\cdots,m_i$ are prediction coefficients of the i-th linear predictor.

The criterion to determine the optimal codeword (reconstruction levels) of a quantizer is to minimize the

This research has been partially sponsored by National Science Council, Taiwan, R.O.C. under contract no. NSC-82-0408-E-007-319.

parametric distortion. $PD^2_{n,E}$ is the parametric distortion produced at the first stage, and can be deduced as follows:

$$\begin{aligned} PD^2_{n,E} &= \sum_{i=1}^{p/2} g^2_{2i}(n)\left(\omega_{2i}(n)-\hat{\omega}_{2i}(n)\right)^2 \\ &= \sum_{i=1}^{p/2}\left\{\eta^2_{2i}(n)\left(\Omega_{2i}(n)-\hat{\Omega}_{2i}(n)\right)^2\right\}+ \\ &\quad \sum_{i=1}^{p/2}\sum_{j=i+1}^{p/2}\left\{2\eta_{2i,2j}(n)\left(\Omega_{2i}(n)-\hat{\Omega}_{2i}(n)\right)\left(\Omega_{2j}(n)-\hat{\Omega}_{2j}(n)\right)\right\} \\ &= \sum_{i=1}^{p/2}\left\{\eta^2_{2i}(n)\left(r_{2i}(n)-\hat{r}_{2i}(n)\right)^2\right\}+ \\ &\quad \sum_{i=1}^{p/2}\sum_{j=i+1}^{p/2}\left\{2\eta_{2i,2j}(n)\left(r_{2i}(n)-\hat{r}_{2i}(n)\right)\left(r_{2j}(n)-\hat{r}_{2j}(n)\right)\right\}, \end{aligned} \quad (9)$$

where

$$\eta^2_{2i}(n)=\sum_{k=1}^{p/2}\left(g^2_{2k}(n)a^2_{ik}\right),$$

$$\eta_{2i,2j}(n)=\sum_{k=1}^{p/2}\left(g^2_{2k}(n)a_{ik}a_{jk}\right). \quad (10)$$

2.3 Second Stage of TSC Scheme

After the even part of LSP parameters is quantized, a third-order two-dimensional prediction is used to estimate the odd part of LSP parameters, i.e., the estimate of $\omega_i(n)$ is based on the neighbors $\hat{\omega}_{i-1}(n)$ and $\hat{\omega}_{i+1}(n)$ in the same frame and also the neighbor $\hat{\omega}_i(n-1)$ of previous frame.

$$\tilde{\omega}_i(n)=\alpha_i\hat{\omega}_{i-1}(n)+\beta_i\hat{\omega}_i(n-1)+\gamma_i\hat{\omega}_{i+1}(n), \quad for \quad i=1,3,\cdots,p-1, \quad (11)$$

where α_i, β_i and γ_i are the prediction coefficients, which can be obtained from a long sequence of speech signals by minimizing the squared prediction error,

$$E_i=\sum_{n=1}^{N} r_i^2(n)=\sum_{n=1}^{N}[\omega_i(n)-\tilde{\omega}_i(n)]^2, \quad for \quad i=1,3,\cdots,p-1, \quad (12)$$

where N is the total frame number of speech signals.

Since $PD^2_{n,O}$ is the parametric distortion produced at the second stage, it is reasonable to determine the codeword (reconstruction levels) at this stage so that $PD^2_{n,O}$ is minimum. $PD^2_{n,O}$ can be expressed in terms of prediction residuals and corresponding reconstructed values as follows.

$$\begin{aligned} PD^2_{n,O} &= \sum_{i=1}^{p/2} g^2_{2i-1}(n)\left(\omega_{2i-1}(n)-\hat{\omega}_{2i-1}(n)\right)^2 \\ &= \sum_{i=1}^{p/2} g^2_{2i-1}(n)\left(r_{2i-1}(n)-\hat{r}_{2i-1}(n)\right)^2. \end{aligned} \quad (13)$$

Finally, the reconstructed LSP vector must satisfy the ordering properties of the LSP parameters to ensure the stability of the short-term filter. The following relationship exists for all n:

$$0<\hat{\omega}_1(n)<\hat{\omega}_2(n)<\cdots<\hat{\omega}_{p-1}(n)<\hat{\omega}_p(n)<\pi. \quad (14)$$

III. QUANTIZATION SCHEMES

In encoding procedure, the residuals are quantized and transmitted every frame. Two quantization schemes will be described and simulated in this paper. They are the scalar quantization (SQ) and the partitioned vector quantization (PVQ).

3.1 Scalar Quantization (SQ)

Each residual r_i has its own optimal reconstruction levels in scalar quantization. The dynamic programming method proposed by Soong and Juang [4] is used to train each globally optimal quantizer. The optimally integer-constrained bit-allocation procedure [5] is adopted to determine the bit assignment.

3.2 Partitioned Vector Quantization (PVQ)

Vector quantization usually performs better than scalar quantization at the cost of much more storage and computational load. Since a single global residual codebook is not realizable, It needs to use two codebooks in PVQ, one for the even part of residuals at the first stage, and the other for the odd part of residuals at the second stage. Each codebook is generated by using the well-known generalized Lloyd method [6]. The order of codebook for even part is equal to or greater by one bit than that for odd part.

IV. EXPERIMENT AND SIMULATION RESULTS

4.1 Speech Database

The training data in the simulation consist of 12141 frames spoken by four speakers (2 males and 2 females). Each speaker uttered 11 sentences. The speech signals were digitized at 8 kHz sampling rate with 16 bits/sample. The order of the LPC model was 10. Table I shows the detailed conditions of the training procedure. In this paper, both inside and outside training speech sequences are tested. For inside test, the test data are randomly selected from the training data provided by 1 male and 1 female speakers. It consists of 4675 spoken frames. An independent test database spoken by different speakers (1 male and 1 female) is used for outside test. From which 4844 frames are selected for this purpose.

The histograms of LSP parameters are depicted in Fig. 1(a). ω_5 and ω_6 have larger variability. The histograms of residuals are shown in Fig. 1(b). Each r_i has a dynamic range much less than that ω_i has. This implies that a lot of bits can be saved when r_i is used instead of ω_i. Furthermore, the more symmetrical and regular shape of r_i's distribution is beneficial to quantization.

TABLE I
EXPERIMENTAL CONDITIONS OF THE TRAINING PROCEDURE

Speakers	2 Males and 2 Females
Sentences	11 sentences/Speaker
Sampling Frequency	8 kHz
Frame Period	10 ms
Window	30 msec Hamming
Analysis Order (p)	10
Bandwidth Expansion	15 Hz
Number of Frames	12141

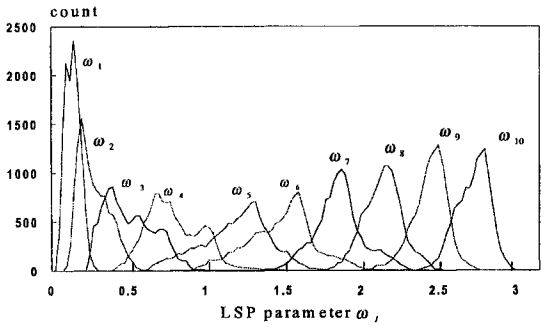

Figure 1(a): Histograms of LSP parameters

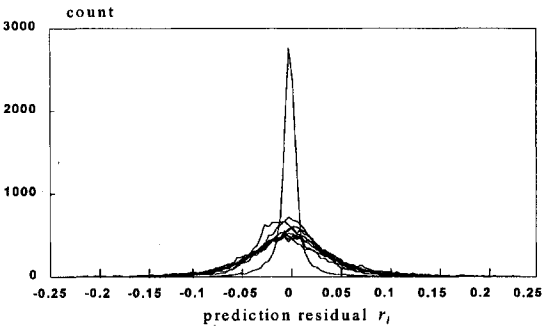

Figure 1(b): Histograms of residuals.

4.2 Spectral Distortion

In this study, the average spectral distortion defined by

$$\overline{SD^2} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\pi}\int_0^{\pi}\left[10\log\frac{S_n(\omega)}{\hat{S}_n(\omega)}\right]^2 d\omega \qquad (15)$$

is used as a major objective measure of performance, where N is the total frame number of test speech.

The average spectral distortion versus bits per frame for the different quantization schemes are given in Fig. 2(a) and Fig. 2(b). Fig. 2(a) is the case of inside test, and Fig. 2(b) is for outside test. Other LSP encoding methods, the forward sequential adaptive quantization method (AQFW) [7], the differential quantization with globally optimal quantizer (DQ) [4], as well as the two-dimensional differential coding scheme (2DdLSP-SQ) [8], are also included for comparison. The 1 dB^2 difference limen (DL) of spectral distortion is labeled by a solid line. In both cases, the TSC method with partitioned vector quantization (TSC-PVQ) outperforms all other LSP encoding methods. It can achieve the DL at 19 bits per frame and a mild degradation can be observed in outside test. Even with the simple SQ scheme (TSC-SQ), a spectral distortion can be obtained at 23 bits per frame.

Besides the average spectral distortion, the percentage of outlier frames is another performance measure. Human ears are very sensitive to such frames which usually cause unacceptable speech degradation. If two coders have the same average spectral distortion, we prefer the coder that has a small number of outlier frames. For comparison, the histograms of the spectral distortion for the DQ (33 bits/frame), AQFW (33 bits/frame), TSC-SQ(23 bits/frame), and TSC-PVQ (19 bits/frame) in outside test are presented in Fig. 3. In order to clarify the comparison, the measure of dB^2 is replaced by dB. There are four measures displayed in each histogram: the mean, the standard deviation, and the percentage of frames greater than 2 dB and 4 dB. The DQ and AQFW schemes with

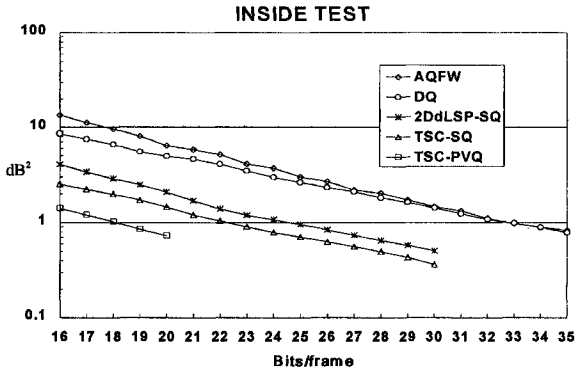

Fig. 2(a): Inside test of bits per frame versus average spectral distortion.

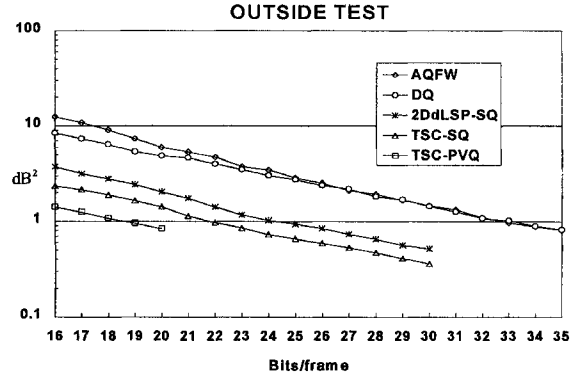

Fig. 2(b): Outside test of bits per frame versus average spectral distortion.

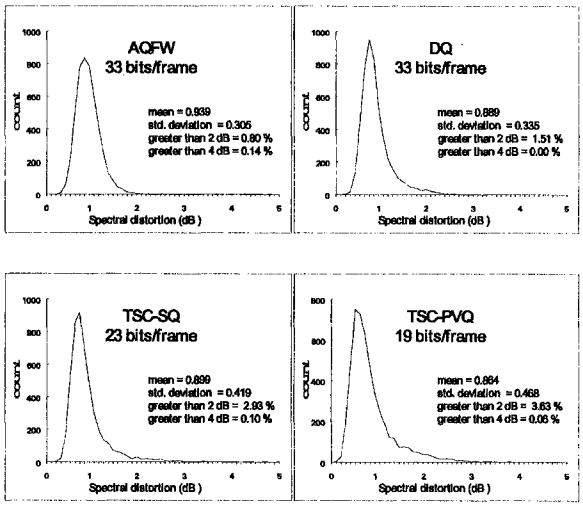

Fig. 3: Histograms of spectral distortion for various quantization schemes.

33 bits per frame give smaller percentage of frames greater than 2 dB as compare to the TSC-SQ and TSC-PVQ schemes due to the reason of using high bit-rate to capture sharp variation of LSP parameters.

4.3 Switched Classifier

Though the TSC-PVQ scheme can reach the DL at 19 bits per frame, there still are 3.63 % (greater than 2 %) outlier frames. In order to reduce the outlier frames down to below 2 %, a variable rate TSC coder was proposed. The idea is to use

less bits for steady frames and more bits for rapidly changing frames. A switched classifier is built for this purpose. There are two quantizers, U_1 and U_2, at the first stage. The bit-rate used in U_2 is larger than that in U_1. Similarly, There are two quantizers, V_1 and V_2, at the second stage and the bit-rate used in V_2 is larger than that in V_1. In the experiment, both U_1 and V_1 have the same bit-rate, so do U_2 and V_2. $SD(U_i, V_j)$ is the spectral distortion calculated by using quantizer U_i at the first stage and quantizer V_j at the second stage. There exist four quantizer combinations to encode a LSP vector. The switched classifier operates as follows:

(1). Calculate $SD(U_1, V_1)$.
If $SD(U_1, V_1) < Threshold$
then transmit side information of quantizer combination (U_1, V_1) and reconstruction indices to decoder, go to the end.
else; go to step (2).

(2). Calculate $SD(U_1, V_2)$ and $SD(U_2, V_1)$.
If $SD(U_1, V_2) < Threshold$ and $SD(U_1, V_2) < SD(U_2, V_1)$, then transmit side information of (U_1, V_2) and the corresponding reconstruction indices to decoder, go to the end.
else;
If $SD(U_2, V_1) < Threshold$ and $SD(U_2, V_1) < SD(U_1, V_2)$, then transmit side information of (U_2, V_1) and the corresponding reconstruction indices to decoder, go to the end.
else; go to step (3).

(3) Calculate $SD(U_2, V_2)$.

Find $(U_{i^*}, V_{j^*}) = \arg \min_{1 \le i, j \le 2} \{SD(U_i, V_j)\}$,

then transmit side information of (U_{i^*}, V_{j^*}) and the corresponding reconstruction indices to decoder, go to the end.

TABLE II
TRANSPARENT QUANTIZATION BY TSC METHOD WITH SWITCHED CLASSIFIER (USING ADAPTIVE SPECTRAL WEIGHTING)

$Bit(U_1, V_1)$	(6, 6)	(7, 7)	(8, 8)
$Bit(U_2, V_2)$	(17, 17)	(17, 17)	(17, 17)
Threshold	1.6 dB^2	1.8 dB^2	2.0 dB^2
$P(U_1, V_1)$	0.552	0.740	0.855
$P(U_1, V_2)$	0.159	0.080	0.044
$P(U_2, V_1)$	0.215	0.135	0.075
$P(U_2, V_2)$	0.074	0.045	0.026
Avg. Bits/Frame	19.423	18.472	18.754
Avg. SD (dB^2)	0.958	0.988	0.906
2-4 dB (%)	0.33	0.39	0.27
> 4 dB (%)	0	0	0

Table II shows three situations that transparent quantization of spectral information can be obtained by using outside test data. The quantizer U_1 is PVQ, so is V_1. The both quantizers U_2 and V_2 are SQ. $Bit(U_i, V_j) = (x, y)$ means quantizer U_i uses x bits at the first stage and quantizer V_j uses y bits at the second stage. $P(U_i, V_j)$ means the occurrence frequency of the quantizer combination (U_i, V_j). The side information of quantizer combination is transmitted by Huffman encoding method. Therefore, (U_1, V_1) needs 1 bit, (U_1, V_2) needs 3 bits, (U_2, V_1) needs 2 bits, and (U_2, V_2) needs 3 bits to encode this side information of quantizer combination. Taking the last column as an example, U_1 and V_1 are 8-bit partitioned codebooks, U_2 and V_2 use 17-bit scalar quantization. The Threshold used in the switched classifier is set to 2.0 dB^2. There are 85.5 % frames encoded by using 17 bits, 4.4 % frames by using 28 bits, 7.5 % frames by using 27 bits, and 2.6 % frames by using 37 bits to transmit the spectral information. The average bit-rate is 18.754 bits per frame, and the average spectral distortion is 0.906 dB^2. There are only 0.27 % outlier frames. This value is much lower than the required threshold of 2 %. There are no frames having spectral distortion greater than 4 dB.

V. CONCLUSION

The TSC method with switched classifier based on KLT and 2D-prediction is proposed in the paper. This scheme takes advantage of both interframe and intraframe correlation of LSP parameters to compress the information of short-time spectral envelope. We have addressed the TSC method, the quantization schemes, as well as the switched classifier to reduce the outlier frames. The simulation results show that the proposed scheme is a very promising one. When frame period is 10 ms, the transparent quantization of LPC information can be achieved at an average bit-rate below 19 bits per frame and the outlier frames down to about 0.27 %.

REFERENCES

[1] K. K. Paliwal and B. S. Atal, "Efficient vector quantization of LPC parameters at 24 bits/frame," *in Proc. ICASSP-91*, pp. 661-664, 1991.

[2] F. Itakura, "Line spectrum representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Amer.*, vol. 57, p. 535(a), p. s35(A), 1975.

[3] F. K. Soong and B.-H. Juang, "Line spectrum and speech data compression," *in Proc. ICASSP-84*, pp. 1.10.1-1.10.4, 1984.

[4] F. K. Soong and B.-H. Juang, "Optimal quantization of LSP parameters," *in Proc. ICASSP-88*, pp. 394-397, 1988.

[5] F.-R. Jean, C.-C Kuo and H.-C. Wang, "Optimal transform coding for speech LSP parameters based on spectral weighted-error criterion," *to appear in Computer Speech and Language*.

[6] Y. Linde, A. Buzo and R. M. Gray, "An algorithm for vector quantizer design," IEEE trans. Commun., vol. COM-28, pp. 84-95, 1980.

[7] N. Sugamura and N. Farvardin, "Quantizer design in LSP speech analysis-synthesis," *IEEE Trans. Select. Areas Commun.*, vol. 6, no. 2, pp. 432-440, 1988.

[8] C.-C Kuo, F.-R. Jean, and H.-C. Wang, "Low bit-rate quantization of LSP parameters using two-dimensional differential coding," *in Proc. ICASSP-92*, pp. I 97- I 100, 1992.