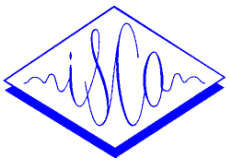


3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

SPECTRAL SENSITIVITY WEIGHTED TRANSFORM CODING FOR LSP PARAMETERS

Fu-Rong Jean, Chih-Chung Kuo, and Hsiao-Chuan Wang

Department of Electrical Engineering
National Tsing Hua University
Hsinchu, Taiwan, 30043, R. O. C.

ABSTRACT

The line spectrum pair (LSP) is one of the most effective representations of the speech short-time spectrum. About 34 bits/frame is needed for direct quantization of LSP parameters to maintain a reasonable accuracy. Based on the spectral-sensitivity-weighted Euclidean distance of LSP parameters, a hybrid TC/DPCM coding of LSP parameters which takes into account the spectral sensitivity weighting is proposed . In addition to the scalar quantization, two vector quantization methods which are multi-stage VQ and partitioned VQ, are considered. With the frame period of 10 ms, the best result shows that the spectral distortion limen of 1 dB^2 can be achieved at 18 bits/frame with in-training test data and 20 bits/frame with out-of-training test data, respectively.

Keywords : *line spectrum pair, spectral distortion, vector quantization.*

1. INTRODUCTION

The line spectrum pair (LSP) is one of the most popular and efficient parameters for representing the short-time spectrum of speech signal. In low bit-rate speech coding, the scalar quantization of each LSP parameter needs totally about 34 bits/frame to maintain a reasonable accuracy, such as FS1016 coder [1].

The spectral distortion (SD^2) is one of the best objective measures for a spectral coding method, but it is too complex to use as the design criterion of the quantization procedure. If a parametric distortion measure is used, it should be closely correlated with the spectral distortion and computationally efficient. The spectral-sensitivity-weighted Euclidean distance of the LSP parameters has been considered as a high fidelity spectral distortion measure [2]. This parametric distortion (PD^2) is defined as

$$PD^2 = \sum_{i=1}^{p} SEN_i^2 (\omega_i - \hat{\omega}_i)^2 \qquad (1)$$

where the ω_i is the i-th LSP parameter and $\hat{\omega}_i$ is its reconstructed value. The spectral sensitivity of the i-th LSP frequency SEN_i is defined as

$$SEN_i = \sqrt{\frac{1}{\pi}\int_0^{\pi} \left| \frac{\partial \log S(\omega)}{\partial \omega_i} \right|^2 d\omega} \qquad (2)$$

The task to design the quantizer algorithm is to minimize the expected value of the weighted Euclidean distance.

2. SPECTRAL SENSITIVITY WEIGHTED TRANSFORM CODING (SSWTC)

From the weighted Euclidean distance measure, we can find that each LSP parameters has its own significance in sense of spectral sensitivity. If we can generate a new vector space where each component of the new vector has the same significance in spectral distortion, the Euclidean distance measure can be applied to the new vectors in quantization procedure. Based on this concept , we introduce a novel LSP encoding method called the spectral sensitivity weighted transform coding which is depicted in figure 1.

Let us consider the LSP, vector, $\mathbf{w}(n)$, obtained from a p-th order all pole model analysis of speech signal at the n-th frame

$$\mathbf{w}(n) = \left[\omega_1(n) \;\; \omega_2(n) \;\; \cdots \;\; \omega_p(n)\right]^T \qquad (3)$$

where $\omega_i(n)$ is the i-th parameter of the n-th LSP vector $\mathbf{w}(n)$. $\bar{\mathbf{w}}$ is the mean vector of $\mathbf{w}(n)$. S is a spectral sensitivity matrix with SEN_i on the diagonal . SEN_i is the average spectral sensitivity associated with i-th LSP parameter. A is an orthogonal transform matrix, that is if

$$\mathbf{A} = \left[\mathbf{a}_1 \;\; \mathbf{a}_2 \;\; \cdots \;\; \mathbf{a}_p\right] \qquad (4)$$

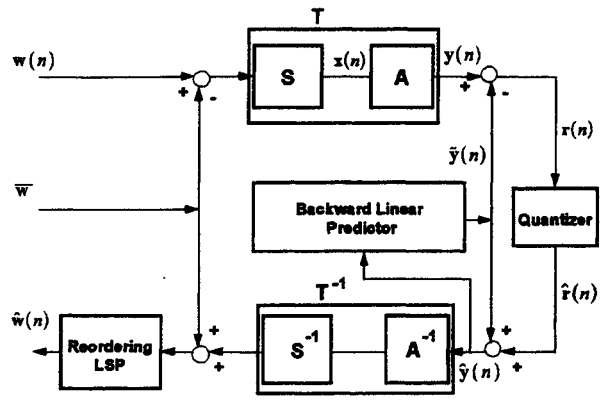


Figure 1: The block diagram of the spectral-sensitivity-weighted transform coding (SSWTC).

This cascade transform is defined by

$$T=AS=\left[SEN_1\cdot a_1 \quad SEN_2\cdot a_2 \quad \cdots \quad SEN_p\cdot a_p\right] \tag{5}$$

It is clear that the i-th column vector of the forward transform kernel T is the i-th column vector of the orthogonal matrix A weighted by the i-th spectral sensitivity SEN_i. An m_i-th order backward linear predictor is used to predict the i-th component of $y(n)$

$$\tilde{y}_i(n)=\sum_{j=1}^{m_i} h_i(j)\hat{y}_i(n-j) \quad for \quad i=1,2,\cdots,p \tag{6}$$

where $\tilde{y}_i(n)$ is the i-th component of the prediction vector $\tilde{y}(n)$, $\hat{y}_i(n)$ is the i-th component of the reconstruction vector $\hat{y}(n)$, and $\{h_i(j)\}$ $j=1,2,\cdots,m_i$ are the prediction coefficients of the i-th backward linear predictor. This prediction process implies no "side information" to be transmitted between encoder and decoder. We need only to quantize the prediction residual vector $r(n)$.

Finally, the reconstruction vector $\hat{w}(n)$ must satisfy the ordering property of the LSP vector to ensure the stability of the short-term filter. The following relationship exists for all n :

$$0<\hat{\omega}_1(n)<\hat{\omega}_2(n)<\cdots<\hat{\omega}_{p-1}(n)<\hat{\omega}_p(n)<\pi \tag{7}$$

We assume that the overload distortion is negligible and each $r_i(n)$ has the same normalized pdf. To minimize the mean of the reconstruction error

$$\sigma_e^2=\frac{1}{p}\sum_{i=1}^{p}\sigma_{e_i}^2=\frac{1}{p}\sum_{i=1}^{p}\left(r_i(n)-\hat{r}_i(n)\right)^2 \tag{8}$$

with the constraint of a given bit assignment

$$R=\frac{1}{p}\sum_{i=1}^{p}R_i=constant \tag{9}$$

is a standard bit-allocation problem discussed in [3, 4]. The optimal solution can be obtained and the following relationship holds

$$\min\{\sigma_e^2\}=\min\left\{\frac{1}{p}\sum_{i=1}^{p}\sigma_{e_i}^2\right\}=\min\left\{\frac{1}{p}\sum_{i=1}^{p}E\left[\left(r_i(n)-\hat{r}_i(n)\right)^2\right]\right\}$$

$$=\min\left\{\frac{1}{p}\sum_{i=1}^{p}E\left[SEN_i^2\left(\omega_i(n)-\hat{\omega}_i(n)\right)^2\right]\right\} \tag{10}$$

because of the variance-preserving property of the orthogonal transform pair. This is indeed the minimization of the expected value of the spectral-sensitivity-weighted Euclidean distance of LSP and its quantized version.

3. QUANTIZATION SCHEMES

Three quantization schemes will be presented and simulated in our experiments.

3.1 Scalar Quantization (SQ)

A scalar quantization scheme is used to design an optimal quantizer Q_i for each prediction residual r_i. The dynamic programming method for globally optimal scalar quantizer design proposed by F. K. Soong and B.-H. Juang [2] is used to train its reconstruction levels. In practice, scalar quantizer usually has the constraint of having to use non-negative integer bits. Here we impose the integer and non-negativity constraints for the number of bits R_i. In the bit allocation procedure, we use the exhaustive search to find $R_1,R_2,\cdots,R_p$ such that the minimization of overall distortions are obtained [4].

3.2 Two-stage VQ (MVQ)

The first stage performs a coarse quantization of the residual vector $r(n)$ using the first codebook. Then, the error vector between the original vector $r(n)$ and the quantized output in the first stage is further quantized using the second codebook. The codewords of these two codebooks are obtained by using the well-known generalized Lloyd method.

3.3 Partitioned VQ (PVQ)

In order to reduce the cost of computation and storage required in two-stage VQ, a high dimensional vector is partitioned into several subvectors to form a partitioned VQ. In this study, the residue $r(n)$ is partitioned into two parts, $r_L(n)$ and $r_H(n)$. A codebook is generated for each part. Then two codebooks are used in this partitioned VQ approach.

4. EXPERIMENTS AND RESULTS

The database used in the simulation consists of 11 sentences spoken by 8 speakers (4 males and 4 females). The speech signals were digitized at 8 KHz sampling rate and 16 bits/sample. We used half of the database as the training data which are described in table 1, the else were used in out-of-training tests.

Table 1 : Experimental conditions of 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
Number of Frames	12161

For Mandarin speech, the average spectral sensitivities of the different LSP parameters are calculated and presented in table 2 .

Table 2: Spectral sensitivities of LSP parameters (in dB/Hz).

SEN_1	0.0220	SEN_6	0.0135
SEN_2	0.0173	SEN_7	0.0123
SEN_3	0.0156	SEN_8	0.0132
SEN_4	0.0120	SEN_9	0.0121
SEN_5	0.0131	SEN_{10}	0.0125

We have examined various data independent orthogonal transform kernels . Based on the same condition of with the third order predictors, the discrete Hartley transform (DHT) provides better performance. The discrete Hartley transform pair [6] is used in our simulation:

$$y_k(n)=\frac{1}{\sqrt{p}}\sum_{i=1}^{p} x_i(n)\left(\cos\frac{2\pi(k-1)(i-1)}{p}+\sin\frac{2\pi(k-1)(i-1)}{p}\right), for \quad k=1,\cdots,p$$

$$x_i(n)=\frac{1}{\sqrt{p}}\sum_{k=1}^{p} y_k(n)\left(\cos\frac{2\pi(k-1)(i-1)}{p}+\sin\frac{2\pi(k-1)(i-1)}{p}\right), for \quad i=1,\cdots,p \quad (11)$$

Now we present the simulation results for the SSWTC method with various quantization schemes described in previous section. Both in-training and out-of-training speech sequence are tested. A speech database which is randomly selected from the training database is used for in-training test. It consists of 4675 frames spoken by 1 males and 1 females. Another database which consists of 4844 frames spoken by different speakers (1 males and 1 females) is selected for out-of-training test.

The average spectral distortion defined by

$$\overline{SD^2}=\frac{1}{N_f}\sum_{n=1}^{N_f}\int_0^{\pi}\left(10\log\frac{S_n(\omega)}{\hat{S}_n(\omega)}\right)^2\frac{d\omega}{\pi} \quad (dB)^2 \qquad (12)$$

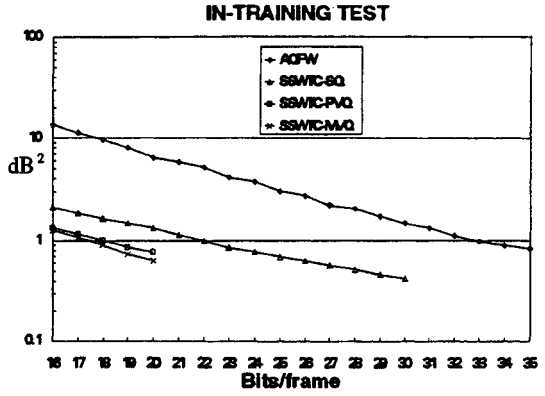
is used as a main objective measure of performance in all experiments. Here, $\hat{s}_n(\omega)$ and $s_n(\omega)$ are the power spectra of the n-th speech frame with and without quantization respectively and N_f is the total number of frames in testing database.

Table 3: The Bit allocation table of scalar quantization.

Bits/frame	Bit assignment
15	3 2 1 1 1 1 1 1 2 2
16	3 2 2 1 1 1 1 1 2 2
17	3 2 2 2 1 1 1 1 2 2
18	3 2 2 2 1 1 1 2 2 2
19	3 2 2 2 1 1 1 2 2 3
20	3 3 2 2 1 1 1 2 2 3
21	3 3 2 2 1 1 2 2 2 3
22	3 3 2 2 2 1 2 2 2 3
23	3 3 2 2 2 2 2 2 2 3
24	3 3 2 2 2 2 2 2 3 3
25	3 3 3 2 2 2 2 2 3 3
26	3 3 3 3 2 2 2 2 3 3
27	4 3 3 3 2 2 2 2 3 3
28	4 3 3 3 2 2 2 3 3 3
29	4 3 3 3 2 2 3 3 3 3
30	4 3 3 3 3 2 3 3 3 3

In scalar quantization, optimal bit allocation from 15 bits per frame to 30 bits per frame is tabulated in table 3. The bits that we assigned to the first codebook of the partitioned VQ equal to the summation of the first five LSP bits . Similarly, the bits allocated to the secondary codebook of the partitioned VQ equal to the summation of the last five LSP bits. In two-stage VQ, the bits assigned to the first codebook is equal to or more than one bit to the secondary codebook.

The average spectral distortions vs. bits/frame for the different quantization schemes are given in figure 2 . Another LSP encoding method, the forward sequential adaptive quantization method (AQFW) [5], is also included for comparison. The 1 dB^2 difference limen (DL) of spectral distortion is labeled by a solid line. For the case of in-training test, the SSWTC method with Two-stage VQ scheme (SSWTC-MVQ) outperforms all other LSP encoding methods. It can achieve the DL at 18 bits/frame. Even with the simple SQ scheme (SSWTC-SQ), a spectral distortion of 1 dB^2 can be obtained at 22 bits/frame. For the case of out-of-training test, a graceful and soft performance degradation can be observed. The SSWTC-MVQ method and the SSWTC method with Partitioned VQ (SSWTC-PVQ) can achieve the DL at 20 bits/frame. The SSWTC-SQ method can achieve DL at 23 bits/frame. It shows that the SSWTC-MVQ or SSWTC-PVQ requires about 4 bits/frame fewer than the SSWTC-SQ scheme. One more bit is needed for out-of-training test as comparing with in-training test in the same spectral distortion.



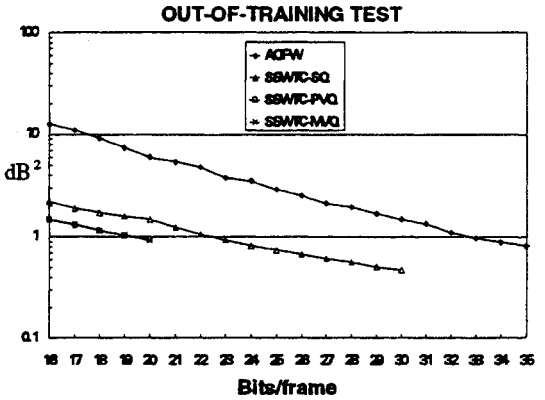


Figure 2: Simulation results of average spectral distortion versus bits/frame in various quantization schemes for both cases of in-training and out-of-training tests.

We implemented a CELP coder similar to FS1016 which is a proposed standard of 4.8 Kbps [1]. We examine the performance of the various LSP encoding methods in FS1016. Table 4 shows the average SEGSNR for different encoding methods. It shows that the SSWTC method performs well in the CELP coder. Especially, the encoding method with MVQ or PVQ scheme can greatly reduce the bits per frame from 34 to 20 and keep the same coded speech quality. The informal listening tests also support this conclusion.

Table 4: Average SEGSNR of CELP coder with various LSP encoding methods.

Method	bits/frame	SEGSNR (dB)
AQFW	34	12.19
SSWTC-MVQ	20	11.98
SSWTC-PVQ	20	11.93
SSWTC-SQ	23	10.88

5. CONCLUSION

The spectral-sensitivity-weighted transform coding (SSWTC) has the following advantages :

- Both strong inter-frame and intra-frame correlation of LSP parameters are effectively used.
- The spectral distortion limen of 1 dB^2 can be achieved at low bit rate.
- It is inhibited to introduce any buffering delay of future frames.
- It is graceful and soft in performance degradation for out-of-training data test. That means a less data sensitivity.

REFERENCES

[1] J. P. Campbell, T. E. Tremain and V. C. Welch, " The proposed federal standard 1016 4800 bps voice coder : CELP," Speech technology, pp.58-64, Apr./May 1990.

[2] F. K. Soong and B.-H. Juang, "Optimal quantization of LSP parameters," IEEE Proc. Int. Conf. Acoust., Speech, Signal Processing (ICASSP), 1988, pp.394-397.

[3] N. S. Jayant and P. Noll, "Digital coding of waveforms," Prentice-Hall, Englewood Cliffs, New Jersey, 1984.

[4] A. Gersho and R. M. Gray, "Vector quantization and signal compression," Kluwer academic publishers, 1992

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

[6] R. N. Bracewell, "The Hartley transform," Oxford university press, New York, 1986.