# Vector Quantization of Speech Parameters in the Time-Frequency Domain

Junchen Du
Department of Electrical Eng.
Polytechnic University
Brooklyn, NY 11201

Seung P. Kim
Department of Electrical Eng.
Polytechnic University
Brooklyn, NY 11201

## Abstract

*We investigate a time-frequency domain vector quantization approach for low bit rate speech coding. The well known CELP standard provides an efficient representation of a speech signal in units of a frame length (30 ms) using an all-pole speech model. For most efficient speech compression, inter and intraframe correlations of speech parameters should be exploited. This observation previously lead to multiframe coding with interpolation [1][2], or LSP prediction [3]. But such linear modeling approaches suffer from the nonlinear speech parameter dynamics. The proposed approach eliminates such problems while exploiting intra and interframe correlations. Computer simulations demonstrate significantly improved compression performance (1.5-2 bits/frame) under a given spectral distortion.*

## 1 Introduction

Code Excited Linear Prediction coder (CELP) has been successfully used in speech coding and adopted as a Federal Standard (FS1016). It is a model-based coding scheme using all-pole filters to represent the spectral envelop of a speech segment or frame. Line Spectral Pairs (LSPs) become the major representation scheme for filter parameters [4][5] because of its excellent properties in terms of filter stability, preservation of minimum phase properties as well as small spectral sensitivity. The mostly used performance measure of parameter encoding is spectral distortion (SD) [4]. Both the average SD and maximum SD represent aspects of quality of a coded modeling (or LPC) filter, hence the quality of a reconstructed speech.

Vector Quantization (VQ) utilizes the correlations within each input block. Since speech parameters tend to have correlations, VQ offers more efficient representation of speech model than the scalar quantization [4]. But the large complexity of VQ hindered its use on compression of speech parameters. It is necessary to use quite a large codebook size in order to achieve a good speech quality with a moderate bit rate. But as well known, it is difficult to train a large codebook, and takes long time to encode and decode the parameters. A solution to this problem has been proposed in [4] through a split VQ where the LSP parameters are grouped into two groups and two smaller codebooks are used independently. It is also well known that correlations exist between speech models in the adjacent frames (interframe correlations). In order to exploit interframe correlations, multiframe coding with interpolation of missing LSP parameters have been investigated [1]. In [3], a linear prediction of LPC parameters using previous frame LPC parameters have been studied. Both approaches however suffer from the fact that changes of LSP parameters are highly nonlinear and linear modeling approaches do not provide significant improvements. Furthermore, in the prediction approach [3], the error in one frame propagates to the following frames. In the interpolation approach, intelligibility is severely degraded even when the parameters are not quantized [1], and introduces long delay (8 frames delay) and large amount of computations.

In this paper, we propose a new scheme based on vector quantization of LSP parameters in the time-frequency domain. The proposed new scheme utilizes both intraframe and interframe correlations of the parameters, but it does not suffer from the problems the previous two approaches encounter. A similar attempt has been made using matrix quantization concept [10]. Our approach is different from matrix quantization in the following aspects: i) LSP parameters in CELP coding is utilized rather than LPC parameters. Hence, the modeling filter stability problem after quantization is removed. ii) In order to reduce the computational complexity, split VQ is applied. Split VQ has been recently investigated by Paliwal and Atal [4], but the



1058-6393/95 $4.00 © 1995 IEEE

coding was based on single frames and does not exploit interframe correlations. iii) In order to further reduce the computational complexity, multi-stage split VQ has been incorporated for a higher bit rate coding.

The paper is organized as follows. Section 2 shows some statistical properties of LPC and LSP parameters which give justifications for using LSP for compression. In Section 3, we describe our new VQ approach in time-frequency domain. We further introduce 2-stage split VQ to reduce computation complexity for higher quality speech coding. In Section 4, we present the simulation results, and compare with other methods [1],[4]. Conclusions are given in Section 5.

## 2 Some Statistical Properties of LPC and LSP Parameters

It is well known that coding of data which represent a given source information is more efficient if the data is less correlated. In other words we can achieve better compression if we use less redundant representation of a source before quantization and compression. This fact is well witnessed by the popularity of transform and subband coding, etc. Since a speech model can be represented by either LPC or LSP parameters, it is also necessary to investigate which is a more suitable representation in terms of data compression. First, we investigate correlations between parameters within a frame. In Table 1, normalized cross-correlations of LPC coefficients are shown and Table 2 shows LSP parameters. About 22,700 speech frames are used from a recording of a FM broadcasting station where two males and one female speakers having discussions. The frame length is 30 msec as defined in FS-1016. LSPi and LPCi represent i-th LSP and i-th LPC coefficients, respectively.

From Table 1 and Table 2 , it can be seen that LSP parameters are less correlated compared with LPC coefficients. Next, auto-correlations of LSPs and LPC coefficients are shown in Tables 3 and 4, respectively. For an efficient compression of a series of frames, it is necessary to exploit correlations between frames. Therefore, it would be better if the parameters reveal the interframe correlations more explicitly. If we examine LSP and LPC autocorrelations across frames, it can be seen that LSP parameters exhibit higher level of autocorrelations than LPC parameters. In the LPC parameters, only the first two parameters show high level of correlation, say, above 0.7 in $\rho(1)$, while in the LSP case, five parameters are above 0.7. Therefore, LSP parameters are more suitable for exploitation of interframe correlations. Other advantages of using LSP parameters such as stability of a modeling filter after parameter quantization are well known and important. The smaller intraframe correlations and larger interframe correlations of LSP parameters are especially attractive when one considers a split-VQ as discussed in the next section.

| | $LSP_2$ | $LSP_3$ | $LSP_4$ | $LSP_5$ | $LSP_6$ |
|---|---|---|---|---|---|
| $LSP_1$ | 0.641 | 0.260 | 0.120 | 0.047 | 0.002 |
| $LSP_2$ | | 0.651 | 0.397 | 0.203 | 0.083 |
| $LSP_3$ | | | 0.706 | 0.313 | 0.287 |
| $LSP_4$ | | | | 0.466 | 0.415 |
| $LSP_5$ | | | | | 0.692 |
| | $LSP_7$ | $LSP_8$ | $LSP_9$ | $LSP_{10}$ | |
| $LSP_1$ | -0.023 | 0.002 | -0.051 | -0.101 | |
| $LSP_2$ | 0.124 | 0.102 | 0.073 | -0.077 | |
| $LSP_3$ | 0.252 | 0.373 | 0.271 | 0.087 | |
| $LSP_4$ | 0.410 | 0.412 | 0.475 | 0.188 | |
| $LSP_5$ | 0.363 | 0.243 | 0.124 | 0.093 | |
| $LSP_6$ | 0.537 | 0.445 | 0.264 | 0.142 | |
| $LSP_7$ | | 0.635 | 0.474 | 0.284 | |
| $LSP_8$ | | | 0.504 | 0.296 | |
| $LSP_9$ | | | | 0.504 | |

Table 1. Cross-correlation of LSPs

| | $LPC_2$ | $LPC_3$ | $LPC_4$ | $LPC_5$ | $LPC_6$ |
|---|---|---|---|---|---|
| $LPC_1$ | -0.893 | 0.457 | -0.360 | 0.353 | -0.313 |
| $LPC_2$ | | -0.732 | 0.445 | -0.391 | 0.402 |
| $LPC_3$ | | | -0.749 | 0.313 | -0.416 |
| $LPC_4$ | | | | -0.612 | 0.437 |
| $LPC_5$ | | | | | -0.733 |
| | $LPC_7$ | $LPC_8$ | $LPC_9$ | $LPC_{10}$ | |
| $LPC_1$ | 0.224 | -0.142 | 0.221 | -0.142 | |
| $LPC_2$ | -0.326 | 0.201 | -0.194 | 0.134 | |
| $LPC_3$ | 0.462 | -0.171 | 0.096 | -0.119 | |
| $LPC_4$ | -0.411 | 0.132 | -0.100 | 0.131 | |
| $LPC_5$ | 0.280 | -0.221 | 0.184 | -0.107 | |
| $LPC_6$ | -0.669 | 0.289 | -0.183 | 0.129 | |
| $LPC_7$ | | -0.631 | 0.192 | -0.035 | |
| $LPC_8$ | | | -0.696 | 0.286 | |
| $LPC_9$ | | | | -0.820 | |

Table 2. Cross-correlation of LPC coefficients

| | $LSP_1$ | $LSP_2$ | $LSP_3$ | $LSP_4$ | $LSP_5$ |
|---|---|---|---|---|---|
| $\rho(1)$ | 0.418 | 0.601 | 0.729 | 0.710 | 0.769 |
| $\rho(2)$ | 0.180 | 0.313 | 0.471 | 0.451 | 0.521 |
| $\rho(3)$ | 0.078 | 0.180 | 0.302 | 0.282 | 0.326 |
| | $LSP_6$ | $LSP_7$ | $LSP_8$ | $LSP_9$ | $LSP_{10}$ |
| $\rho(1)$ | 0.731 | 0.706 | 0.674 | 0.620 | 0.547 |
| $\rho(2)$ | 0.461 | 0.442 | 0.413 | 0.371 | 0.288 |
| $\rho(3)$ | 0.263 | 0.246 | 0.226 | 0.210 | 0.142 |

Table 3. Auto-correlation of LSPs

| | $LPC_1$ | $LPC_2$ | $LPC_3$ | $LPC_4$ | $LPC_5$ |
|---|---|---|---|---|---|
| $\rho(1)$ | 0.769 | 0.738 | 0.663 | 0.600 | 0.598 |
| $\rho(2)$ | 0.497 | 0.463 | 0.365 | 0.319 | 0.338 |
| $\rho(3)$ | 0.280 | 0.253 | 0.215 | 0.174 | 0.201 |
| | $LPC_6$ | $LPC_7$ | $LPC_8$ | $LPC_9$ | $LPC_{10}$ |
| $\rho(1)$ | 0.565 | 0.575 | 0.583 | 0.593 | 0.510 |
| $\rho(2)$ | 0.298 | 0.332 | 0.353 | 0.346 | 0.258 |
| $\rho(3)$ | 0.158 | 0.205 | 0.243 | 0.221 | 0.134 |

Table 4. Auto-correlation of LPC coefficients.

## 3 Vector Quantization of LSP Parameters in Time-Frequency Domain (TFVQ)

Vector quantization(VQ) considers a set of data as an input unit and take advantage of the fact that such higher dimensional input data usually do not span all possible combinations. Therefore, such input data tends to have highly nonuniform distribution in the input data space. By utilizing an optimized nonuniform quantizer, an efficient representation of source

codebooks are used where three of them are used for the first stage of 3-split VQ and the last is used for VQ of the residual error. In our simulation, an identical codebook size is used for all of them. It will be possible to improve the performance by using an optimal bit-allocation algorithm among these codebooks, and the result will reported later.

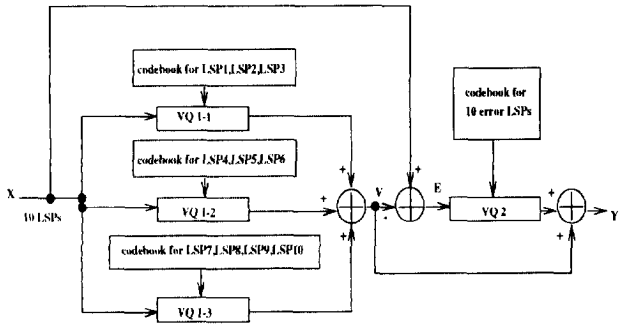


Fig.2 Two-stage TFVQ Scheme

## 4 Simulation Results and Discussions

In our simulation, the speech sample is obtained from a FM radio station, where three – two males and one female – speakers have discussions. The lowpass filtered speech is digitized at sample rate 8 Khz and we use 16 bits to quantize each sample. The frame size is 30 ms as defined in the Federal Standard 1016 [8]. Spectral analysis is performed for each frame by open-loop, 10-th order autocorrelation LPC analysis with no preemphasis and a 15 Hz bandwidth expansion and using a 30 ms Hamming Window. There are 20,000 frames in the training data. The evaluation uses 2,500 frames out side of the training speeches but from the same three speakers.

The performances of Atal's 2-split VQ (Atal VQ) [4], VQ-based interpolation (VQI) [1], and our proposed split time-frequency VQ (TFVQ) are compared. Depending on the desired bit rates, either 2-split TFVQ shown in Fig.1-a) or 3-split TFVQ shown in Fig. 1-b) are implemented.

Figure 3 shows the average spectral distortions (defined in Eq.(1)) of these three methods depending on the bit rates. The proposed TFVQ achieves about 1.5-2 bits more compression compared with Atal VQ at the same distortion level. It also outperforms VQI which is aimed for very low-bit rate coding. VQI [1][2] uses Atal VQ as bases for interpolation. Atal VQ is very difficult to handle coding with more than 26 bits/frame because of the large complexity for a large codebook size. Due to the same reason, it is very difficult for VQI to use more bits per frame. Even though we can imagine using more bits, from Fig. 3, we can

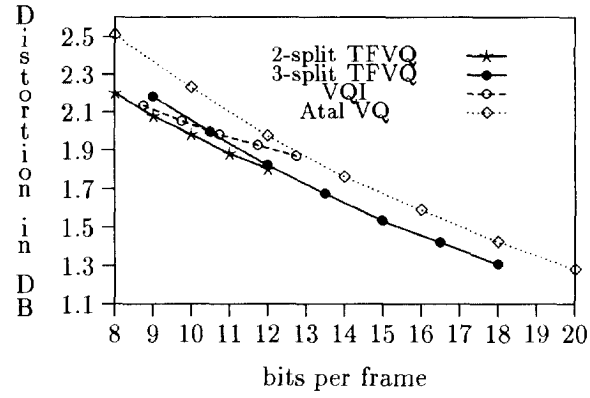


Figure 3. Average spectral distortion per frame

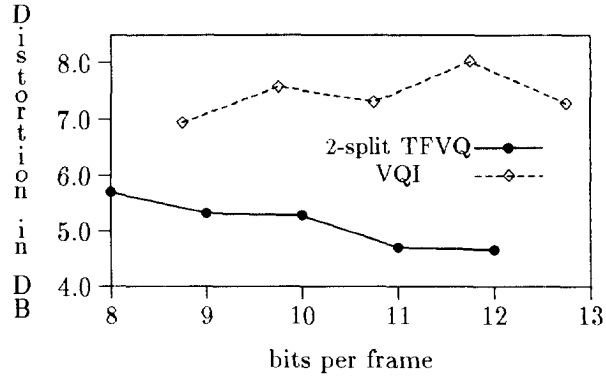


Figure 4. Maximum spectral distortion per frame

predict by extrapolation of the curve that the performance will be very poor. It should be also noted that the performance of 2-split is better than 3-split at low-bit rate where less than 12 bits/frame are used.

In a low bit rate speech coding, maximum spectral distortion as well as the percentage of frames with a distortion larger than 3 dB are generally used to indicate the quality of the synthesized speech. In Figure 4, maximum spectral distortions are compared between 2-split and VQI methods. In Figure 5 the percentages of frames with spectral distortion greater than 4 dB and 3 dB are shown. From Figures 4 and 5, we can see that both the maximum SD and outer percentage in 2-split TFVQ is much smaller than VQI. An important observation is the fact that as the bit rate increases, both maximum SD and outer percentage in 2-split TFVQ monotonically decreases, while in VQI, those values fluctuate around a constant level. This observation is consistent with the fact that the interpolated frames (4 out of 8 frames) causes some reduction in intelligibility even when the non-interpolated frames parameters are represented by full precisions[1]. Hence distortion does not improve even as bit rate is increased.

data is possible. If the input vector size is increased, the nonuniformity of input data distribution becomes more significant, hence more efficient representation (i.e., higher compression) is possible. But the complexity of the system increases exponentially, requiring larger codebook size, larger set of data for quantizer (codebook) design, longer decoding time due to search from a larger codebook.

One of the compromises to reduce the complexity of the system is to decompose the high dimensional input data space into nonoverlapping subspaces recursively, which is usually referred as tree-structured VQ(TSVQ). TSVQ is a suboptimal quantizer since nonuniformity of input data distribution is explored locally.

If we interpret 10 LSP parameters in a frame as one input vector, and assume that a representation of such a vector would require 24 bits[1] the codebook size will be approximately $2^{24} \approx 16 \times 10^6$ which is too large large for practical applications. Recently, a split VQ [4] has been proposed which separates 10 LSPs into two blocks - lower 4 LSPs and higher 6 LSPs. It reduces the input vector into two smaller sizes, hence reduces the codebook size to $2^{12} + 2^{12} = 8K$ approximately (assuming that 2.4 bits are used for each LSP on the average). Due to the nonoverlapping nature of the split, intraframe correlations between the two groups of parameters are not exploited. In the LSP parameter case, it has been shown that such intraframe correlations are smaller for LSP parameters than LPC. Hence LSP parameters are suitable for such a split VQ. The split VQ presented in [4] does not utilize the interframe correlation. Since speech signals are strongly correlated in the time-frequency domain, a better result can be achieved if the splitting is performed in the time-frequency domain so that more correlated data are grouped together.

### 3.1 Split TFVQ

In our first approach, we split LSP coefficients as suggested in [4], but each group in a frame is combined with the corresponding group in the adjacent frame. Therefore, input vectors are defined in a time-frequency domain as shown in Fig.1-a). The number of split is determined by the trade-off between the complexity of the system and desired quality. If one wants to increase the quality of parameter coding without increasing system complexity significantly, more split is desirable using a moderate number of bits for each groups. For example, in our investigation, we used two-split for the design of a coding system utilizing 6 to 12 bit/frame, and three-split for 13 to 18 bits/frame. In the three-split, the first block contains the first 3 LSPs, the second block the middle 3 LSPs and the third block the last 4 LSPs of every adjacent frames as shown in Fig 1-b). These blocks of data in the time-frequency domain are treated as independent vectors and separate codebooks are designed. Since each vector dimension is reduced, smaller codebook sizes can be used even for a higher bit-rate coding system. We used LBG algorithm [7] in our codebook designs. The distance measure is a modified version used in [4] $d(f, \hat{f}) = \sum_{i=1}^{10} c_i w_i (f_i - \hat{f}_i)^2$, where $f_i$ and $\hat{f}_i$ are the i-th original and reconstructed (codebook) LSPs, the weight $w_i = [P(f_i)]^{0.15}$, where $P(f_i)$ is the power spectrum of the frame's modeling filter at frequency $f_i$, and $c_i$ is an additional weighting terms which are 1.0 except for $c_9 = 0.9$ and $c_{10} = 0.7$.

| Frame n | Frame n+1 |
|---|---|
| LSP1-1 | LSP2-1 |
| LSP1-2 | LSP2-2 |
| LSP1-3 | LSP2-3 |
| LSP1-4 | LSP2-4 |
| LSP1-5 | LSP2-5 |
| LSP1-6 | LSP2-6 |
| LSP1-7 | LSP2-7 |
| LSP1-8 | LSP2-8 |
| LSP1-9 | LSP2-9 |
| LSP1-10 | LSP2-10 |

a) 2-split TFVQ

| Frame n | Frame n+1 |
|---|---|
| LSP1-1 | LSP2-1 |
| LSP1-2 | LSP2-2 |
| LSP1-3 | LSP2-3 |
| LSP1-4 | LSP2-4 |
| LSP1-5 | LSP2-5 |
| LSP1-6 | LSP2-6 |
| LSP1-7 | LSP2-7 |
| LSP1-8 | LSP2-8 |
| LSP1-9 | LSP2-9 |
| LSP1-10 | LSP2-10 |

b) 3-split TFVQ

Fig.1 Block structures for 2-split and 3-split TFVQ

### 3.2 Two-stage split TFVQ

For high quality speech coding, one needs to use more bits to encode LSPs. The Split TFVQ presented in the previous section has difficulty in encoding LSPs with more than 18 bits/frame (or more precisely, 36 bits/two-frame since two frames are coded at the same time) because of large codebook size, and hence training and computation. A well-known suboptimal tree-structured VQ (TSVQ) coding scheme is not attractive for exploiting inter/intra-frame correlations. Another suboptimal VQ known as multi-stage VQ has been described in [9]. This approach is more suitable here to handle the situation. The second stage of VQ can be advantageously used for further exploitation of intra-frame correlations, which was not possible due to split of LSP parameters in the first stage. In the tree-structured VQ, such an improvement is not possible. The overall structure of the proposed 2-stage TFVQ is shown in Fig.2. In the proposed coding scheme, four

[1] In order to achieve "transparent" quantization of LPC parameters (meaning less than 1 dB loss in log spectral distortion measure) for low bit rate coders[4], a scaler quantization would need about 32-40 bits. The number 24 is a rough estimate considering a performance improvement by using VQ.

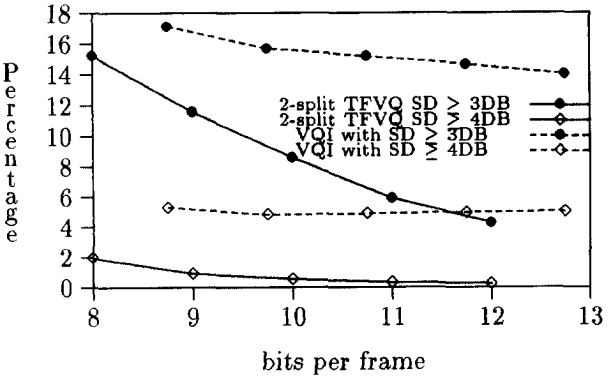


Figure 5. Percent of frames with SD greater than 3 DB and 4 DB

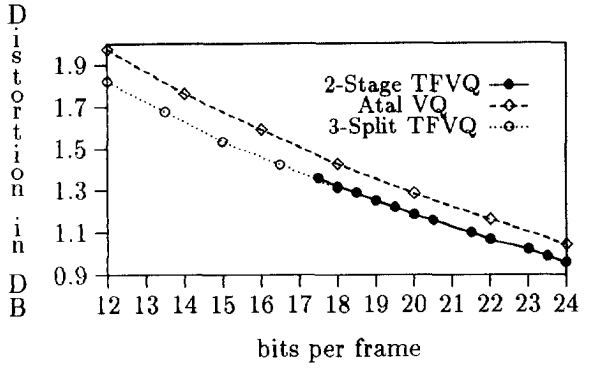


Figure 6. Average Spectral Distortion for high bit rate coding

Fig 6. shows the simulation results of our second approach – 2-stage split TFVQ(2S-TFVQ) aimed at higher bit rate. It is interesting to see that the performance of two-stage TFVQ coding is almost coincide with the extrapolated performance curve of the 3-split TFVQ. Compared with Atal's split-VQ [4], our method requires 1.5 bits less per frame under a given SD measure. This fact demonstrates that 2S-TFVQ provides a better complexity performace tradeoff for higher quality speech coding than previous appraoches.

## 5 Conclusions

In this paper, we introduced a new LSP parameter coding scheme based on Time-Frequency domain Vector Quantization (TFVQ) exploiting both interframe and intraframe correlations. This approach is inspired by split-VQ [4], which reduces codebook size by separate coding of LSP parameters. The proposed scheme does not suffer from the problems which occur in interpolation and prediction approaches, such as error propagation and inherent error which does not decrease even with a increased bit rate. It has been demonstrated that the proposed approach improves the performance by 1.5 - 2 bits/frame compared to previous approaches. The optimal split of LSP parameters in the time-frequency domain and optimal bit allocation will further improve the performance of the algorithm.

For further research, one can use more refined vector quantizer depending on the properties of speech data. For example, two set of vector quantizers can be optimized for voiced and unvoiced speeches separately and used with a voiced/unvoiced discriminator. The approach can be further expanded for use with more refined speech classification in the time-frequency domain.

## References


[1] D. P. Kemp, J. S. Collura and T. E. Tremain, "multi-frame coding of LPC Parameters at 600 - 800 bps , " Proc. IEEE ICASSP-91 , pp. 609 - 612. Toronto, 1991.

[2] J. M. Lopez-Soler and Nariman Farvardin, "A Combined Quantization-Interpolation Scheme for Very Low Bit Rate Coding of Speech LSP Parameters", Proc. IEEE ICASSP-93, pp.II-21 to II-24, Minneapolis, 1993.

[3] E. Erzin and A. E. Cetin, "Interframe Differential Vector Coding of Line Spectrum Frequencies", Proc. IEEE ICASSP-93, pp.II-25 to II-28, Minneapolis,1993.

[4] K. K. Paliwal and B. S. Atal, "Efficient Vector Quantization of LPC Parameters at 24 bits/frame," Proc. IEEE ICASSP-91, pp. 661-664. Toronto, 1991.

[5] F. K. Soong and B. H. Juang, "Line Spectrum Pair (LSP) and Speech Data Compression" . Proc. ICASSP, pp. 1.10.1-1.10.4,1984.

[6] A. H. Gray, Jr. and J. D. Markel, "Distance Measure for Speech Processing", IEEE Trans. on ASSP, Vol. ASSP-24, No.5, pp. 380-391, Oct. 1976.

[7] Y. Linde, A. Buzo and R. M. Gray, " An Algorithm for Vector Quantizer Design," IEEE Trans. Commun., vol. COM-28, pp. 84-95, Jan. 1980.

[8] "Federal Standard 1016, Telecommunications: Analog to Digital Conversion of Radio Voice by 4,800 bits/second Code Excited Linear Prediction (CELP)",National Communications System, Office of Technology and Standards, Washington, DC 20305-2010, 14 February,1991.

[9] Y. Tanaka and T. Taniguchi, " Efficient Coding of LPC Parameters Using Adaptive Prefiltering and MSVQ with Partially Adaptive Codebook", Proc. IEEE ICASSP-93,pp. II-5 to II-8, Minneapolis, 1993.

[10] C. Tsao and R. M. Gray, "Matrix quantizer design for LPC speech using the generalized Lloyd algorithm," IEEE Trans. Acoust., Speech, Signal Proc., Vol. ASSP-33, pp. 537-545, June 1985.