PREDICTIVE SVD-TRANSFORM CODING OF SPEECH WITH ADAPTIVE VECTOR QUANTIZATION

E. Masgrau, José A. Rodríguez-Fonollosa, Josep R. Mallafré

Dpt. T.S.C. E.T.S.I. Telecomunicacion. UPC
Apdo. 30002. 08080-Barcelona. SPAIN
E-mail:masgrau@tsc.upc.es

ABSTRACT

In this work we propose a Predictive Transform Coder using Singular Value Descomposition (SVD) and Adaptive Vector Quantization. The LPC excitation is obtained by a weighting VQ quantization of the SVD transform of the prediction residual. The orthogonal representation of the excitation provided by the SVD transform permits a fine quantization of the perceptually more important components. The excitation vector is segmented in subvectors which are coded by means and VQ product code. Some dynamic bit allocation approachs based on the relative importance of the singular values are described and these do not require side information. The integration of the usual perceptual noise shaping is very easy and, also, a simple noise shaping based on the direct reduction of the singular values spread is presented. The LPC coefficients are quantized by an Adaptive Multistage VQ (AMSVQ) approach developed by the authors [3]. The previous results obtained show a very good performance at the 7-8Kb/s range.

I. INTRODUCTION

In LPC speech coding it is possible to obtain an exact copy of the speech by using the LPC residual as excitation signal of the LPC filter. The major problem consists in to obtain an accurate representation of the excitation signal by using a limited number of bits. The reproduction of the excitation waveform can need as many bit rate as the original speech signal. The efficient solutions try to preserve in the coding only the more perceptually important characteristics of the excitation. The main objective is not to produce a signal that is physically identical to the original speech but to produce one that sounds identical [1]. Several popular LPC-based speech coders respond to this approach. The Residual Excited LP Coding (RELP), the Multi-Pulse (MP) and the Vector Excited Coding (VXC) reduce drasticly the possible excitation waveforms and they obtain fairly high quality for the synthetic speech by using a perceptual weighting criterion in the selection of the excitation sequence.

Atal did propose [1] a efficient representation of the LPC residual based in the Singular Value Descomposition of the impulse response matrix of the LPC filter. In this approach, the excitation sequences are represented as a linear combination of the right singular vectors of the LPC filter matrix. The SVD of this matrix defines an orthogonal transform over the speech and excitation spaces. The amplitude of each orthogonal component of the speech signal equals the amplitude of the corresponding component of the excitation weighted by its singular value. The reconstructed speech is obtained by a linear combination of the singular vectors whose weights are the transformed components of the speech signal. The orthogonal characteristic of this representation can allow to discern which components of the excitation signal are perceptually important. Also, the aproximately bandpass characteristic of the singular vectors provides a spectral interpretation of the reconstruction process of the speech. The large inherent complexity of the SVD -based representation is a big problem at present, but fast structured algorithms for SVD may be available at medium term. The great number of applications of the SVD in several signal processing fields (noise cancellation, array processing, signal coding, spectral analysis, ...) and its interesting properties have multiplied the efforts in this way [2].

* Work supported by PRONTIC grant number 105/88

II. THE SVD REPRESENTATION OF THE LPC SYNTHESIS

In this work we follow the SVD formulation presented in [1]. Thus, a frame of N samples y(n) of original speech depurated by the ringing of the previous excitation frames can be expresed in terms of the excitation signal x(n) as:

$$\mathbf{y}(n) = \mathbf{H}\,\mathbf{x}(n) \qquad (1)$$

where $\mathbf{y}(n)$ and $\mathbf{x}(n)$ are column vectors and the NxN $\mathbf{H}$ matrix is defined from of the impulse response h(n) of the LPC filter. The matrix $\mathbf{H}$ admits the SVD representation

$$\mathbf{H} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T \qquad (2)$$

where $\mathbf{U}$ and $\mathbf{V}$ matrix contain the left and right singular vectors of $\mathbf{H}$, respectively. The $\mathbf{D}$ is a diagonal matrix containing the ordered singular values d_i. From expressions (1) and (2) and omitting the temporal notation n by simplicity, we can write:

$$\mathbf{y} = \mathbf{U}\,\mathbf{D}\,\mathbf{V}^T\,\mathbf{x} = \mathbf{U}\,\mathbf{D}\,\mathbf{g} = \sum_{i=1}^{N} d_i\, g_i\, \mathbf{u}_i \qquad (3)$$

where $\mathbf{g} = \mathbf{V}^T\,\mathbf{x}$ is a column vector containing the g_i SV transformed component of the excitation signal $\mathbf{x}$. Also, we can define $\mathbf{z} = \mathbf{U}^T\,\mathbf{y} = \mathbf{D}\,\mathbf{g}$ and the expression (4) can be rewritten as:

$$\mathbf{y} = \mathbf{U}\,\mathbf{z} = \sum_{i=1}^{N} z_i\, \mathbf{u}_i \qquad (4)$$

The vector $\mathbf{z}$ contains the SV transformed components of the speech signal $\mathbf{y}$ and $z_i = d_i\, g_i$, that is to say, the SV transformed components of the signal z_i are equal to the corresponding excitation or residual transformed component g_i multiplied by its associated singular value. Therefore, any error made in the quantization of a particular component of the excitation have influence only in the corresponding component of the speech. The perceptually less important components z_i (and its corresponding g_i excitation components) need a lower number of bits for its quantization. Also, the components associated

CH2977-7/91/0000-3681 $1.00 © 1991 IEEE

with smaller singular values yield less influence in the speech signal.

An interpretation of the singular values can be obtained by expressing the SVD of the autocorrelation matrix of the LPC filter. From the expression (2) we can write:

$$\mathbf{R}_{hn} = \mathbf{H}^T \mathbf{H} = \mathbf{V}\, \mathbf{D}^2\, \mathbf{V}^T \qquad (5)$$

or

$$\mathbf{D}^2 = \mathbf{V}^T \mathbf{R}_{hn}\, \mathbf{V} \qquad (6)$$

where the matrix $\mathbf{V}$ contain the eigenvectors of the autocorrelation matrix, and the diagonal matrix $\mathbf{D}^2$ contains the eigenvalues d_i^2 of the $\mathbf{R}_{hn}$ matrix. It is well known that if $N \rightarrow \infty$ (N enough large), the $\mathbf{R}_{hn}$ matrix present a circulant structure and the Fourier Transform diagonalizes this matrix type. Thus, the d_i^2 eigenvalues tend to aproximate the spectral envelope of the LPC filter, is that to say, of the speech signal. This spectral interpretation of the d_i singular values provides a great insigth in the signal reconstruction process from its orthogonal components and it will shown very useful later.

III. THE PREDICTIVE SVD-TRANSFORM CODER

The efficient representation of the excitation shown in the preceding section suggests a SVD-based predictive-transform coding system. In the figure 1 is shown the scheme of a such coder. The decoder scheme is direct of this figure.

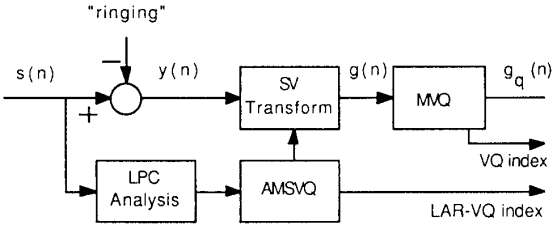

Figure 1. Scheme of the Predictive SVD-Transform Coder

First, a 10^{th} order LPC analysis is carried out over the original speech signal by using an overlapping Hamming window. This signal window is segmented in non-overlapping frames of N samples (N=20-40). Previously to the prediction process, each signal frame is depured by the ringing of the previous excitation sequences. These are the true quantized excitation used in the decoder. Likewise, the LPC synthesis filter uses the quantized LPC coefficients available in the decoder. Thus, the ringing component of the signal is exactly recovered. The impulse response matriz $\mathbf{H}$ is obtained from the quantized LPC filter and its SVD is calculated one time per LPC analysis interval. Then, the SV transform of the prediction residual is obtained by using the inverse of the expresion (3). Thus

$$\mathbf{g} = \mathbf{D}^{-1}\, \mathbf{U}^T \mathbf{y} = \mathbf{D}^{-1}\, \mathbf{z} \qquad (7)$$

The g_i components associated to zero singular values correspond to zero z_i components, and they are considered zero values. An alternative approach consists in the calculation of the prediction residual x(n) by inverse LPC filtering and subsequent SV-transform.

The vector $\mathbf{g}$ is vector quantized using a weighting distortion measure. The distortion measure in the speech domain is the usual MSE :

$$J = (\mathbf{y}-\mathbf{y}_q)^T\, (\mathbf{y}-\mathbf{y}_q) \qquad (8)$$

By using the expression (1) and (3) or (7) we can write

$$J = (\mathbf{x}-\mathbf{x}_q)^T\, \mathbf{H}^T\, \mathbf{H}\, (\mathbf{x}-\mathbf{x}_q) = (\mathbf{x}-\mathbf{x}_q)^T\, \mathbf{V}\, \mathbf{D}^2\, \mathbf{V}^T\, (\mathbf{x}-\mathbf{x}_q)$$

finally resulting in :

$$J = (\mathbf{g}-\mathbf{g}_q)^T\, \mathbf{D}^2\, (\mathbf{g}-\mathbf{g}_q) \qquad (9)$$

Thus, the quantization of the transformed components g_i must take into account its associated square singular values since this value represents its relative importance in the reconstructed $\mathbf{y}_q$ signal. The large dimension of the SV transformed excitation $\mathbf{g}$ (20-40 samples) requires a product vector quantization approach or multi-VQ(MVQ). The overall vector is segmented in subvectors of fixed or variable length. Each subvector is then vector quantized independently, taking advantage of the obvious descomposable characteristic of the distortion measure (9). The bit assignment of the available bits among the defined subvectors is carried out from the amplitude distribution of the associated singular values, assigning more bits to the subvectors corresponding to the larger singular values. The fine quantization of the excitation needed to obtain a high quality of the synthetic speech requires a dynamic bit allocation approach and the details will studied in the next section. The overall bit flow transmitted from the coder includes the codewords index of the MVQ product code of the excitation and the bits used in the LPC coefficients quantization. These are vector quantized using an adaptive multistage VQ (AMSVQ) scheme developed by the authors [3] which has been shown very efficient in similar context as CELP and MP coders. The AMSVQ approach is described in section VI. The LPC quantization is the only side information transmitted since the bit allocation is obtained from the SVD transform reproducted in the decoder from the LPC coefficients.

The decoder carries out the inverse coding process, obtaining the quantized signal $y_q(n)$ from the $\mathbf{g}_q$ vector by using the expresion (3). The ringing of the previous frames is obtained from the LPC filter and it is added to the $y_q(n)$ component. As it was said for the coder side, the reconstruction process can be made by obtaining $x_q(n)$ and subsequent LPC filtering.

IV. BIT ALLOCATION

The minimization of the weighting criterion formulated in (9) leads to aproximately white or flat quantization noise shape in the speech domain. A vector quantization of the overall excitation vector should provide this effect but the segmentation in subvectors and the corresponding VQ product code used requires a non-uniform bit allocation. The rate-distortion theory provides the optimal bit allocation [4] of the available M_b bits, asuming a flat spectrum for the transformed excitation $\mathbf{g}$:

$$R_i = R + \beta_i + \frac{1}{2}\log 2 \frac{\left(\prod_{j=1}^{k_i} d_{ij}^2\right)^{1/k_i}}{\left(\prod_{j=1}^{m}\prod_{h=1}^{k_i} d_{jh}^2\right)^{1/N}} \qquad (10)$$

where, $R = M_b/N$, R_i is the average bits/sample assigned to the i^{th} subvector, k_i is the dimension of the i^{th} subvector, m is the subvector number, d_{ij} is the associated singular values to the j^{th} component of the i^{th} subvector and β_i take into account the advantages provided by the dimensionality of the i^{th} subvectors. The expression (10) is only based in the singular values and, therefore, the bit allocation does not require side information. By simplicity, we take $\beta_i = 0$ and the bit assignment becomes as in the scalar case, and the VQ advantages are not taken into account. An alternative bit allocation approach to avoid the β_i estimation problem [5] consists in to define the subvector lengths so that the energies of the associated singular values $E_i = \sum_{j=1}^{k_i} d_{ij}^2$ are identical, (is that to say, $E_i \cong \frac{E}{m}$ where E is the overall SV energy) and the number of bits/subvector is then taken uniform.

V. NOISE SHAPING

The flat quantization noise shape leads to a MSE distortion but it is not perceptually the optimal solution. Generally, a noise shaping of the quantization distortion provides a superior subjective quality of the synthetic speech. The inclusion of this noise shaping effect is made by using a frequency weighting distortion defined by

$$J_W = (\mathbf{y}-\mathbf{y}_q)^T \mathbf{W}^T \mathbf{W} (\mathbf{y}-\mathbf{y}_q) \qquad (11)$$

where the weighting matrix **W** contain the impulse response of the perceptual filter $W(z)=A(z)/A(z/\gamma)$, used usually in CELP and others Vector Excited Coders (VXC). By using the expresion (1), the formula (11) leads to

$$J_W = (\mathbf{x}-\mathbf{x}_q)^T \mathbf{H}^T \mathbf{W}^T \mathbf{W} \mathbf{H}(\mathbf{x}-\mathbf{x}_q) = (\mathbf{x}-\mathbf{x}_q)^T \mathbf{H}_W^T \mathbf{H}_W(\mathbf{x}-\mathbf{x}_q) \qquad (12)$$

where $\mathbf{H}_W$ matrix contains the impulse response of the $H_W(z) = 1/A(z)\ W(z) = 1/A(z/\gamma)$; that is to say, the new impulse response is $h_W(n)=h(n)\ \gamma^n$, and $0<\gamma<1$.

Making the SVD of the $\mathbf{H}_W$ matrix we obtain

$$J_W = (\mathbf{g}-\mathbf{g}_q)^T \mathbf{D}_W^2 (\mathbf{g}-\mathbf{g}_q) \qquad (13)$$

where $\mathbf{g} = \mathbf{V}_W^T \mathbf{x}$ and $\mathbf{V}_W$ contains the singular vectors of the $\mathbf{H}_W$ matrix and $\mathbf{D}_W$ matrix contains its singular values.

The distortion defined by the expresion (13) is used for the codebooks design and the signal coding. The noise shaping leads to a SVD transform presenting a less singular values spread. This way, the relative importance of the different g_i components is more uniform and, therefore, the bit assignment is more uniform too. Thus, the signal-to-noise accross the different codebooks achieves a more constant value.

The preceding ideas and the spectral interpretation of the square singular values pointed in the section II suggests an alternative noise shaping approach. It consist in to weight directly the singular values in the distortion formula (9) by an exponent less than unity. Thus, the new weighting distortion is defined by:

$$J_W = (\mathbf{g}-\mathbf{g}_q)^T \mathbf{D}^{2\gamma} (\mathbf{g}-\mathbf{g}_q) \qquad (14)$$

Clearly, the less singular values spread obtained are equivalent in both approachs. The formula approach present a minor drawback in the decoding or synthesis process. The original LPC matrix **H** must used in the speech reconstruction and the synthesis formula (3) must be replaced by

$$\mathbf{y} = \mathbf{U}\ \mathbf{D}\ \mathbf{V}^T\ \mathbf{V}_W\ \mathbf{g}_q \qquad (15)$$

and, therefore, two SVD should be required. Clearly, the LPC filtering of the $\mathbf{x}_q = \mathbf{V}_W\ \mathbf{g}_q$ is prefered.

VI. THE LPC COEFFICIENTS QUANTIZATION

In the LPC coefficients quantization an Adaptive Multistage VQ (AMSVQ) developed by the authors has been used [3]. The MSVQ method has always been seen as a suboptimal VQ scheme with reduced complexity and storage. It is adequate for the VQ of the large dimension vector as the LPC coefficients vector. The Adaptive MSVQ updates the codewords of each stage from its quantization error that is available in the output of the next stage. The update approach makes use of the time correlation between adjacent LPC vectors. The details of AMSVQ approach can be seen in [3]. The AMSVQ of the LAR-LPC (Log-Area-Ratio) coefficients has shown a very good performance when it has been integrated in a CELP and a Multipulse coders.

VII. CODEBOOK DESIGN

The LBG algorithm based on a large training data base has been used for the codebook design of the multi VQ. The excitation vectors in the data base are segmented in subvectors and the each codebook is designed with a standard LBG algorithm. In the dynamic bit allocation case, several codebook sizes has been considered for each subvector. An high independence between the singular values and the corresponding excitation signal has been observed. It permits to design all of the codebook of different sizes with a common cluster with a little perfomance loss. Thus, the density of training vectors per centroid is always holded very high. The centroid calculation take into account the weighting distortion (9), (13) or (14) if noise shaping is considered) resulting in

$$c_i = \frac{\sum_{l=1}^{L} d_{il}^2 g_{il}}{\sum_{l=1}^{L} d_{il}^2} \qquad (16)$$

where c_i is th i^{th} component of the centroid and L is the number of the training subvectors in the corresponding cluster.

VIII. EMPIRICAL RESULTS

In this section we present some previous results obtained by the proposed coder. The tests was carried out at the bit rate range of 7-8Kb/s. The proposed system present sa large tradeoff between performance and complexity. For example, an increase of the speech frame duration leads to a better profit of the vector quantization properties, a more efficient bit allocation approach and a reliable noise shaping. On the contrary, a large vector dimension leads to an high complexity and the codebook size must be limited to a relative low value. Then, the bit allocation can be often trangressed leading to a performance loss. In these previous results, the complexity has been the major criterion on the choice of the parameter magnitudes. Thus, the frame or vector length is 25, the number of subvectors is 4 and fixed subvector lengths are 2, 3, 10 , 10 in increasing singular values order. These subvector lengths have been aproximately optimized in the fixed bit allocation case, and these have been holded in the dynamic bit allocation case. Defining the subvector complexity as $C_i = k_i.2^{Mb_i}$, this magnitude has been limited by $C_i < 3100$. So, the two subvectors of largest singular values allow a maximum bit assignment of 10 bits.

In the LPC analysis the speech signal was not preemphasized and the autocorrelation method with a Hamming window of 205 samples (25,625ms) was used to obtain 10 LPC coefficients every 175 samples or 7 speech frames. The overlapping is of 30 samples. The log-area-ratio (LAR) are quantized with the AMSVQ approach and the corresponding quantized LPC coefficients are obtained and they are used in the signal coding.

The training data base consists in an high number of spanish utterances. A balanced sample of good and bad speakers (male and female) utter a mixed ensemble of phrases with high contents of unvoiced sounds. In order to classify these utterances by its predictivity, all of these were coded by the CCITT-ADPCM 32 Kb/s. They range from 20 dB to 30 dB in SEGSNR, with an average SEGSNR about 25 dB. A small ensemble of different utterances and phrases is used as test or outside material. The average SEGSNR of this ensemble ranges about 25 dB in SEGSNR, too. It is a SNR low value, indicating a discreet average predictivity of the data speech. Firstly, a fixed bit allocation case is simulated. The bit allocation is calculated from expression (11). For a bit number of 21 (6.72 Kb/s), an assignment of 8, 8, 5, 0 bits is found for the subvectors of dimension 2,3,10,10 respectively. Thus, the last subvector is not sent. For this case and using non-quantized LPC coefficients, the objetive quality of the synthetic speech results SNR=11,9dB and SEGSNR = 10 dB. The subjective quality can be considered as very good communication. Due to the great vector per centroid ratio used in the codebook design, the inside and the outside quality is very aproximately identical. The loss of almost 2dB in SEGSNR is due to the bad coding of the low energy frame and it is noticeable in the listening test. The use of the dynamic bit alloeation approach defined by the formula (10) we report an improvement of the speech quality. The signal to noise results are SNR = 12,5dB and SEGSNR=10,8dB and the subjective quality is higher than in he fixed bit case.

The perceptual quality of the speech can be improved by the use of a noise shaping approach. We have tested the noise shaping defined by the expresion (13) for several values of the γ parameter. A good result is obtained for γ values about 0.9. It is obtained a lower SNR values but the subjective quality is improved. The introduction of the LPC quantization by means the AMSVQ approach reduces lightly the coder performance. If the AMSVQ-27bits [3] is used, the SNR measures are decreased about 0.5dB and it is not observed any noticeable change in the listening test. If the AMSVQ-19 bits [3] is used, the SNR measures are decreased about 1dB and an ligth audible distortion is present. The first case is a fully-quantized coder with bit rate about 8kb/s (6.72+1.23Kb/s). The second case ranges about 7,5Kb/s (6.72+0.87Kb/s).

A drastic reduction of the bit rate requires the use of larger dimension frame what should permit a more efficient profit of the VQ properties and, overcoat, a more efficient bit assignment and noise shaping. Of course, the use of a long-term predictor should provide some great improvements. At present we are tested a scheme working with a frame length of 40 samples, the dynamic bit allocation method with variable subvector length indicated in section IV and a noise shaping scheme.

IX. CONCLUSIONS

In this paper a Predictive SVD-Transform Coder is proposed. The orthogonal representation of the speech signal provides a great insigth in the perceptual importance of the LPC excitation components. Thus, some bits assignment schemes based on the singular values distribution are presented. All of these do not need to transmit any side information. Likewise, some noise shaping approachs are proposed in order to improve the perceptual speech quality. The noise shaping effect is obtained by means of a reduction of the singular values spread. Some previous results are described and some preliminary versions of the Adaptive Predictive Transform Coder are presented. These results point out a good perfomance of this coding scheme, and it is hoped that the consideration of the aditional improvements such as the use of larger vector dimensions, variable subvectors lengths and the inclusion of a long-term predictor (usual in equivalent coders as CELP and Multipulse) provides high quality at rates below 0.75 bits/sample.

REFERENCES

[1] B.S. Atal. "A model of LPC excitation in terms of eigenvectors of the autocorrelation matrix of the impulse response of the LPC filter". Proc. of ICASSP'89, pp. 45-48. Glasgow. 1989.

[2] F. Deppretere (Ed.). "*SVD and Signal Processing. Algorithms, Applications and Architectures*". Elsevier Science Publ. Co. 1988.

[3] J.A. Rodríguez-Fonollosa et al. "Robust LPC vector quantization based on Kohonen's design algorithm". Proc. od EUSIPCO'90, pp. 1303-1306. Barcelona. 1990.

[4] V. Cuperman. "On adaptive vector transform quantization for speech coding". IEEE Trans. on Communication. Vol. 37, No. 3, pp. 261-267. March 1989.

[5] I. Trancoso, B.S. Atal. "Efficient search procedures for selecting optimum innovation in stochastic coders". IEEE Trans. on Ac., Speech, and Signal Processing. Vol. 38 No. 3, pp 385-395. March 1990.