COMPARISON OF THE MODIFIED HERMITE TRANSFORMATION WITH OTHER UNITARY TRANSFORMATIONS IN A PTC SCHEME

Victoria E. Sanchez, Jose C. Segura, A.M. Peinado,
Juan M. Lopez and Antonio J. Rubio

Dpto. de Electronica y Tcno. de Computadores
Facultad de Ciencias, Univ. de Granada
Granada (SPAIN)

ABSTRACT

Recently a Predictive Transform Coding (PTC) scheme has been proposed. This is a transform coding scheme with a strong link to the LPC model of speech production. In this paper several unitary transformations are studied within this scheme. These are the Discrete Cosine Transform, a unitary transformation resulting from applying the Singular Value Decomposition to the impulse response matrix of the LPC filter, the identity transformation and the recently developed Modified Hermite Transformation. We determine the number of parameters needed in this scheme for each transformation, in order to have a high quality synthesis and make both objective and subjective measures.

1. INTRODUCTION

In the last years residual speech coders have had a great development. In these kind of coders the speech signal is represented by the LPC filter and by the LPC residual as the excitation. Different representations of the LPC residual lead to different schemes, such as multipulse, CELP and others which attempt to represent the residual in a simpler manner [3].

Recently a unified framework for LPC excitation representation in residual speech coders has been presented [3]. Within this framework a new scheme has been proposed called Predictive Transform Coding (PTC), which is a transform coding scheme with a strong link with the LPC model of speech production.

In this scheme we are going to study the performance of several unitary transformations, among them the recently developed Modified Hermite Transformation, and to determine the number of parameters needed for the excitation in order to have a high quality synthesis.

This work has been supported by CICYT (project TIC 88-0774).

2. VECTORIAL EXPRESION OF THE LPC EXCITATION

Let's consider the excitation $x(n)$ of the LPC filter as a linear combination of a subset of vectors $\mathbf{L}$ taken from a given set $\mathbf{V}$ of M vectors of length N [3]

$$\mathbf{x} = \sum_{i=1}^{l} b_i \mathbf{v}_i, \quad \mathbf{v}_i \in \mathbf{V}, \quad l < N \tag{1}$$

and in matrix form

$$\mathbf{x} = \mathbf{Lb} \tag{2}$$

We have for the synthesized signal $\hat{s}(n)$

$$\hat{s}(n) = \sum_{k=0}^{n} h(n-k)x(k) + r(n) \tag{3}$$

where $h(n)$ is the impulse response of the LPC filter, $x(n)$ the excitation and $r(n)$ is the contribution to the present frame of the excitation in the previous frames. The error is given by

$$e(n) = s(n) - r(n) - x(n)*h(n) \tag{4}$$

Weighting $e(n)$ with a filter of the form

$$W(z) = \frac{1 + \sum_{k=1}^{p} a_k z^{-k}}{1 + \sum_{k=1}^{p} a_k \gamma^k z^{-k}} = \frac{A(z)}{A(z/\gamma)} \tag{5}$$

we can tolerate larger errors in the formant regions than in the in-between formant regions. Finally we have for the weighted mean squared error

$$E_w = \sum_{n=0}^{N-1} \left([f(n) - x(n)]*h_w(n)\right)^2 \tag{6}$$

where $f(n)$ is the signal resulting from passing $s(n)-r(n)$ through the filter $A(z) = 1/H(z)$ and $h_w(n)$ is the impulse response of the weighted LPC filter. Substituting (1) in (6) and expressing it in matrix form

$$E_w = (\mathbf{f} - \mathbf{Lb})^t \mathbf{H}_w^t \mathbf{H}_w (\mathbf{f} - \mathbf{Lb}) \tag{7}$$

Let's now minimize E_w with respect to the excitation. For a given subset $\mathbf{L}$ the coefficients $\mathbf{b}$ that minimize E_w are given by [3]

$$\mathbf{b}_m = (\mathbf{L}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{L})^{-1} \mathbf{L}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{f} \tag{8}$$

and replacing (8) in (7) we have the minimum value of E_w for a given $\mathbf{L}$

$$E_w^m = \mathbf{f}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{f} - [(\mathbf{f}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{L})$$

$$(\mathbf{L}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{L})^{-1} (\mathbf{L}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{f})] =$$

$$\mathbf{f}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{f} - \Delta E_w^m \tag{9}$$

As $\mathbf{V}$ is known, we still have to determine which subset $\mathbf{L}$ of $\mathbf{V}$ is the one that minimizes E_w^m. From (9) it is clear that it will be the one that maximizes ΔE_w^m.

So the excitation is characterized by the indexes of the vectors of $\mathbf{V}$ that belong to $\mathbf{L}$ and by the values of the coefficients $\mathbf{b}_m$.

Let's take now $\mathbf{H}_w^{-1}\mathbf{T}$ as $\mathbf{V}$, being $\mathbf{T}$ a unitary matrix. As $\mathbf{L}$ is a column submatrix of $\mathbf{V}$ then

$$\mathbf{L}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{L} = \mathbf{I}_l \tag{10}$$

where $\mathbf{I}_l$ is the identity matrix of order $l \times l$. Substituting in ΔE_w^m we have

$$\Delta E_w^m = (\mathbf{f}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{L})(\mathbf{L}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{f})$$

$$= \| \mathbf{L}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{f} \|^2 \tag{11}$$

We will maximize ΔE_w^m by choosing as $\mathbf{L}$ those vectors of $\mathbf{V}$ whose indexes correspond to the l biggest magnitude elements of the vector

$$\mathbf{q} = \mathbf{V}^t \mathbf{H}_w^t \mathbf{H}_w \mathbf{f} =$$

$$\mathbf{T}^t \mathbf{H}_w \mathbf{f} = \mathbf{T}^t \mathbf{y} \tag{12}$$

where $\mathbf{y}$ is the signal $s(n)-r(n)$ weighted by the filter $W(z)$. The coefficients $\mathbf{b}_m$ are given by the values of the elements of $\mathbf{q}$ chosen.

As it can be seen in (12) we are applying the unitary transform $\mathbf{T}^t$ to the weighted speech signal $s(n)-r(n)$, and considering equal to zero the smallest magnitude elements as common in conventional transform coding. However this scheme is different because the speech signal is weighted by the filter $W(z)$ and it is considered the contribution of the previous frames to the present frame. Due to this link with the LPC model of speech production, this scheme has been called Predictive Transform Coding (PTC) [3].

We are interested in studying the performance of different unitary transforms and , specially, in determining how many elements can be considered equal to zero in this new scheme for each transformation without degrading speech quality.

3. SEVERAL UNITARY TRANSFORMS

We are going to study four different unitary transforms: the identity transform, the unitary transform $\mathbf{U}$ resulting from applying the Singular Value Decomposition (SVD) to the matrix $\mathbf{H}_w$, the discrete cosine transform (DCT) and, finally, the recently developed [2] Modified Hermite Transformation (MHT).

3.1 Identity transform

Making $\mathbf{T} = \mathbf{I}$ results in

$$\mathbf{q} = \mathbf{T}^t \mathbf{y} = \mathbf{y} \tag{13}$$

So in this case we just take the biggest magnitude elements of the weighted speech signal, and reconstruct the signal by passing these elements through the inverse weighting filter $1/W(z)$ and adding the signal $r(n)$.

3.2 Unitary transform resulting from SVD

Let's weight (3) and express it in matrix form

$$\mathbf{y} = \mathbf{s}_w - \mathbf{r}_w = \mathbf{H}_w \mathbf{x} \qquad (14)$$

If now we apply the Singular Value Decomposition to the matrix $\mathbf{H}_w$

$$\mathbf{H}_w = \mathbf{U}\mathbf{D}\mathbf{K}^t \qquad (15)$$

where **U** and **K** are unitary matrixes and **Σ** is a diagonal matrix. Making **T** = **U** we finally have

$$\mathbf{q} = \mathbf{U}^t \mathbf{y} = \mathbf{D}\mathbf{K}^t \mathbf{x} \qquad (16)$$

3.3 Discrete Cosine Transform

The Discrete Cosine transform of a data sequence $s(n)$, $n = 0,1,\ldots,(N-1)$ is given by [1]

$$c_s(0) = \frac{\sqrt{2}}{N} \sum_{n=0}^{N-1} s(n) \qquad (17)$$

$$c_s(k) = \frac{2}{N} \sum_{n=0}^{N-1} s(n) \cos\frac{(2n+1)k\pi}{2N} \qquad (18)$$

$$k = 1, 2, \ldots, (N-1)$$

This transform has been traditionally used due to its close performance to that of the Karhunen-Loeve transform.

3.3 Modified Hermite Transformation

This unitary transformation has been recently developed by [2] from the binomial and discrete hermite families. It is defined by the unitary matrix A

$$A(r,k) = \frac{[B(0,k)B(0,r)]^{1/2}}{2^{N/2}} H(r,k) \qquad (19)$$

where

$$B(0,k) = \binom{N}{k}$$

$$B(r,k) = \binom{N}{k} \sum_{m=0}^{r} (-2)^m \binom{r}{m} \frac{k^m}{N^m}$$

$$H(r,k) = \frac{B(r,k)}{\binom{N}{k}}$$

$$k = 0, 1, \ldots, N-1$$

$$r = 0, 1, \ldots, N-1$$

It has also been developed an algorithm for the MHT which is very easy to implement. We are interested in evaluating the performance of this new transform in a PTC scheme.

4. EXPERIMENTAL RESULTS

For performance evaluation we used four Spanish sentences pronounced by two male speakers and two female speakers. The sample was low-pass filtered at 4 kHz cut-off frequency and digitized by a 12 bit A/D converter at 8 kHz sampling.

The length of the analysis frame was considered of 15 msec, 120 samples, and it was divided into 3 sub-blocks of 40 samples each. Synthesis filter order was established to be 10 and the value of $\gamma = 0.8$.

We have studied the performance of the four transformations in three different cases:

A- We consider different from zero 5 values in each sub-block.

B- We consider different from zero 10 values in each sub-block.

C- We consider different from zero 15 values in each sub-block.

In all the cases without quantization.

For objective evaluation we have used the Segmental SNR. In figures 1a and 1b the SNRseg is plotted for male and female sentences respectively. We evaluate the four transformations at the three different cases A, B and C.

For subjective evaluation we have used the Mean Opinion Score (MOS) scale with five categories, ranging from 1 (Unacceptable) to 5 (Excellent). Opinion rating was made by ten listeners over the four Spanish sentences. In figure 2 we have the MOS for the four transformations at the three different cases A, B and C.

The results obtained show a close objective performance of the DCT and of U. Their SNRseg is 1.5 db approximately better than that of the MHT, being the performance of the identity transformation quite worse. As far as the subjective measure is concerned we get a MOS of 3 to 4 for DCT, U and MHT if we consider different from zero

10 to 15 elements. A score of 4.0 on the MOS scale signifies high-quality and 3.5 is an acceptable quality for telephone communication. As it can be seen the MHT also has a worse subjective performance with respect to the DCT and U, however this difference diminishes as the number of elements different from zero increases, being the performance of these three transforms quite close for around 15 elements.

There is another point to take into account when evaluating the different transforms and it is the amount of calculation involved. The quickest to implement is, of course, the identity transform but it gives worse results. U implies a large amount of calculation as we have to do the SVD of the matrix $\mathbf{H}_w$ for every frame. For the DCT and the MHT there are fast algorithms developed, being the one developed for the MHT easier to implement than any of those developed for the DCT [2] to the authors knowledge. So the MHT is more advantageous in this aspect.

The results obtained are without quantization of the different parameters, what would evidently introduce a certain amount of degradation over this results.

5. CONCLUSIONS

We have studied the performance of four unitary transforms in a predictive transform coding (PTC) scheme. The identity transformation is very easy to implement but it gives poorer results. The DCT and U have a better objective and subjective performance over the MHT, but for around 15 elements different from zero, the subjective performance of these three transforms is very close, with a MOS of about 4 without quantization. As the MHT is easier to implement, this transform can be a good election when implementation reasons dominate in the development of PTC coders.

REFERENCES

[1] Ahmed N., Natarajan T., Rao K.R., *"Discrete Cosine Transform"*, IEEE Trans. on Compt., January 1974.

[2] Haddad R.A. and Akansu A.N., *"A New Orthogonal Transform for Signal Coding"*, IEEE Trans. on ASSP, vol. 36, No. 9, September 1988.

[3] Ofer E., Malah D. and Dembo A., *"A Unified Framework for LPC Excitation Representation in Residual Speech Coders"*, ICASSP 1989

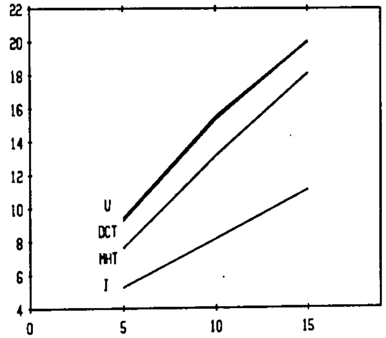

Figure 1a.- SNRseg(db) for cases A,B,C for male sentences

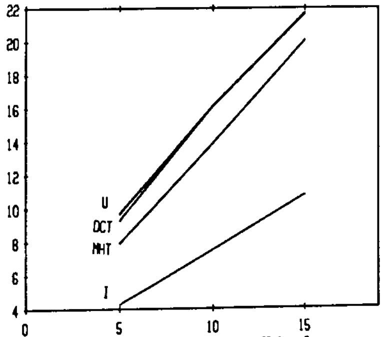

Figure 1b.- SNRseg(db) for cases A,B,C for female sentences

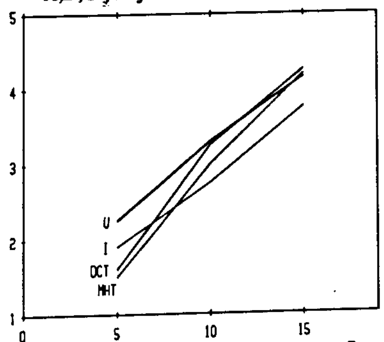

Figure 2.- Mean Opinion Score (MOS) for cases A,B,C