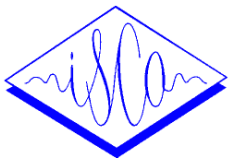

3rd International Conference on
Spoken Language Processing (ICSLP 94)
Yokohama, Japan
September 18-22, 1994

IMPROVING CELP VOICE QUALITY BY PROJECTION SIMILARITY MEASURE

Miguel A. Ferrer-Ballester[†], *Aníbal R. Figueiras-Vidal*[††]

[†] ETSI Telecomunicación, Univ. Las Palmas, Campus de Tafira, 35017 Las Palmas, Spain.
Tel/Fax: +34 28 451250 / 451243 E-mail: mferrer@cibeles.teleco.ulpgc.es
[††] GPSS-DSSR, ETSI Telecom-UPM, Ciudad Universitaria, 28040 Madrid, Spain.
Tel/ Fax: +34.1.3367226 / 3367350 E-mail: weruaga@gtts.upm.es

ABSTRACT

It is well-known the limited quality of the reconstructed speech by Code-Excited Linear Prediction (CELP) speech coder at low bit-rate (2400-8000 bps). The major source of audible distortion has been attributed to an inaccurate degree of periodicity of the voiced speech signal. In the present paper we alleviate this drawback by modifying the MSE distance measure, which is not able to capture the periodicity adequately. The new error criterion proposed is a projection similarity measure, which computes a projection distance of original onto coded perceptually weighted voice on a point to point basis. The improvement of the quality of the speech reconstructed by the CELP with projection distance measure has been checked by subjective A-B test. This result emphasizes the perceptual importance of an adequate description of the pitch-pulse waveform of the original (uncoded) LP residue by the CELP-coded one.

I. INTRODUCTION

Speech coding algorithms operating at rates between 2.4 and 8.0 Kb/s have been significantly improved during the last decade [1]. Among the current coders operating at these rates, this paper pays attention to the Code-Excited Linear Prediction (CELP) [2], a type of linear-prediction (LP) speech coding algorithm based on analysis-by-synthesis techniques.

The intelligibility of the speech reconstructed by this type of coders is high; however, a drawback for many practical applications is the increasingly unnatural speech quality with decreasing bit rate.

This unnatural speech quality is mainly caused by three types of distortion, which are usually described as noise, reverberation, and tonal artifacts [4]. All three are closely connected with the periodicity of the voiced speech signal. Then, it can be said that the major source of audible distortion is mainly caused by an inaccurate degree of periodicity of the voiced speech signal [4].

Some solutions, which change the CELP-coder structure, have been proposed in order to allow obtaining a highly periodic excitation; for instance, Grazow *et al.* have proposed in [5] to synthesize periodic speech using single-pulse excitation, and Wang and Gersho [6] have coded each frame taking into account the local phonetic content. Other approaches to avoid the problem are based on a sinusoidal representation of the speech signal [7]. All the above schemes include voiced/unvoiced decision.

In this paper, in order to improve the speech periodicity, we pay attention to the traditional objective measure: the signal-to noise ratio (SNR). It is well known that the SNR is not able to capture the periodicity adequately; in fact, even in waveform coders such as CELP it has been found advantageous to modify the reconstructed voiced speech signal to have a higher level of periodicity, although this results in a lower instantaneous SNR [8].

Considering the above reason, we have modified the MSE distance measure of the CELP by a projection distance, which computes a projection distance of original onto coded perceptually weighted voice on a point to point basis. This approach has the advantage of improving the quality of the final CELP-speech neither changing the CELP-coder structure nor including the voice/unvoiced decision. Additionally, the computational load is not significantly increased.

The organization of this paper is as follows: in the next section we describe the CELP speech coder algorithm. Section III presents the proposed projection distance measure. In section IV we give the graphical, subjective and objective results. After it, the conclusions in section V close the paper.

II. THE CELP SPEECH CODER DESCRIPTION

In this section we describe briefly the CELP, following the notation of [8], as a necessary introduction to the application of the projection distance measure to the CELP.

The CELP-coder procedure (a block diagram is depicted in Fig. 1), is as follows: speech signal $s(n)$ is divided into frames of length L and subframes of length N (usually $L/N=4$). Each subframe is filtered by a pole-zero, noise-weighting linear filter to obtain $X(z)=S(z)A(z)/W(z)$, where $x(n)$ is the target signal for the coding process, $A(z)=1+a(1)z^{-1}+...+a(p)z^{-p}$ (with $p=10$) is the standard LP polynomial corresponding to the current frame, and $W(z)=1+a'(1)z^{-1}+...+a'(p)z^{-p}$ is a modified polynomial obtained from $A(z)$ by shifting the zeroes towards the origin in the z-plane, i.e., by using the coefficients $a'(i)=a(i)\gamma^i$ with $0<\gamma<1$ (typical value: $\gamma=0.8$).

The coder attempts to synthesize, at every frame, a signal $y(n)$ as close to the target signal $x(n)$ as possible, usually in a Mean-Square-Error (MSE) sense. The synthesis algorithm is based on the following equations.

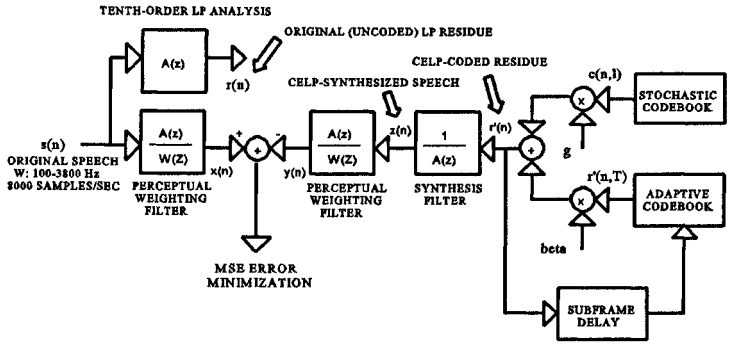

Fig.. 1 - Block diagram of a CELP analysis by synthesis speech coder.

$$r'(n,T)=r'(n-kT) \;;k=int(\frac{n}{T}+1) \;;1\le n\le N \qquad (1)$$

$$y(n)=y_0(n)+\beta r'(n,T)*h(n)+gc(n,l)*h(n) \;;1\le n\le N \qquad (2)$$

where * denotes the convolution operation, $h(n)$ is the impulse response of the $1/W(z)$ filter in the range [0,...,N-1], $y_0(n)$ is the initial state response (without any input), and $r'(n)$ is the past coded LP excitation.

The other variables are the parameters that code the LP residue, and are divided in two groups. The first group, called adaptive contribution, is formed by pitch gain β and pitch delay T. Note that $r'(n,T)$ is a codebook of vectors (as many as possible values of T) which is updated every subframe: it is called the adaptive codebook. The second group, called stochastic contribution, is formed by the stochastic vector gain g and the stochastic codebook label l. Each vector $c(n,l)$ of the stochastic codebook contains a white Gaussian sequence of zero mean and unit variance. The possible values of these parameters are contained in several tables: the size of these tables determines the number of bits available to the system for synthesizing the coded voice.

A key issue in CELP coding is the strategy of selecting a good set of parameters from the various codebooks to minimize the distance between the target and the synthesized signal. An exhaustive search, although possible in principle, is prohibitively complex. Therefore, sub-optimal procedures are used. The usual strategy is to separate pitch parameters T and β from stochastic parameters g and l, and to select the two groups independently. T and β are found first and, after it, the best g and l are found. Three steps are required:

1. Find best T and β according to eq.(3):

$$T^*,\beta^*=\underset{T,\beta}{argmin}\left(\sum_{n=1}^{N}[x(n)-y_0(n)-\beta r'(n,T)*h(n)]^2\right) \qquad (3)$$

2. Calculate the resulting error signal with eq.(4):

$$d(n)=x(n)-y_0(n)-\beta^* r'(n,T^*)*h(n) \;;1\le n\le N \qquad (4)$$

3. Finally, the coder tries to find g and l which best match $d(n)$ according to eq.(5):

$$g^*,l^*=\underset{g,l}{argmin}\left(\sum_{i=1}^{N}[d(n)-gc(n,l)*h(n)]^2\right) \qquad (5)$$

After getting the parameters, we synthesize the CELP voice by the means of

$$r'(n)=\beta^* r'(n,T^*)+g^* c(n,l^*) \;;1\le n\le N \qquad (6)$$

$$\sum_{i=0}^{P} a_i z(n-i)=r'(n) \qquad (7)$$

where $z(n)$ is the CELP synthesized voice.

III. THE PROJECTION DISTANCE MEASURE

The limited quality of the MSE criterion in the reconstruction of the periodic signal could be seen from the function of the codification-error energy of the CELP-coded residue, which is defined in eq.(8) as $E_{RCE}(n)$ and depicted in Fig. 2.

$$E_{RCE}(n)=\sum_{m=-M}^{M}[r(n+m)-r'(n+m)]^2 \;;1\le n\le N \qquad (8)$$

where $r(n)$ is the original (uncoded) LP residue, $r'(n)$ is the CELP-coded residue, N the subframe length. We have found that four is a good value for M.

In this figure we can see how the error is periodic, that is, there are some aspects of the periodicity that the MSE criterion is not able to describe correctly.

The contribution of this paper consists on a new distance criterion: the projection similarity measure described in [9]. The projection similarity measure computes a projection distance of one curve onto another on a point by point basis. This projection distance can be summed for all points N of the pattern in an L_p-norm fashion to form a distance measure as described by

$$d^r_{proj}(x,y)=\frac{1}{N}\sum_{n=1}^{L}|Proj_{y(n)}x(n)|^r \qquad (9)$$

This process is illustrated in mode detail in Fig. 3.

Fig. 2 - Waveform of the original LP residue (continuous line) overlapped with the normalized codification-error energy of the CELP-coded residue (broken line): $E_{RCE}(n)$.

This distance measure can be made symmetric and thus become a true metric by using,

$$d(x,y)=\frac{1}{2}(d_{Proj}(x,y)+d_{Proj}(y,x)) \qquad (10)$$

The term $Proj_{y(n)}x(n)$ can be computed as follows. Let for each point n the curve $y(n)$ is given by (x_1, y_1) and a slope s estimated by first differences of adjacent points. The point n on the curve $x(n)$ is given by (x_2,y_2). We wish to find the projection distance of point (x_2,y_2) onto a line defined by (x_1,y_1) and the slope given by $s=tan\alpha$. This can be determined by shifting the coordinate system by $(-x_1,-y_1)$ and then rotating it by $-\alpha$. The transformed coordinate y'_2 will be the projection distance. This result in a formula for the projection distance of

$$Proj_{y(n)}x(n)=\frac{(y_2-y_1)-(x_2-x_1)|s|}{\sqrt{1+s^2}} \qquad (11)$$

In the case of a point by point distance $x_1=x_2=n$ and $y_1=y(n)$, $y_2=x(n)$, thus simplifying the distance to

$$Proj_{y(n)}x(n)=[x(n)-y(n)]\cos\alpha \qquad (12)$$

which is just the coordinate distance scaled by a rotation. This measure proved to reflect visual similarity of two patterns better than the previously mentioned measure.

In the application of this measure to the CELP, we have used $r=2$ (see eq.(9)) and only in the adaptive contribution (for the stochastic contribution we have used the traditional MSE distance measure). Then, the eq.(3) becomes to:

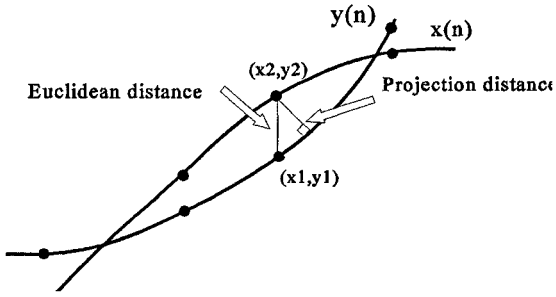

Fig. 3- Projection distance measure.

$$T^*,\beta^*=\underset{T,\beta}{argmin}\left(\sum_{n=1}^{N}\{[x(n)-y_0(n)-\beta r'(n,T)*h(n)]p^2(n)\}^2\right)$$

with $p(n)=(\cos\alpha_1+\cos\alpha_2)/2$, α_1 the angle of the slope of $x(n)$ on point n, and α_2 the angle of the slope of $y(n)$ on point n. Then, after select T, we will calculate β as follows:

$$\beta=\frac{\sum_{n=1}^{N}p(n)x(n)y(n)}{\sum_{n=1}^{N}p(n)y(n)y(n)} \qquad (14)$$

The computation loads increase of the CELP with projection criterion is less than 10% with respect to standard CELP.

The choice of this way of modifying the error measure has the following advantages: 1) it does not need additional bits, 2) the additional computation charge is very little, and 3) it does not modify the CELP-coder structure. Also, in the informal perceptual test, the quality of the unvoiced frames remains unchanged. In the same way, the new distance measure does not modify the CELP robustness against additive noise.

IV. RESULTS

In this section we present results of the projection distance measure CELP-coder in comparison with the 4800 bps standard CELP-coder [3]. The parameters have been coded in the same way for both schemes. The results are classified in three subsections: 1) graphical views of the improvement of the projection distance measure CELP-coder, 2) perceptual comparisons, and 3) objective measures.

4.1 Graphical comparisons

A comparison between the $E_{RCE}(n)$ (see eq.(8)) functions obtained from the standard CELP and the projection distance measure CELP can be seen in Fig. 4. This figure shows how the final objective has been reached: to smooth the function of the codification-error energy of the CELP-coded residue. Consistently, the projection distance measure CELP-coder will reproduce the original periodicity better than the MSE CELP-coder.

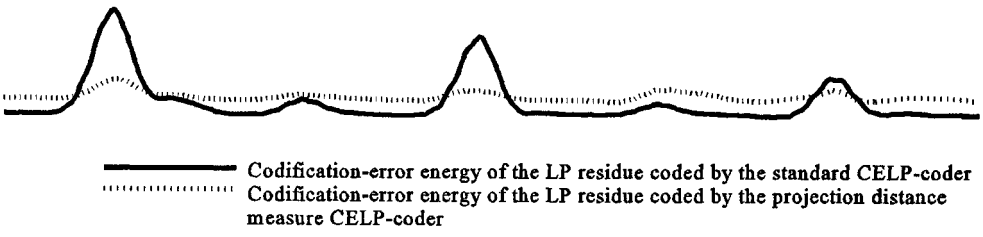

Fig. 4 - Comparison by overlapping between the codification-error energy of the LP residue coded by standard CELP (continous line) and projection distance measure CELP (broken line) coder.

The penalty of the above improvement is an increase of the error of codification in unvoiced frames (see Fig. 4) of the projection distance CELP speech coder.

3.2 Subjective comparison

In order to obtain a subjective comparison, we have generated two versions of coded speech: one from a 4800 bps standard CELP, and the other from the 4800 projection distance measure CELP. The coding procedures have been the same for both schemes. Then, the difference between the two coded versions of speech can be attributed here to the different distance measure.

Subjective quality evaluation has been done here through informal A-B or pair comparison test [10] using 10 listeners. Six Spanish sentences (spoken by three male and three female speakers recorded form different sources: microphone, AM and FM radio station) have been used for evaluation. Each comparison is done between the two coded versions of a sentence. All possible A-B pairs are generated and presented in a randomized order. Listeners' task has been to prefer either one or the other of the two-coded versions, or to indicate no preference. Results from these informal tests have shown that 80% indicated projection distance measure CELP preference, 5% standard CELP preference, and 15% indifference. From this, we have concluded that, in this experiment, the 4800 bps WMSE CELP-coder is preferred over the 4800 bps standard CELP.

3.3 Objective results

The objective evaluation has been done with the segmental signal-to-noise ratio (SEGSNR) [10]. As expected, the SEGSNR obtained with the projection distance measure has been less (about 3 dB) that the SEGSNR offered by MSE standard CELP (about 8.0 dB). A reason of the above is that the projection distance measure CELP does not search to minimize the MSE distance.

V. CONCLUSIONS

We have introduced a CELP with a new distance measure (projection distance measure) to improve the degree of periodicity in the reconstructed CELP speech signal. The new distance measure smooths the energy of the error of codification of the CELP-coded residue, which has maxima in pitch-pulse environments.

The subjective improvement of the reconstructed projection distance measure CELP-coder speech quality has been checked by A-B test. This result emphasizes the importance of the correct description of the original LP-residual pitch-pulse waveform by the CELP-coder. This improvement is due to the better selection (in a perceptual sense) of the codevectors from the stochastic and adaptive codebooks that the projection distance criterion allows. This advantage does not require additional bits nor a significant increase of the computational load.

REFERENCES

[1] N. Jayant, "Signal Compression: Technology Targets and Research Directions", IEEE Journal on Selected Areas in Communications, vol. 10, no.5, 796-818, 1992.

[2] B.S. Atal and M.R. Schroeder, "Stochastic Coding of Speech at Very Low Bit Rate", Proc. Int. Conf. on Communications, pp. 1610-1613, Amsterdam, 1984.

[3] J.P.Campbell, T.E.Tremain and V.C.Welch, "The Federal Standard 1016 4800 bps CELP Voice Coder: Digital Signal Processing", Academic Press, 1991.

[4] W.B.Kleijn, "Continuous Representation in Linear Predictive Coding", Proc. Int. Conf. ICASSP-91, 210-204, Toronto, Mayo 1991.

[5] W.Grazow, B.S.Atal, K.K.Paliwal and J.Schroeter, "Speech Coding at 4 kb/s and Lower Using Single Pulse and Stochastic Models of LPC excitation", Proc. Int. Conf. ICASSP-91, 217-220, Toronto, Mayo 1991.

[6] S.Wang and A.Gersho, "Improved Phonetically-Segmented Vector Excitation Coding at 3.4 Kb/s,", Proc. Int. Conf. ICASSP-92, vol I, 349-351, 1992.

[7] D.W.Griffin and J.S.Lim, "Multiband Excitation Vocoder", IEEE Trans. on Acoustics, Speech and Signal Processing 36 (8), 1223-1235, 1988

[8] Y.Shoham, "Constrained-Stochastic Excitation Coding of Speech at 4.8 kb/s: Advances in Speech Coding", Edited by B.S.Atal, V.Cuperman and A.Gersho; 339-348, Kluwer Academic Publishers, 1991.

[9] W.A.Kittel,M.H.Hayes, "Symbolic representation of process monitoring signals", Signal Processing, vol. 29, no. 1, pp. 93-106, oct. 1992.

[10] N.Kitawaki, "Quality Assessment of Coded Speech: Advances in Speech Signal Processing", Edited by S.Furui and M. Sondhi; 357-385, Marcel Dekker, 1992.