# S12.8

## FRACTAL EXCITATION SIGNALS FOR CELP SPEECH CODERS

*Petros Maragos* and *Kenneth L. Young*

Division of Applied Sciences, Harvard University, Cambridge, MA 02138.

**ABSTRACT**: This paper presents a class of random signals as a new excitation codebook for stochastic predictive speech coders. These signals, known as *fractional noises*, depend on a single parameter $0 < H < 1$ that controls the low- or high-pass trend of their power spectrum $1/\omega^{2H\pm1}$. Preliminary experiments have shown that their performance improves by limiting $H$ to a subinterval of $(0,1)$ and becomes better than that of the standard Gaussian codebook as the codebook bit rate decreases from 0.25 down to 0.1 bits per excitation sample. Their parametric nature allows efficient coding by quantizing $H$. Further, with no search of the codebook but at a certain loss of speech quality, a suboptimum excitation can be estimated with an $H$ that matches the fractal characteristics of the speech residual.

## 1 Introduction

Stochastic *code-excited linear predictive (CELP)* coders [1,2] and their generalization [3] are an important class of predictive speech coders that can produce high-quality speech at low bit rates about 4.8 Kbits/sec. In CELP analysis, the pre-emphasized speech signal S[n] is filtered by the spectral envelop prediction error filter $1 - A(z) = 1 - \sum_{k=1}^{P} \alpha_k z^{-k}$ to produce the *prediction error signal*

$$U[n] = S[n] - \sum_{k=1}^{P} \alpha_k S[n-k] \tag{1}$$

Then, $U[n]$ is filtered by the pitch prediction error filter $1 - B(z) = 1 - \beta z^{-\gamma}$ to produce the *excitation* signal

$$E[n] = U[n] - \beta U[n-\gamma] \tag{2}$$

The synthesized speech is produced by exciting the synthesis filter $1/[1 - A(z)][1 - B(z)]$ with $E[n]$ as input. The parameters of the predictors $A(z), B(z)$ are found in some open-loop or closed-loop optimization procedure using weighted mean squared error criteria. During each short-time frame, the optimum excitation sequence is found by searching among all $K$ possible sequences (the "codewords") of a *codebook*, where the optimization criterion is the closeness between synthesized and original speech. The standard approach has been so far to use a codebook of $K = 1024$ white Gaussian sequences; henceforth, we refer to them as the "Gaussian" codebook.

This work was supported by the National Science Foundation under Grant MIPS-86-58150 with matching funds from Bellcore, Sun, TASC, and Xerox, and in part by the ARO under Grant DAALO3-86-K-0171.

Two important research problems in stochastic CELP coders are: i) to find useful classes of stochastic excitation signals, and ii) to develop efficient procedures that reduce the search for finding the optimum excitation sequence. In this paper we present our research results from using two new classes of parametric stochastic signals for exciting CELP-type speech coders. These two classes are finite-length sequences from random realizations of *fractional Brownian motion (FBM)* and *fractional Gaussian noise (FGN)*. (See [4] for their properties and application in modeling random phenomena.) In continuous time, FBMs are *fractal* signals because their fractal dimension exceeds their topological dimension; the FGNs are time-derivatives of FBMs. Our motivations for introducing FBMs and FGNs as excitation signals are twofold: 1) To model the random roughness (fragmentation) of speech residuals based on their fractal characteristics. (Note that the measurement of fractal characteristics can be also applied to the speech signal itself with applications to speech analysis [6].) 2) To find a more structured (e.g., parametric) class of signals for stochastic codebooks exciting speech coders. For example, by tuning the parameter of these fractal signals to the fractal characteristics of the speech residual signals we developed a method that produces a suboptimum excitation sequence without searching the codebook.

We tested our ideas by exciting the implemented CELP coders with 3 classes of random signals: a) The standard Gaussian codewords as in [1,2]. b) The FBMs, and c) The FGNs. Preliminary experiments are reported on comparing these 3 classes of excitation signals with respect to several viewpoints.

## 2 Modeling Speech Residuals with Fractional Noises

The continuous-time FBMs with parameter $0 < H < 1$ are time-varying random signals with stationary, Gaussian-distributed, and statistically self-similar increments. The FGNs also depend on the same single parameter $H$ because they are defined as the (generalized) time-derivatives of FBMs; they are stationary random signals with Gaussian distributed amplitudes. The power spectral density of FBMs is $PS_{FBM}(\omega) \propto 1/|\omega|^{2H+1}$, while that of FGNs is $PS_{FGN}(\omega) \propto 1/|\omega|^{2H-1}$. Hence, an efficient algorithm [5] to synthesize an FBM is to create a random sampled spectrum whose average magnitude is $1/|\omega|^{H+0.5}$ and its random phase is uniformly distributed over $[0, 2\pi]$. In our experiments we synthesized and then transformed sampled random spectra $1/\omega^{\Delta}$ via a 1024-point inverse FFT to obtain FBM sequences (by set-



CH2847-2/90/0000-0669 $1.00 © 1990 IEEE

ting $\Delta = 2H+1$) or FGN sequences (for $\Delta = 2H-1$) from which we retained the first $N << 1024$ samples, where $N$ is the length of the codewords. FGNs could possess either a low-pass (for $H > 0.5$) or a high-pass (for $H < 0.5$) trend in their spectrum. FBMs can only have a low-pass trend in their spectrum; in the time domain the roughness of FBMs increases as $H$ decreases. Figure 1 shows several FBMs and FGNs for different $H$'s. For $H = 0.5$ FBMs and FGNs reduce to the standard Brownian motion and white Gaussian noise, respectively. The fact that FGNs and FBMs have Gaussian-distributed amplitudes correlates well with the same property of the standard Gaussian codewords; the latter was motivated by the nearly Gaussian distribution of the prediction error samples [2].

Let $X[n]$ be a finite-duration speech prediction residual sequence; i.e., it could correspond to $E[n]$ for voiced speech and either $U[n]$ or $E[n]$ for unvoiced speech. One of our goals in this work is to model $X[n]$ by using a single realization from FBM or FGN so that searching the whole codebook could be avoided. Next we outline three possible approaches:

*I. Mean Squared Error Method:* Given a collection of $K$ FBM or FGN signals $f_H[n]$ of the same duration as $X[n]$, where $H$ assumes any out of $K$ distinct values in the interval $(0,1)$ or some subinterval, the problem is to find an amplitude $G$ and an $H$ such that the error

$$\mathcal{E}(G,H) = \sum_n (X[n] - G f_H[n])^2$$

is minimized, where $\sum_n = \sum_{n=0}^{N-1}$. Setting $\partial\mathcal{E}/\partial G = 0$ yields that $G = G(H) = (\sum_n X[n] f_H[n])/(\sum_n f_H^2[n])$. Replacing this value of $G$ in $\mathcal{E}$ yields an error

$$\mathcal{E}^*(H) = \sum_n X^2[n] - \frac{(\sum_n X[n] f_H[n])^2}{\sum_n f_H^2[n]}$$

Then, by exhaustive search we can find first which $H$ minimizes $\mathcal{E}^*$, and then find $G$. However, such an approach is computationally expensive because we need to search over all $K$ codewords $f_H$.

*II. Power Spectrum Method:* The power spectrum of FBMs or FGNs obeys the law

$$\log PS(\omega) = -\Delta \log(\omega) + \textit{constant}$$

where $\Delta = 2H + 1$ for FBMs and $\Delta = 2H - 1$ for FGNs. Hence by fitting a straight line using linear regression to the plot of the data $(\log PS(\omega), \log \omega)$ and finding its slope, we can estimate $H$. This approach, however, suffers from the problems of power spectrum estimation. Thus, the variance of the spectral values varies a great deal and so does the slope $\Delta$, which makes it very unrobust. In addition, as Fig. 2 shows, the estimated $H$ using this spectrum method can even be outside its correct range $(0,1)$.

*III. Fractal Dimension Method:* The fractal dimension [4] of a signal quantifies the geometrical roughness of its graph. The fractal dimension of an FBM signal with parameter $H$ is $D = 2 - H$. Thus, if we are to model $X[n]$ with an FBM, we can estimate the fractal dimension $D(X)$ of $X$ and set $H = 2 - D(X)$. To model $X[n]$ with an FGN, we can *integrate* $X$ to produce a signal $Y[n]$, which is an FBM-like version of $X$, compute the fractal dimension $D(Y)$ of $Y$ and set $H = 2 - D(Y)$. We have found this method to be much more robust than the spectrum method for modeling either speech residuals or FGN signals. Fig. 2 shows a speech residual $X$ from voiced speech, and two FGNs to model $X$ with their $H$ parameter estimated by the fractal dimension and spectrum method. The measurement of the fractal dimension was done by using the method employed in [6] to measure the fractal dimension of speech signals. Namely, for each scale $\varepsilon = 1, 2, 3...$, morphological dilations and erosions of the signal $X$ by structuring elements of size $\varepsilon$ were used to create upper and lower envelops of $X$. These envelops create a cover around the signal. Integrating the difference signal between these two envelops yields the *area* $Cov(\varepsilon)$ of this cover. Fitting a straight line (by using linear regression) on the data $(\log Cov(\varepsilon)/\varepsilon^2, \log 1/\varepsilon)$ yields the fractal dimension as the slope of this line.

## 3 CELP Coder Operation

Two types of coders were implemented in this research: i) the standard CELP coder [1,2], where the parameters of both the short-delay predictor $A(z)$ and the long-delay predictor $B(z)$ are found in an open-loop optimization (i.e., during the analysis phase), and ii) the more general CELP coder of [3], where the pitch predictor parameters $\beta, \gamma$ are found in a closed-loop optimization, i.e., in an analysis-by-synthesis procedure. In both coders the excitation is found via a closed-loop synthesis procedure. The CELP coder of [3] yielded a significant gain in performance (since the parameters of $B(z)$ are optimized in a closed-loop) with an attendant increase in processing time. Due to its superior performance, for the rest of this paper we focus only on this more general CELP coder whose general operation is abstracted in Fig. 3.

The original speech was sampled at 8 KHz and pre-emphasized with the filter $1 - 0.4z^{-1}$. Ten LPC parameters $\{\alpha_k\}$ were calculated every 10 msec using the autocorrelation method in a 20 msec Hamming window centered over the current frame. The coefficients thus derived constitute $A(z)$.

The determination procedure for the pitch predictor is straightforward in an analysis-by-synthesis scheme. First, the ringing from the previous frames, calculated by running the inverse LPC predictor $\frac{1}{1-A(z)}$ on the previous frame of $U[n]$ with zero new input, is subtracted from the original signal $S[n]$. This new signal is called $R_p[n]$. $R_p[n]$ is also run through a (perceptually-important) weighting filter

$$W(z) = \frac{1 - A(z)}{1 - A(0.8^{-1}z)}$$

to hide noise under the formants. The pitch predictor, then, assumes a filter of the form $\beta z^{-\gamma}$. By testing all values $40 <= \gamma < 160$, an inverse pitch predictor with gain $\beta = 1$ is run on zero excitation sequence input to arrive at a candidate for the current frame of $U[n]$. This frame is run through $\frac{1}{1-A(z)}$ and $W(z)$ to arrive at a candidate $R_{p,\gamma}[n]$. This signal can be correlated with $R_p[n]$ to arrive at the optimum $\beta$ and the mean squared error. The optimum $\gamma$ with the lowest mean squared error is chosen, along with its corresponding optimum $\beta$.

Adapting the codewords in the excitation sequence is needed to capture the time-varying characteristics of the speech residuals. The codewords are selected in a manner similar to that of the pitch predictor parameters. First, all ringing effects and the contributions of the pitch predictor are subtracted from $S[n]$, leaving the yet-unmodeled portion of the speech. This signal is run through $W(z)$ and called $R_c[n]$.

The signal $R_c[n]$ is compared to the signals created from a fixed codebook (FGN, FBM, or Gaussian) of $K$ = 1024, 64, or

16 codewords. Each codeword represents 5 msec of the excitation sequence. For the FGN and FBM codewords, which were sorted by their $H$ parameter, three different schemes were tried in the implementation of the codeword search. 1) *Full search*: In the first approach, all codewords were test-matched with $R_c[n]$ by filtering each through $\frac{1}{1-A(z)}$ and $W(z)$, then computing the optimum scaling factor $G$ and mean squared error. The codeword $C[n]$ with the lowest mean squared error is chosen, along with its $G$.

2) *Limited search:* A second approach was to model $R_c[n]$ with an FGN or FBM by estimating its $H$ and then greatly limit the amount of time spent in the search. Specifically, the $H$ of the $R_c[n]$ signal was measured with the fractal dimension method, and only a window consisting of 10% of the codebook centered around the measured $H$ was searched. This scheme offers the possibility of greatly decreased computation time with many of the advantages of a large codebook in quality performance.

3) *No search:* The third approach did not implement a codebook search at all, using the single codeword corresponding to the estimated $H$ value (modulo the quantization error in representing the $H$ range with $K$ distinct values).

Figure 4 shows 20 msec of typical speech waveforms generated during the operation of the CELP coder.

## 4 Results and Discussion

In the experiments, several FGN and FBM codebooks were tried, each with a different size $K$ and range of $H$. (All the predictor parameters were unquantized.)

For $K = 1024$ codewords (0.25 bits per excitation sample), the Gaussian and the FGN codebooks with $H$ in [0.1,0.9] performed similarly under full search. The performance was judged both via informal subjective listening tests and segmental SNRs on coded speech from male and female speakers. The speech synthesized by FBMs sounded "smoother" (more low-pass) than that synthesized by FGNs, which is expected since FBMs are moving averages of FGNs.

Figure 5 shows the $H$ of the FGN signal that best matches (using the fractal dimension method) the closed-loop excitation signal for a typical speech signal. As shown there, most of the time $H$ does not use its full range $(0,1)$. Hence, for a fixed $K$, we can allocate the $K$ FGN codewords in a smaller subinterval of the $H$ range with an increase in the coder's performance. Thus, for the same $K$, the FGNs with $H$ limited in [0.3,0.7] performed better than the Gaussian codewords. This difference in performance was very small for $K = 1024$, increased for $K = 64$ (0.15 bits per excitation sample), and became more pronounced for $K = 16$ (0.1 bits per excitation sample).

For FGN and FBM codebooks, the limited search and no-search approaches performed worse than the full search. However, the no-search approach shows some promise for future research, because it is fast, the chosen single $H$ is suboptimum (as opposed to a random choice) since it matches the fractal characteristics of the speech residual, and the synthesized speech is more than intelligible.

Overall, the performance of FGNs degrades more gracefully than that of the Gaussian codebook. Further, the FGNs and FBMs are more structured classes of excitation signals for various reasons: 1) Their total dependence on a single parameter $H$ with bounded range allows their efficient coding by just quantizing $H$. 2) While keeping the codebook bit rate fixed, we can limit $H$ to a subinterval (chosen to match certain statistics of speech classes) of $(0,1)$. Thus, the $H$ parameter can be sampled more densely and increase the performance of the codebook. 3) At a certain loss of speech quality, we can estimate a single (suboptimum) $H$ that best matches the speech residual and thus avoid searching the codebook. 4) We can view both FBMs and FGNs as a combined class of parametric excitation signals with power spectrum $1/\omega^{\Delta}$, with $\Delta \in (-1,3)$. The standard Gaussian codewords correspond only to $\Delta = 0$ since their spectrum is flat. Hence, by varying $\Delta$ we obtain an additional degree of freedom in coding the speech excitation. 5) By limiting $\Delta$ to subintervals of its full range, we can simulate various degress of smoothness (large $\Delta$) or roughness (small $\Delta$) in the synthesized speech.

**Acknowledgement:** We wish to thank R. Rose at Lincoln Labs for many useful discussions on CELP coders.

## References

[1] B. Atal and M. Schroeder, "Stochastic coding of speech signals at very low bit rates", *Proc. ICC '84*, 1984.

[2] M. Schroeder and B. Atal, "Code-Excited Linear Prediction (CELP)", *Proc. ICASSP '85*, Tampa, 1985.

[3] R. Rose, "The Design and Performance of an Analysis-by-Synthesis Class of Predictive Speech Coders", *Ph.D thesis*, Georgia Inst. Tech., 1988.

[4] B. Mandelbrot, *The Fractal Geometry of Nature*, NY: Freeman, 1982.

[5] R. Voss, "Fractals in nature: From characterization to simulation", in *The Science of Fractal Images*, H.-O. Peitgen and D. Saupe, Eds, Springer-Verlag, 1988.

[6] P. Maragos, "Analysis of Image-like Information in Speech Signals using Mathematical Morphology", *Proc. Conf. Electronic Imaging*, Boston, Oct. 1989.

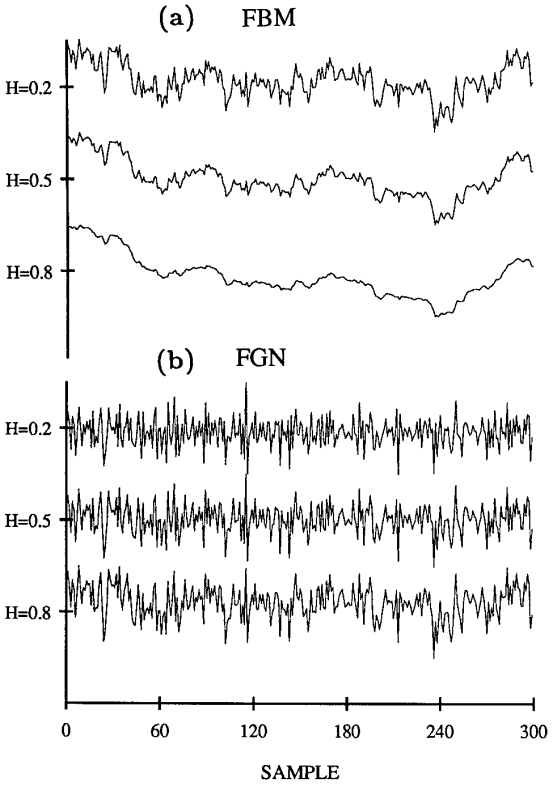


**Figure 1.** (a) FBM, and (b) FGN signals.

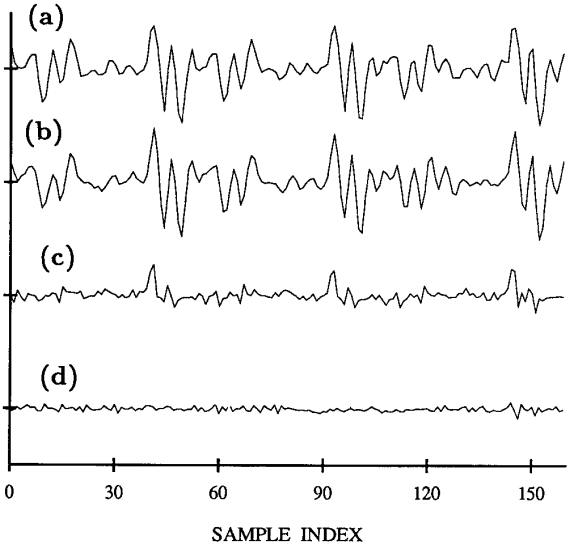


**Figure 4.** (a) Original speech. (b) Synthesized speech ($K$ = 1024 FGNs, $H \in [0.1, 0.9]$). (c) Prediction error $U[n]$. (d) Excitation $E[n]$.

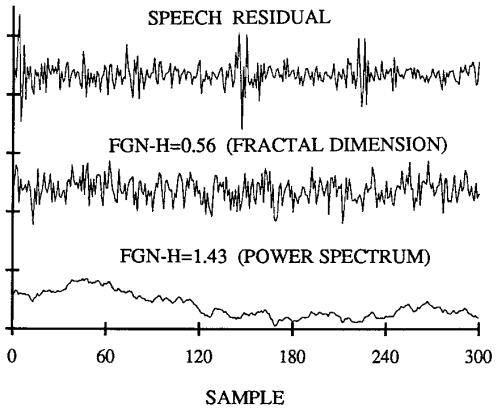


**Figure 2.** Modeling a speech residual signal by an FGN.

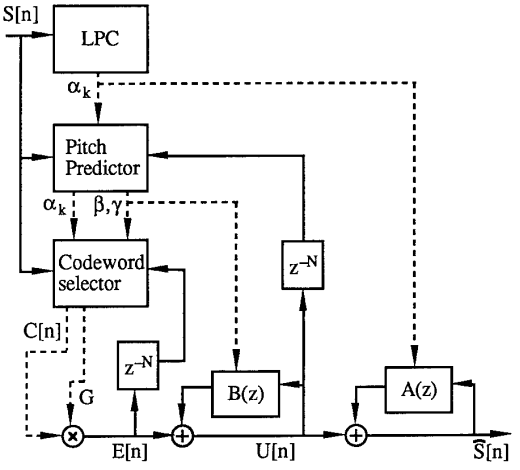


**Figure 3.** Analysis-by-synthesis procedure during encoding. (Solid line shows signal flow, whereas dotted line shows flow of parameters.)

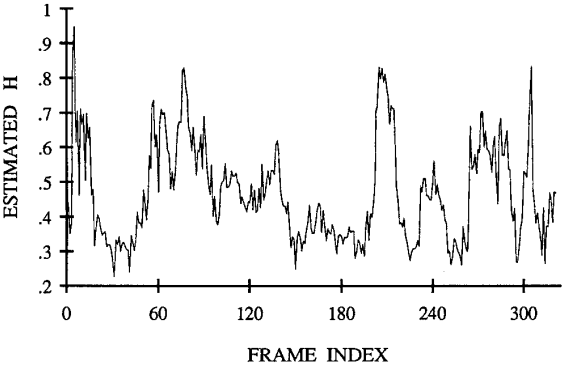


**Figure 5.** Variation of estimated $H$.