# SPEECH MODELING AND VOICED/UNVOICED/MIXED/SILENCE SPEECH SEGMENTATION WITH FRACTIONALLY GAUSSIAN NOISE BASED MODELS

*Sh. Oveisgharan, M. B. Shamsollahi*

Department of Electrical Engineering, Sharif University of Technology, Iran

## ABSTRACT

The ARMA filtered fractionally differenced Gaussian Noise (FdGn) model and a new AR Filtered FdGn Added up model are applied to speech signal and performance of their parameters on speech Unvoiced/Voiced/Mixed/Silence classification is evaluated against Zero Crossing Rate (ZCR) feature. For parameter estimation of AR filtered FdGn model two methods were applied: iterative Maximum Likelihood (ML) method of Tewfik [2] and a new computationally efficient Linear Minimum Square Error (LMSE) algorithm. Also for parameters estimation of new Added up model two approaches were implemented: an Expectation-Maximization (*EM*) based approach and an iterative MSE approach. The described models and methods were applied to speech signal and also its real Cepstrum. The performance of described models on V/U/M/S speech classification was obtained based on *J1* parameter in this order: Added up model on real Cepstrum of speech, Filtered FdGn model on real Cepstrum of speech (LMSE method), Filtered FdGn model on speech (LMSE method), ZCR, and Filtered FdGn model on speech (Tewfik method).

## I. INTRODUCTION

In the most commonly used model of speech production, speech signal is decomposed into a time varying filter component and an excitation component [13]. The excitation is represented by the superposition of two sources: periodic pulse train produced by vibration of the vocal cords and white Gaussian noise produced by forcing air past some constriction in the vocal tract. But this model lacks of the ability to observe the long term dependencies observed in speech signal because of its *ARMA* statistical model [2]. In other words it has been observed that by assumption of white noise excitation we can not interpret long term dependencies in speech signal.

To solve this problem, *FBM* (Fractionally Brownian Motion) model is introduced by Mandelbrot and Ness[1]. In contrast with ARMA models which are characterized by correlation function that decay exponentially with the lag, FBM signals with $1/f$ -type spectra have a correlation function that decreases hyperbolically fast with the lag $k$ as $k^{\alpha}$ [2].

Because FBM is a non-stationary model, its derivative called Fractionally differenced Gaussian noise (FdGn) is applied to speech. But the FdGn model seems not to discover the short term dependencies of speech signal which could be found out by poles and zeros of ARMA model. Hosking ([6]) solved this problem by presenting the FdGn AR filtered model. In this model the Gaussian excitation source is not assumed to be white at all, but it can also have weak dependencies between far samples.

In this paper we applied the FdGn AR filtered model of Hosking to speech signal. In order to compute model's parameters, we used the iterative method of Tewfik ([2]) which searches for a local optimal in log likelihood surface of model parameters. Because of high computational complexity of obtaining Log likelihood value, we present a new fast method based on LMSE of $O(n \log n)$ complexity. As we observed, this method not only runs so faster, but also obtains better results in speech V/U/M/S segmentation.

Finally, a new AR filtered FdGn added up model is presented which is an AR filtered of summation of a number of independent FdGn signals. In the case of speech signal which is the superposition of the environmental white noise and signal of consecutive sounds, this model seems to be able to discriminate between noise and two or more sounds contributed in composition of speech signal. We present two methods for parameters estimation of added up model. The first method uses *EM* approach to search for the optimal local point in the joint surface of log-likelihood model parameters. We also applied another iterative method based on *MSE* fitting. We applied the added up model on Real Cepstrum of speech signal and obtained best results in V/U/M/S speech segmentation.

## II. FDGN ARMA FILTERED MODEL

As we observe in figure 2.1, the output of an ARMA filtered FdGn model is built of a composition of ARMA filtering and an FdGn filtering on white noise. As figure 2.1 suggested, the FdGn process can be defined as $(-d)^{th}$ fractional difference (or summation) of discrete time white Gaussian noise ($-0.5 \le d \le 0.5$):

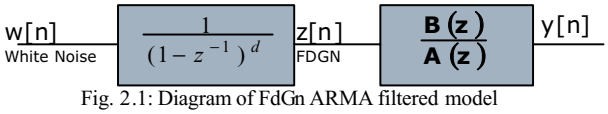

Fig. 2.1: Diagram of FdGn ARMA filtered model

$$w_d[n] = \sum_{k=0}^{\infty} \binom{-d}{k} (-1)^k w[n-k] \quad (2.1)$$
$$= \sum_{k=0}^{\infty} c[k]\, w[n-k]$$

where $w[n]$ is white Gaussian noise. The continuous form of FdGn also can be defined as derivative of FBM:

$$x_{FdGn}(t) = \frac{d}{dt} x_{FBM}(t) \quad (2.2)$$

where FBM is defined as a zero mean Gaussian process with the following correlation property :

$$B_H(0) = 0, E\left\{(B_H(t) - B_H(s))^2\right\} = \alpha(t-s)^{2H} \quad (2.3)$$

$H$ is the Hurst parameter and is related to $d$ with equation $d = H - 0.5$. In the case of $H = 0.5$ general Brownian motion would be obtained. Equation 2.3 implies the value of variance and autocorrelation matrix of FBM process:

$$\sigma^2_{B_H(t)} = E\left\{(B_H(t) - B_H(0))^2\right\} = \alpha t^{2H}$$
$$R_H(t,s) = \frac{\alpha}{2}\left((t-s)^{2H} - t^{2H} - s^{2H}\right) \quad (2.4)$$

In [5] it was proved that variance of FdGn with maximum Log-Likelihood value is equal to:

$$\hat{\sigma}^2 = \frac{1}{N} x^T R_1^{-1}(d)\, x \quad (2.5)$$

In [5] also an algorithm was presented to evaluate the Log-Likelihood values of FdGn by complexity of $O(N^2)$. So for ML estimation of $\sigma, d$ of FdGn process, it is enough to search in $[-0.5, 0.5]$ for a value of $d$ by a heuristic method like Maximum descent algorithm and then apply equation 2.5 to estimate variance parameter with ML. Now we review Tewfik's iterative approach and a new LMSE approach for FdGn AR Filtered parameters estimation.

### II.A. Tewfik's Iterative Approach

In this algorithm an initial value for parameters of the ARMA filter and Hurst parameter and variance is selected. Then in *k'th* step, the observed vector, $y[n]$, is filtered by inverse of ARMA filter whose parameters are $a_k(i), b_k(j);\ 1 \le i \le P, 1 \le j \le Q$ to obtain $z_k(n)$ and then according to Maximum Likelihood algorithm described in [5] we find the new values $H_{k+1}$ and $\sigma^2_{k+1}$ for FdGn model. Then we apply the inverse of FdGn filter with respect to $H_{k+1}$ and $\sigma^2_{k+1}$ to observed input vector, $y[n]$, to obtain $x_k[n]$ and then according to Levinson's algorithm, we find new values $a_{k+1}(i), b_{k+1}(j);\ 1 \le i \le P, 1 \le j \le Q$ of parameters of ARMA model. The algorithm will stop at *k'th* step, if the difference norm of vector parameters between two consecutive steps becomes less than a predefined value. In mathematical word:

$$\left\|(a_{k+1}(i), b_{k+1}(j), H_{k+1}, \sigma_{k+1}) - (a_k(i), b_k(j), H_k, \sigma_k)\right\| < \varepsilon \quad (2.1.1)$$

No theoretical proof on computational complexity of this method has been found yet ([2]).

### II.B. New LMSE Algorithm

In this algorithm we apply fundamental property 2.3 of FBM process in order to estimate the parameters of ARMA filtered FdGn process.

First of all, given the observed vector we implement the Levinson's algorithm on it in order to estimate the parameters of ARMA filter. Then according to the diagram of figure 2.1, we implement the inverse of ARMA filter in order to estimate the FdGn signal $z[n]$.

Now for parameter estimation of FdGn signal, $z[n]$, we estimate the original FBM process, $b[n]$, from which the FdGn signal $z[n]$ is obtained by a summation on $z[n]$:

$$b[n] \overset{\Delta}{=} \sum_{m=1}^{n} z[m] \quad (2.2.1)$$

By applying equation 2.3 to equation 2.2.1 we have:

$$\log\left(E\left\{(b[m]-b[n])^2\right\}\right) = \log(\alpha) + 2H \times \log(m-n) \quad (2.2.2)$$

So for estimation of Hurst parameter, we estimate the expectation in 2.2.2 and then we apply a MSE line on its log-log curve. For best estimation of Hurst parameter, we can apply MSE line just in an interval $[n_{\min}, n_{\max}]$ in domain of $f[n]$ defined in equation 2.2.3.

$$f[n] \overset{\Delta}{=} \log(E\{(b[m]-b[m-n])^2\}) \cong$$
$$\log\left(\frac{1}{N-n}\sum_{m=n+1}^{N}(b[m]-b[m-n])^2\right)$$
$$= \log\left(\frac{1}{N-n}(g[n]-2p[n])\right);\ \textit{where}: \quad (2.2.3)$$
$$p[n] = \sum_{m=n+1}^{N} b[m]b[m-n];$$
$$\begin{cases} g[n] = g[n-1] - b^2[n] - b^2[N+1-n] & : n > 0 \\ g[0] = 2\sum_{m=1}^{N} b^2[m] \end{cases}$$

As we observe in equation 2.2.3, the expressions of $g[n]$ is computable by an *O(N)* recursive manner where $N$ is the number of samples in observed vector. For computation of $p[n]$ we write:

$$p[n] = \sum_{m=n+1}^{N} b[m]b[m-n] \Rightarrow \quad (2.2.4)$$
$$p[n] = \Pi_N[n]\, IFFT_{2N}\left\{FFT_{2N}\{b[n]\}^2\right\}$$

where $\Pi_N[n]$ is the rectangular window of size *N*. According to 2.2.4, we observe $p[n]$ is computable in $O(N \log N)$. So computational complexity of the algorithm is $O(N \log N)$ which is so faster than Tewfik's method.

## III.NEW AR FILTERED FDGN ADDED UP MODEL

In many physical applications and situations the observed signal is a summation of so many independent signals. Especially it is the case when we are working with natural signals like speech. In many situations the sound which is pronounced is caused by the combination of some consecutive letters ([8]) and hence if each letter alone has a special Hurst parameter, we expect the coming speech signal to have a combination of these Hurst parameters. This fact motivated us to define stochastic model below.

Let us $\{x_1[n], x_2[n],..., x_K[n]\}$ be a set of independent FdGn signals with Hurst parameters $\{H_i\}_{i=1}^K$ and variance of $\{\alpha_i^2\}$. Now we define FdGn added up random process $y[n]$ as:

$$y[n] = \sum_{i=1}^{K} x_i[n] \tag{3.1}$$

and AR Filtered FdGn added up model as AR filter of $y[n]$. Equation 3.1 implies that $y[n]$ is also a zero mean Gaussian signal with covariance matrix as:

$$R_y[n] = \sum_{i=1}^{K} \alpha_i^2 R_{H_i}[n] \tag{3.2}$$

where $R_{H_i}[n]$ is covariance matrix of a normal FdGn with corresponding Hurst parameter of $H_i$. Here we present two iterative methods in order to estimate parameters $\{\alpha_i^2\}$ and $\{H_i\}$ of model according to an observed vector. In both of methods we assume the coefficients of AR filter were estimated by applying Levinson's algorithm and so we have estimated FdGn added up signal $y[n]$.

### III.A. Iterative Expectation Maximization Estimator

A robust ML estimator could be applied in $\{H_i\}$ and $\{\alpha_i^2\}$ estimation by a search in the space of $[0,1]^K R_+^K$ and using equation 3.2. Here we apply an Expectation Maximization (*EM*) method in order to decrease dimension of ML search from *2R* to 2. For details and proof of EM you can refer to [9] and [10].

Defining a complete vector $\{x_i[n]\}_{i=1}^K$ from observed incomplete vector $y[n]$ with relation (3.1), expected value of log likelihood of complete vector with respect to parameters $\{\alpha_i, d_i\}$ is computable as:

$$\begin{aligned} &Q(x_1[n],..., x_K[n]) = \\ &= E\{\log(\ p(x_1,..., x_K | \{\alpha_i, d_i\}))\,|\,\{\alpha_i^k, d_i^k\}, y[n]\} \\ &= \sum_{j=1}^{K} Q(x_j[n]) \end{aligned} \tag{3.1.1}$$

where $\{\alpha_i^k, d_i^k\}$ is $k'th$ estimation of parameters. As we observe in equation 3.1.1, for maximization of $Q(x_1[n],...,x_K[n])$ we can maximize each of $Q(x_j[n])$ independently. Hence complexity for each iteration of EM would be linear with $K$. For estimation of $Q(x_j[n])$ we have:

$$\begin{aligned} &Q(x_j[n]) = E\{\log(\ p(x_j | \{\alpha_i, d_i\}))\,|\,\{\alpha_i^k, d_i^k\}, y[n]\} \\ &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\log|R(\alpha_j, d_j)| \\ &\quad -\frac{1}{2} Trace\left\{R^{-1}(\alpha_j, d_j) E\{x_j x_j^T \,|\,\{\alpha_i^k, d_i^k\}, y[n]\}\right\} \end{aligned} \tag{3.1.2}$$

Now for computation of correlation matrix of $x_j[n]$ by knowledge of $(\{\alpha_i^k, d_i^k\}, y[n])$, we first obtain its mean value:

$$\begin{aligned} &\mu_{x_j|y}(\{d_i^k, \alpha_i^k\}) = (\alpha_j^k)^2 R(d_j^k). \\ &\left[\sum_{i=1}^{K} (\alpha_i^k)^2 R(d_i^k)\right]^{-1}.y \end{aligned} \tag{3.1.3}$$

Now for computation of correlation matrix we have:

$$\begin{aligned} &E\left\{x_j x_j^T \,|\,\{\alpha_i^k, d_i^k\}, y\right\} = \\ &\mu_{x_j|y}^2(\{\alpha_i^k, d_i^k\}) + (\alpha_j^k)^2 R(d_j^k) - (\alpha_j^k)^4 R(d_j^k). \\ &.\left[\sum_{i=1}^{K} (\alpha_i^k)^2 R(d_i^k)\right]^{-1}.R(d_j^k) \end{aligned} \tag{3.1.4}$$

Note that expression 3.1.4 just needs to be computed once for each iteration of EM algorithm. For maximization of $Q(x_j[n])$, we can apply an iterative manner like maximum descent gradient method.

### III.B. Iterative Minimum Square Error Estimator

By definition of:

$$\begin{aligned} b[n] &\stackrel{\Delta}{=} \sum_{m=1}^{n} y[m] \\ &= \sum_{i=1}^{K}\sum_{m=1}^{n} x_i[m] = \sum_{i=1}^{K} b_i[n] \end{aligned} \tag{3.2.1}$$

we observe summation of observed vector is equal to summation of some independent FBMs with corresponding $\{H_i\}$ and $\{\alpha_i^2\}$ parameters. So According to equation 2.3, we can write:

$$\begin{aligned} r[m] &\stackrel{\Delta}{=} E\left[(b[n] - b[n-m])^2\right] \\ &= \sum_{i=1}^{K} \alpha_i^{\,2} m^{2H_i} = \sum_{i=1}^{K} \alpha_i^{\,2} m^{\beta_i} \end{aligned} \tag{3.2.2}$$

We apply equation 2.3.3 again in order to estimate function $r[m]$ by the complexity of $O(\mathrm{N}log\mathrm{N})$. Now we present an iterative algorithm based on 3.2.2 and 2.2.2 for model's parameters estimation. The algorithm starts with an initial estimation of parameters: $\{\alpha_i^{(0)}, \beta_i^{(0)}\}$ which satisfies equation 3.2.2 for just some supposed small scales. Then in *m'th* level we estimate expression 2.3 for each scale and all values of *i=1,..,K*:

$$r_i[n] \stackrel{\Delta}{=} E\left[(b_i[j] - b_i[j-n])^2\right] \cong \hat{r}_i^{(m)}[n]$$

Table 1. Numerical Result of each method on best initial parameters: degree of AR filter, interval,..

| Method | Tewfik method with AR(10) | ZCR with AR(1) | LMSE AR(11) on interval [10,300] | Real Cepstrum LMSE on interval [10,470] | Real Cepstrum MSE Added up(3) on [1,600] |
|---|---|---|---|---|---|
| *J1 value* | 0.35 | 0.53 | 1.74 | 2.18 | 3.47 |

$$\hat{r}_i^{(m)}[n] \stackrel{\Delta}{=} r[n] - \sum_{j=1}^{i-1} \left(\alpha_j^{(m+1)}\right)^2 |n|^{\beta_j^{(m+1)}} - \sum_{j=i+1}^{K} \left(\alpha_j^{(m)}\right)^2 |n|^{\beta_j^{(m)}} = \quad (3.2.3)$$

$$\begin{cases} \hat{r}_{i-1}^{(m)}[n] + \left(\alpha_{i-1}^{(m+1)}\right)^2 |n|^{\beta_{i-1}^{(m+1)}} - \left(\alpha_i^{(m)}\right)^2 |n|^{\beta_i^{(m)}} : i > 1 \\ \hat{r}_K^{(m-1)}[n] + \left(\alpha_K^{(m)}\right)^2 |n|^{\beta_{i-1}^{(m)}} - \left(\alpha_i^{(m)}\right)^2 |n|^{\beta_i^{(m)}} : i = 1 \end{cases}$$

Then an MSE line is fit to log-log curve of $\hat{r}_i^{(m)}[n]$ in order to obtain new estimations $\beta_i^{(m+1)}$ and $\alpha_i^{(m+1)}$. The condition of stopping the algorithm at *M'th* level is the same as algorithm described in II.A. As we observe in equation 3.2.3, the algorithm takes a recursive $O(N)$ manner in order to obtain $\hat{r}_i^{(m)}[n]$ according to its previous values. So each iteration of algorithm has the complexity of $O(KN)$. We couldn't proof convergence of proposed algorithm but our simulations proved so higher speed of convergence of this method than that of EM method.

For estimation of the number of independent components, *K*, we apply our algorithm for small values of *K* initially and increase it until a value where two estimated Hurst parameter $H_i$ and $H_j$ are found for which difference is less than a predefined value.

## IV. EXPERIMENTAL RESULTS

For performance evaluation of models and methods explained in this paper, a database of *Persian* speech files recorded from a male speaker, sampled at *8KSample/Sec* and quantized by *16 bits* were used in speech *V/U/M/S* segmentation. The evaluator of each model's performance in speech segmentation was $J_1$ parameter defined as follow:

$$J_1 = trace(W^{-1}B) \quad (4.1)$$

where *W* is within class scatter matrix and *B* is between class scatter matrix. For implementing models, speech signal was divided into *20msec* frames in order to take the AR filter on it. Numerical results proved the best window size for taking FdGn model is *60msec* (*480* samples) except for Added up model which is *75msec*(*600* samples). We also apply all the models on Real Cepstrum of speech signal. The results are presented in Table 1.

As we observe, the best result was obtained by an Added up model of *3* independent FdGn signals on Real Cepstrum of *600* samples speech signal(*3.47*). FdGn AR filtered model obtained worse result than ZCR feature by Tewfik's method(*0.35*) but obtained better result by LMSE approach(*1.74*). LMSE method even obtained better results (*2.18*) on Real Cepstrum. Intervals [*10,300*], [*10,470*], [*1,600*] in Table 1 represent best region found for MSE line fitting on the curve in MSE based algorithms.

## V. CONCLUSION

In this paper, FdGn based models and their applications in speech segmentation were discussed. A Fast $O(N \log N)$ LMSE approach on parameters estimation of AR filtered FdGn model was presented and was observed to obtain so better results than iterative ML based approach in speech V/U/M/S segmentation. A new AR filtered FdGn Added up model was presented and two iterative algorithms were discussed for model's parameters estimation. MSE iterative method on parameters estimation of this model obtained best results on speech V/U/M/S segmentation and so suggests this model to be a acceptable model in speech processing problems.

## REFERENCES


[1] B. B. Mandelbrot and W. Vanness, "*Fractional Brownian Motion , Fractional Noises and Applications*" *SIAM* Rev., vol. 10,no. 4, pp.422-437, Oct. 1968.

[2] A. H. Tewfik, et. Al., "*Signal Modeling with Filtered Discrete Fractional Noise Processes*", IEEE Trans. On Signal Processing,pp. 2839-2849, Vol. 41,Sept. 1993.

[3] A. Langi and W. Kinsner, "*Consonant Characterization using Correlation Fractal Dimension for Speech Recognition*", IEEE WESCANEX, pp. 208-213, 1995.

[4] P. Maragos, K. L. Young, "*Fractal Excitation Signals For CELP Speech Coders*", IEEE Trans., pp. 669-672, Feb. 1990.

[5] A. H. Tewfik, et. Al., "*Maximum Likelihood Estimation of the parameters of Discrete Fractionally Differenced Gaussian Noise Process*", IEEE Trans. On Signal Processing,pp. 2977-2989, Vol. 41,Oct. 1993.

[6] J. R. M. Hosking, "*Fractional Differencing*", *Biometrica*, vol. 68, no. 1, pp. 165-176, 1981.

[7] S. Fekkai, et. Al., "*Fractal Dimension Segmentation: Isolated Speech Recognition*", Inst. Of Electrical Engineers, pp. 1-4, 2000.

[8] D. J. Tritton, "Physical Fluid Dynamics", Oxf. U. P., 1988.

[9] A.P.Dempster,N.M. Laird, D.B. Rubin, "*Maximum -likelihood from incomplete data via the EM algorithm*". J. Royal Statist. Soc. Ser. B., 39, 1977.

[10] R. Redner and H. Walker, "*Mixture densities, maximum likelihood and the EM algorithm*", SIAM Review, pp. 26-29, 1984.

[11] T. J. Thomas, "*A Finite element model of Fluid Flow in the vocal tract*", Computer Speech and Language, pp. 131:151, 1986.

[12] E. L. J. Bohez, et. Al., "*Fractal Dimension and Iterated Function System(IFS) For Speech Recognition*", Electronics Letters, pp. 1382-1384, Vol. 28, July 1992.

[13] S. E. Levinson and D. B. Roe, "*A Perspective on Speech Recognition*", IEEE Communications Magazine, pp. 28-34, Jan. 1990.