FRACTAL MODELING OF SPEECH SIGNALS

Jacques LEVY VEHEL, Khalid DAOUDI & Evelyne LUTTON
INRIA, Rocquencourt
B.P. 105, 78153 Le Chesnay CEDEX, France.
e-mail: jlv@bora.inria.fr, daoudi@bora.inria.fr, lutton@bora.inria.fr

In this paper, we present a method for speech signal analysis and synthesis based on IFS theory. We consider a speech signal as the graph of a continuous function whose irregularity, measured in terms of its local Hölder exponents, is arbitrary. We extract a few remarkable points in the signal and perform a fractal interpolation between them using a classical technique based on IFS theory. We thus obtain a functional representation of the speech signal, which is well adapted to various applications, as for instance voice interpolation.

1. INTRODUCTION

We are interested here in the general problem of speech signal synthesis. The precise setting of our work is the following: the French CNET(Centre National d'Etudes en Télécommunication) has developed a synthesis algorithm, PSOLA[1], based on the concatenation of diphones. The latter are obtained by segmentation of longer acoustic units, the logatomes, previously recorded by a human operator. Although this method gives very good results, there subsists the problem of the dictionary construction: three months are necessary for the recording and the segmentation of the 1200 logatomes needed in French. Each time a new voice is being created, one has to go through a complex and time consuming process. A solution to this problem is to use existing dictionaries to create new "voices". A simple idea is be to perform interpolations between corresponding logatomes of two dictionaries coming from two different voices in such way that at each step of the interpolation, the signal remains a logatome. A first step towards this goal is to obtain a robust functional representation of a logatome. This is the particular problem we address hereafter.

2. FRACTAL INTERPOLATION

2.1 Introduction

Speech sounds are in some cases produced by turbulence phenomena[2]. It is well known that several aspects of turbulence are multifractal. This and other theoretical considerations[3], as well as the general aspect of the signals (see fig. 1), motivate the use of a multifractal approach for speech signal modeling.

There are two main ideas in our approach : the first one is that there exists in a speech signal a set of remarkable points, where the acoustic information is specially relevant (for instance the points that define the waveform). These points we call tag points. The second idea is that the singularities (Hölder exponents) at each point of the signal play a fundamental role in the determination of the voice texture. A simple and efficient tool for controlling both tag points and singularities is fractal interpolation using IFS.

2.2 Spectrum of singularity of self affine functions

The general setting is the following:

Given a set of data points $\{(x_i, y_i) \in [0\,;1] \times [a\,;b], i = 0, 1, ..., N\}$, consider the IFS given by the N contractions $w_n (n = 1, ..., N)$ defined on $[0\,;1] \times [a\,;b]\,(-\infty < a < b < +\infty)$, by:

$$w_n(x, y) = (L_n(x)\,; F_n(x, y))$$

where L_n is the contraction that maps $[0\,;1]$ to $[x_{n-1}\,; x_n]$ and $F_n : [0\,;1] \times [a\,;b] \to [a\,;b]$ is a contraction function with respect to the second variable such that :

$$F_n(x_0, y_0) = y_{n-1}\,; F_n(x_N, y_N) = y_n$$

It is well known that the attractor of this IFS is the graph of some continuous fractal function f which interpolates the data points[4].

We derive here an expression for the spectrum of singularity for a particular case of such IFS. For general definitions and properties of the multifractal analysis of functions, see[5,6].

We focus on the particular case of self-affine functions with equidistant interpolation points.

We recall defintion of the local Hölder exponent of a function:

Definition 1 *A function f is said to be of Hölder exponent $\alpha(t_0) > 0$ at point t_0 iff:*
i) for every real γ such that $0 < \gamma < \alpha(t_0)$:

$$\lim_{h \to 0} \frac{|f(t_0 + h) - P(t - t_0)|}{|h|^\gamma} = 0$$

ii) for every real $\gamma > \alpha(t_0)$

$$\limsup_{h \to 0} \frac{|f(t_0 + h) - P(t - t_0)|}{|h|^\gamma} = +\infty$$

where P is a polynomial whose degree is less than $\alpha(t_0)$.

Let $S_i \, (1 \leq i \leq m)$ be affine transformations represented in matrix notation with respect to (t, x) by:

$$S_i \begin{pmatrix} t \\ x \end{pmatrix} = \begin{pmatrix} 1/m & 0 \\ a_i & c_i \end{pmatrix} \begin{pmatrix} t \\ x \end{pmatrix} + \begin{pmatrix} (i-1)/m \\ b_i \end{pmatrix}$$

We suppose $0 \leq t \leq 1$ and $1/m < c_i \leq 1$. Let f be the function whose graph is the attractor of the IFS defined by the S_i's (with usual conditions an a_i and b_i to ensure the continuity of f).

Proposition 1 *Let $0.i'_1...i'_k...$ be the base-m expansion of a real $t \in [0\,;1[$,and $i_j = i'_j + 1$, then:*

$$\alpha(t) = \liminf_{k \to +\infty} \frac{\log(c_{i_1}...c_{i_k})}{\log(m^{-k})}$$

In the case of multiple expansions, the one yielding the lower $\alpha(t)$ has to be taken.

proof:

This proof is an adaptation of the classical computation of the box dimension of self affine curves (see for instance [7])

Let t_0 be a real in $[0\,;1[$ whose base-m expansion is $0.i'_1...i'_k...$ and $I_{i_1...i_k}$ be the interval of reals whose base-m expansion begins with $0.i'_1...i'_k$ where $i_j = i'_j + 1$. Since the abscissa of $S_{i_1} \circ ... \circ S_{i_k}(t, x)$ is $tm^{-k} + (i_k - 1)m^{-k} + ... + (i_1 - 1)m$ for every (t, x), then $F_{|_{I_{i_1...i_k}}} = S_{i_1} \circ ... \circ S_{i_k}(A)$, wich is a translation of $T_{i_1} \circ ... \circ T_{i_k}(A)$ where T_i is the linear part of S_i. We can easily see that the matrix representing $T_{i_1} \circ ... \circ T_{i_k}$ is

$$\begin{pmatrix} m^{-k} & 0 \\ m^{1-k}a_{i_1} + m^{2-k}c_{i_1}a_{i_2} + ... + c_{i_1}c_{i_2}...c_{i_{k-1}}a_{i_k} & c_{i_1}c_{i_2}...c_{i_k} \end{pmatrix}$$

If we note $a = \max |a_i|, c = min(c_i), r = \frac{a}{1-(mc)^{-1}}$ we have $|m^{1-k}a_{i_1} + m^{2-k}c_{i_1}a_{i_2} + ... + c_{i_1}c_{i_2}...c_{i_{k-1}}a_{i_k}| \leq rc_{i_1}...c_{i_k}$, so that if s is the height of the rectangle containing A, then $F_{|_{I_{i_1...i_k}}}$ is contained in the rectangle whose height is $(r + s)c_{i_1}...c_{i_k}$. Thus for every h such that $0 < |h| < m^{-k}$, we have $|f(t_0 + h) - f(t_0)| \leq r_1 c_{i_1}...c_{i_k}$ where $r_1 = r + s$ this yields:

$$|f(t_0 + h) - f(t_0)| \leq |h|^{r_1/m^{-k}} |h|^{\log(c_{i_1}...c_{i_k})/\log(m^{-k})}$$

Let $\beta(k) = \frac{\log(c_{i_1}...c_{i_k})}{\log(m^{-k})}$,$\beta = \liminf_{k \to +\infty} \beta(k)$ and consider a real γ such that $0 < \gamma < \beta$. Then

$$\frac{|f(t_0 + h) - f(t_0)|}{|h|^\gamma} \leq C(h, k)|h|^{\beta(k) - \gamma}$$

where $C(h, k) = |h|^{r_1/m^{-k}} \to 1$ when $h \to 0$.

There exists a real ϵ_0 and an integer K such that for every $k > K$ we have $\beta(k) - \gamma > \epsilon_0 > 0$. Since $h \to 0$ is equivalent to $k \to +\infty$, we have:

$$\lim_{h \to 0} \frac{|f(t_0 + h) - f(t_0)|}{|h|^\gamma} = 0$$

On the other hand, if q_1, q_2 and q_3 are three non-colinear points choosen from $S_1(p_1), ..., S_1(p_m), p_m$, where p_1 and p_m are the fixed points of S_1 and S_m then $S_{i_1} \circ ... \circ S_{i_k}(A)$ contains the points

$(t_j, f(t_j) = S_x(q_j) = S_{i_1} \circ ... \circ S_{i_k}(q_j)(j = 1, 2, 3)$. The height d_k of the triangle with these vertices is at least $dc_{i_1}...c_{i_k}$ where d is the vertical distance from q_2 to $[q_2\,; q_3]$. Suppose that $f(t_1) \leq f(t_3) < f(t_2)$ (the other cases are handeled similary). In this case we have two possibilties for t_0:
i) $f(t_0) \leq f(t_3)$ then: $|f(t_0) - f(t_2)| \geq f(t_2) - f(t_3) \geq d_k/2$
ii) $f(t_0) \geq f(t_3)$ then: $|f(t) - f(t_1)| \geq f(t_3) - f(t_1) \geq d_k/2$

Thus, for every $t \in I_{i_1...i_k}$ and for every integer k, there exists a real $h_k, |h_k| < m^{-k}$ such that: $|f(t + h_k) - f(t)| \geq c_{i_1}...c_{i_k} d_k/2$
Using the same arguments as when the inequality is reversed we prove, if $\gamma > \beta$ and $r_1 = d/2$ that:

$$\frac{|f(t_0 + h_k) - f(t_0)|}{|h_k|^\gamma} \geq C(h_k, k)|h_k|^{\beta(k)-\gamma}$$

$C(h_k, k) = |h_k|^{r_1/m^{-k}} \to 1$ when $k \to +\infty$ and ther exists a subsequence $\sigma(k)$ such that $|h_{\sigma(k)}|^{\beta(\sigma(k))-\gamma} \to 0$ when $k \to +\infty$ because there exists a real ϵ_1 and an integer K such that for every $k > K$ we have $\beta(\sigma(k)) - \gamma < \epsilon_1 < 0$. We deduce that

$$\limsup_{h \to 0} \frac{|f(t_0 + h) - f(t_0)|}{|h|^\gamma} = +\infty$$

for every $\gamma > \beta$, and the proof is complete. $\triangle$

Using this proposition, it is easy to deduce the spectrum $(\alpha, F(\alpha))$ of singularity of f. The proof is analogous to the one for multinomial measures.
Corollary 1 *With the same notations as above, and assuming that the proportion $\phi_i(t)$ of $(i-1)$'s in the base-m expansion of t exists for each i we have:*

$$\alpha(t) = -\sum_{i=1}^{m} \phi_i(t) \log_m c_i \;\; ; \;\; F(\alpha) = -\sum_{i=1}^{m} \phi_i \log_m \phi_i \;\; ; \;\; \tau(q) = -\log_m \sum_{i=1}^{m} {c_i}^q$$

Remark :
Using the relation $\dim_B \, graph\, f = 1 - \tau(1)$ we recover the classical result[7] :

$$\dim_B \, graph\, f = 1 + \log_m \sum_{i=1}^{m} c_i$$

3. APPLICATION :

We explain briefly the outline of the method. It involves two steps: we must first determine the interpolation points, which will control the general shape of the attractor, and then compute the interpolation functions themselves, which will take care of the local singularity.

3.1 Determination of the interpolation (tag) points

For vowels, the tag points will be given by the pitch marking of the signal. The principle is based on the observation of the autocorrelation function (ACF). In the first step, pitch values are founded by observing maxima of the ACF. In the second step, temporal marks are placed on the waveform according to the estimation of the pitch lags and the maxima of the waveform.

In the case of consonants, we compute the Hölder exponents at sharp variation points of the original signal using the wavelet transform maxima method[8]. We then consider the sets consisting of points that have the same singularities. We finally choose from these sets the one which has maximal cardinality (notice that we work on discrete data), and we use a procedure to ensure that the points will be equally spaced.

3.2 Computation of the interpolation functions

We tried two methods: the first one is adapted for the case of affine functions, where we can use results of section (2.2). The second is more general, but more time consuming.

Matching of the local singularities:

Before presenting the method, let us remark that the determination of tag points and their corresponding Hölder exponents fix the values of c_1 and c_m. Indeed, it is easy to see that the singularity at the interpolation points is $\inf(-\log_m c_1\,; -\log_m c_m)$. We take $c_1 = c_m$ in order to have the same singularity on both sides of the interpolation points.

The tag points being determined, we consider other points in the signal where we know that the estimation of the Hölder exponents is robust (the number of these points is normally greater than the one of tag points). By computing their base-m expansion, we come up with an overdetermined linear system in $\log_m c_i$'s. The solution of the latter gives the c_i's whose corresponding attractor fits the singularities the best. The results on both vowels and consonants are disappointing. The reason seems to lie in the fact that affine functions induce too much undesired high singularities in the attractor.

Genetic algorithm method:

The second method uses a genetic optimization algorithm[9]. We compute the contraction factors so that the corresponding attractor approximates the original signal the best, in the sense of the L^2 norm.

In this case we may use general functions F_n's (with notation of section (2.2)) since no knowledge of the singularities is needed. Fig. 1 displays a result obtained on the vowel /a/ when the F_n's are chosen to be of sinusoidal type. Results on consonants are comparable.

3.3 Discussion of results

The satisfactory results we obtain in the sinusoidal case may be justified by the fact that vowels have a pseudo-periodic shape which is well recovered by sinusoidal functions. For consonants, the reason may be that, with this type of functions, we do not generate a "large" number of high singularities in the attractor.

In conclusion, IFS defined by functions of sinusoidal type seem to be able to give functional representations of speech signals. However, other type of functions could probably be used depending on the signal being studied.

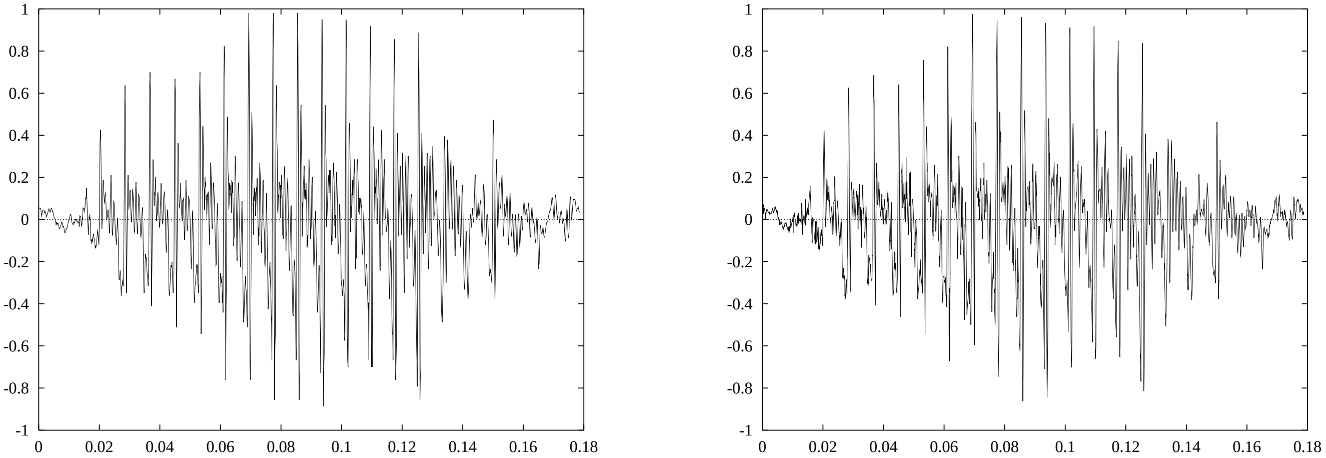

Fig. 1 vowel /a/ signal (on the left) and its reconstruction by sinusoidal IFS (on the right).

REFERENCES

1. D. Bigorgne and O. Boeffard, ICASSP (1993).
2. Calliope, *la parole et son traitement automatique*, Masson, 1989.
3. J. Lévy Véhel and K. Daoudi, Control of local singularities using multifractals, preprint, Technical report, INRIA, France, 1994.
4. M. Barnsley, Constructive Approximation (1985).
5. A. Arneodo, J.-F. Muzy, and E. Bacry, Multifractal formalism for fractal signals, Technical report, Paul-Pascal Research Center, France, 1992.
6. S. Jaffard, C.R. Acad. Sci. Paris , 19 (1992), T. 315, Série I.
7. K. Falconer, *Fractal Geometry*, John Wiley and Sons, 1993.
8. S. Mallat and S. Zhong, IEEE Tr on PAMI **14**, 710 (1992).
9. E. Lutton and J. Lévy Véhel, Fractal'93, London (1993).