Nonlinear Analysis and Synthesis Of Speech

Florin Grigoras and Vasile Apopei
Institute for Information Science
Copou 8, Iasi 6600
ROMANIA

email: flg@iit.iit.tuiasi.ro;vapopei@iit.iit.tuiasi.ro

Horia-Nicolai Teodorescu
Technical University of Iasi
Copou 11, Iasi 6600
ROMANIA

email: hteodor@etc.tuiasi.ro

Abstract: This paper reports on recent developments of a research on nonlinear methods in speech analysis and on flexible modelling of speech production. The final aim of our research is to bring together speech analysis and synthesis, in order to better understand the underlying processes of speech and to address specific applications.
Several techniques used in the analysis of dynamic nonlinear systems, are applied in order to investigate some of the short-term nonlinear characteristics of speech signal production. The research starts from the physiological evidence that the phonatory system is a time-varying nonlinear system, moreover that nonlinear processes are present in speech production, such as turbulent flow.
Before the main analysis, the speech signal is decomposed into two parts: a low-dimensional, almost linear part and a high -dimensional nonlinear part, respectively. For the latter, the largest Lyapunov exponent and fractal dimensions (capacity, correlation and information) are computed.
The synthesis tool is based on a fuzzy model of speech signal production, that implements knowledge about the speech production apparatus.

Keywords: speech production, nonlinear and nonstationary dynamic processes, Lyapunov exponents, fuzzy modelling, source-filter paradigm.

Florin Grigoras was born in 1959. He holds a MSc. degree in Electronics from Technical University of Iasi, Romania. He is now a senior researcher at the Institute for Information Science Iasi. Currently, he is completing doctoral studies in chaotic fuzzy systems at the same University. His main interests are fuzzy systems, nonlinear dynamics, chaos, digital signal processing.

Horia-Nicolai L. Teodorescu is a Corresponding Member of the Romanian Academy, a senior fellow of the Institute for Information Science of the Romanian Academy (Iasi), a Professor at Technical University of Iasi, Romania, and currently a Visiting Professor at University of South Florida, USA.
Professor Teodorescu wrote more than 150 papers, authored, co-authored, edited or co-edited more than 20 volumes, and holds 19 patents. He is Editor -in- Chief of the journals *Fuzzy Systems – Reports and Letters* and *Int. J. Chaos Theory and Applications*, a Co-Editor of *Fuzzy Economic Review*, and a member of the Editorial Board of the international journals *Fuzzy Sets and Systems, The Journal of Grey Systems, and BUSEFAL – Bulletin for Studies and Exchange of Fuzziness and its Applications*. He is a Senior Member IEEE, and got several honorary titles, including "Eminent scientist" of Fuzzy Logic Systems Institute, Japan, and he was awarded the Honorary Medal of a higher school in Barcelona, Spain.

Vasile Apopei was born in 1959. He received a MSc. degree in Electrical Engineering from Technical University of Iasi, Romania. He has been with the Institute for Information Science since 1984, holding a senior research engineer position. His research interests cover the areas of fuzzy systems, automatic control, digital signal processing, CAD.

1. Introduction

The understanding of how humans produce, hear and recognise speech signals is important in computer science, communication and medicine. But several mechanisms in speech production, perception and understanding are still insufficiently understood.

The concern for involving nonlinear techniques in analysing intricate aspects of speech production has become significant [1- 11]. Also, there are reported researches on nonlinear processes in sounds production by several musical instruments, mainly those with reeds [12]. These processes are known to undergo a dynamics similar to speech generation (at the level of vocal cords, as well as at the level of the velum).

The present research has started four years ago, with a view at verifying the underlying hypotheses:

I. to analyse nonlinear characteristics of speech sounds;

II. to determine the possibility of recognising phonemes by their nonlinear features (parameters), or to improve the phonetic recognition and speech synthesis by using such features;

III. to determine the possibility of identifying subjects' voices by using nonlinear parameters of some specified utterances;

IV. to determine how nonlinear parameters are influenced by the (health) state of the subject (neurological disorders related to speech and hearing, health state of vocal cords, larynx and of other parts of the vocal tract); based on these potential findings, to derive new diagnosis tools.

The first aim (I) is based on the rather natural hypothesis that nonlinear processes occur in speech production due to the following reasons:

• Turbulent air flow produced in the vocal tract, when either vowels are pronounced, or consonants are generated; at least for siflant and plosive

consonants, this hypothesis looks to be rather a datum.

• Nonlinear neuro-muscular processes should occur at the level of vocal cords and of larynx.

• Nonlinear couplings could be produced, during speech generation, between different parts of the vocal tract, due to (synchronous) neuro-muscular commands.

• As the neuro-muscular response to stimuli is known to be (highly) nonlinear (by physiological evidence, i.e. by directly testing the response to electrical stimuli), the nonlinear character of vocal production can be considered, from the beginning, as a nonlinear process. This way, the essential question asked by researches based only on output data from the system: "is the system nonlinear?" will be answered.

• As with physiological evidence for (nonlinear) coupling between nonlinear elements in the phonatory system, i.e. there is a feedback over a nonlinear system, one is motivated to expect some chaotic behaviour during speech production.

The next aims (II, III, IV) are related to the first hypothesis, complemented by the hypotheses that nonlinear processes play a great part in speech production, such that their relevance is high enough.

This research is intended to apply nonlinear analysis methods to speech signal. At present we deal only with simple phonemes (individual vowels). The phonemes are denoted: /a/, /e/, /i/, /o/, /u/. Further work should be done to get the research extended to larger statistics, on the one hand, and to extend the work to semi-vowels and consonants, on the other hand.

2. The Speech Production Mechanism

Speech production (see [13], [14], [21]) is accomplished by the human vocal / articulatory apparatus. Speech sounds are produced either by the quasi-periodic vibration of vocal cords (for voiced sounds), or by turbulence at some constriction point on the vocal tract (that is larynx, pharynx, oral and nasal cavities). For the voiced sounds, the pitch (i.e. the base frequency) is controlled by the vocal cords tension and from the lungs air pressure.

The vocal tract is delimited by hard and soft tissue structures, that may be considered "fixed" (the hard palate and teeth) or "movable" (referred as articulators). Most of the variation in the vocal upper tract shape is due to the primary articulators, namely tongue, lips, lower jaw and velum.

It is essential to stress the indirect coupling that exists between various articulators, by means of muscles connecting them. The couplings are realised by muscles connected at various points on larynx, hyoid and other anatomic elements (see Figure 1).

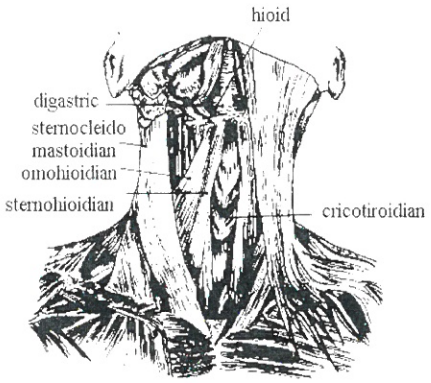

Figure 1. Muscular Coupling Between the Parts of the Vocal Tract

The specific resonance characteristics of the vocal tract are mainly dependent on the length of the vocal tract and its shape, i.e. the cross sectional area profile from vocal cords to lips. The shape of the vocal tract is controlled by neuro-muscular mechanisms involving some 20 muscles.

The source sound energy is reinforced through the vocal tract resonators, at some particular frequencies. These frequencies are commonly called resonant frequencies, poles, or formant frequencies.

Spectral Fourier analysis, applied to the acquired speech signal, reveals the underlying contribution of the excitation (as a harmonic structure of the pitch frequency) and of the filtering characteristic of the upper vocal tract (a slowly varying frequency function). The phenomena involved in speech production are studied with the help of non-invasive techniques and/or by matching a presumed model, in order to obtain a good perceptual quality of the synthesised speech.

The main trends in speech modelling are: articulatory, acoustic, stochastic (Hidden Markov

Models), and nonlinear (Neural Networks, wavelets, etc.) modelling.

Some of the most relevant models of the speech production mechanism belong to acoustic modelling, that is based on the acoustic theory of speech production [3]. According to this theory, speech waveform is considered to be the output of a resonant network (namely the vocal tract filter) that is excited by sound sources placed at the glottis. The main sections of the speech production mechanism, namely the voice source, vocal tract and radiation effects, are likely to be linearly modelled in a non-coupled manner following a source - filter arrangement. The assumption that the source and the filter can be separately modelled probably holds for most of the cases. However, this assumption is questionable for low frequencies, because the nonlinear coupling may produce damping of the first formant. It is also disputable for unvoiced speech (excitation is due to turbulence originating at constrictions on the vocal tract itself) [3, p. 14].

There are two main techniques emerging from the acoustic modelling framework: time-domain techniques (e.g. linear prediction) and frequency-domain techniques (cascade and parallel formant synthesisers).

The usual, linear model of the vocal tract has a frequency characteristic with peaks at frequencies corresponding to the resonant frequencies of the vocal tract.

It is known (see for instance [21]) that the vocal tract filter can be approximated by an all-pole filter, i.e. its transfer function can be represented as a rational function, whose parameters vary "relatively slowly" in time, for a given speech sound. Thus, samples of the speech signal over relatively short time intervals are formed when an excitation (a series of glottal pulses) is applied to that all-pole filter.

The effect of radiation on the lips can be linearly modelled by a polynomial function in z variable.

All the systems in such models should be time-varying, with their parameters changing in accordance to the sound to be produced.

For voiced speech, the excitation can be approximated by a pulse train in which the pulses appear according to the instantaneous pitch rate. If a single pitch period is analysed at a time, an analysis known as "pitch synchronous analysis", only one pulse occurs somewhere in the period.

The classic linear model is satisfactory only as a first approximation of the overall nonlinear process of speech production, and only for short time frames, on which the signal is quasi-stationary.

3. Variability and Nonlinearity in Speech Signal

The variability of the speech signal originates in the specific dynamics of the articulatory apparatus. We emphasise the well -known fact that the phonatory system is a time-varying system, and consequently speech signal is nonstationary (not only second order, but higher orders too). A large class of nonstationary - and, as pointed out below - nonlinear processes are involved in speech production.

Globally, the speech signal is a nonstationary signal. Below, we discuss the "variability" at the level of a single vocalic phoneme, in the central part of it, that could be supposed "repeatable" on a frame translation basis. A long-time variability (nonstationarity), due to the change of the spoken phonemes, is beyond the scope of this discussion. Neither is of interest here the variability in pronunciation of the same utterance or of phonemes, at different moments of time, or the variability in the frame of a single phoneme at the level of the starting and terminal parts of it (its connection to the previous and subsequent phonemes).

Taking the source -filter model as a basis, the variability of voiced speech can be roughly explained by considering the time variation of:

- glottal pulse train (shape and pitch, both being nonlinear functions in the natural case);
- central frequencies, bandwidths and relative amplitudes of the formants;
- couplings between the "source" and the "filter" parts, if accurately modelled.

As a consequence, we assert that, for the case of vocalic phonemes, the nonstationarity can be produced:

i) as a consequence of the time-dependent nature of the system which generates speech (the phonatory system), or

ii) as a consequence of the nonlinear and dynamic nature of the speech source (glottis),

or

iii) as a consequence of both time-varying and nonlinear nature of the phonatory system: the glottal generator has a nonlinear dynamics and the upper tract is continuously changing in time.

To differentiate between these cases on short time series, such as those from the middle part of vowels, is a challenging task. Besides, to our knowledge, a fundamental theory of time-dependent nonlinear systems is missing almost completely, as a methodology does. It is also likely that one cannot discriminate between cases (i) and (ii). So, the results in this research are limited by such restrictions of validity. If the nonlinearity can still be determined in such a case, by using classic methods from the theory of nonlinear systems, then this research hopefully proves (see below) that there are nonlinear processes in speech production. The main justification for using the nonlinear analysis lies in the physical and physiological evidence for nonlinear processes in speech production. However, it seems to be impossible now to trace a boundary between the role played by nonlinear processes and that played by nonstationarity processes in speech production (iii).

It turns out that generally, only for simplifying purposes the speech signal is considered to be a locally stationary signal (frames of 5-40 ms, depending on some prerequisites).

The natural vocal signal never repeats itself, even in the case of constantly uttered vowels. The variability is easily noted by monitoring: the zero crossing rate, the pitch, the shape of time domain signal, time domain envelope, variation of the central frequencies and of the bandwidths of formants etc. Statistical tests for nonstationarity can also be used to the same end. In our opinion, the speech signal may be regarded as a concatenation of nonlinear regimes, i.e. a mixture of nonlinear and nonstationary processes.

4. Portraying the Dynamics of Speech Signal

Drawing attractors of the acquired speech signal gives an intuitive picture of its dynamics. The phoneme attractors were constructed with the help of phase- space reconstruction of the time series. Commonly used maps for graphic representations in the phase -space are:

$$x = x[k];\ y = f(x[k-1])) \qquad (1)$$

$$x = x[k];\ y = x[k-1];\ z = x[k-2] \qquad (2)$$

The above mentioned maps work well in the case of systems of difference equations of relatively low order (lower than 3). Given the fact that the speech signal is the projection of a dynamic process with several degrees of freedom, we preferred to define the phase-space in the following manner:

$$\begin{aligned} x &= x[k] \\ y &= x[k+1] - x[k-1] \\ z &= x[k+1] - x[k] \end{aligned} \qquad (3)$$

where x[k], k=1, ..., N, denotes the samples of the vocal signal.

Figures 2 through 7 illustrate the plots of y(x), y(z), z(x) corresponding to vowels, as mentioned above. In these Figures, signals from two speakers (denoted by I and II) are exemplified.

The diagrams are composed of large windings, corresponding to lower spectral components (larger amplitude), as well as of small windings, corresponding to higher spectral components (smaller amplitude).

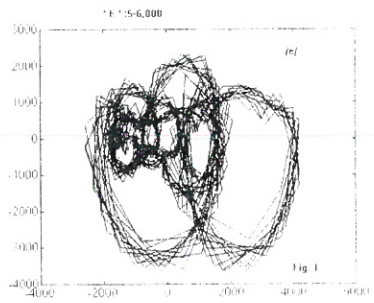

Figure 2. /e/ Vowel Attractor, 1000 Samples (80 msecs.)

Figure 2 represents the /e/ vowel attractor, built-up using 1000 samples (80 msecs., @12 Ksamp/sec). Using a larger time window, one can see that the attractor's tube has a larger section, filling the phase space and tending to cover the small windings (Figure 3: the /e/ vowel attractor,

drawn using 5000 samples, i.e. 400 msecs.). Overall shape is still distinguishable.

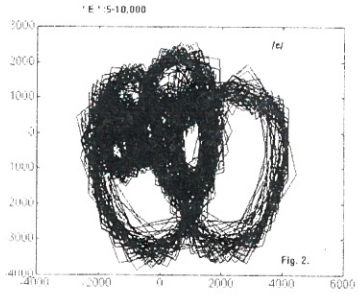

Figure 3. /e/ Vowel Attractor, 5000 Samples (400 msecs)

A comparison between two speakers uttering the /a/ vowel (500 samples, 40 msecs.) is presented in Figure 4 and a comparison between /a/ attractors drawn for different segments of the same voice file, 2000 samples apart (160 msecs.) is illustrated in Figure 5.

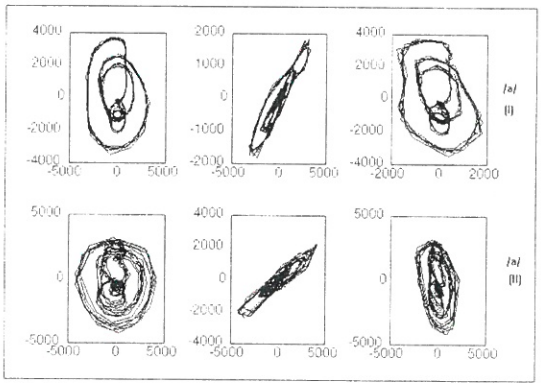

Figure 4. A Comparison Between Two Speakers Uttering the /a/ Vowel (500 Samples, 40 msecs.) Plots for x,y,z

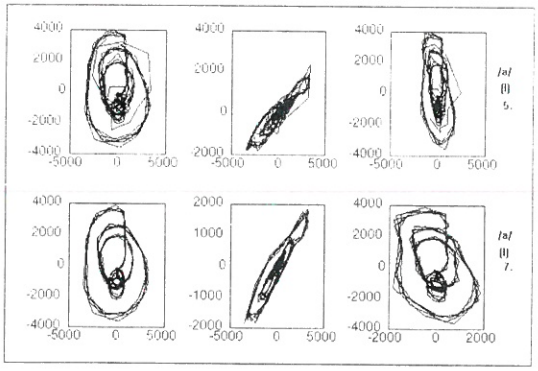

Figure 5. A Comparison Between /a/ Attractors Drawn for Different Segments of the Same Voice File, 2000 Samples Apart (160 msecs.) Showing Variability

Two phonemes uttered by the same speaker may be compared, as in Figure 6 (/a/ and /o/), as well as phonemes produced by two speakers (Figure 7, /a/ and /e/).

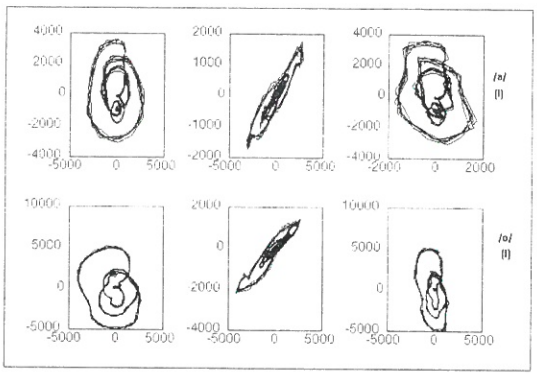

Figure 6. A Comparison Between Two Phonemes (/a/ and /o/) Uttered by the Same Speaker

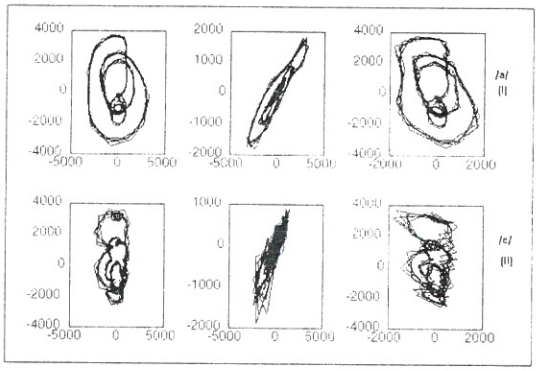

Figure 7. A Comparison Between Two Different Phonemes, Uttered by Two Speakers (/a/ and /e/)

The following comments result directly from the Figures:

a) The existence of relative stationarity periods is visually confirmed: attractors are well individualised for a time series length corresponding to 20-40 ms; although the utterance was intended to be stationary, significant variability does exist: larger time periods result in a diffuse-like aspect of the attractor (Figure 3), but still preserving a global aspect. Local variability, for very short time periods, say one or two quasi-periods, can be easily noted. For instance, in Figure 7, lower row: the signal has a rather noise-like pattern for certain speakers, affected by some disfunctions (e.g. inducing involuntary tremolo).

b) This approach is not enough to perform a phoneme or speaker classification / recognition only by looking at the attractor, but may be seen as a part of a virtually more complex analytic method. As to be shown, drawing the attractors generated the suggestion for a new method of nonlinear speech signal processing: an original

technique for decomposing and analysing the speech signal.

c) Drawing the attractors is a better way to figure out the dynamics of the system. The existence of large and small windings, of some local irregularities, suggests the idea of decomposing the speech signal into two parts:

- a "smooth variant" that follows roughly the significant windings, accounting for a so -called "main" part of the signal, keeping its "salient" properties that carry basic information contained in the speech signal; an intuitive example is low rate speech coding.
- the difference resulting from extracting the "smooth variant" from the acquired signal; we grant this derived signal to contain most of the "local", short-term variability of the speech signal.

Other techniques for obtaining the attractors are also known: time delay (Takens) [26] and SVD (singular value decomposition [27]) embedding. Our method is quite similar to the first one, while the second has the disadvantage of eliminating high order relevant dynamic features (namely the ones that we are looking for).

5. The Nonlinear Analysis Method

Testing the relevance of various complexity parameters as descriptors for nonlinear processes in speech production was one of the purposes of our research. According to the dynamics at hand, different measures of complexity show different sensitivities in contrasting processes from the same class. The maximal Lyapunov exponent is found to serve well as a descriptor for nonlinear processes in speech production.

According to our analysis method, short successive frames of speech signal are explored, in order to substantiate a nonlinear measure for the locally quasi-stationary regimes. Simple phonemes (namely vowels), uttered in some specific way ("constant" amplitude and pitch on long enough periods), were analysed.

The acquired signal is considered to be a monodimensional projection of a nonlinear dynamics. The projection carries information from the originating dynamics. Although this information is not expected to be exhaustive (i.e. to fully describe the phenomena involved), applying several nonlinear techniques may substantiate specific features of the speech signal.

Our study introduces, for the speech signal, the computation of a relative divergence rate for specifically derived time series, previously reconstructed in a high dimensional state space. The resulting measure leads to a new description of some varying nonlinear phenomena present in the speech signal.

5.1 The Principles

For complex systems that are difficult to describe by governing equations, the phase- space reconstruction technique has become widespread. Consequently, some geometric parameters that characterise the phase -space may be computed: dimension (several different definitions are known) and Lyapunov exponents [15-20]. These measures globally characterise the underlying dynamics, making it possible to obtain some basic information about the analysed system.

The phase -(or state) space is the set of variables that (hopefully, when no analytic form is available) adequately describes the evolution of the system. The attractor of the dynamics in the state -space is a geometrical object, the limit set which the dynamics converges to after transient extinction.

The Lyapunov spectrum is defined by:

$$\lambda_i = \lim_{v \to \infty} \frac{1}{v} \log[\, \mathrm{eig} \prod_{s=0}^{v} J(s) \,] \qquad (4)$$

$i = 1 \,..\, l$ (l = Lyapunov dimension)

where J is the Jacobian of the system as the generic point s moves around the attractor. Lyapunov exponents describe the average rate of exponential divergence / convergence of the (adjacent) trajectories, in a set of orthonormal directions within the embedding space. For the calculus based on time series, one uses some successive neighbourhoods, comprising neighbour points of the current point on the trajectory, in order to compute tangent maps Ts of the application associated with the dynamics. The divergence / convergence rate is computed using pairs of points on the trajectory, and iteratively (after n steps) the current pair is replaced by another one. Using a recurrence relation:

$$T_{s+k} Q_{k-1} = Q_k R_k \qquad (5)$$

(with k = 2,3,...,m, m = working embedding dimension, $m \leq l$), the Lyapunov spectrum follows from a QR decomposition:

$$\lambda_i = \lim_{v \to \infty} \frac{1}{v} \sum_{j=1}^{v} \log\left(R_j\right)_{ii} \qquad (6)$$

As it is well -known, the attractor of a dynamic system may be: a fixed point, a limit cycle, an *i*-torus or a chaotic attractor. A simple circumstance is that when the parameters of the system do not change with time. This does not occur for the vocal production system, so we have strong reasons to believe that the attractors (in fact a whole set of "local" attractors, that we attempt to reveal) change correspondingly.

In the classical models for speech production and analysis / recognition, the nonlinearity aspects of the vocal signal are dealt with either using various noise models, or referring to turbulence phenomena or even to the variability of vocalic phonemes [13], [14]. As at least turbulence is a nonlinear process, it is appropriate (and desirable) to apply principles and methods specific to dynamic nonlinear systems. According to the technique developed by us and described below, one computes the maximal Lyapunov exponent for some nonlinear time series originating in the acquired vocal signal.

5.2 The Method

The vocal signal was acquired using a 16 bit resolution A/D interface, at a sampling rate of 22050 samples/sec. The resulting time series {x[k], k=1..N}, representing the acquired vocal signal, are stored as PCM files (pulse coded modulation format). As introduced in Section 4, the shapes of the attractors for the acquired speech suggests further analysis by performing a certain signal decomposition, as we describe in this Chapter.

Other methods for speech signal decomposition are known to be effective. Mainly, they aim to evidence a deterministic part and a stochastic one. One approach (see [6]) is a version of the Harmonic + Noise Model [Stylianou, Laroche, Moulines, 1995] that implements the deterministic part as a sum of harmonically related sinusoids with time-varying harmonic amplitudes and linear phase. Although the HNM is granted to improve speech synthesis, we state its lack of natural basis. There is no proof that pure noise may arise from the dynamics of the speech production apparatus, although low level random components may be present in the acquired speech signal. More likely, the "nonperiodic" part is the result of the turbulence of the air flow and of the nonlinear, nonstationary dynamics in speech production apparatus. Theory and practice of the nonlinear dynamic systems show that the "output" of such systems is a highly irregular signal, eventually containing harmonic and random components too. In our method we prefer not to extract the "maximum" of the periodic part, but to separate a basic part (easily linear - modelable, carrying the minimum amount of perceptually relevant information), and a residue containing nonlinear dynamics that is subject to in -depth analysis. In our approach, we choose to define the decomposition of the speech signal on a perceptual basis. This is because we try to identify and describe the phenomena that add perceptual specificity to a basic, nonspecific voice signal (e.g. band-limited, or low order LPC - linear predictive coding).

We take into consideration the hypothesis that the speech production apparatus evolves in a nonlinear manner during the production of phonemes. Consequently, this should induce a specific pattern on the acquired speech signal, viewed as a monodimensional projection of the overall dynamics. Short -time nonlinear processes should be identified with typical methods of nonlinear dynamics analysis. Moreover, the results obtained should be in good concordance with the established results in the classical theory of speech. Therefore, we compare our results with some classical characteristics of the analysed speech signal. Since we try to emphasize some aspects also related to the nonstationarity of speech, it is most advisable to consider the characteristics of pitch as a reference carrying information about the vocal tract dynamics.

In our view, an approach to signal decomposition that deserves consideration is computing (followed by its extraction) of a synchronous mean of the signal or a low order LPC model:

$$\tilde{x}(k) = \frac{1}{3}\left(x(k-1) + (k) + x(k+1)\right), \quad k = 1..N, l \in \mathbb{N} \qquad (7)$$

The time series to be analysed is obtained by subtracting from the primary voice signal x[k] its own "smoothed variant" $\tilde{x}[k]$. There results the *difference signal*, obtained by a synchronous (sample by sample) extraction:

$$d(k) = x(k) - \tilde{x}(k), \quad k = 1..N \qquad (8)$$

Estimates of the correlation dimension [18], [19], [20], [25] for such time series are centred around 5.5, on a scale between 1 and 10, which reflects a high dimensional underlying process.

The final processing of the signal was done by using various length frames (p = 128 .. 2048 samples) from the initial time series: $\{x_j\}_{j=q1..q2} \subseteq \{x[k], k=1< q1,..q2 <... N\}$, $p = q_2 - q_1$. The subseries were used as input for the maximal Lyapunov exponent computing algorithm. Finally, time characteristics are drawn for the whole file, making a parallel with pitch period characteristic $\tilde{T_0}$ (expressed in the number of samples) of the voice signal.

The algorithm for the calculus of a Lyapunov exponent has two parameters: embedding dimension D and the number of steps *n* until a new pair of points is taken. Another parameter for the overall method is the length of the time subseries, *p* (number of samples). The lower limit of D parameter is forced by the estimated embedding dimension and is taken to be D = 6. The influence of the other two parameters is fully explained in the next Section.

We used three different implementations of the known algorithms for computing Lyapunov exponent [23], [24], [25]. The source code in [23] was recompiled to allow a larger number of occupied boxes. Comments about computational outputs corresponding to the use of different implementations follow in Section 6.

Figures 8 through 13 (time diagrams) illustrate the most typical cases of all those analysed. The lower and upper bounds of the parameter $\tilde{T_0}$ and of the computed value for L are on the same vertical with the corresponding symbol; *t* stands for time.

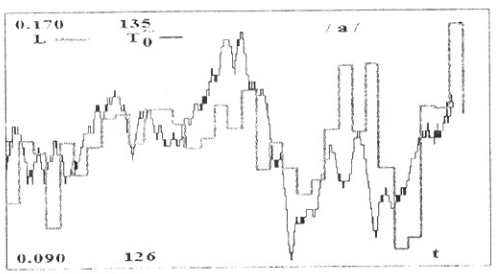

Figure 8. Phoneme /a/, Characteristics of Pitch Period $\tilde{T_0}$ and of Maximal Lyapunov Exponent L (D=6, n=3, p=2048), Time Length of Series: T=3 sec.

We have considered 3 sec. speech waveforms, save for Figure 10 that presents the case $p = 512$ (1.5 msecs. length of the waveform). Run tests for n = 3, 5, 7 are carried out (*n* = number of steps, as previously introduced).

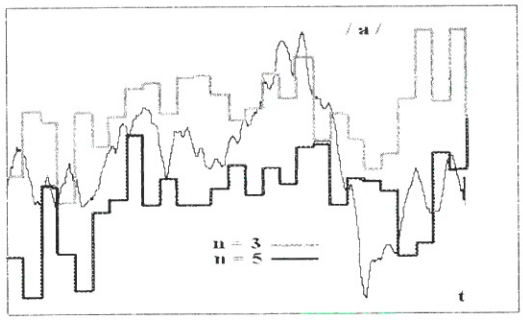

Figure 9. Phoneme /a/, Characteristics of Pitch Period $\tilde{T_0}$ and of Maximal Lyapunov Exponent L (D=6, p=2048, n=3 -Upper and n=5 -Lower Characteristic of L, Slightly Downwards Translated); Only the Time Evolution Is Illustrated

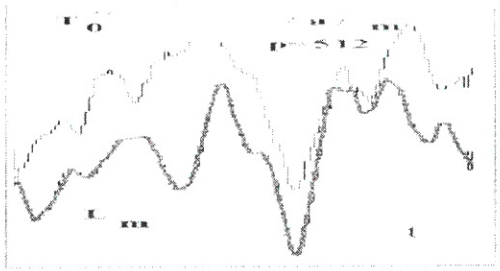

Figure 10. Phoneme /a/, for p=5 12 (Adjacent Series Are Taken With A Superposition of 128 Points and Finally Both $\tilde{T_0}$ and L Are Three Times Averaged Using Relation (7), Each Time Starting From Both Ends); T=1.5 ms.

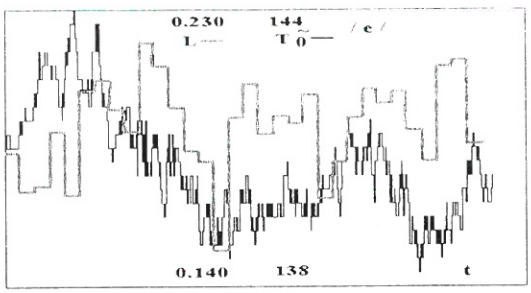

Figure 11. Phoneme /e/, $\tilde{T_0}$ and L, n=3, p=2048, T=3 sec

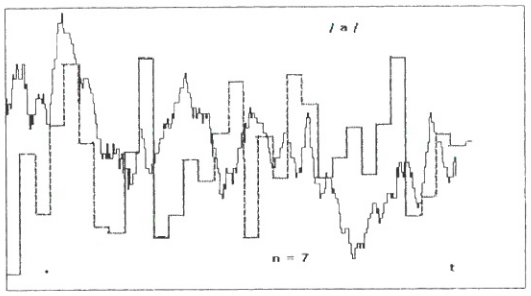

Figure 12. Phoneme /a/$_2$: An Example of Inappropriate Setting of n Parameter; n=7 (Too High)

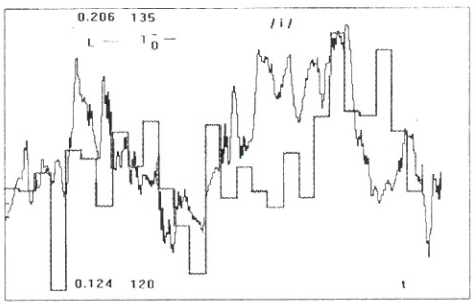

Figure 13. Phoneme /i/, T_0 and L, n=3, p=2048, T=3 sec

The pitch characteristic was computed by an interactive algorithm that is, for this type of signal, completely error free (no simplifying hypotheses are made), resulting in the "true" value of the pitch period measured in a number of samples. The algorithm is based on the fact that every quasi-period (T_0) of the time-domain speech signal has a major peak that can be detected by a threshold routine. It picks the largest peak within a time window of length equal to some presumed mean pitch period. In order to validate the calculus, the results are carefully inspected visually, for the whole length of the time series. The method is quite cumbersome, but it is effective, since we are interested in the "true" pitch characteristic, not in a long term average. Anyway, this calculus is intended to be done only at the stage of understanding the contribution of pitch variation to natural speech production.

6. Discussion

As there is strong medical (physiological) evidence that the system which generates speech (the phonatory system) is a nonlinear system, moreover as there is physical evidence that sound waves are produced - partly - by turbulence, our opinion is that the question: "does the speech signal include chaotic components?" has an a priori answer (Yes). To conclude, the problem is to analyse whether this component is strong enough (with respect to noise and to the harmonic model), and to derive its characteristics.

The results shown in the above Figures confirm, based on Lyapunov exponents, that nonlinear processes are present in speech production, as expected.

The behaviour of both parameters p and n shows a good match of our method with the classical theory, namely speech and dynamic nonlinear systems domains. When suitable parameters are chosen, the analysis of the local attractor for the quasi-stationary nonlinear regime is successfully done. As a consequence, the L characteristic for the whole x[k] series has a coherent evolution in comparison with other features of the analysed signal (here, pitch period). When the parameters do not conform to the classical theory, bad results are obtained, as one may anticipate.

The evidence that arises from the analysed cases is that there exists a significant correlation between the pitch period T_0 and the maximal Lyapunov exponent L (see Figures 8 through 10). The computed value of the maximal Lyapunov exponent L, for the derived time series, follows in the same manner the altering (growth or decay) of T_0. The dependency relation is not expressed by a multiplying constant, because, as we suppose, the computed value for L is also influenced by other phenomena, for instance the formantic structure altering. It is easy to note (Figure 10) that L better follows T_0 when relative variation of T_0 may be considered significant (large slopes); we suppose that noise affects L value in the case of smaller T_0 variations, but a rough tracing may still be noted. For our future work, we intend to preprocess the analysed series by a suitable filter, in order to cancel the noise and still preserve high order components we try to characterise. Another error source is the specific algorithm implementation [15 - 17], based on ergodic and computational hypotheses, that do not unconditionally fit every particular case studied .

The optimum value for the p parameter is within the range of 256 to 512 samples (a time interval of 11-20 ms), as in Figure 10. This certifies the coherence of the method (as to the lengths of the analysed frames), and assesses its reliability. It

shows that best results appear for a duration of the subseries equal to the classical stationarity period for the speech signal.

We may conclude that the best embedding of nonlinear phenomena (that is the optimum modelling of the nonlinear system) is done when the length of the time series, as fed into the maximal Lyapunov exponent computation algorithm, is close to the stationarity period for the analysed signal.

When applying series averaging (as in Figure 10) the resemblance between the characteristics of T_0 and those of L becomes even more obvious. Our method, at the present stage, makes only a rough computation by taking adjacent or half-overlapped series. As a consequence, there still results a small divergence of the nearby values. The averaging of the resulting L series reveals the almost "true" characteristic, as expected for a speech signal (that is basically slowly variable).

The optimum value for n is $n = 3$ (see Figure 9: for $n = 5$ the L series only roughly follows the T_0 characteristic, and for $n = 7$ the relevance (i.e. convergence) of the calculus is completely lost, as in Figure 12). This behaviour is as expected. The computation algorithm for the Lyapunov exponent is also based on the assumption that a small value for n ($n = 1$, $n = 2$) does not allow a proper evaluation of the exponent (the system has not really evolved). A high value for n (here, $n > 5$) implies that the divergence / convergence of the trajectory is lost.

We do not attempt to make a direct and uncertain connection between the computed values for L, in the cases described in this paper, and the computed values of Lyapunov exponents for chaotic systems. We only state the adequacy of the introduced measure for analysing continuously varying nonlinear systems like human vocal apparatus.

Our previous studies showed that a mean value of L roughly defined a weak clustering of vowels [9] (see Figure 14), but the spelling context deeply influenced the quality of the clustering. A similar remark is made in [1], concerning computed values for the fractal dimension of speech: absolute estimated values are not so important as their average ranges and relative differences in the context of time scale, specific discrete algorithm, speaking state.

It is also worth mentioning that we used several implementations of different algorithms and we concluded that (partially) distinct results might appear. Some values in the L time series may differ for the same x[k] series. We suspect algorithm's lack of capacity to cope with the entire specific feature set of the time series to be the source of error. A careful analysis of time series' compliance with algorithm's parameters is needed in order to obtain significant results.

Furthermore, notice that L is not a weak measure for the pitch. Tests were made using low-pass filtered (0..300 Hz) speech signal and synthetic "chirp" signal. The computed values for L were (as expected with respect to the theoretical basis of the algorithm) very close to zero and had very low dispersion, testifying an almost periodic signal. We had also tested the sensitivity of the present method to vocal signal amplitude variation and found no significant influence for a variation of 75%.

The analysis was done using signal acquired from five speakers. Here we have shown only the most representative cases.

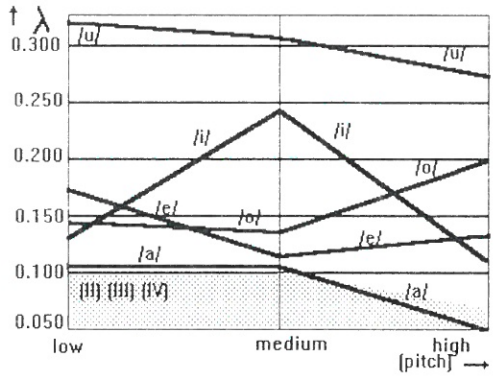

Figure 14. Mean Value for L Defines A Weak and Context - Sensitive Clustering of Vowels (the Case of Three Different Pitches)

More details on the nonlinear analysis of speech signal, according to the method described above, can be found in [28] and [29], where the first part of the paper is directly derived from.

7. The Fuzzy Model of Speech Production

7.1 General Considerations

Fuzzy modelling is a continuously growing field due to its capability of dealing with both high nonlinearities and (partial) uncertainty in the knowledge about the system.

Generally, when starting with modelling simple processes (here, vowels), the prospected model has to be flexible enough to take over the next stages of development (consonants, co-articulations). Fuzzy systems fit well in this strategy of building a final model.

A plethora of nonlinear models for the speech production mechanism is known. Some of them have major drawbacks.

The model we propose is based on the acoustical theory of speech production, and implements a set of rules that is derived from knowledge on the process of phonation.

7.2 The Model

The glottal wave is considered as a succession of pulses, each one described by some parameters that may vary from one pulse to another: amplitude (vertex); "adduct", "abduct" and "closed" phases of the pulse. This parametric representation allows good adaptation of the speaker style by changing the corresponding pulse shape. It is implemented by computing samples with the following function (To = opening, Tc = closing, Tp = pitch durations, Figure 16):

$$x(t) = \begin{cases} \frac{1}{2}\left[1 - \cos\left(\frac{\pi t}{T_O}\right)\right] & 0 \le t \le T_O \\ \cos\left(\frac{\pi\left(t - T_O\right)}{2T_C}\right) & T_O \le t \le T_O + T_C \\ 0 & T_O + T_C < t < T_P \end{cases} \quad (9)$$

The spectrum of the glottal pulse is a continuously falling characteristic.

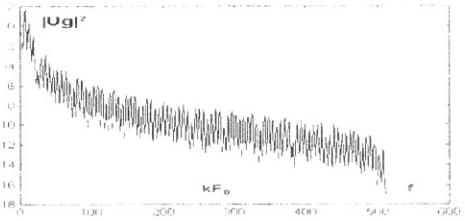

Figure 15. The Spectral Characteristic of Excitation

The upper vocal tract is considered as a resonant filter that is excited by the sound sources on the glottis. The resonant frequencies (formants) are denoted by F1 to F5.

We interpreted the source-filter theory in the sense of exploiting some specific properties of the Fourier transform *F(x)*: convolution and linearity (superposition of the effects). Based on the fact that the output (i.e. speech) signal spectrum is obtained by multiplying the source and the filter spectrums (that is convolution in time domain), we chose to consider a 'local spectrum'. So, the output spectrum is being computed step by step, for every difference of samples in the glottal pulse, finally summing up all these 'local effects'. For our purpose, a good indicative of the 'local spectrum' is the discrete derivative of the sampled glottal pulse, namely the difference between two successive samples. This may characterise the 'local' contribution, corresponding to a reduced segment / portion of the glottal excitation, to the final output spectrum, computed as a sum of all successive contributions of this kind.

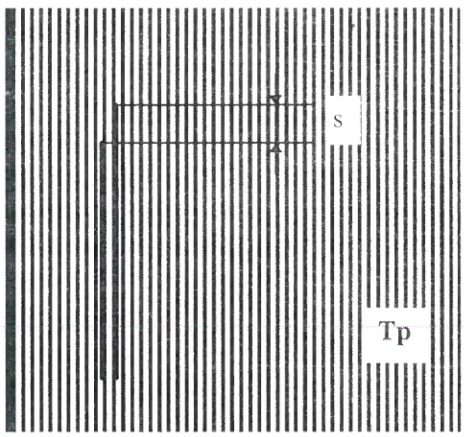

Figure 16. The Local Slope of the Glottal Pulse

The frequency bandwidth of the synthetic voice signal spectrum is closely related to the overall shape of the glottal pulse: a high derivative implies a wide bandwidth, that is wide output spectrum, while a small derivative implies a small bandwidth. Implementing this mechanism makes production of natural-like variability possible, along the same (voiced) phoneme by modifying the parameters of the glottal pulse train.

The upper formants are excited at lower levels of energy, due to rapidly falling edge of the excitation spectrum. Therefore we assume that the presence of various spectral subbands should be weighted correspondingly.

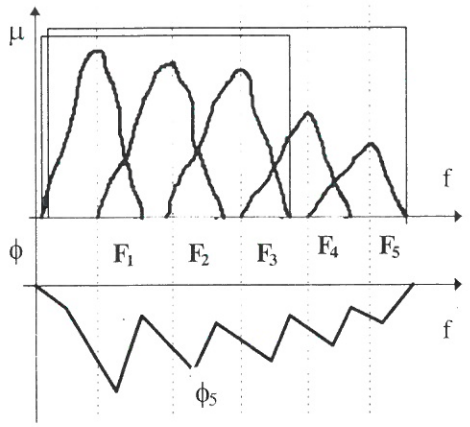

Figure 17. Formants (resonators) Spectres and Overall Phase Diagram for the Case of Five Formants

A fuzzy model is completely defined when the following are specified:

- the rule system
- the input membership functions: IMF and the fuzzification operation
- the output membership functions: OMF and the defuzzification procedure
- the inference method

The model is divided into two parts: the model for the excitation (glottal system) and the one for the upper tract. The input membership functions are defined as in Figure 18:

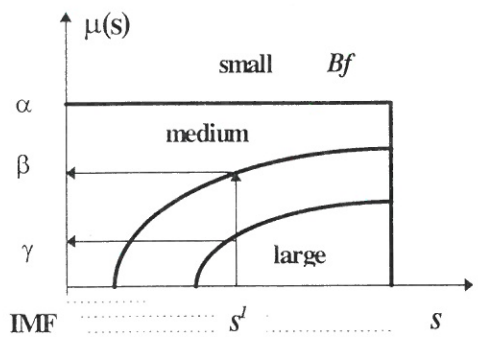

Figure 18. Input Membership Functions Definitions

Denoting the local slope of the glottal pulse by *s* and the output frequency bandwidth by B_f, the rule system for the excitation is:

IF *s* is small THEN ***Bf*** is small

IF *s* is medium THEN ***Bf*** is medium

IF *s* is large THEN ***Bf*** is large

By 'small' Bf we mean three formants, four for 'medium', and five formants for 'large' *Bf*.

For the upper vocal tract, the rule system implements rules like:

IF ***Bf*** is small, THEN OMF is **O**s

IF ***Bf*** is medium, THEN OMF is **O**me

IF ***Bf*** is large, THEN OMF is **O**lg

Os, Ome and Olg are output membership functions that graphically describe the amplitude spectral characteristic of the desired vocal signal. They are piece-wise functions and are tuned in order to give a good perceptual quality of the synthesised signal.

The inference rules are:

IF *s* is **medium**,

THEN **O** is **small** with the degree of trust α

and is **medium** with the degree of trust β

and is **large** with the degree of trust γ

the effective slope is s'

THEN: **O** is

$$\mu_O\left(s^1,f\right)=\mu_{s'\to small}\left(s^1,f\right)+\mu_{s'\to med}\left(s^1,f\right)= \tag{10}$$

$$\alpha\mu_O^{small}(f)+\beta\mu_O^{med}(f)+\gamma\mu_O^{lg}(f)$$

where $\alpha > \beta > \gamma$; f denotes the frequency. These parameters result from the fuzzification of the input (the current slope value, s^1). They are influenced by the way the IMF are designed. In fact, α , β, γ implement the contribution of the low, medium and high frequency components in the final output voice signal.

The reasoning type is FITA ("first infer then aggregate").

The defuzzification operation (obtaining crisp output, i.e. the desired signal) is done by building a discrete spectrum, according to the formula:

$$y = y_{small} + y_{me} + y_{lg} = \sum_{i=1}^{N} \{\alpha \mu_O^{small}(f) + \beta \mu_O^{me}(f) + \gamma \mu_O^{lg}(f)\} * \cos(2\pi k_i F_0 + \varphi_i) \quad (11)$$

The relation is a time domain convolution. The operation * is a vector product, with index *i* (denoting the harmonics of the fundamental frequency F_0, that is the "pitch frequency"). N is the number of harmonics taken into account in the synthesis process. We take $N = Fs / F_0$, with Fs denoting the sampling frequency. The phase characteristic is obtained from a classical model and is illustrated in Figure 17. It illustrates the phase variation along a series of coupled resonances.

The synthesis tool includes a GUI (Graphic User Interface) that allows manual tuning of spectral characteristic (amplitude and phase) or importing from preprocessed acquired vocal signal.

In contrast to the frequent practice to use noise in voice synthesisers, we prefer to use chaotic signals. Recent research assesses the presence of chaotic processes in speech production, so it is obviously a good choice to use chaotic modulation signal to give naturalness to synthesised speech (see Figure 19).

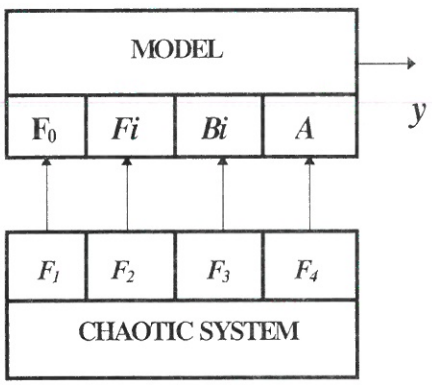

Figure 19. The Chaotic Modulation Principle

The parameters that may be modulated are: pitch frequency F_0, central frequencies and bandwidths of the formants, and the amplitude envelope of the time - domain signal.

8. Conclusions

In this paper we described a nonlinear analysis method for speech analysis that made use of Lyapunov maximal exponent computation and, finally, we introduced a new framework for speech production modelling, proved to be flexible, and able to accommodate the recent trend in nonlinear speech analysis. The knowledge acquired by analytic techniques is transferred to the production model and perceptual analysis is consequently done.

This research is starting from the physiological evidence that nonlinear processes occur in the mechanisms of speech production. So, the classical problem of researches starting from the output data of a system, namely: "is the system producing the data nonlinearly, or not?" is avoided. In contrast, one can say that speech is a process of controlling a (highly) nonlinear system to produce data with some specific coherence. What is needed in this case is to determine

- how the nonlinear "incoherence" is mastered,
- how large is the part of remaining incoherence, and
- how the nonlinear part of the speech signal can be used to assess the state of the subject and to identify the subject, taking into account that "mastering the incoherence" is dependent on the subject, and on her/his (health, emotional, etc.) state.

The analysis proved that, even in the case of constantly uttered vowels, coherent high order nonlinear processes might be outlined using specific techniques. The new introduced measure L is found to be in a close relation with other classical measures able to describe vocal signal, e. g. pitch period.

The following conclusions may be drawn.

1) Nonlinear processes are present in speech production.

2) As nonlinearity and nonstationarity coexist in the process of speech production, the speech process should be considered as a sequence of

nonlinear regimes that compose a nonstationary signal.

For the next stage of research, we intend to:

- improve the method by refining the analysis (adaptive length of the series, adding new parameters to the L estimation algorithm);
- increase the voice database;
- analyse other significant signals; other case studies are pending: vowels uttered with different "constant" pitches; "constant" amplitude and varying pitch, continuously or in three or four distinct steps of height;
- apply an algorithm for analysing the contribution of formants; a reliable elicitation of formant information is rather difficult, but the use of effective interactive algorithms is reported (see [2]).

The usefulness of this research is two-fold: the introduction of a new method for the characterisation of nonlinear phenomena in speech, and knowledge acquisition by this method may be used towards a more natural-like speech synthesis.

One of the direct applications of the nonlinear analysis of speech signals is in speech synthesis. It is generally accepted that synthetic voice suffers of unnaturalness. The reasons reside for a great part in the lack of variability of the synthetic speech, and various solutions have been proposed (e.g. [10], [11]). However, other causes of unnaturalness may also exist. We suppose that the lack of natural - like nonlinear processes in synthetic speech production is one major deficiency of the present speech synthesisers.

Also a large area of interest for the investigation of speech is medicine (see for instance [22]). We believe such methods could be used in assessing the state of the vocal tract and the related pathology. For instance, studies on the nonlinear aspects in Parkinsonian voice, as reported in [21], could benefit from applying the above methodology .

The subject identification based on subjects' voices by using nonlinear parameters for specified utterances is another application that calls for a minute investigation of the reported phenomena.

Acknowledgment

H.-N. Teodorescu is grateful for the partial support enabled by a grant from Research Funds, Switzerland (Virtual Reality in Rehabilitation). Fl. Grigoras acknowledges the partial support of a grant (GAR 111/1996) of the Romanian Academy, covering the initial stage of the research reported in this paper.

The authors would like to thank Professor Takaya Miyano for his most encouraging comments and very helpful suggestions.

REFERENCES

1. MARAGOS, P., **Fractals Aspects of Speech Signals. Dimensions and Interpolation**, Proceedings of IEEE ICASSP-91, Toronto, Canada, May 1991, pp. 417 - 420.

2. MARAGOS, P., QUATIERI, TH. F. and KAISER, J.F., **Speech Nonlinearities, Modulations, and Energy Operators,** Proceedings of IEEE ICASSP91, Toronto, Canada, May 1991, pp. 421 - 424.

3. BURROWS, T.-L., **Speech Processing with Linear and Neural Network Models**, Ph.D Thesis, Cambridge University, Cambridge, England, 1996, ftp://svrftp.eng. cam. ac .uk/pub/reports/burrows_thesis. ps.Z

4. BANBROOK, M. and MCLAUGHLIN,S., **Speech Characterisation by Nonlinear Methods**, IEEE TRANS. ON SPEECH AND AUDIO PROCESSING , 1996.

5. BANBROOK, M. and MCLAUGHLIN, S., **Dynamical Modelling of Vowel Sounds As A Synthesis Tool**, Proceedings of ICSLP'96, Fourth International Conference on Spoken Language Processing, 1996, CD-ROM proceedings: http: // www. asel. udel. edu / icslp/ cdrom /icslp96.htm

6. STYLIANOU, Y., **Decomposition of Speech Signals into A Deterministic and A Stochastic Part**, Proceedings of ICSLP'96, Fourth International Conference on Spoken Language Processing, 1996, CD-ROM proceedings: http: // www. asel. udel. edu / icslp/ cdrom /icslp96.htm

7. MIYANO, T., **Are Japanese Vowels Chaotic ?**, Proceedings of IIZUKA '96, International Conference on Soft Computing, Iizuka, Japan, September 30 October 5, 1996, pp. 634-637, World Scientific, Singapore, 1996.

8. KUBIN, G., **Synthesis and Coding of Continuous Speech with the Nonlinear Oscillator Model**, Proceedings of ICASSP-96, IEEE, 1996.

9. TEODORESCU, H.N. and GRIGORAS, F., **Nonlinear Techniques in Speech Signal Analysis**, Proceedings of International Conference on Intelligent Technologies in Human-Related Sciences, ITHURS '96, Leon, Spain, July 5-7 1996, pp. 293-298.

10. TEODORESCU, H.N., **Making Speech Synthesisers Noise-adaptable**; ELECTRONIC ENGINEERING, July 1987, p. 23.

11. TEODORESCU, H.N., CHELARU, M., SOFRON, E. and ADASCALITEI, A., **Adaptive Speech Synthesis**, in Digitale Sprach-verarbeitung - Prinzipien und Anwendungen, VDE Verlag, Berlin (W), 1988, pp. 183 - 188.

12. RODET, X., **Models of Musical Instruments from Chua's Circuit with Time Delay,** IEEE TRANS. ON CIRC. AND SYST. - II: ANALOG AND DIGITAL SIGNAL PROCESSING, Vol. 40, No. 10, October 1993, pp. 696 - 701.

13 C. Rowden (Ed.) **Speech Processing**, MCGRAW-HILL, 1992.

14. BOITE, R. and KUNT, M., **Traitement de la parole,** PRESSES POLYTECHNIQUE ROMANDES, Lausanne, 1987.

15. ECKMANN, J.-P., OLIFFSON KAMPHORST, S., RUELLE, D. and CILIBERTO, S., **Lyapunov Exponents From Time Series,** PHYSICAL REVIEW A, Vol. 34, No. 6, December 1986, pp. 4971-4979.

16. ECKMANN, J.-P. and RUELLE, D., **Ergodic Theory of Strange Attractors**, REVIEWS OF MODERN PHYSICS, Vol. 57, Part I, July 1985, pp. 617-656.

17. WOLF, A., **Quantifying Chaos with Lyapunov Exponents**, in A. Holden (Ed.) Chaos, PRINCETON UNIVERSITY PRESS, Princeton, 1986, pp. 273-290.

18. FARMER, J.D., OTT, E. and YORKE, J.A., **The Dimension of Chaotic Attractors**, PHYSICA 7D, 1983, pp. 153- 180.

19. GRASSBERGER, P. and PROCACCIA, I., **Measuring the Strangeness of Strange Attractors**, PHYSICA 9D, 1983, pp. 189-208.

20. GRASSBERGER, P., **Estimating the Fractal Dimensions and Entropies of Strange Attractors**, in A. Holden (Ed.) Chaos, PRINCETON UNIVERSITY PRESS, Princeton, 1986, pp.291-311.

21. YAIR, E. and GATH, I., **High Resolution PoleZero Analysis of Parkinsonian Speech**, IEEE TRANS. BME, Vol. 38, No. 2, February 1991, pp. 161167.

22. TEODORESCU, H.N., BUCHHOLTZER, L., CHELARU, M. and TEODORESCU, L., **A Laryngeal Prothesis Based On Perilaryngean Reflexes,** Proceedings of 9th International EMBS Conference IEEE, Vol. 4, IEEE, Boston, USA, 1987, pp. 2114 - 2115.

23. WOLF, A., **Lyapunov Exponent Software and Documentation** (free software), Department of Physics, The Cooper Union, NY, USA, http://www.users.interport.net/~wolf/chaos/cha os.httn

24. SANTIS, **Signal Analysis Time Series Processing** (free software), Laboratory of Biomedical Systems Analysis, Institute of Physiology at the University of Aachen, http Te WWW.Physiology.Aachen.DE/santis/

25. SPROTT, J. C. and ROWLANDS, G.,**Chaos Data Analyser**, Version 1.0 / 1992, PHYSICS ACADEMIC SOFTWARE, American Institute of Physics, USA, 1992 .

26. TAKENS, F., **Dynamical Systems and Turbulence**, Vol. 898, Lecture Notes in Mathematics, SPRINGER- VERLAG, Berlin, 1981, pp. 366-381.

27. BROOMHEAD, D.S. and KING, G.P., **Nonlinear Phenomena and Chaos**, in A. Hilger (Ed.) On the Qualitative Analysis of Experimental Dynamical Systems, MALVERN SCIENCE SERIES, Bristol, UK,1986, pp. 113-144.

28. GRIGORAS, F. TEODORESCU, H.N. and APOPEI, V., **Analysis of Nonlinear and Nonstationary Processes in Speech Production**, Proceedings of WASPA International IEEE Workshop, 1997.

29. TEODORESCU, H.N., GRIGORAS, F. and APOPEI, V., **Nonlinear Processes in Speech Production**, INTERNATIONAL J. CHAOS THEORY AND APPLICATIONS, Vol. 2, No. 2, 1997, pp. 35-52.