Characterization of phonemes by means of correlation dimension

PACS REFERENCE: 43.25.TS (nonlinear acoustical and dinamical systems)

Martínez, F.; Guillamón, A.; Alcaraz, J.C.

Departamento de Matemática Aplicada y Estadística (Universidad Politécnica de Cartagena)

C/ Doctor Fleming, s/n

Cartagena (ESPAÑA)

Tel: 0034968325586

Fax: 0034968325633

E-mail: f.martinez@upct.es , a.guillamon@upct.es

ABSTRACT

Speech can be considered as being generated by mechanical systems with inherently nonlinear dynamics. Recently, several techniques are applied in the analysis of dynamic nonlinear systems in order to evidence and analyze some short-time nonlinear characteristics of speech production. In order to obtain an estimation of the system complexity, in this paper we study the correlation dimension (D_2) corresponding to a recorder speech signal. Our study has been developed over frames of *N=512* points sampled at *16 KHz*, which corresponds to *32 ms* of recorder signal, because this frame selection guarantee the stationary hypothesis. The correlation dimension (D_2) has been calculated using the procedure given by Grassberger and Procaccia (1983), [8], and it shows significant differences between the results obtained over vowel and nasal consonants (low dimension) and fricatives sounds (high dimension).

Characterization of phonemes by means of correlation dimension

PACS REFERENCE: 43.25.TS (nonlinear acoustical and dinamical systems)

Martínez, F.; Guillamón, A.; Alcaraz, J.C.

Departamento de Matemática Aplicada y Estadística (Universidad Politécnica de Cartagena), Antiguo Hosital de Marina, C/ Doctor Fleming, s/n. 30202. Cartagena(ESPAÑA).
Tel: 0034968325586. Fax: 0034968325633. E-mail: f.martinez@upct.es , a.guillamon@upct.es

ABSTRACT

Speech can be considered as being generated by mechanical systems with inherently nonlinear dynamics. Recently, several techniques are applied in the analysis of dynamic nonlinear systems in order to evidence and analyze some short-time nonlinear characteristics of speech production. In order to obtain an estimation of the system complexity, in this paper we study the correlation dimension (D_2) corresponding to a recorder speech signal. Our study has been developed over frames of *N=512* points sampled at *16 KHz*, which corresponds to *32 ms* of recorder signal, because this frame selection guarantee the stationary hypothesis. The correlation dimension (D_2) has been calculated using the procedure given by Grassberger and Procaccia (1983), [8], and it shows significant differences between the results obtained over vowel and nasal consonants (low dimension) and fricatives sounds (high dimension).

1. INTRODUCTION

Recent suggestions that speech production may be a nonlinear process have sparked great interest in the area of nonlinear analysis of speech, see [1, 2, 3, 4, 5]. These authors assume the rather natural hypothesis that nonlinear processes occur in speech production, due to: turbulent air flow produced in the vocal tract; nonlinear neuro-muscular processes that should occur at the level of vocal cords and the larynx; nonlinear coupling, during speech production, between different parts of the vocal tract. Consequently, as opposed to the classical models of the vocal tract, it is possible to reconstruct a model of the vocal tract for a given configuration from the time series of the uttered signal in that configuration. Such an approach is important because the model thus obtained can be directly used for prediction purposes.

Several algorithms have been proposed for the computation of characteristic invariant measures from an experimental time series. The correlation dimension, D_2, yields a lower boundary for the degrees of freedom a signal possesses, and in this sense might be regarded as a measure of the complexity of dynamical system, [6]. In this paper we use the correlation

dimension to investigate the number of independent variables required to model the vocal tract over different Spanish sounds.

For the estimation of this nonlinear measure (D_2) we applied the reconstruction procedure proposed by Takens [7] to each speech signal frame by embedding the signal into a *d-dimensional* phase space. The correlation dimension was computed applying a modified version of the Grassberger and Procaccia algorithm, [8]. This measure was calculated for a variety of Spanish voiced sounds (vowels and nasals) and unvoiced sounds (fricatives) in natural speech.

The remainder of this paper is organized as follows. In Section 2, the techniques used in this work for estimating the correlation dimension are commented. In Section 3, we show the experimental results. Finally, conclusions based on the results are presented in Section 4.

2. METHOD

Taking into account that the speech production mechanism is nonlinear dynamic system, the methods for modeling speech production with nonlinear characteristics are required. In order to study the dynamics is necessary to embed the data into a dimension that is achieved using time delay embedding. In the reconstructed phase space the speech signal $\left(x_k\right)_{K=1}^{N}$ is represented by the set of vectors $X_i=(x_i , x_{i+t} , \ldots , x_{i+(d-1)t})$ (for $i=1,2 \ldots, N-(d-1)t$), where t is the *reconstruction delay* parameter and d is the *embedding dimension*.

The *Taken's embedding theorem* gives only a sufficient condition, but not necessary, for the embedding dimension selection, d, ($d \geq 2n+1$), where n is the *fractal dimension*, [7]. Thus the *correlation dimension*, D_2, can be considered as lower bound of the *fractal dimension* [6]. For a speech time series given, if the value of D_2 is computed over the original signal, using Taken's embedding theorem, the lower limit of d is fixed at: $d>2D_2+1$.

The estimation of the reconstruction delay parameter t is obtained by means the autocorrelation function. In this sense, we look for the time at which the *autocorrelation function* $C(t)$ has the first zero, which makes the coordinates linearly uncorrelated

$$C(t) = \frac{1}{N}\sum_{i=1}^{N-t}(x_i - \bar{x})(x_{i+t} - \bar{x}) \quad (1)$$

where $\bar{x}$ is the arithmetic means.

As it can be seen in Figures 1- 3 a plot of $C(t)$ versus τ for examples of vowels, nasals and fricatives sounds in order to obtain an estimation of the reconstruction delay parameter t. The optimal values for t parameter over the speech signal studied are $t\sim7$ for the vowel sound */a/*, $t\sim17$ for the nasal sound */m/* and $t\sim2$ for the fricative sound */s/* respectively.

In the practice, one does not know a priori the dimension of the dynamical system, and the embedding dimension is unknown too, parameters that are necessary for the phase space reconstruction. Thus, the dimensional estimate is computed for increasing embedding dimensions until the dimensional estimate stabilizes, as it can see in Figures 4-8.

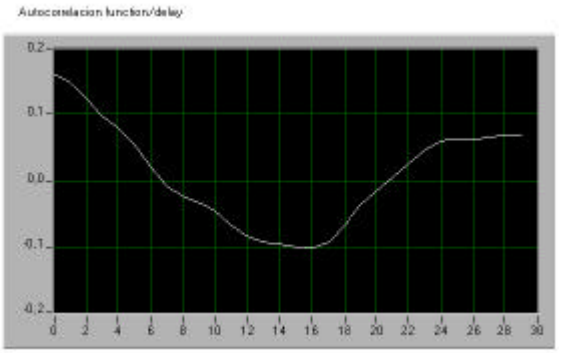

Figure 1. Plot of $C(t)$ versus t for the vowel sound */a/.*

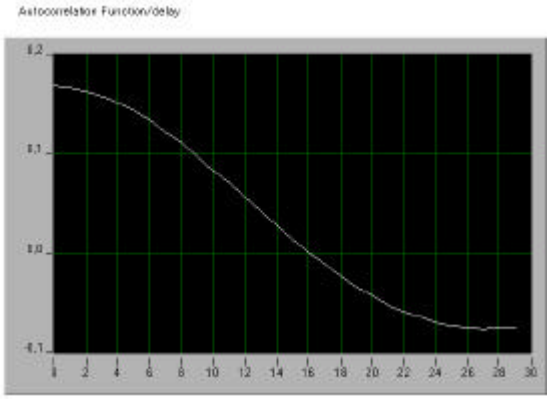

Figure 2. Plot of $C(t)$ versus t for the vowel sound */m/*

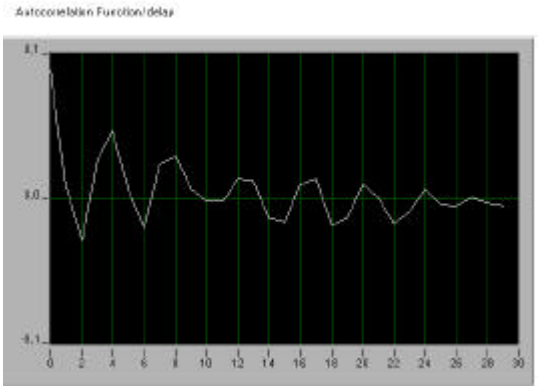

Figure 3. Plot of $C(t)$ versus t for the vowel sound */s/*

The Grassberger and Procaccia algorithm estimates the *correlation dimension* by examining the scaling properties of the *correlation integral,* $C_d(\varepsilon)$, [6]

$$C_d(\varepsilon) = \lim_{N\to\infty} \frac{1}{N^2} \sum_{i,\, j=1, i\neq j}^{N} q\left(\varepsilon - \left\| X_i - X_j \right\|\right) \tag{2}$$

where q is the Heaviside function (i.e. $q(x)=0$ *if* $x\leq 0$ and $q(x)=1$ *if* $x>0$), d is the embedding dimension, X_i and X_j represent reconstructed vectors time series from speech signal $\left(x_k\right)_{K=1}^{N}$ and the norm used is:

$$\left\| X_i - X_j \right\| = \max_{1\leq k\leq d} \left\{ X_{i+(k-1)t} - X_{j+(k-1)t} \right\} \tag{3}$$

from [7,8], we use that, for a small ε, the correlation integral should tend to zero, as some power of ε, i.e.

$$C_d(\varepsilon) \sim \varepsilon^{D_2} \quad (\varepsilon \to 0) \tag{4}$$

Using (4), we obtain the *correlation dimension, D_2,* by means of:

$$D_c = \lim_{\varepsilon \to 0} \frac{\log C_d(\varepsilon)}{\log \varepsilon} \tag{5}$$

It should be noted that the choice of embedding dimension is not critical to this implementation which relies on producing results for a range embedding dimensions usually in accordance with Taken´s theorem.

As seen in Figures 4-5, the graph of dependence of D_2 versus ε clearly shows a long plateau for a range of length scaling (*x-axis)* and a range of embedding dimension. In this paper, the estimation values of the correlation dimension were automatically extracted from the averaged values of the plateau for a suitable range of embedding dimension.

3. RESULTS

This section shows the experimental results obtained during this study. The correlation dimension was calculated for a variety of voiced sounds (nasals and vowels) and unvoiced sounds (fricatives). It is not possible to reproduce the full results for each voiced sounds in the space available and consequently the full results are only shown for some representative examples. Concretely the voiced sounds considered in our research are:

- The vowel */a/*, as a Spanish word *"tomaré"*
- The vowel */e/*, as a Spanish word *"mantel"*
- The vowel */i/*, as a Spanish word *"picota"*
- The vowel */u/*, as a Spanish word *"nube"*
- The vowel */m/*, as a Spanish word *"mañana"*
- The nasal */n/*, as a Spanish word *"ana"*
- The fricative */s/*, as Spanish word *"salten"*

Our experiments have been computed over recorder signals from a speech Spanish database (AHUMADA Database), [9]. We based our calculations on *N=512* data points sampled at *16 kHz* (*32 ms* duration) in order to guarantee the stationary hypothesis

For each sound analyzed, the reconstruction delay τ were calculated as described above. Table 1 the achieved results for all the sounds.

	/a/	*/e/*	*/i/*	*/u/*	*/m/*	*/n/*	*/s/*
τ	*7*	*11*	*14*	*14*	*17*	*15*	*2*

Table 1. Reconstruction delay τ for each sound studied.

Figures 4 and 5 show the plot of D_2 versus for the sounds */a/, /e/, /n/* and */s/* respectively. In these Figures, D_2 was estimated considering a reconstruction delay τ in accordance with Table 1 and setting a suitable rang of embedding dimension. In each Figure

the line represent the value of plateau that is used to extract the correlation dimension.

Table 2 shows the results obtained for the estimation of the correlation dimension, for each sound analyzed, by means the procedure developed in this paper.

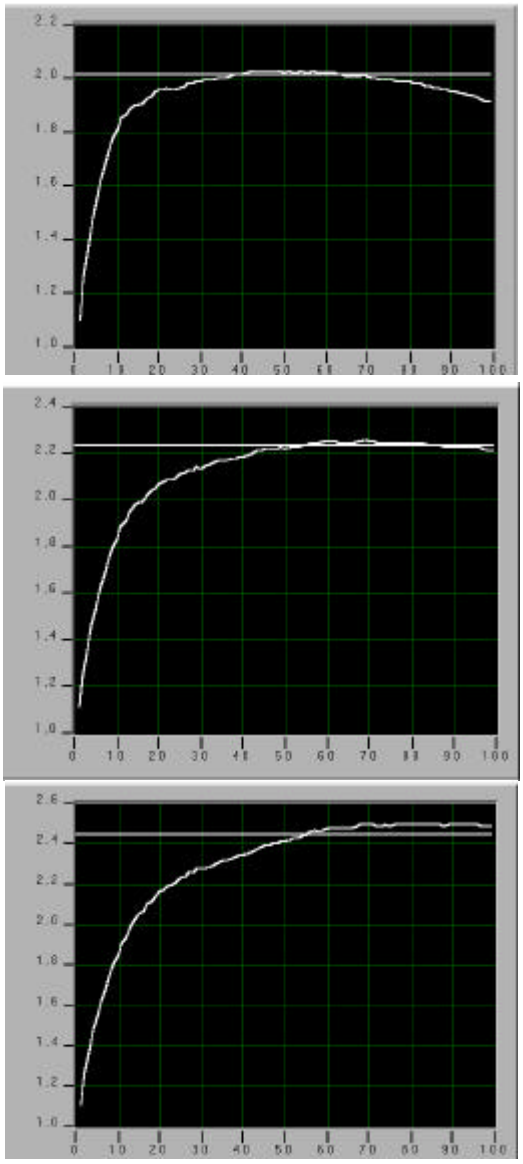

Figure 4. Plot of D_2 versus t for vowel /a/ for a range of embedding dimension: (a) $d=5$, (b) $d=6$ and (c) $d=7$.

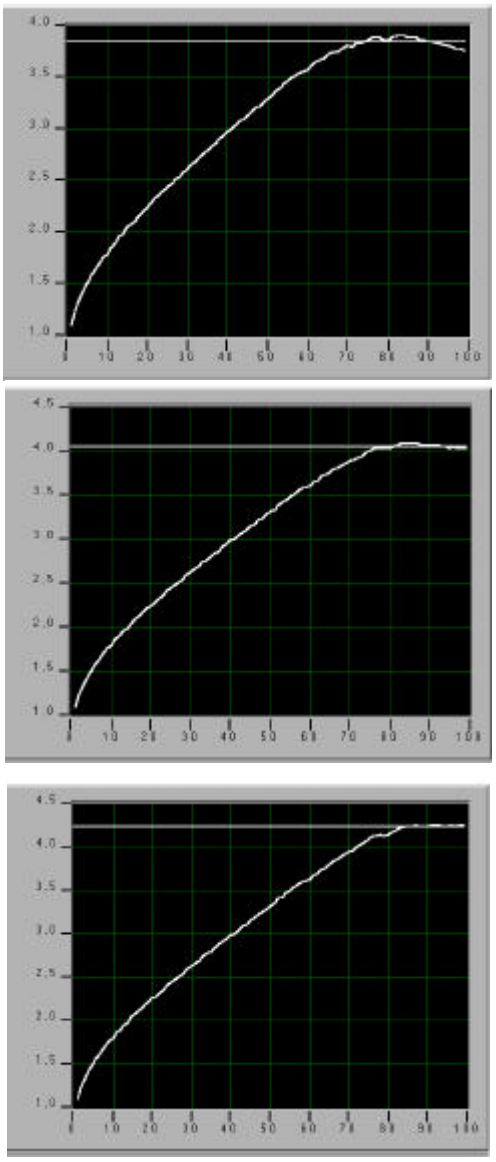

Figure 5. Plot of D_2 versus t for fricative /s/ for a range embedding dimension: (a) $d=5$, (b) $d=6$ and (c) $d=7$.

Remark.- In order to detect the plateau to estimate the correlation dimension, we used the method proposed in [2] (*d=n+1)* with initial values of *the embedding dimension parameter d*.

	/a/	*/e/*	*/i/*	*/u/*	*/m/*	*/n/*	*/s/*
D_2	*2.3*	*2.2*	*1.8*	*1.9*	*1.8*	*1.7*	*4.2*
Range of d	*[5,7]*	*[5,7]*	*[3,5]*	*[3,5]*	*[3.5]*	*[3,5]*	*[11,13]*

Table 2. Correlation dimension for each sound studied.

4. CONCLUSIONS

In this paper, we show a fast a easily implemented method for the estimation of the correlation dimension. The proposed procedure offers similar performances to that of the classical methods and the results obtained are comparable with those achieved in previous studies, [1, 2].

The results obtained in this research confirm the initial conjecture about the dimensionality of the different sounds studied. Vowels and nasals sounds are low-dimensional (i.e. $D_2<3$). Fricatives, particularly the sound */s/*, have correlation dimension $D_2>4$. Consequently, trajectories of fricatives require greater number of independent variables to model as shown by their higher dimensions.

The calculation of the correlation dimension on a variety of vowels sounds presented here suggest that this parameter may be linked to the manner of articulation. In this sense, we suggest that the vowels produced with high articulation have lower dimensionality than others produced by low articulation.

REFERENCES

[1] M. Banbrook and S. McLaughlin, “Is speech chaotic? Invariant geometrical mesures for speech data”, *IEEE Colloquium on Exploiting Chaos in Signals Processing,* 8/1-8/10, 1994.

[2] A. Kumar and S.K. Mullick, “Attractor dimension, entropy and modeling of speech time series", *Electronics Letters*, Vol.26, No.21, 1990, pp. 1790-1791.

[3] H.N. Teodorescu, F. Grigoras and V. Apppei, “Nonlinear and nonstacionary Processes in Speech Production”, *Int. J. of Chaos Theor. and Appl,* Vol. 5, No. 3, 1996, pp. 1453-1457.]

[4] M. Banbrook and S. McLaughlin P., “Speech Characterization and Synthesis by Nonlinear Methods”, *IEEE Transactions on Speech and Audio Processing*, Vol. 7, No 1, 1999, pp. 1-17.

[5] T. Montero, F. Martínez, A. Guillamón and J.C. Alcaraz, “Caracterización de señales del habla natural mediante la Entropía de Renyi”, *XIX Congreso Anual de la Sociedad Española de Ingeniería Biomédica,* 2001.

[6] P. Grassberger and I. Procaccia, “Characterization of the strange attractors”, *Phys. Rev. Lett.*, Vol. 5, 1983, pp. 346-349.

[7] F. Takens, “Detecting Strange Attractorsin Turbulence”, Lectures Notes in Mathematics 898, Springer, Berlín, 1981, pp. 366-381.

[8] P. Grassberger and I. Procaccia,“Estimation of Kolmogorov entropy from a chaotic signal”, *Phys. Rev. A,* Vol. 28, 1983, pp. 2591-2593.

[9] R. García and J.L. Gómez, *Base de Datos para la Identificación y Clasificación de Locutores*, Universidad Politécnica de Madrid, 1997.