Identificación de Señales Verbales en el Espacio de Fase Reconstruido

José A. Brito
Universidad de Los Andes, Postgrado en Computación,
Mérida, Venezuela, 5101
jabbmv@cantv.net

Wladimir J. Rodríguez
Universidad de Los Andes, Postgrado en Computación,
Mérida, Venezuela, 5101
wladimir@ula.ve

y

Flor E. Narciso
Universidad de Los Andes, Departamento de Computación,
Mérida, Venezuela, 5101
fnarciso@ula.ve

Abstract

In this paper we describe the use of Multilayer Perceptron Array for learning and classifying speech signals, using characteristic vectors of reconstructed dynamics. First, we consider the phonatory system as a black-box, where the only available data is its output: the speech signal. Theoretically, if reconstruction of system dynamics is properly made, geometric structures or attractors outlined in the space are topologically equivalent to original, and inaccessible, structures. This is a way of accessing underlying dynamics, and is the starting point for two kinds of experiments: classification of vowels and digits, with Venezuelan Spanish voices. Results verify positively that characteristics vectors extracted from underlying dynamics hold discriminative power for distinguishing between classes of speech signals. Besides, neural networks are able to generalize using this kind of data.
Keywords: Speech signals classification, reconstructed dynamics, pattern recognition, non linear dynamics, neural nets, SpeechDat.

Resumen

Este artículo se describe el uso de arreglos de redes neuronales de retropropagación para el aprendizaje y clasificación de señales verbales, usando vectores de características de la dinámica reconstruida. Primero, se considera el sistema fonatorio como una caja negra, donde la única data disponible es la salida: la señal verbal. Teoreticamente, si la reconstrucción de la dinámica del sistema es correcta, las estructuras geométricas o atractores del espacio son topologicamente equivalentes a las estructuras originales inaccesibles. Esta es una forma de acceder a la dinámica subyacente, y es el punto de partida para dos tipos de experimentos: clasificación de vocales y dígitos, con voces en español venezolano. Los resultados verifican positivamente que los vectores de características extraidos de la dinámica subyacente tiene poder discriminatorio para distinguir entre clases de señales verbales. Además, las redes neuronales son capaces de generalizar usando este tipo de datos.
Palabras claves: , Clasificación de señales verbales, espacio de fases reconstruido, reconocimiento de patrones, dinámica no lineal, redes neuronales, SpeechDat.

1. INTRODUCCIÓN

Recientemente, ha surgido el interés por la parametrización no lineal de señales verbales a partir de la reconstrucción de la dinámica del sistema fonatorio [5, 6]. En principio, se trata de resolver un problema de clasificación utilizando un enfoque cualitativo sobre perfiles del proceso físico subyacente [11, 12], en este caso, la fonación. Por el contrario, las técnicas convencionales de análisis recurren a la hipótesis de linealidad en la emisión verbal, aunque existen numerosas objeciones a dicho proceder. Por ejemplo, en el popular modelo **fuente-filtro** de generación de voz, la fuente de excitación es la turbulencia desarrollada en el mismo tracto vocal, por lo que en este caso la fuente natural no corresponde con el modelo. Además, un modelo lineal va a tener dificultades para ajustarse a la alta variabilidad de la señal. Por lo tanto, al final, estos modelos son solo una aproximación del sistema. El enfoque presentado aquí se basa en un arreglo de redes neuronales de retropropagación para el aprendizaje y clasificación de las señales verbales, usando vectores de características del espacio de fase reconstruido.

Se realizaron dos tipos de experimentos: vocales y dígitos. En ambos casos las señales fueron extraídas de la base de datos de voces venezolanas SpeechDat [4]. Por lo que este trabajo es el primero que emplean técnicas no lineales con voces venezolanas. En cada experimento se define dos corpus de voces: C_E y C_P, consistentes de señales de entrenamiento y prueba de la red neuronal respectivamente.

2. ESPACIO DE FASE RECONSTRUIDO

En el caso de los sistemas no lineales, con datos incompletos, la extracción de información nueva a partir de los datos resulta más difícil que en la contraparte lineal [2]. Si el sistema en cuestión es altamente complejo (eg., el sistema fonatorio), pero sólo una de sus propiedades (eg., la señal verbal) está al alcance de algún sensor, los procedimientos de análisis tradicionales resultarán muy limitados. Como alternativa, la reconstrucción del espacio de estados permite recuperar la dinámica de un sistema no lineal a partir de una única serie de tiempo [1]. En concreto, las trayectorias forman en este espacio unas estructuras geométricas denominadas atractores. Naturalmente, el espacio reconstruido no equivale completamente a la dinámica interna del sistema, pero bajo ciertas restricciones teóricas, preserva la topología de la misma. Esto permite que las conclusiones obtenidas en la dinámica reconstruida resulten válidas también en la verdadera e inaccesible dinámica interna (caja negra) [1, 7, 9]. Además, el espacio de fase reconstruido facilita la detección de estructuras que en la serie de tiempo podrían pasar desapercibidas.

A continuación se describe la obtención del espacio de fase reconstruido. Considérese un conjunto de muestras uniformemente espaciadas de una única variable, como la señal verbal S_v. El espacio de fase reconstruido es una representación multidimensional de la señal contra versiones demoradas de sí misma (subseries). En términos más formales, el espacio de fase reconstruido se forma mediante la definición de vectores $\mathbf{V}_n$ en $\Re^m$:

$$\mathbf{V}_n = \{S_v[n], S_v[n+\tau], ..., S_v[n + (m - 1)\tau]\} \qquad (1)$$

o

$$\mathbf{V}_n = \{S_v[n], S_v[n-\tau], ..., S_v[n - (m - 1)\tau]\} \qquad (2)$$

donde $S_v[i]$ es el valor de la señal en tiempo i (una muestra). Por su parte, m y τ son los parámetros fundamentales de reconstrucción conocidos como dimensión embebida y retardo respectivamente. El teorema de Takens [9], que relaciona el espacio de fase reconstruido con la verdadera dinámica interna del sistema, expresa que dados m y τ suficientes, la dinámica real y el espacio de fase reconstruido resultan topológicamente idénticos. Pruebas preliminares con el método de entropía diferencial [3] arrojaron valores bajos para m y τ sobre el corpus de vocales. De esta forma, se fijó $m = 2$ y $\tau = 3$ en los subsiguientes experimentos. La Figura 1 muestra el espacio de fase reconstruido para una vocal arbitraría, con una malla de 100 bloques sobre el plano. Por el otro lado, los dígitos constituyen señales muy complejas debido a su superior riqueza fonética, y consecuentemente, una representación espacial sencilla no es posible. Un enfoque algo diferente, discutido en la próxima sección, se utilizará para los dígitos.

Note que para cada eje en la figura corresponde al intervalo [-1,+1], esto se logra por medio de la normalización de la señal verbal: cada muestra se divide por $max(abs(S_v))$. Este paso, aunque trivial, es importante ya que permite que los bloques B_i sea de dimensión fija. Estrictamente, cada bloque es un cuadrado, y su área $4 / r$ (sin dimensiones), donde r es el total de bloques sobre el plano. En la figura 1, $r = 100$, y por lo tanto el área de cada bloque es 0.04.

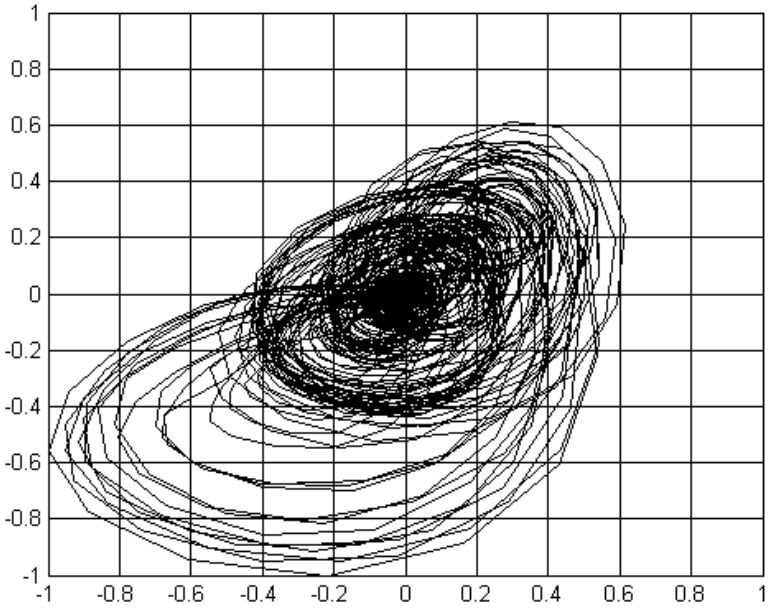

Figura 1. Espacio de fase reconstruido para una vocal en español venezolano

3. EXTRACCIÓN DE CARACTERÍSTICAS

Para las vocales, el vector de características V_C^V viene dado por la densidad espacial de cada bloque $B_i(1 <= i <= r)$. Así, V_C^V consta, en principio, de *r* elementos. En cada B_i la densidad espacial se computa como sigue:

$$spatialDensity(B_i) = \frac{|B_i|}{|S_v|} \qquad (3)$$

Adicionalmente, para dotar de robustez a la clasificación, V_C^V se extiende con los *r* elementos resultantes de aplicar el análisis anterior sobre la serie de tiempo $S_v^d = S_v[i + 1] - S_v[i]$. De esta forma, se incorpora al vector la rapidez de variación en la señal, mediante una aproximación con las primeras diferencias.

En el caso de los dígitos, el análisis descrito se aplica sobre tramas superpuestas proporcionadas, sin los *r* elementos de S_v^d. Para nuestros efectos, una trama es una secuencia de muestras, o en otras palabras, es una subsecuencia de S_v. En una señal dada, las tramas siempre poseen la misma longitud, pero entre señales, el tamaño de la trama puede variar. Precisamente se denominan proporcionadas porque su tamaño es una proporción del de la señal. La señal se divide en *np* segmentos de igual tamaño, *np* par. Siendo *LSeg* la longitud de cada uno de los *np* segmentos (es decir, *LSeg = Longitud(S_v) / np*), entonces la longitud de la trama *LTr* se establece en $2 \times LSeg$. La primera trama inicia con la primera muestra de la señal. De allí en adelante, una trama inicia en la muestra central de la trama que le antecede y se extiende *LTr* muestras. Por ende, las *np-1* tramas en una señal inician respectivamente en las muestras

$$i \text{ x } LSeg + 1 \ (0 \le i \le np - 2)$$

Posteriormente, se aplica en cada trama un análisis similar al de las vocales. De esta forma, el análisis de las tramas contribuye al vector de características de los dígitos, V_C^D con $(np-1) \times r$ elementos.

4. ARREGLO DE REDES PERCEPTRONICAS MULTICAPA

Sea *n* la cantidad de categorías a reconocer (en el caso de las vocales, *n=5*, y con los dígitos, *n=10*). Luego, para cada experimento, el clasificador consiste en un arreglo $[R_1R_2...R_n]$ de *n* redes perceptrónicas multicapa R_i. Así, se asocia una red con cada categoría. Una vez definido el corpus C_E, puede procederse con la sesión de entrenamiento del clasificador. Básicamente, R_i ($1 \le i \le n$) se entrena con todas las señales *j* ($1 \le j \le cardinalidad(C_E)$) pertenecientes a C_E. A todas las entradas de entrenamiento en las que se verifique *categoría(j) = categoría(i)*, se les asocia una salida igual a 1; en el otro caso, la salida es 0.

Posteriormente, al momento de clasificar, la señal de entrada se caracteriza y el vector resultante se administra a cada una de las *n* redes. La red con la salida más alta determina la categoría en la que se clasifica la señal.

Las redes neuronales usadas en el clasificador constan de tres capas. La cantidad de unidades en la entrada dependerá del tamaño del vector de características, sea V_C^V o bien V_C^D. Luego, si dicho vector incorpora p componentes, entonces se dispondrá de p unidades de entrada. Por ejemplo, si se particiona el espacio de estados en *100* bloques, y se calcula la densidad de puntos en cada uno, V_C^V, y por ende la capa de entrada de la red, constará de *200* elementos. Por su parte, en la capa oculta se ubican 5 unidades, y en la capa de salida se coloca una sola unidad. Las funciones de activación son sigmoides logarítmicas, a excepción de la neurona de salida, que emplea una función de transferencia lineal. Finalmente, el algoritmo usado en el entrenamiento de las redes es Levenberg-Marquardt [8]. Este es un algoritmo avanzado para la optimización no lineal, y suele converger al mínimo error más rápidamente que la retropropagación clásica, aunque su consumo de memoria resulta notoriamente elevado.

5. RESULTADOS

Para cada experimento a realizar, se construyó un MPA. Se estableció *r = 100*, y con los dígitos se fijó *np* en *6*. A fin de verificar el funcionamiento de los clasificadores, se les proporcionó como entrada las señales en los corpora de entrenamiento, obteniendo el 100% de exactitud en el reconocimiento. Ensayos dependientes del hablante arrojaron tasas promedio de reconocimiento del 92% y 65.5%. Por otro lado, con las pruebas independientes del hablante, más interesantes, se obtuvo, para las vocales, un reconocimiento promedio de 55% (ver tabla 1), y para los dígitos, una tasa de 66% (ver tabla 2).

Tabla 1. Matriz de confusión para vocales independientes del hablante

	a	**e**	**i**	**o**	**u**	**%**
a	8	2	2	8	0	40.00
e	0	8	9	3	0	40.00
i	0	4	12	2	0	60.00
o	2	2	0	14	2	70.00
u	0	1	3	3	13	65.00
						55.00

Tabla 2. Matriz de confusión para dígitos independientes del hablante.

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**%**
0	11	0	0	0	0	2	0	7	0	0	55.00
1	0	12	2	4	0	0	0	0	1	1	60.00
2	0	0	13	3	1	0	0	0	0	3	65.00
3	2	0	2	14	0	1	0	1	0	0	70.00
4	1	0	0	0	17	0	0	2	0	0	85.00
5	1	0	0	0	0	11	0	8	0	0	55.00
6	2	1	0	4	1	0	9	1	0	2	45.00
7	3	0	0	0	1	2	0	14	0	0	70.00
8	0	0	1	0	1	0	0	0	18	0	90.00
9	0	3	0	2	0	1	0	1	0	13	65.00
											66.00

Obsérvese cómo los valores en las diagonales principales de las matrices de confusión confirman la tendencia de los MPA a clasificar correctamente las señales de entrada. En el caso de las vocales, la variabilidad entre las señales independientes del hablante deteriora la exactitud de reconocimiento. De forma interesante, en los dígitos no acontece así, quizás porque estas señales incluyen más información que las vocales, y el MPA logra capturarla.

Algunos estudios han abordado previamente la caracterización de la densidad de los atractores en el espacio de estados, para clasificar señales verbales. Por ejemplo, en [10] se emplea un clasificador bayesiano, con una exactitud promedio de 34.49% en las vocales, si bien se trata de vocales inglesas. En [5, 6] se utiliza el espacio de información difuso, con un reconocimiento perfecto, pero con un corpus de sólo seis señales. Comparando con técnicas en el dominio de la frecuencia, se tiene el trabajo de Maldonado [4], el cual procede sobre corpus de dígitos venezolanos, con tasas de reconocimiento superiores al 90%. Empero, este último trabajo utiliza aproximaciones analíticas ya establecidas, como coeficientes ceptrales y modelos ocultos de Markov.

6. CONCLUSIONES

Considerando que este análisis esta basado completamente en el dominio del tiempo, las ratas de reconocimiento son bastantes buenas. Sin embargo, se necesita realizar más investigaciones para determinar el efecto de un análisis dimensional mayor $m > 5$. Cuando se utilizan más de dos dimensiones, la caracterización se hace más difícil. Excepto por el entrenamiento de las redes neuronales, las técnicas expuestas no requieren de recursos computacionales muy altos. Por lo que habría que preguntarse si un análisis de altas dimensiones es conveniente, tomando en cuenta la precisión de las técnicas en el dominio de frecuencias. Este tipo de investigación es necesaria debido que el aprendizaje de más fenómenos seguramente necesitaran más atractores.

Finalmente, existe una información discriminante en la dinámica reconstruida, y las redes neuronales, son capaces de generalizar lo que distingue entre las clases de señales. Se encontró que la conjunción entre las características de la dinámica reconstruida y las redes neuronales contiene el suficiente poder discriminatorio para clasificar las señales de voz, aunque es necesario continuar con las investigaciones.

Referencias

1. H. Abarbanel, R. Brown, J. Sidorowich, y L. Tsimring, "The analysis of observed chaotic data in physical systems", *Reviews of Modern Physics*, vol. 65, No. 4, 1993.
2. E. Bradley, "Time series analysis", in *Intelligent Data Analysis: An Introduction*, Springer, 1999.
3. T. Gautama, D. Mandic, y M. Van Hulle, "A differential entropy based method for determining the optimal embedding parameters of a signal", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, 2003.
4. J. L. Maldonado, *Tratamiento y reconocimiento automático de señales de la voz venezolana*, Disertación doctoral, Universidad de Los Andes, 2003.
5. W. Rodríguez, H.-N. Teodorescu, F. Grigoras, A. Kandel y H. Bunke, "A fuzzy information space approach to speech signal non-linear analysis", *International Journal of Intelligent Systems*, vol. 15, No. 4, pp. 343-363, 2000.
6. W. Rodríguez, "Similarity of Dynamical Systems", Ph.D. Thesis, University of South Florida, 1998.
7. T. Sauer, J. A. Yorke, y M. Casdagli, "Embedology", *Journal of Statistical Physics*, vol. 65, pp. 579-616, 1991.
8. Shepherd, A. J. *Second-Order Methods for Neural Networks*, Springer-Verlag, 1997.
9. F. Takens, "Detecting strange attractors in turbulence", *Dynamical Systems and Turbulence*, Warwick, 1980.
10. J. Ye, M. T. Johnson, y R. J. Povinelli, "Phoneme Classification using Naive Bayes Classifier in Reconstructed Phase Space", *10th IEEE Digital Signal Processing Workshop*, 2002.
11. F. Zhao, "Extracting and Representing Qualitative Behaviors of Complex Systems in Phase Space", *Artificial Intelligence*, vol. 69, pp. 51-92, 1994.