

Complejidad Lempel-Ziv: robustez frente a *spikes* y pérdida de muestras en registros EEG

A. Molina Picó¹, E. Cirugeda Roldan¹, D. Cuesta Frau¹

¹ Departamento de informática de sistemas y computadores (Disca), Universidad politécnica de Valencia (Campus de Alcoy), Alicante, España, {anmopi, ecirugeda}@giica.com, dcuesta@disca.upv.es

Resumen

Los registros biológicos aparecen generalmente enmascarados por artefactos extrínsecos al sistema de generación o altamente contaminados por ruido. Los métodos tradicionales o de inspección visual resultan insuficientes cuando se desea realizar una caracterización exhaustiva de los mismos.

Los *spikes* se consideran impulsos interferentes en señales biomédicas y se caracterizan por ser fácilmente detectables pero difícilmente corregibles.

Este trabajo caracteriza la influencia de los *spikes* en la complejidad de Lempel-Ziv (LZC) sobre señales electroencefalográficas y se estudia cómo afecta la eliminación de los mismos, induciendo una pérdida de muestras en la señal.

Los resultados presentados muestran como para ocurrencias de *spikes* con probabilidad de hasta 50% el efecto introducido en LZC es asumible y para probabilidades de aparición superiores hasta un 80% se recomienda su eliminación.

1. Introducción

Los métodos de análisis lineal o las técnicas de inspección visual utilizadas generalmente para el análisis y la caracterización de las señales biomédicas y en particular de los electroencefalogramas (EEG), son a menudo insuficientes. Estas técnicas presentan habitualmente un coste computacional alto o son incapaces de revelar las dinámicas que, a veces, puede aparecer contaminadas por artefactos ruidosos [1].

Estas señales pueden contener de manera frecuente valores atípicos en forma de impulsos de corta duración. Estos impulsos reciben el nombre de *spikes*. Los *spikes* pueden deberse a fallos electrónicos del sensor, interferencias externas o a errores en el procesamiento de la señal. Aunque pueden ser corregidos o eliminados, este proceso puede eliminar características importantes de la señal [2-4].

Los métodos de análisis no lineal basados en medidas de entropía o complejidad se consideran más adecuados para el análisis de registros biológicos al no hacer asunciones implícitas inherentes a las señales que analizan [5-7]. La complejidad Lempel-Ziv (LZC) es una técnica de análisis no lineal que evalúa la aleatoriedad de secuencias finitas en función del número de patrones distintos existentes en una señal y la codificación simbólica [8].

Este trabajo presenta una caracterización de la influencia de los *spikes* en la estimación de LZC sobre señales EEG y su aplicabilidad en la separación entre sujetos patológicos y sanos.

2. Métodos

2.1. Complejidad Lempel-Ziv (LZC)

LZC es una técnica de análisis no lineal propuesta por Lempel y Ziv en [8]. Esta medida relaciona el número de patrones distintos que se encuentran a lo largo de una secuencia. La utilidad de la medida como ayuda al diagnóstico ha quedado demostrada en diversos entornos como la neurología [9], cardiología [10] y genética [11].

La base del algoritmo para el cálculo de LZC radica en la transformación de una serie temporal $x(n)$, $n=1:N$ en una secuencia simbólica $P = \{s(1), s(2), \dots, s(n)\}$. De acuerdo a [8, 9, 12, 13] se establece que una codificación binaria es adecuada para aplicaciones fisiológicas. Para realizar la codificación, es necesaria la definición de umbral TH de forma que éste asocia a un valor numérico de $x(n)$ un símbolo $s(i)$ [14]:

$$s(i) = \begin{cases} 0 & x(i) < TH \\ 1 & x(i) \geq TH \end{cases} \quad (1)$$

Se considera que la secuencia P está compuesta por distintas palabras y cada palabra se compone de distintos símbolos. La secuencia P se recorre en busca de distintas palabras. Por cada palabra nueva, se incrementa en una unidad el contador de complejidad $c(n)$. Este proceso presenta los siguientes pasos [13]:

- Definiciones:
 - S, Q : subsecuencias de P y
 - SQ : concatenación de S y Q .
 - $SQ\pi$: secuencia SQ tras eliminar el último carácter.
 - $v(SQ\pi)$: vocabulario que contiene todas las secuencias diferentes de $SQ\pi$.
- Inicialización de las variables:
 - $c(n) = 1$
 - $S = s(1)$
 - $Q = s(2)$
 - $SQ\pi = s(1)$,
- En general, $S = \{s(1), s(2), \dots, s(r)\}$ y $Q = s(r+1)$, siendo r el índice del símbolo analizado. Si Q pertenece al diccionario $v(SQ\pi)$, con $SQ\pi = \{s(1), s(2), \dots, s(r)\}$, entonces Q es una subsecuencia.
- Se actualiza la subsecuencia $Q = \{s(r+1), s(r+2)\}$, y se comprueba si pertenece a $v(SQ\pi)$ o no.
- Repetir el paso anterior hasta que Q no pertenezca a $v(SQ\pi)$. Entonces, se puede decir que $Q = \{s(r+1),$

$s(r+2), \dots, s(r+i-1)\}$ no es una subsecuencia de $SQ\pi$ y se incrementa $c(n)$ en una unidad.

6. Se actualizan las secuencias S y Q . $S = \{s(1), s(2), \dots, s(r+i)\}$ y $Q = \{s(r+i+1)\}$.
7. Todos estos pasos se repiten hasta llegar al último carácter $r=n$. En un último paso, se aplica una normalización del contador $c(n)$ por el factor $n/\log_a(n)$. El objetivo de esta normalización es obtener un valor de complejidad independiente de la longitud de la serie temporal.

2.2. Generador de Spikes

El método utilizado para generar los *spikes* se basa en [2, 15], donde los *spikes* se definen como un fenómeno interferente que ocurre frecuentemente en señales biomédicas y afecta considerablemente a la entropía aproximada (ApEn) y muestral (SampEn).

La ocurrencia de *spikes* en las señales se genera a partir de un proceso aleatorio binomial $B(N, p_s)$, donde N es la longitud del tren y p_s su probabilidad de aparición.

La amplitud de los *spikes* se define mediante una variable aleatoria uniforme $U(-3K, 3K)$, siendo K la amplitud pico-pico de la secuencia $x(n)$.

Todos los *spikes* generados se definen como impulsos de muy corta duración y, por tanto, se les asigna una duración fija de una muestra. El tren de *spikes* se define matemáticamente, como:

$$s(n) = \sum_k a_k \delta(n - n_k) \quad (2)$$

Donde a_k representa la amplitud y n_k la posición de cada *spike*. Se realizan 50 simulaciones de trenes de *spikes* sintéticos y se superponen a cada una de las secuencias EEG originales. La figura 1 representa dos realizaciones de *spikes* superpuestos a un registro EEG perteneciente a un paciente epiléptico con $p_s=10\%$.

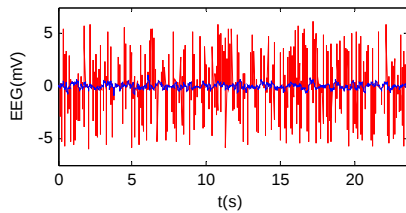


Figura 1. Registro EEG perteneciente a un sujeto epiléptico (azul) contaminado con spikes ($p_s=0.1$) (rojo).

2.3. Pérdida de muestras

El método de pérdida de muestras consiste en eliminar las muestras que contienen los *spikes* generados sintéticamente. El ratio de eliminación coincide con p_s y se incrementa a intervalos del 10% hasta alcanzar el 90% de las muestras suprimidas.

2.4. Métodos de evaluación

A continuación se presentan los métodos de evaluación considerados

2.4.1 Test de segmentación

El test de segmentación considerado es el Mann-Whitney U-Test (UT) que se trata de la versión no paramétrica del test de Student. El test indica si se puede asumir la hipótesis que dos muestras independientes provienen de distribuciones de igual mediana. Si la probabilidad asociada al estadístico es inferior a 0.05, la hipótesis de que ambos conjuntos siguen una distribución idéntica no es asumible y se considera que son separables.

2.4.2 Coeficiente de correlación cruzada

El coeficiente de correlación cruzada (CC) es una medida de la similitud existente entre dos conjuntos de datos. Se define a partir de las matrices de covarianza y autocovarianza de los vectores. Si el valor obtenido es superior a 0.8 se puede asumir que los resultados son suficientemente similares y ofrecen validez clínica [16].

2.4.3 Curvas ROC

Las curvas ROC son una representación gráfica que indica la capacidad de discriminación de los métodos de análisis entre clases de sujetos. La gráfica representa la *sensibilidad* frente a *1-especificidad*. La sensibilidad indica la probabilidad de clasificar correctamente a un sujeto enfermo y la especificidad la probabilidad de clasificar correctamente a un sujeto sano.

El área bajo la curva (AUC) es un parámetro estadístico derivado de la curva ROC que permite clasificar de manera aproximada el rendimiento del método diagnóstico. Un valor próximo a la unidad ($AUC \approx 1$) indica que los conjuntos de datos no se solapan y el método permite discriminarlos perfectamente. Sin embargo, un valor $AUC \approx 0.5$ indica que el método no es capaz de segmentar entre los dos conjuntos.

3. Materiales

Los registros EEG considerados se clasifican en 5 clases (A–E), estando cada una de ellas formada por 100 segmentos de un único canal de 23.6 segundos de duración ($F_s=173.61\text{Hz}$) a partir de un único registro de larga duración preprocesado para eliminación de artefactos externos [17], es de libre distribución y se puede descargar en <http://epileptologie-bonn.de>.

Las clases A y B consisten en registros tomados de sujetos sanos. Los voluntarios se encontraban despiertos en un estado de relajación con los ojos abiertos (A) y cerrados (B). Las clases C, D y E pertenecen a 5 pacientes epilépticos que han conseguido tener un completo control de los ataques. Los segmentos en la clase D han sido registrados en la zona epileptogénica, mientras que los pertenecientes a la clase C se han registrado en el hipocampo, que se encuentra en el hemisferio opuesto del cerebro. Si bien las clases C y D contienen solo actividad medida durante intervalos libres de ataques, la clase E está compuesta únicamente por actividad convulsiva. Los ataques epilépticos representan un estado patológico de la actividad cerebral, caracterizados por descargas sincronizadas de grandes grupos de neuronas [8].

4. Resultados

4.1. Registros originales sin spikes

Los resultados obtenidos indican que los registros EEG pertenecientes a sujetos de control son más irregulares y complejos que los pertenecientes a sujetos epilépticos. En la tabla 1 se muestran las separabilidad (UT (p)), la capacidad de clasificación a partir de la AUROC y el coeficiente de correlación cruzada (CC) de las series con *spikes* respecto al caso sin *spikes* ($p_s=0$). En ella se observa como en ausencia de *spikes*, LZC es capaz de clasificar correctamente ambos grupos (AUROC ≈ 1). El test de separabilidad UT es también positivo y se rechaza con total probabilidad la hipótesis de que ambos grupos tienen una misma mediana.

4.2. Influencia sobre LZC al superponer spikes

En la tabla 1 se expresan los resultados del test de separabilidad UT, el AUROC que da una idea de la capacidad clasificatoria del método y los coeficientes de correlación cruzada que indican la robustez de la medida.

p_s	UT (p)	AUROC	CC
0	0.001	0.994 ± 0.001	1.00
0.1	0.001	0.991 ± 0.001	0.98
0.2	0.001	0.988 ± 0.001	0.97
0.3	0.001	0.983 ± 0.001	0.95
0.4	0.001	0.972 ± 0.001	0.91
0.5	0.001	0.944 ± 0.002	0.85
0.6	0.001	0.868 ± 0.003	0.70
0.7	0.001	0.718 ± 0.004	0.43
0.8	0.001	0.576 ± 0.005	0.16
0.9	0.001	0.514 ± 0.005	0.02

Tabla 1. Parámetros de evaluación estadísticos obtenidos al superponer spikes sintéticos con probabilidad p_s .

El UT indica que se puede asumir separabilidad entre los sujetos de control y patológicos si la probabilidad obtenida en el test es $p < 0.01$. Puede observarse como esto se cumple incluso cuando el 90% de las muestras se encuentran contaminadas.

El AUC indica la capacidad discriminatoria de LZC entre los dos grupos para distintas cantidades de *spikes* superpuestos. Los valores obtenidos en la tabla indican que LZC tiene buen rendimiento (AUROC > 0.75) para tasas de *spike* inferiores al 65%.

La figura 2 muestra la evolución de LZC al superponer *spikes* con distinta probabilidad p_s . La complejidad aumenta de forma logarítmica con p_s . Se observa como la influencia de los *spikes* sobre la LZC de señales patológicas es superior a la que estos tienen en la LZC de sujetos control ya que inicialmente LZC de patológico es inferior a LZC de control y la distancia entre LZC de ambos grupos se ve disminuida conforme aumenta p_s .

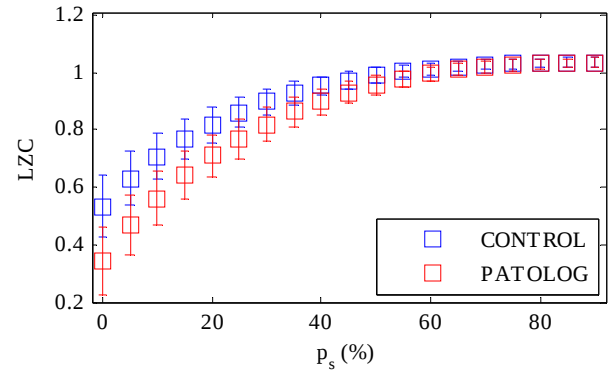


Figura 2. LZC ($\mu \pm 2\sigma$) al superponer spikes con distinta probabilidad p_s a los registros EEG pertenecientes a sujetos de control (azul) y patológicos (rojo).

4.3. Influencia sobre LZC al eliminar los spikes

En este apartado se muestran los resultados obtenidos al eliminar las muestras de las señales contaminadas que contienen *spikes* y aplicar LZC.

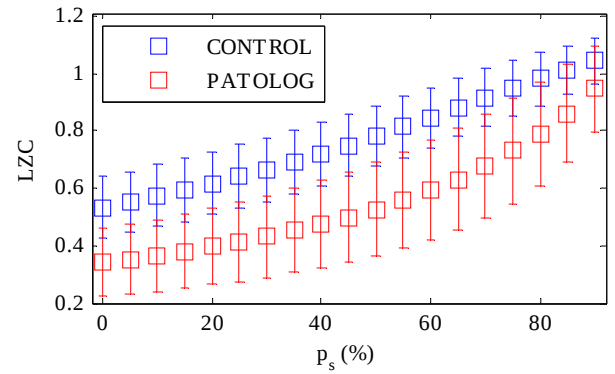


Figura 3. LZC ($\mu \pm 2\sigma$) al eliminar las muestras con los spikes superpuestos sintéticamente con probabilidad p_s en los registros de control (azul) o patológicos (rojo).

La evolución de LZC según la cantidad de *spikes* eliminados se representa en la figura 3. Puede observarse como la complejidad aumenta casi de manera lineal en ambos grupos al eliminar las muestras contaminadas. Estos resultados coinciden con los parámetros estadísticos mostrados en la tabla 2.

p_s	UT (p)	AUROC	CC
0	0.001	0.994 ± 0.001	1.00
0.1	0.001	0.995 ± 0.001	1.00
0.2	0.001	0.995 ± 0.001	0.99
0.3	0.001	0.996 ± 0.001	0.99
0.4	0.001	0.996 ± 0.001	0.99
0.5	0.001	0.996 ± 0.001	0.98
0.6	0.001	0.995 ± 0.001	0.97
0.7	0.001	0.992 ± 0.001	0.94
0.8	0.001	0.980 ± 0.002	0.90
0.9	0.001	0.891 ± 0.004	0.74

Tabla 2. Parámetros de evaluación estadísticos obtenidos al eliminar los spikes superpuestos con probabilidad p_s .

El test de separabilidad UT indica que la complejidad obtenida en ambos grupos es distinta y LZC permite discernir entre sujetos de control y patológicos con una probabilidad de error inferior al 1%.

5. Discusión

LZC es capaz de segmentar entre los registros de la base de datos estudiada pertenecientes a sujetos de control y epilépticos. Los valores obtenidos coinciden con estudios previos [6, 18, 19] en los que se asocia una mayor complejidad a sujetos sanos. Los registros EEG se componen principalmente de ritmos de baja frecuencia, asociando los ataques epilépticos a la generación de ritmos más regulares y con mayor frecuencia.

Los *spikes* suponen un incremento en la complejidad debido a la aparición de nuevas subsecuencias en el diccionario $v(SQ\pi)$. Pese a estas variaciones, LZC es capaz de discernir y clasificar entre los dos grupos de pacientes hasta con tasas de *spikes* del 50%, límite que se considera válido por el coeficiente de correlación según [16]. La justificación a este buen comportamiento se encuentra en el propio método: el hecho de convertir la secuencia numérica a símbolos binarios provoca que se pierda resolución en la señal, pero aumenta la robustez frente a *spikes*.

La eliminación de muestras provoca un incremento lineal en la complejidad. Este incremento no se debe tanto a una variación en el contador de complejidad, sino a un decremento en el tamaño de las secuencias. Pese a este sesgo introducido por la longitud de las señales, LZC es capaz de separar y clasificar entre los sujetos de control y los patológicos incluso al aplicar un diezmado del 90%. Sin embargo, los coeficientes de correlación obtenidos solo aportan validez clínica a un diezmado inferior al 80%.

6. Conclusión

En este estudio se ha analizado la capacidad discriminatoria y clasificatoria de LZC en registros EEG contaminados con *spikes*. La medida es capaz de discriminar perfectamente entre sujetos de control y epilépticos en ausencia de *spikes*. Esta capacidad permanece hasta llegar a una tasa de *spikes* del 50% (la mitad de las muestras de la señal). Si las señales contaminadas contienen entre un 50 y un 80% de *spikes*, éstos deben ser detectados y eliminados mediante un diezmado. Si la cantidad de *spikes* es mayor, los registros deben ser descartados.

Agradecimientos

Este trabajo ha sido financiado por el MICINN a través de los proyectos de investigación TEC2008-05871 y TEC2009-14222.

Referencias

- [1] Motamedi-Fakhr S, Moshrefi-Torbati M, Hill M, Hill CM, White PR. Signal processing techniques applied to human sleep EEG signals-a review. *Biomed Sig Process Control*, vol 10, sup 0, 2014, pp 21-33.
- [2] Molina A, Cuesta D, Aboy M, Crespo C, Miro P, Oltra S. Comparative study of approximate entropy and sample entropy robustness to *spikes*. *Artific Intell Med*. vol 53, sup 2, 2011, pp 97-106.
- [3] Chen Z, Ivanov PC, Hu K *et al*, Effect of nonstationarities on detrended fluctuation analysis. *Phys. Rev. E*, vol 65, sup 4, 2002, pp 041107-1-15.
- [4] Niebur E, Generation of synthetic spike trains with defined pairwise correlations. *Neural comput*, vol 19, sup 7, 2007, pp 1720-1738.
- [5] Rényi A, On measures of entropy and information. In *Proc. Fourth Berkeley Symp. Math. Statist. Prob*, 1961, University of California Press, Berkeley, 1961. pp 547-61
- [6] Pincus S, Gladstone I, Ehrenkranz R, A regularity statistic for medical data analysis. *J Clin Monit Comput*, vol 7, 1991, pp 335-45.
- [7] Pincus S, Goldberger AL, Physiological time-series analysis: what does regularity quantify?, *Am J Physiol*, vol 266, sup 4 pt.2, 1994, pp H1643-56.
- [8] Lempel A, Ziv J. On the complexity of finite sequences. *IEEE Trans inform theory*. Vol. 22, sup. 1, 1976, pp. 75-81.
- [9] Radhakrishnan N, Gangadhar BN. Estimating Regularity in Epileptic Seizure Time-Series Data: A Complexity Approach. *IEEE Eng Med Biol*, vol 17, 1998, pp 89-94 (ISSN: 1932-8737).
- [10] Zhang XS, Zhu YS, Thakor NV, Wang ZZ. Detecting ventricular tachycardia and fibrillation by complexity measure. *IEEE Trans Biomed Eng*, vol 46, sup 5, 1999, pp 548-555.
- [11] Gusev VD, Nemytikova LA, Chuzhanova NA. On the complexity measures of genetic sequences. *Bioinformat*, vol.15, sup 12, 1999, pp 994-999.
- [12] Wu X, Xu J. Complexity and brain function. *Acta Biophys Sinica*. Vol. 7, pp 103-106.
- [13] Zhang XS, Roy RJ, Jensen EW. EEG complexity as a measure of depth of anesthesia for patients. *IEEE Trans. Biomed. Eng*. Vol 48, sup 3, 1999, pp 1424-1433.
- [14] Nagarajan R. Quantifying physiological data with Lempel-Ziv complexity –certain issues, *IEEE Trans. Biomed. Eng*. Vol. 49, sup11, 2002, pp. 1371-1373.
- [15] Chen Z, Ivanov PC, Hu K, Stanley HE. Effect of nonstationarities on detrended fluctuation analysis. *Phys. Rev. E*, Vol 65, sup. 4, 2002.
- [16] Shimizu K. Telemedicine y mobile communication. *Eng Med Biol Mag*, vol 18, sup 4, 1999, pp 32-44.
- [17] Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, Elger CE. Indications of nonlinear deterministic and finite dimensional structure in time series of brain electrical activity: Dependence on recording region and brain state. *Phys Rev E*. Vol 64, 2001, pp 1-8.
- [18] Lake DE, Richman JS, Griffin MP, Moorman JR. Sample entropy analysis of neonatal heart rate variability. *Am J. Physiol. Regul. Integr. Comp. Physiol*, Vol. 283, Sup. 3, 2002, pp. 789-797.
- [19] Lundelin K, Vigil L, Bua S, Gomez I, Honrubia T, Varela M. Differences in complexity of glycemic profile in survivors and nonsurvivors in an intensive care unit: a pilot study. *Crit. Care Med*, Vol.38, Sup.3, 2010, pp. 849-854.