19. Matemática aplicada

Comparación de la señal de vibración de la piel en el cuello respecto de otras señales asociadas al aparato fonador para la estimación de la frecuencia fundamental.

Cuaranta, Marina Candelaria; Schlotthauer, Gastón
mericuaranta@gmail.com; gschlotthauer@gmail.com;
Facultad de Ingeniería
Universidad Nacional de Entre Ríos

Resumen

En este trabajo realizamos una comparación entre distintas señales asociadas al aparato fonador: señal de voz, señal de electroglotograma (EGG) y señal de vibración de la piel en el cuello (VPC); aplicando sobre ellas diferentes métodos de detección de la frecuencia fundamental (F0). El objetivo fue estudiar a la señal de VPC como alternativa a las de voz o EGG en la tarea de estimar la frecuencia fundamental, con vistas a que pueda ser utilizada en un sistema de adquisición portátil para monitoreo de la actividad vocal. Utilizamos 83 señales de cada tipo, correspondientes al registro de la vocal /a/ sostenida. Como estimadores de la F0 implementamos diferentes algoritmos de corte clásico: función de autocorrelación (AC), periodograma (PG) y de detección de eventos (DE). Medimos el desempeño de los algoritmos implementados a través de la raíz del error cuadrático medio (RECM), el cual no muestra una diferencia significativa para dos de los tres métodos en su aplicación a las distintas señales. Concluimos en función de esto que resulta factible el uso de la señal de VPC para extraer información confiable sobre la frecuencia fundamental.

Palabras clave: Frecuencia fundamental, voz, electroglotograma, vibración de piel en cuello.

Introducción

Este trabajo se desarrolló en el Laboratorio de Señales y Dinámicas no Lineales (LSyDNL) de la Facultad de Ingeniería de la Universidad Nacional de Entre Ríos. El trabajo consiste en una comparación entre distintas señales asociadas al aparato fonador: señal de voz, señal de electroglotograma (EGG) y señal de vibración de la piel en el cuello (VPC); aplicando sobre ellas diferentes métodos de detección de la frecuencia fundamental (F0) producto de la vibración cuasi – periódica de la glotis durante los fonemas sonoros.

Llevar a cabo esta comparación tiene como principal motivación evaluar el desempeño de la señal de VPC en su uso para el cálculo de la F0, en relación con las señales de voz y EGG, dado que posee algunas importantes ventajas en cuanto a su adquisición en un sistema portátil que permita la evaluación objetiva de la actividad vocal: (1) el sistema de registro es mucho más económico que un electroglótografo (Baken & Orlikoff, 2000) y (2) resulta menos susceptible a ruidos del ambiente que un micrófono frente a la boca que capture la señal de voz.

La organización del informe es la siguiente: en primer lugar se plantea brevemente el objetivo de este trabajo; luego se describen las señales utilizadas y los métodos implementados; posteriormente están presentados los resultados obtenidos junto a una discusión sobre ellos y por último, se exponen las conclusiones sobre el trabajo realizado.

Objetivos

El objetivo de este trabajo fue estudiar a la señal de VPC como alternativa a las de voz y EGG en la tarea de estimar la frecuencia fundamental de la voz, con vistas a que pueda ser utilizada en un sistema de adquisición portátil para monitoreo de la actividad vocal.

Materiales y Métodos

El LSyDNL cuenta con una base de datos de registros simultáneos de señales asociadas al proceso de fonación: señal de voz, señal de electroglotograma (EGG) y señal de vibración de la piel en el cuello (VPC).

La base de datos se construyó con la participación de 83 sujetos distintos, para cada uno de los cuales se registraron señales de vocales sostenidas (5), conjuntos de vocales (2), transiciones continuas entre vocales (2) y periodos de habla continua (2). Todas las señales tienen una frecuencia de muestreo de 50 kHz y una resolución de 16 bits.

Para este trabajo utilizamos únicamente las señales correspondientes al registro de la vocal /a/ sostenida, ya que la práctica clínica ha demostrado que el uso de vocales sostenidas es práctica y suficiente para una evaluación general de la actividad vocal (Tsanas et al., 2014).

A continuación haremos una breve descripción de los tres tipos de señales utilizadas, haciendo hincapié en las características prácticas a los fines de este trabajo.

De forma más o menos simplificada diremos que la señal de voz es una representación de la onda sonora producida por el aparato fonador que es capturada por un micrófono. Puede decirse que los sonidos del habla son el resultado de la excitación acústica del tracto vocal, cuyas características varían constantemente (Rufiner, 2009). La excitación típica de los fonemas sonoros, como los correspondientes a las vocales,

es producto de una vibración de las cuerdas vocales que modelan el flujo de aire proveniente de los pulmones, lo que da como resultado la generación de pulsos cuasi-periódicos (Rufiner, 2009). Los intervalos de vibración glótica o su recíproco, la frecuencia glótica, son los responsables de la frecuencia fundamental observable en este tipo de fonemas.

La señal de electroglotograma (EGG) es producto de una técnica llamada electroglotografía, propuesta en 1957 por Fabre. En términos generales, esta técnica consiste en la medición de la impedancia del tejido de la zona del cuello a la altura de la glotis mediante la inyección de una corriente de alta frecuencia y baja amplitud. Debido a las diferencias de impedancias entre el tejido blando y el aire que circula por la laringe, se registra una variación de la amplitud de la señal que sigue de alguna forma el movimiento oscilatorio de la glotis. (Baken & Orlikoff, 2000)

Por último, a través de un sensor de presión puede sensarse la vibración de la piel del cuello. Estas vibraciones son un reflejo del choque mecánico que se produce en el cierre de la glotis y de las variaciones de la presión subglótica (Askenfelt, Gauffin, & Sundberg, 1974), de mayor frecuencia que las primeras (Neumann, Gall, Schutte, & Miller, 2003). Entonces, en estas señales se podrá observar también el comportamiento cuasi-periódico de la vibración de la glotis.

En la Figura 1 se muestra un ejemplo de cada una de estas señales, donde se puede observar claramente el comportamiento cuasi-periódico de las mismas que permite extraer de ellas una frecuencia fundamental (F0).

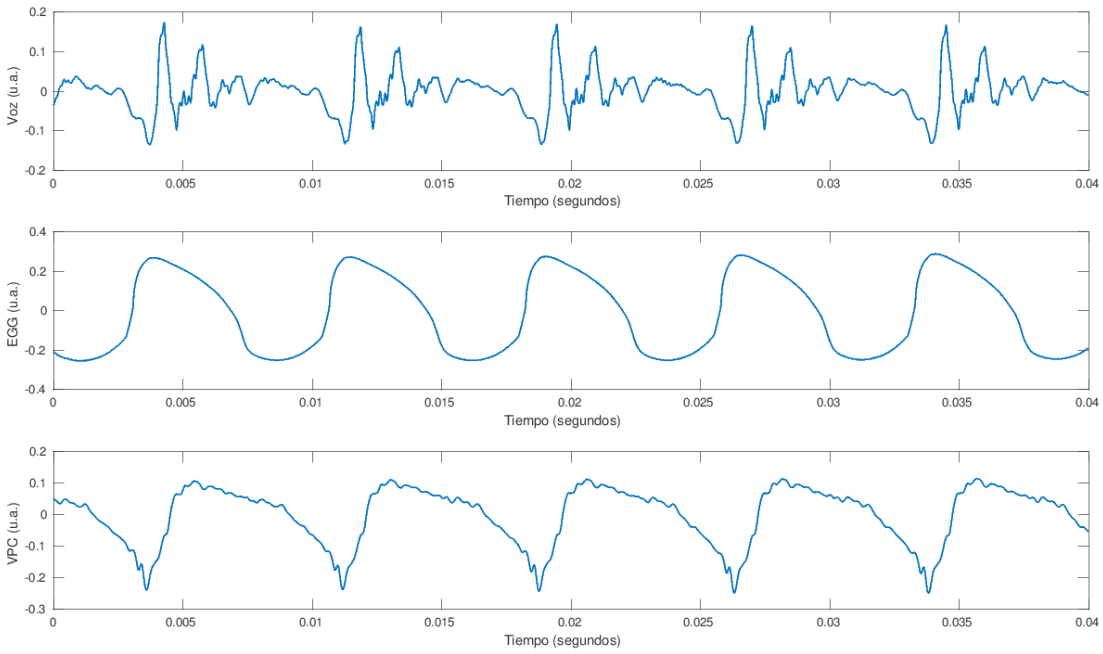

Figura 1: señales de Voz, EGG y VPC en función del tiempo

Para estimar la F0 de las distintas señales recién expuestas, se implementaron diferentes algoritmos de corte clásico que se pueden englobar en tres grupos: función de autocorrelación (AC), periodograma (PG) y de detección de eventos (DE).

Empezaremos por hacer una distinción básica entre los mismos, para luego abocarnos a la descripción individual de cada uno y sus particularidades para cada señal.

Los dos primeros (AC y PG) trabajan dividiendo la señal en segmentos o *frames* de una longitud fija, para cada uno de los cuales se obtiene un único valor promedio de F0. Por ello, pueden englobarse bajo el nombre de Métodos de Análisis de Corto Plazo. Con esto, la elección de la longitud de los segmentos se convierte en un punto clave del algoritmo: debe ser lo suficientemente pequeña para que la F0 pueda ser considerada constante dentro de sus límites, y debe ser lo suficientemente grande para permitir la medición de la F0 (Hess, 2008). Para la mayoría de los algoritmos un segmento de longitud aceptable es aquel que contienen dos o tres períodos fundamentales completos (Hess, 2008). Luego, en este trabajo consideramos una longitud de 10 milisegundos. Recomendamos la discusión planteada en (Rabiner, 1977) respecto de la importancia de adaptar la longitud del segmento según el hablante, con la intención de poder aplicar un criterio semejante en trabajos futuros.

Por otro lado, los algoritmos del tercer grupo (DE) trabajan directamente sobre el dominio temporal de la señal, de manera que periodo a periodo buscan detectar una característica o evento. Este evento se utiliza como marcador para generar una secuencia de límites de periodos (Hess, 2008). Luego, la F0 se calcula como la inversa de la diferencia temporal entre dos marcadores consecutivos. Evidentemente, el evento a detectar va a depender de la señal con la que se esté trabajando, por lo que esta clase de algoritmos implican un pre-procesamiento diferente para cada una de las señales para así modificar su estructura temporal y resaltar una característica en particular.

Para mayor claridad, cada método será explicado en una subsección. El proceso de desarrollo de los algoritmos ha sido el punto clave del desarrollo de este trabajo y amerita su atención.

Función de autocorrelación (AC)

La AC de un *frame* $s_f(n)$ de señal se define como:

$$\emptyset(\tau) = \frac{1}{N} \sum_{m=0}^{N-1-\tau} s_f(n+m)\, s_f(n+m+\tau),$$

donde *N* es la longitud del *frame* y τ recibe el nombre de retardo o *lag*.

Si $s(n)$ es una señal periódica, la AC exhibe un pico en el valor correspondiente al periodo fundamental promedio (Hess, 2008). El algoritmo de detección de F0

basado en la AC que desarrollamos tiene como objetivo principal hallar ese pico. Los pasos para realizar esta tarea son:

1. Dividir la señal en segmentos de longitud fija de 10 milisegundos.
2. Multiplicar cada segmento por una ventana rectangular.
3. Calcular la AC sin sesgo para cada segmento.
4. Detectar el máximo de la función de AC entre los valores de retardo correspondientes a 75 Hz y 600 Hz.

Este método tiende a fallar cuando es implementado sobre la señal de voz debido a la posible presencia de una primera formante de frecuencia cercana a la fundamental (Schlotthauer, 2010). Para evitar este inconveniente se opta por incluir un pre-procesamiento no lineal sobre las señales de voz denominado *center clipping* (Schlotthauer, 2010)(Rabiner, 1977).

Periodograma (PG)

Este algoritmo hace uso de la transformada discreta de Fourier, la cual nos permite observar la señal en un dominio frecuencial. Para un segmento $s_f(n)$ de señal, con transformada discreta de Fourier $S_f(k)$, el periodograma se define como:

$$P(k) = \frac{1}{N} \left| S_f(k) \right|^2,$$

donde *N* es la longitud del segmento.

Sobre el PG sólo queda determinar el pico correspondiente a la frecuencia fundamental. Nuevamente enlistamos los pasos realizados en este algoritmo:

1. Dividir la señal en segmentos de longitud fija de 10 milisegundos.
2. Multiplicar cada segmento por una ventana de Hamming, la cual fue seleccionada tras la prueba sucesiva de las funciones ventana más comunes, enlistadas y analizadas en (Schlotthauer, 2010).
3. Calcular el PG para cada segmento, implementando la transformada discreta de Fourier a través de la transformada rápida de Fourier.
4. Detectar el primer pico del PG entre los valores de frecuencia correspondientes a 75 Hz y 600 Hz.

Detección de eventos (DE)

Como ya se mencionó, este grupo de algoritmos trabaja con un pre-procesamiento diferente para cada señal.

Para la señal de voz se aplicó un filtrado inverso sencillo, utilizando coeficientes de predicción lineal (LPC) para obtener los parámetros del tracto vocal y siguiendo las recomendaciones realizadas en (Alzamendi, 2016) para la aplicación de filtros de pre-énfasis y radiación de los labios. Este filtrado nos permitió obtener una estimación de la función glótica, la cual presenta picos negativos a una frecuencia igual a F0,

como puede observarse en la Figura 2. Luego, estos picos se convierten en los eventos a detectar.

Para la señal de EGG se trabajó con su derivada (DEGG), el cual se muestra en la Figura 3. El DEGG presenta un pico prominente en los instantes de cierre de las cuerdas vocales, evento que ocurre una vez por período para vocales sostenidas y que por tanto permite marcar los límites de un periodo fundamental. Por ser el EGG la forma de sensado más apropiada para evaluar la actividad glótica, se consideró la F0 estimada a partir de esta señal y con este método como el "patrón verdad" sobre el cual comparar los distintos grupos método - señal. Por esta misma razón, los marcadores aquí obtenidos fueron corroborados uno por uno de forma visual.

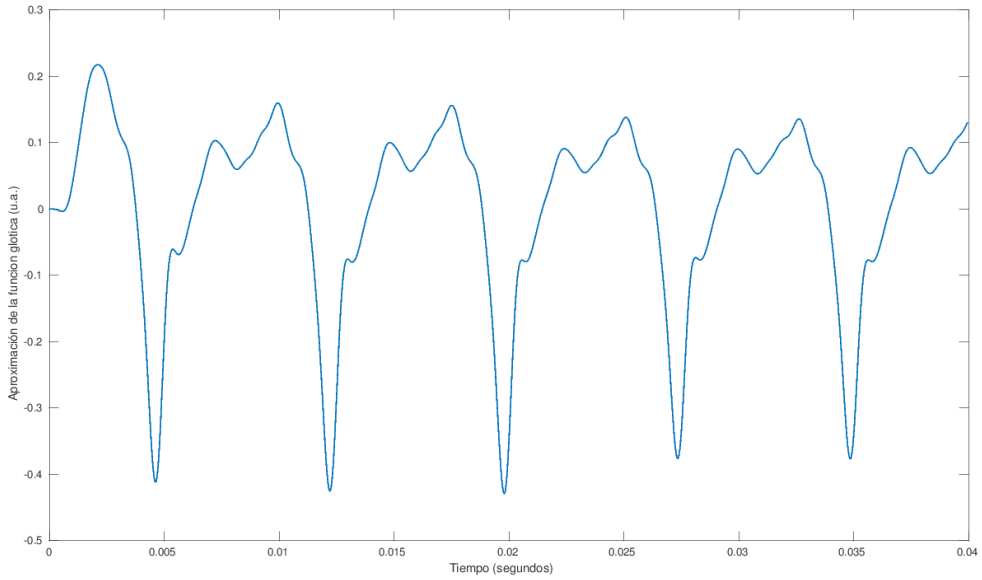

Figura 2: aproximación a la función glótica por medio de un filtrado inverso.

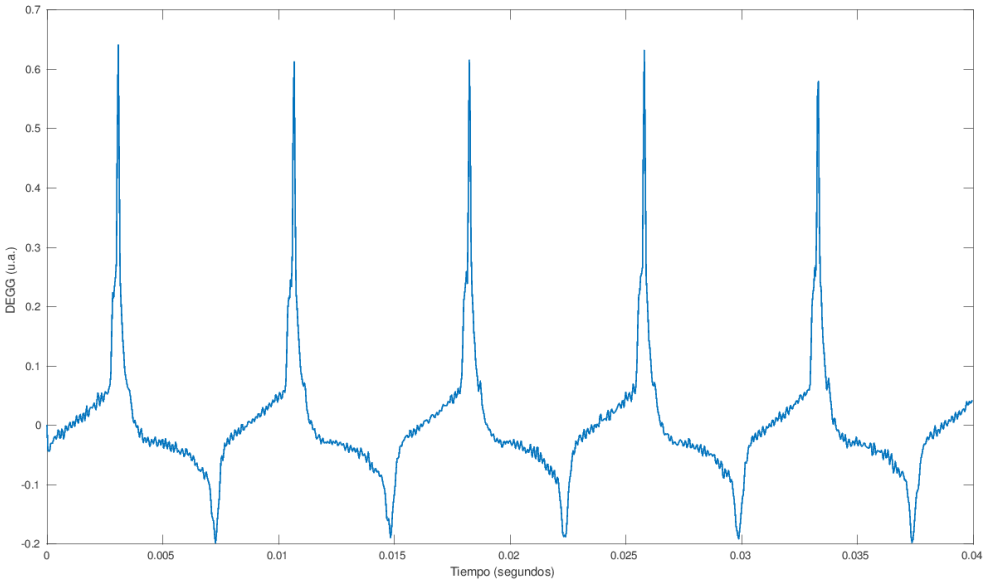

Figura 3: derivada de la señal de EGG (DEGG).

En el caso de la señal de VPC, no encontramos antecedentes que realizaran una modificación sobre la estructura temporal similar a las vistas para voz y EGG. Por lo tanto, simplemente se optó por un filtrado pasa - bandas con frecuencia de corte inferior en 50 Hz y frecuencia de corte superior en 800 Hz, el cual se implementó mediante dos filtros sucesivos – uno pasa altos y otro pasa bajos – tipo FIR. Con esto se obtiene una señal que suele presentar únicamente dos picos por periodo, los cuales serán los eventos a detectar por el método.

Resultados y Discusión

Para analizar el desempeño de los algoritmos implementados se utilizó como medida de error la raíz del error cuadrático medio (RECM), utilizada en (Tsanas et al., 2014) y (Christensen & Jakobsson, 2009) y calculada como:

$$RECM = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$$

Donde $\hat{y}_i$ es la estimación del patrón verdad, y_i la estimación del método a evaluar y *N* es el número de estimaciones para una señal.

Para realizar este cálculo, es evidente que y y $\hat{y}$ deben tener la misma cantidad de elementos. O, en otras palabras, que todos los métodos deben generar el mismo número de estimaciones de F0. Esto no ocurre en una primera instancia por las ya explicadas diferencias entre los algoritmos que trabajan por secciones y aquellos que trabajan por detección de eventos. Para afrontar este problema, optamos por realizar una interpolación mediante *splines* cúbicas utilizando la base de tiempo de $\hat{y}$.

Método – señal	RECM promedio [Hz]
AC – EGG	1,4523
AC – Voz	1,1827
AC – VPC	1,5076
PG – EGG	1,5704
PG – Voz	1,7546
PG – VPC	2,5243
DE – Voz	3,2390
DE – VPC	7,8464

Tabla 1: valores medios del RECM para cada par método – señal.

En la Tabla 1 se pueden observar los valores promedio del RECM para cada uno de los grupos método – señal. Observando solo este indicador, podemos decir que todos los grupos método – señal muestran un desempeño aceptable para una estimación de la F0. También se puede notar un claro desvío de la media para el grupo “DE – VPC” respecto de los demás. Este desvío queda aún más evidenciado con la realización de un test de comparaciones múltiples, el cual nos arroja la Figura 4, que muestra con distintos colores aquellos grupos que tienen medias significativamente distintas. También mediante este test podemos observar que los demás grupos no son estadísticamente separables a través de nuestra medida de error.

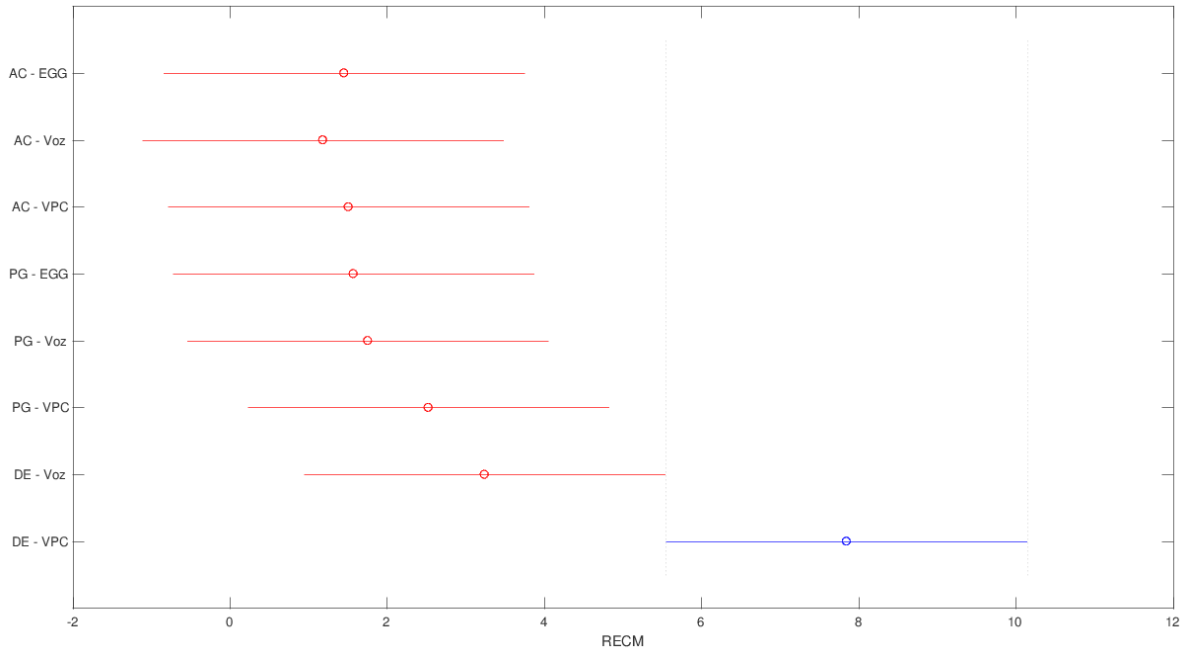

Figura 4: test de comparaciones múltiples.

Para analizar mejor los resultados obtenidos es útil plasmar los datos en un gráfico de cajas como el de la Figura 5. Es importante aclarar aquí que el gráfico fue limitado en el eje de las ordenadas para poder tener una mejor visión de los datos. Esto fue necesario debido a la presencia de valores atípicos u *outliers* muy altos que presentan algunos de los grupos, a pesar de que el 50% de los valores – es decir, aquellos contenidos dentro de las cajas – se encuentren dentro de rangos similares. Esto puede ser un indicador de inconvenientes en la robustez de los métodos implementados en relación con las distintas señales. Pueden encontrarse testimonios en la bibliografía (Rabiner, 1977) de que los detectores de F0 basados en la AC destacan por su robustez.

Es interesante destacar que el grupo "DE - VPC" muestra la menor mediana en simultáneo con el mayor valor promedio para nuestra medida de error. Información que podría hablarnos de una muy buena exactitud en el método para aquellas situaciones en las que no comete grandes equivocaciones.

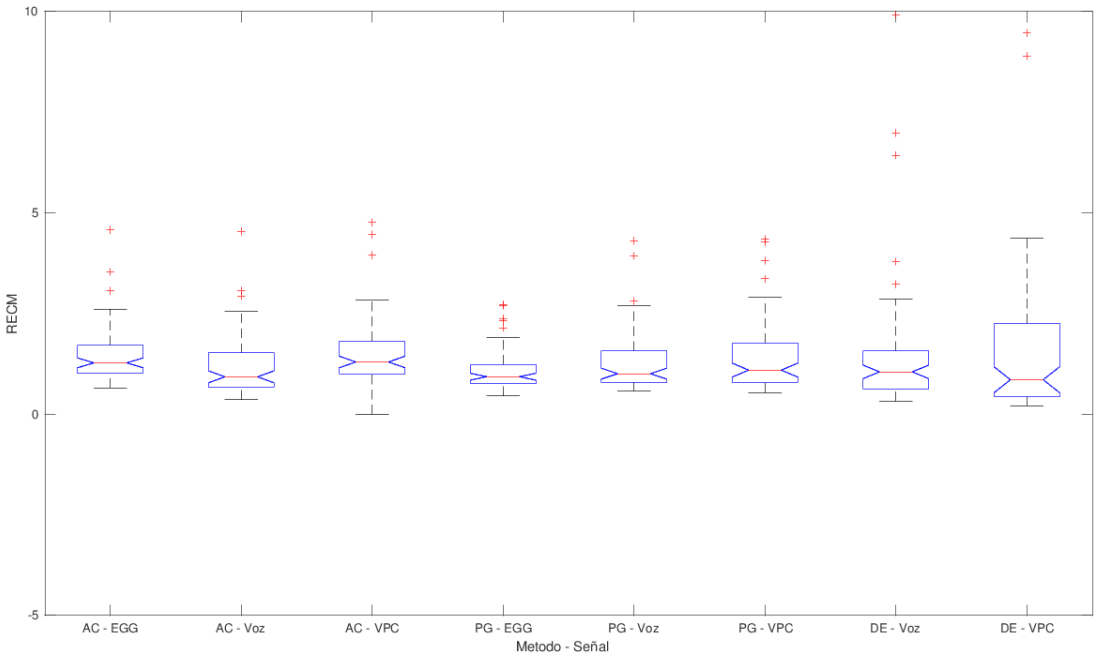

Imagen 5: gráfico de cajas.

Conclusiones

Como respuesta al objetivo principal de este trabajo nos encontramos satisfechos con los resultados obtenidos, dado que para dos de los tres métodos implementados no hay una diferencia significativa entre su aplicación a las distintas señales. Concluimos en función de esto que resulta factible el uso de la señal de VPC para extraer información confiable sobre la frecuencia fundamental, pudiendo ser utilizada en un sistema portátil para monitoreo de la actividad vocal.

Consideramos que es importante seguir trabajando en mejorar los métodos implementados en trabajos futuros, en especial el de detección de eventos sobre la señal de VPC, con vistas a mejorar la robustez de este método que muestra indicios de una buena exactitud.

Por sobre todas las cosas, y dada la calidad de alumna de grado de la autora, valoramos carácter formativo de la realización de este trabajo.

Bibliografía

Alzamendi, G. A. (2016). *Modelado estocástico de la fonación y señales biomédicas relacionadas*.

Askenfelt, A., Gauffin, J. A. N., & Sundberg, J. (1974). A C O M P A R I S O N OF C O N T A C T M I C R O P H O N E AND ELECTROGLOTTOGRAPH FOR THE M E A S U R E M E N T OF V O C A L F U N D A M E N T A L FREQUENCY, 258–273.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical Measurement of Speech and Voice* (2nd ed.).

Christensen, M. G., & Jakobsson, A. (2009). Multi-Pitch Estimation,

(February).

Hess, W. J. (2008). Pitch and Voicing Determination of Speech with an Extension Toward Music Signals. In *Springer Handbook of Speech Processing* (pp. 181–212). https://doi.org/10.1007/978-3-540-49127-9_10

Neumann, K., Gall, V., Schutte, H. K., & Miller, D. G. (2003). A new method to record subglottal pressure waves: Potential applications. *Journal of Voice*, *17*(2), 140–159. https://doi.org/10.1016/S0892-1997(03)00037-7

Rabiner, L. R. (1977). On the Use of Autocorrelation Analysis for Pitch Detection. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *25*(1), 24–33. https://doi.org/10.1109/TASSP.1977.1162905

Rufiner, H. L. (2009). *Análisis y modelado digital de la voz* (1a ed.). Santa Fe, Santa Fe, Argentina.

Schlotthauer, G. (2010). *Análisis de señales con descomposición empirica en modos y aplicaciones a la señal de voz*.

Tsanas, A., Zañartu, M., Little, M. A., Fox, C., Ramig, L. O., & Clifford, G. D. (2014). Robust fundamental frequency estimation in sustained vowels: Detailed algorithmic comparisons and information fusion with adaptive Kalman filtering. *The Journal of the Acoustical Society of America*, *135*(5), 2885–2901. https://doi.org/10.1121/1.4870484

Financiamiento

Este trabajo se realizó en el contexto del PID 6171 “Procesamiento, análisis y modelado de señales biomédicas: un enfoque integrador”, dirigido por Gastón Schlotthauer.