Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos

Gabriel A Alzamendi[†,§], Gastón Schlotthauer[†,§], Hugo L Rufiner[‡,§] y María E Torres[†,‡,§,*]

[†] Laboratorio de Señales y Dinámicas no Lineales; Facultad de Ingeniería; Universidad Nacional de Entre Ríos
[‡] Centro de I+D en Señales, Sistemas e Inteligencia Computacional; Facultad de Ingeniería y Ciencias Hídricas; Universidad Nacional del Litoral
[§] Consejo Nacional de Investigaciones Científicas y Técnicas

E-mail: [*] metorres@santafe-conicet.gov.ar

Resumen. La señal de voz normal presenta irregularidades intrínsecas necesarias para que se perciba "natural". Cuando existen patologías estas irregularidades aumentan volviéndose más evidentes, incluso para un oído no entrenado. Los parámetros acústicos que las cuantifican son útiles en la práctica médica para caracterizar la voz y detectar patologías. Aquí se propone un modelo para la síntesis de voz irregular que permite fijar dos parámetros acústicos, habitualmente empleados en la práctica médica, relacionados con las perturbaciones instantáneas en la amplitud y el periodo fundamental: Shimmer y Jitter. Se genera la señal glótica artificial a partir de un tren de pulsos equi-espaciados, modificando la amplitud y periodo de cada pulso y aplicando a la señal resultante un filtro lineal autorregresivo equivalente al del tracto vocal, obteniendo así una señal de voz sintética. Se desarrollaron modelos para la perturbación de la amplitud y del periodo a partir de métodos estadísticos sencillos. Mediante algoritmos de predicción lineal se generó el filtro del tracto vocal usando voces reales. Se generó un conjunto de señales y se analizó el desempeño del modelo. Las señales sintetizadas resultaron morfológicamente similares a las voces reales, aunque aún distan de percibirse como naturales. Los valores de las medidas de Shimmer y Jitter obtenidos coincidieron mayoritariamente con los valores teóricos. Sin embargo, se observó que el Jitter se aleja del comportamiento ideal para valores pequeños debido a la frecuencia fundamental y a la naturaleza temporal discreta de las señales sintetizadas. Los resultados sugieren que el modelo desarrollado es útil para generar voces artificiales, tanto sanas como patológicas, para un amplio rango de valores de los indicadores de Shimmer y Jitter.

1. Introducción

A lo largo del tiempo, el estudio y modelado de los mecanismos intervinientes en la generación de la voz ha sido un campo en constante investigación que ha abarcado diversas áreas de las ciencias y puntos de vistas inter-disciplinarios, debido a la gran complejidad y diversidad de elementos que participan. Los ejes principales en los que se centra son el análisis de las estructuras anatómicas y los fenómenos involucrados en el proceso del habla, considerando su comportamiento dinámico y relaciones estructurales [1, 2]. Los avances alcanzados han permitido el desarrollo de nuevas técnicas y métodos empleados en campos muy diversos. En [3] se detallaron métodos para el reconocimiento de hablantes, considerando además sus posibles aplicaciones. En [4, 5] se analizaron estrategias para mejorar la calidad de las voces artificiales

y su uso como interfaz hombre-máquina. Se han presentado diversas técnicas desarrolladas recientemente destinadas al modelado, acondicionamiento, síntesis, compresión y transmisión de señales de voz [1, 2, 6, 7].

Existe en la bibliografía una gran variedad de modelos ideados con el fin de analizar e imitar el proceso de generación del habla. Estos difieren entre sí en cuanto a las estrategias y métodos seleccionados en el desarrollo del modelo, así como en las aplicaciones consideradas. En [8, 9] se emplearon sistemas mecánicos con elementos lineales y no lineales para modelar el movimiento oscilatorio de las cuerdas vocales. En [10] se estudió el uso de conductos semi-rígidos de área variable como analogía del aparato fonador. En [11] se analizó el comportamiento del flujo del aire desarrollado en el tracto vocal desde el punto de vista de la mecánica del continuo y técnicas de elementos finitos. En [12] se simuló la generación de voz aplicando la teoría de señales y sistemas lineales discretos. Además, es común encontrar, en determinadas aplicaciones, modelos híbridos desarrollados asociando diferentes métodos, conseguiéndose así modelos más versátiles aunque más complejos.

Recientemente, ha surgido en la medicina un especial interés en la modelización del habla, gracias a que su empleo en el estudio y síntesis de voces patológicas ha permitido desarrollar un mayor entendimiento sobre las etiologías y alteraciones presentes en los diferentes trastornos [13, 14]. Sin embargo, en contraposición a esta necesidad del ámbito de la salud, el desarrollo de un modelo que comprenda el mayor número de patologías posibles es una meta difícil de alcanzar, debido a la falta de conocimiento o consenso acerca de todos los aspectos involucrados en las diferentes alteraciones del aparato fonador. Más aún, se ha demostrado que incluso en voces sanas se presentan irregularidades y que éstas son las responsables del grado de naturalidad con que se perciben [14, 15].

En la práctica médica es habitual el empleo de parámetros acústicos que, en conjunto con el análisis perceptual y los estudios específicos, permiten al especialista caracterizar la voz de un individuo y determinar la presencia de patologías. Entre otros, mencionamos la frecuencia fundamental, el rango de frecuencia fonatoria, varias medidas de la perturbación en la frecuencia y la amplitud, índices de turbulencias y ruido presentes en la voz [14]. El *Shimmer* y el *Jitter* son los parámetros más empleados para cuantificar las alteraciones instantáneas en la amplitud y la frecuencia, respectivamente. Se ha encontrado que son muy útiles para caracterizar los diferentes tipos de voz y que son especialmente sensibles a los diversos trastornos [15, 16, 17].

El presente trabajo tiene como finalidad proponer y desarrollar un modelo sencillo para la síntesis de voz basado en parámetros acústicos de interés en la práctica médica. En particular, se centrará la atención en las medidas de *Shimmer* y de *Jitter*, considerando tanto voces sanas como patológicas. La estructura de este artículo es la siguiente: en la sección 2 se desarrolla el modelo propuesto, se explica la metodología de trabajo y se detallan los materiales necesarios; en la sección 3 se muestran y analizan los resultados alcanzados y, por último, en la sección 4 se presentan las conclusiones obtenidas y trabajos futuros en esta línea.

2. Materiales y métodos

En este trabajo proponemos un método para la síntesis de voz basado en el modelo del aparato fonador denominado *fuente-filtro*. Este enfoque presenta varias ventajas: posee un marco teórico sencillo, existe abundante material bibliográfico al respecto y ha demostrado ser útil en una gran variedad de aplicaciones [1, 6]. El modelo se inspira en la fisiología del aparato fonador y el proceso mediante el cual se genera el habla. En este proceso, el flujo de aire proveniente de los pulmones es modificado por la acción de las cuerdas vocales generando pulsos regulares, denominados pulsos glóticos (PG). Estos son transmitidos acústicamente a lo largo del tracto vocal (TV), el cual está formado por conductos semirígidos de área variable, y en su trayecto son modificados, dando como resultado la señal de voz propiamente dicha (para mayor información referirse a [2, capítulo 2]). Analizando este proceso desde el punto de vista de la teoría de las

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos"
Memorias del XVIII Congreso Argentino de Bioingeniería (SABI 2011), set, 2011.

señales discretas, se puede decir que una señal correspondiente a los PG, denominada fuente glótica (FG), es modificada o filtrada por el TV dando como resultado la señal de la voz. A continuación, se estudiarán cada uno de los componentes del modelo.

2.1. Fuente glótica

La morfología de la FG considerada en el modelo depende del tipo de voz que se desee analizar o generar. En particular, en esta aplicación se considerará únicamente la síntesis de vocales sostenidas, las cuales presentan una morfología regular y un comportamiento semi-periódico para el caso de voces sanas. Este tipo de emisión es el más utilizado en los estudios acústicos. Considerando estas propiedades, proponemos aquí generar la FG a partir de un tren de pulsos con amplitud y periodo variables representada por:

$$u[n] = \sum_{i=1}^{I} A_i \; \delta \left[n - \sum_{j=1}^{i} P_j \right], \tag{1}$$

donde A_i y P_j son la amplitud y periodo de cada uno de los pulsos [1, 6]. El valor $\frac{1}{P_j}$ determina la frecuencia instantánea (F_0) del pulso. La principales ventajas de (1) son que permite: i) lograr la regularidad y periodicidad necesaria para la aplicación y ii) modificar los valores de A_i y P_j a voluntad, logrando así introducir alteraciones controladas en la señal de voz.

En diversos trabajos se han propuesto alternativas para simular los fenómenos de *Jitter* y de *Shimmer*. Por ejemplo, en [18, 19] se le adicionó a los términos A_i y P_j ruido aleatorio de varianza determinada, obteniéndose así la FG perturbada. En [20] se consideró que el periodo fundamental es alterado por rudio con memoria, donde el ruido adicionado depende de los valores pasados de P_j. Sin embargo, estos modelos no permiten preestablecer los valores de los índices de perturbación. Aquí proponemos generar voces artificiales, a partir de modelos estadísticos, donde sea posible fijar los valores de *Jitter* y *Shimmer* de la señal. Para ello es necesario obtener una relación entre estas medidas y los parámetros de la FG.

Se define como *Shimmer*[15] a las alteraciones instantáneas presentes en las amplitudes de la señal de voz, considerando dos pulsos sucesivos. La medida más empleada es la *razón de Shimmer porcentual* ($Shimmer_{\%}$) y está definida a partir de la siguiente ecuación [15]:

$$Shimmer_{\%} = 100 \frac{\frac{1}{N-1} \sum_{i=1}^{N-1} |A_{i+1} - A_i|}{\frac{1}{N} \sum_{i=1}^{N} A_i}, \tag{2}$$

donde A_i es la amplitud para el pulso i-ésimo y N es la cantidad de pulsos presentes en la señal.

Recibe el nombre de *Jitter* la fluctuación o perturbación que presentan dos periodos contiguos en la señal de voz [15]. De las diversas medidas para cuantificarlo existentes, la más utilizada es la denominada *razón de Jitter porcentual* ($Jitter_{\%}$), dada por [15]:

$$Jitter_{\%} = 100 \frac{\frac{1}{N-1} \sum_{j=1}^{N-1} |P_{j+1} - P_j|}{\frac{1}{N} \sum_{j=1}^{N} P_j}, \tag{3}$$

donde P_j es el periodo para el pulso j-ésimo y N es la cantidad de periodos presentes en la señal.

Suponemos aquí que la variación en las amplitudes y en los periodos de los pulsos de la FG son independientes entre sí, lo que permite hacer uso de la ecuación (1). Además, suponemos que las series A_i y P_j presentan comportamiento Gaussiano siendo sus distribuciones $\mathcal{N}(A_0, \sigma_A)$ y $\mathcal{N}(P_0, \sigma_P)$ respectivamente, donde los términos A_0 y P_0 corresponden a los valores medios y los σ_A y σ_P corresponden a los desvíos estándares respectivos. Esta hipótesis ha sido empleada

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos"
Memorias del XVIII Congreso Argentino de Bioingeniería (SABI 2011), set, 2011.

anteriormente con resultados muy satisfactorios, tanto en el análisis de la dinámica de la señal de voz [21] como en la clasificación entre voces sanas y patológicas [13].

Trabajando a partir de las distribuciones, se generan las series $\Delta A_i = A_{i+1} - A_i$ y $\Delta P_j = P_{j+1} - P_j$, observándose que las mismas poseen distribuciones dadas por $\mathcal{N}(0, \sqrt{2}\sigma_A)$ y $\mathcal{N}(0, \sqrt{2}\sigma_P)$ respectivamente. Se tiene así que la serie temporal de los valores absolutos $|\Delta A_i| = |A_{i+1} - A_i|$ posee un comportamiento hemi-Gaussiano y distribución de la forma:

$$\begin{cases} \mathcal{N}(0, \sqrt{2}\sigma_A), & \text{si } |\Delta A_i| = 0; \\ 2\,\mathcal{N}(0, \sqrt{2}\sigma_A), & \text{si } |\Delta A_i| > 0; \\ 0, & \text{cualquier otro caso.} \end{cases} \tag{4}$$

Se puede demostrar que el valor esperado de $|\Delta A_i|$ se encuentra determinado por:

$$E\{|\Delta A_i|\} = \int_0^\infty \frac{2|\Delta A_i|}{(4\pi\sigma_A^2)^{1/2}} e^{\left(\frac{-|\Delta A_i|^2}{4\sigma_A^2}\right)} = \frac{2\sigma_A}{\sqrt{\pi}}. \tag{5}$$

Por teoría estadística, sabemos que para el caso $N \to \infty$ se cumple que $\frac{1}{N-1}\sum_{i=1}^{N-1} |A_{i+1} - A_i|$ converge a $E\{|\Delta A_i|\}$ y $\frac{1}{N}\sum_{i=1}^{N} A_i$ converge a A_0. Finalmente, reemplazando (5) en la ecuación (2) se obtiene:

$$\sigma_A = \frac{\sqrt{\pi} A_0 Shimmer_{\%}}{200}. \tag{6}$$

De manera similar, se demuestra en el caso del periodo que:

$$\sigma_P = \frac{\sqrt{\pi} P_0 Jitter_{\%}}{200}. \tag{7}$$

De (6) y (7), se deduce que para sintetizar vocales con periodo fundamental P_0 y amplitud media A_0, con valores de $Shimmer_{\%}$ y $Jitter_{\%}$ establecidos a voluntad, se requieren valores de A_i y P_j a partir de ruido Gaussiano aleatorio con medias A_0 y P_0 y desvíos estándares σ_A y σ_P, respectivamente.

2.2. Tracto vocal

Las propiedades de filtrado del TV se pueden representar a partir de un modelo lineal autorregresivo donde la señal de voz en un instante dado depende de sus valores pasados y del valor de la FG en ese instante [1, 2, 6]. Se lo representa mediante una ecuación en diferencias de la forma:

$$s[n] = -\sum_{k=1}^{K} a_k\, s[n-k] + G\, u[n], \tag{8}$$

donde $s[n]$ es la señal de la voz, $u[n]$ es la FG y los a_k son los *coeficientes de predicción lineal* (LPC). Los modelos autorregresivos poseen como ventajas: *i*) su simplicidad, *ii*) su fácil implementación y *iii*) la existencia de métodos rápidos y eficientes [1, 6, 12]. Los algoritmos LPC se basan en hallar los parámetros que minimicen el error cuadrático entre la voz real y la sintetizada. Como resultado, se obtienen los coeficientes a_k y G de manera tal que el filtro obtenido es estable e invertible y la señal de voz sintetizada se considera una buena estimación de la señal real.

Aplicando la transformada-$\mathcal{Z}$ a ambos lados de la ecuación (8) se puede analizar el comportamiento del sistema en el dominio frecuencial, obteniéndose la función de transferencia del sistema:

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 + \sum_{k=1}^{K} a_k\, z^{-k}} = \frac{G}{A(z)}, \tag{9}$$

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos"
Memorias del XVIII Congreso Argentino de Bioingeniería (SABI 2011), set, 2011.

Tabla 1. Valores medio, máximo y mínimo de $Shimmer_{\%}$ y $Jitter_{\%}$ para individuos con voces sanas y patológicas, correspondientes a la base de datos analizada. Se observa que las voces patológicas poseen valores superiores en comparación con las de los individuos sanos.

Población	Parámetro Acústico	Media (DE*)	Valor Máximo	Valor Mínimo
Voces Sanas	$Shimmer_{\%}$	2,205 (0,924)	4,802	0,963
	$Jitter_{\%}$	0,615 (0,437)	2,529	0,175
Voces Patológicas	$Shimmer_{\%}$	7,103 (5,027)	31,296	1,230
	$Jitter_{\%}$	2,539 (2,838)	21,322	0,212

* DE: Desvío Estándar.

donde G es un término constante, $S(z) = \mathcal{Z}\{s[n]\}$, $U(z) = \mathcal{Z}\{u[n]\}$ y $A(z) = 1 + \sum_{k=1}^{K} a_k\, z^{-k}$ respectivamente. Los polos de la función de transferencia toman valores reales o complejos conjugados y su localización determina la respuesta frecuencial del filtro TV.

Se denomina *error de predicción* o *residuo* al producto del filtrado de la señal de habla a partir del filtro TV inverso. Se considera que esta señal representa el comportamiento de la FG real y su aplicación en la síntesis de voz permite mejorar las características acústicas y perceptuales de la señal generada [1, 6].

2.3. Señales reales

Para el desarrollo de este trabajo se empleó la base de datos brindada por la empresa *Kay Elemetrics*, desarrollada por el *Massachusetts Eye and Ear Infirmary (MEEI) Voice and Speech Lab*. La misma consta de grabaciones de vocales */a/* sostenidas de 53 individuos con voces sanas y 654 que presentan voces alteradas debido a una variedad de patologías de origen orgánicos, neurológicos, psicogénicos o causadas por traumatismos. Estas señales fueron usadas para obtener los coeficientes LPC necesarios para modelar el TV, según la sección 2.2. Cada grabación cuenta con su correspondiente información clínica, reunida a partir de diferentes estudios y de la opinión de especialistas. En la tabla 1 se muestran los valores medio, máximo y mínimo de $Shimmer_{\%}$ y $Jitter_{\%}$ de la población analizada y se puede observar que, en comparación, las voces patológicas presentan valores más elevados y una mayor dispersión de los parámetros acústicos.

Las señales originales de la base de datos se grabaron en un ambiente controlado con frecuencias de muestreo (F_m) de 50 KHz o 25 KHz y con una resolución de 16 bits. A fines de homogeneizarlas, se las filtró aplicando un filtro pasa-bajo con frecuencias de corte de $9,5$ KHz y se las muestreó nuevamente a una F_m de 20 KHz. Para cada una de éstas, se aplicó un filtro de pre-énfasis, se obtuvieron las series de los coeficientes LPC de la señal, aplicando ventanas de *Hamming* de 20 [ms] de longitud y $98,75\,\%$ de superposición, y se calculó el residuo correspondiente. En esta aplicación se trabajó con 22 coeficientes LPC. En la figura 1 se pueden apreciar 25 [ms] de una vocal real (izquierda) y su correspondiente residuo (derecha) para el caso de una voz sana perteneciente a un individuo de sexo masculino.

2.4. Síntesis de voces

Se generó un tren de pulsos de amplitud unitaria y periodo fundamental P_0, con $P_0 = 1/F_0$. Se alteraron independientemente la amplitud y el periodo de cada pulso, teniendo en cuenta lo descrito en la sección 2.1, para obtener los valores de $Jitter_{\%}$ y $Shimmer_{\%}$ deseados. Para darle más naturalidad a la vocal sintetizada, se tomó un periodo del residuo y se generó la FG como resultado de la convolución entre el tren de pulsos perturbado y el residuo. Se generó la

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos"
Memorias del XVIII Congreso Argentino de Bioingeniería (SABI 2011), set, 2011.

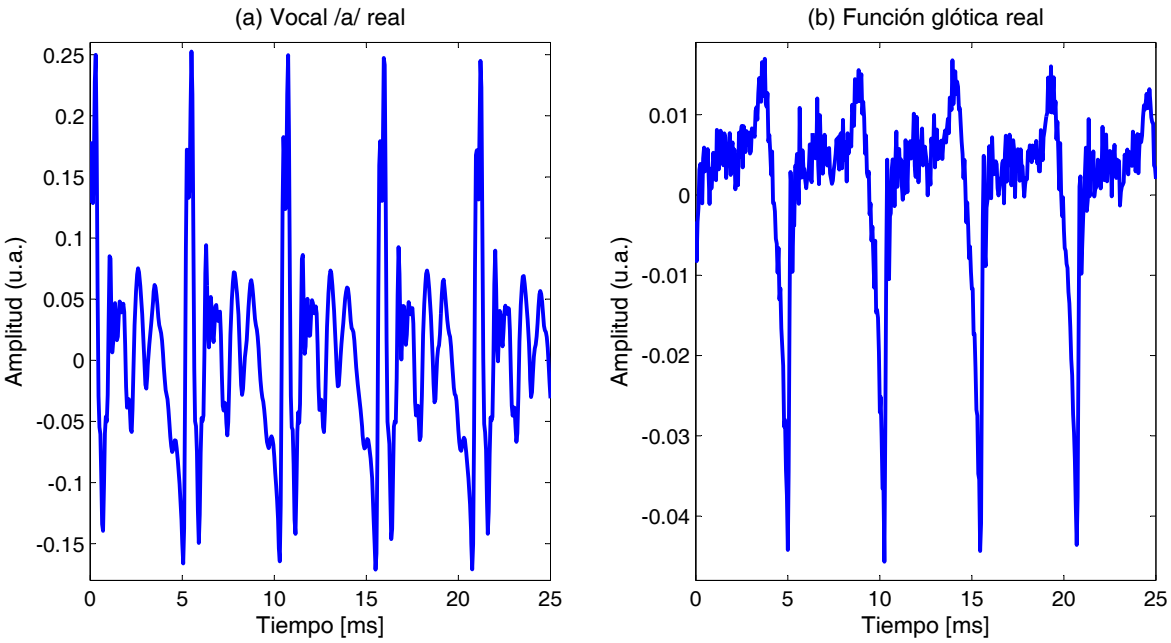

Figura 1. Señal de voz para una voz sana, correspondiente a un individuo de sexo masculino ($F_0 = 189,295[Hz]$, $Jitter_{\%} = 0,269\,\%$ y $Shimmer_{\%} = 1,826\,\%$). a) Vocal */a/* sostenida, b) Residuo correspondiente.

señal de voz conforme a lo explicado en las secciones 2.2 y 2.3. En la figura 2 se muestran 25 [ms] de una vocal sintetizada (izquierda) aplicando el procedimiento y con una F_m de 50 [KHz], junto con su correspondiente función glótica (derecha). A fines comparativos, la señal se sintetizó considerando la información clínica y la voz real del individuo analizado en la figura 1.

Se sintetizó un conjunto de señales tomando diferentes valores de $Jitter_{\%}$ y $Shimmer_{\%}$. Los rangos entre los que se trabajó son $0,00 \leq Jitter_{\%} \leq 3,00$ y $0,00 \leq Shimmer_{\%} \leq 5,00$, siendo el paso en cada caso de 0,05. Los extremos se tomaron en función de los valores mínimos y máximos de cada parámetro acústico correspondiente a voces sanas, tomados de la base de datos (ver tabla 1). Como se explicó anteriormente, se consideró que las perturbaciones son independientes entre sí, permitiendo de este modo modificar un parámetro acústico sin influir en el otro. Además esto posibilita generar y analizar el comportamiento de un conjunto de señales diferentes para un mismo valor en uno de los parámetros acústicos, simplemente variando el otro a voluntad.

3. Resultados

Comparando las figuras 1 y 2, se puede observar que las señales sintetizadas aplicando el modelo propuesto son visualmente comparables a las reales. En el caso de los residuos, la similitud se debe a la convolución del tren de pulsos con un periodo del residuo real y esa es la razón de que la señal sintetizada sea más regular que la real. Por el contrario, la morfología de la vocal sintetizada difiere considerablemente de la real, no así su regularidad. Esto último se debe a que el filtro TV únicamente simula aproximadamente el espectro de amplitud del tracto vocal, no garantizando la obtención de una replica de la señal original [1]. El aplicar el filtro de pre-énfasis y la convolución del tren de pulsos con el residuo mejora la calidad de la señal sintetizada. Sin embargo, ésta aún dista de percibirse auditivamente natural.

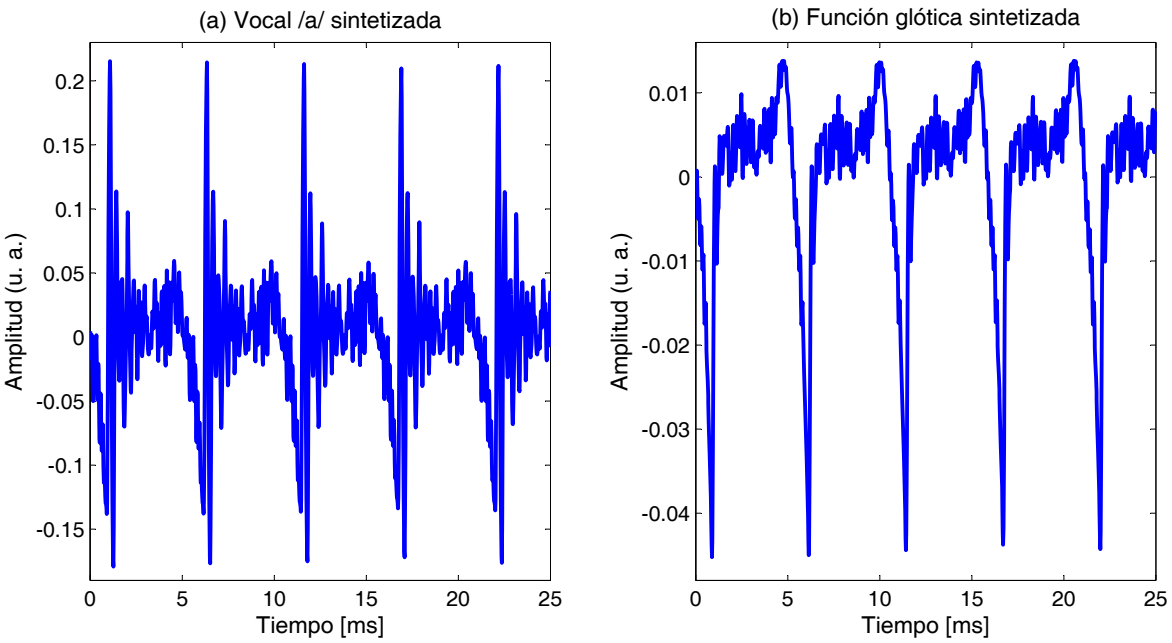

Figura 2. Señal de voz sintetizada considerando una voz sana, correspondiente a un individuo de sexo masculino ($F_0 = 189,295[Hz]$, $Jitter_{\%} = 0,269\,\%$ y $Shimmer_{\%} = 1,826\,\%$). a) Vocal */a/* sostenida, b) Residuo correspondiente.

Con el objeto de analizar el comportamiento del modelo, se estudió su desempeño considerando un conjunto de señales sintetizadas a partir de una F_m de 50 [KHz]. Para cada una de ellas, se calculó su $Jitter_{\%}$ y $Shimmer_{\%}$ aplicando las ecuaciones (2) y (3) respectivamente. Considerando cada parámetro acústico por separado, se tomaron aquellas señales correspondientes al mismo valor teórico del parámetro y se obtuvieron medidas estadísticas de ese conjunto. En la figura 3 se muestran en las ordenadas los valores de $Shimmer_{\%}$ (izquierda) y $Jitter_{\%}$ (derecha) obtenidos en función de sus valores teóricos correspondientes, en las abscisas. Se representa en línea continua azul el valor medio encontrado, en líneas continuas grises el desvío estándar de la familia de señales y en línea punteada roja el valor teórico. Los coeficientes de correlación encontrados para cada una de las curvas son de 0,999986 para el $Shimmer_{\%}$ y 0,999939 para el $Jitter_{\%}$ respectivamente. Se observa que en ambos casos la media acompaña bien a los valores teóricos sobre gran parte de los valores analizados. Además, se encontró que al aumentar la magnitud de las perturbaciones lo mismo ocurre con la dispersión de los valores reales de los parámetros. Esto último es de esperase por la naturaleza estadística del modelo y es útil para modelar las irregularidades encontradas en las señales de voz [15].

En el caso particular del $Jitter_{\%}$, se obtuvo que el modelo aquí propuesto se aleja ligeramente del comportamiento ideal para valores menores a 0,2. Esto se puede apreciar en la figura 3 y se debe principalmente a la naturaleza discreta de la señal sintetizada. Considerando la ecuación (3) para valores pequeños de $Jitter_{\%}$, se observa que la dificultad radica en el cálculo de $|P_{j+1} - P_j|$ debido a que la capacidad de reconocer como diferentes dos periodos sucesivos depende exclusivamente del periodo de muestreo empleado. Para valores del parámetro cada vez menores, se dificulta distinguir los pulsos diferentes lo que ocasiona que el valor obtenido se encuentre por debajo del valor teórico y, además, se llega a un punto para el cual todos los periodos parecen iguales por lo que el valor cae súbitamente a cero.

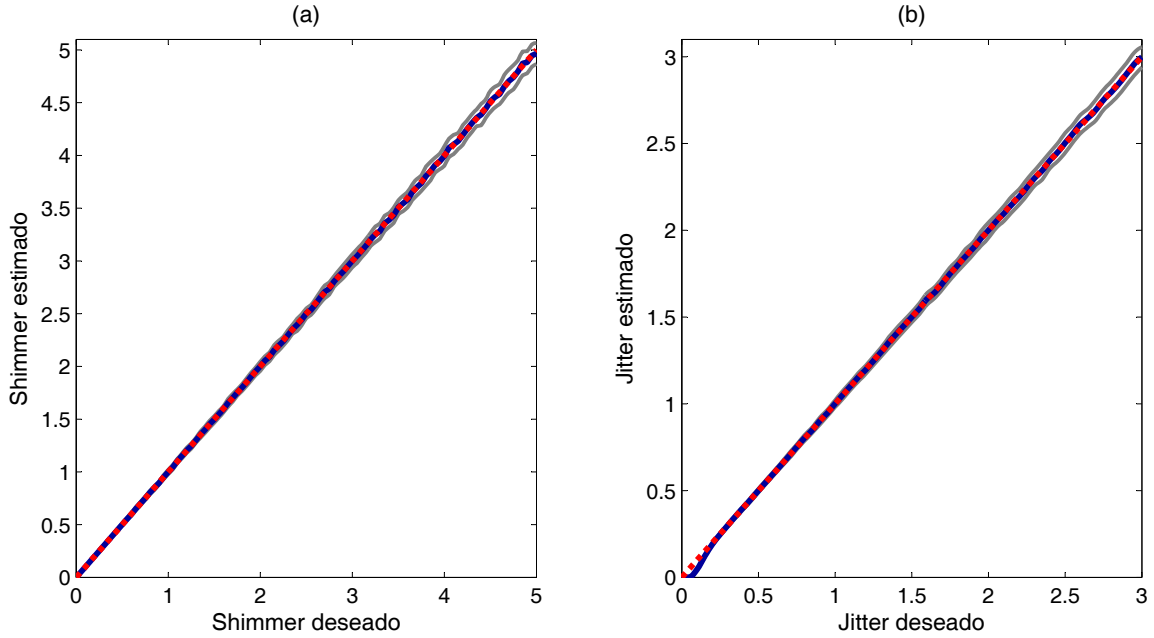

Figura 3. Parámetros acústicos calculados en función de los valores teóricos. En línea azul continua se representa el valor medio, en líneas grises continuas el desvío estándar y en línea punteada roja el valor teórico. a) $Shimmer_{\%}$, b) $Jitter_{\%}$.

Para corroborar esta hipótesis, se repitió el experimento variando la F_m del conjunto de señales. En la figura 4 se muestra el comportamiento del $Jitter_{\%}$ para voces artificiales con F_m de 35 (línea celeste), 50 (línea azul), 75 (línea verde) y 100[KHz] (línea negra) comparándolo con los valores teóricos (línea punteada roja). Se encontró que al aumentar la F_m se mejora el desempeño del modelo propuesto, aunque aumenta su costo computacional. Cabe destacar que el comportamiento para $F_m = 100$[KHz] cerca del origen está fuertemente determinado por la poca cantidad de puntos considerados, lo que ocasiona que su morfología sea muy similar a la curva teórica. Por lo tanto, si se aumentan los puntos analizados su comportamiento se alejará del ideal.

De la ecuación (3) se desprende que el $Jitter_{\%}$ depende también del P_0 de la voz analizada. Al aumentar la F_0, disminuye el P_0 y esto ocasiona que el valor obtenido se aleje del teórico para rangos de $Jitter_{\%}$ mayores. En la figura 5 se muestra este fenómeno para el caso de voces artificiales sanas correspondientes a un hombre con $F_0 = 189, 295$[Hz] (línea continua azul) y a una mujer on $F_0 = 230, 323$[Hz] (línea continua verde), comparándolos con los valores teóricos (línea punteada roja).

Si bien este comportamiento parecería ser un falla del modelo, al apreciar los valores de la tabla 1 se observa que el menor valor encontrado es 0,175. Sin embargo, al estudiar la base de datos se observa que las mayoría de los valores se encuentran por encima de este último, lo que muestra que el modelo se puede aplicar a la síntesis de voces sanas y patológicas (bajo la consideración de $Jitter_{\%}$ mayor a 0,2). De ser necesario sintetizar señales con $Jitter_{\%}$ menores a 0,2, se debe trabajar con una F_m acorde.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos"
Memorias del XVIII Congreso Argentino de Bioingeniería (SABI 2011), set, 2011.

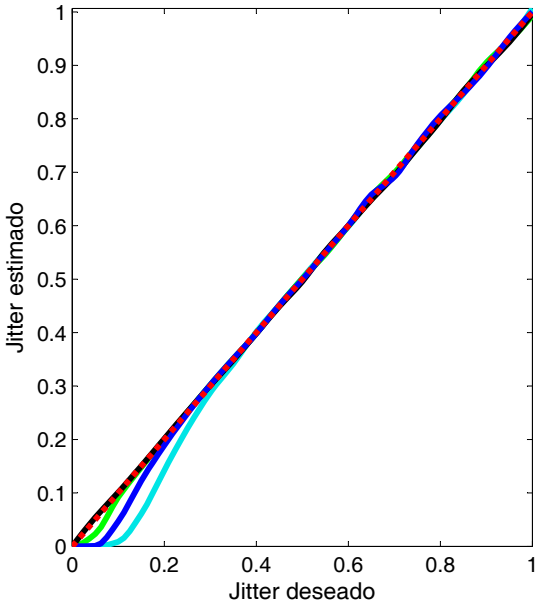

Figura 4. Valor de $Jitter_{\%}$ para señales con diferentes F_m. Se representa en celeste $F_m = 35$[KHz], en azul $F_m = 50$[KHz], en verde $F_m = 75$[KHz] y en negro $F_m = 100$[KHz].

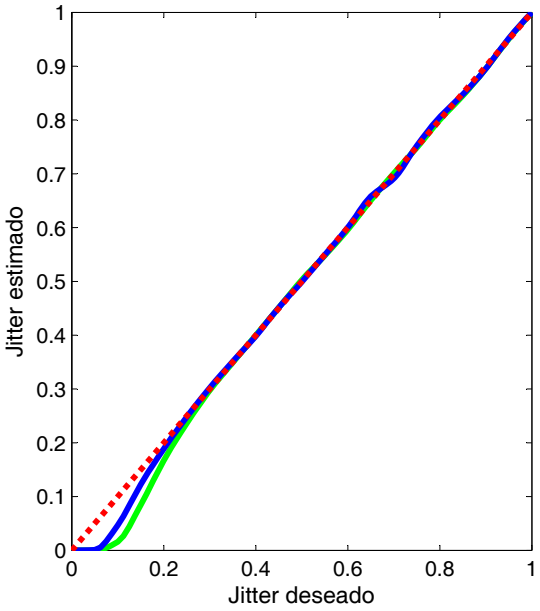

Figura 5. Valor de $Jitter_{\%}$ para señales con diferentes F_0. En azul se presenta la voz de un hombre con $F_0 = 189,295$[Hz] y en verde la voz de una mujer con $F_0 = 230,323$[Hz].

4. Conclusiones y trabajos futuros

En este artículo se propuso un modelo para la generación de voces artificiales con perturbaciones controladas, basado en la representación fuente-filtro del aparato fonador. Los parámetros acústicos considerados fueron el *Shimmer* y el *Jitter* y en función de ellos se desarrollaron sendas reglas para la modificación de la amplitud y el periodo de la FG, permitiendo obtener los valores deseados. Este modelo se aplicó a la síntesis de vocales sostenidas, generadas a partir de señales de voz reales, obteniéndose señales artificiales con comportamiento similar a las reales.

Se analizó el desempeño para un conjunto de señales sintetizadas y se mostró que los parámetros obtenidos se correspondieron con los valores teóricos, para casi la totalidad del rango considerado. También, se observó que al aumentar las perturbaciones lo mismo sucedió con la dispersión de las medidas calculadas, hecho que se corresponde con la marcada irregularidad de las voces reales que poseen valores elevados en sus parámetros acústicos. Además, se encontró que para valores muy pequeños de *Jitter* el desempeño del modelo se aleja ligeramente del comportamiento ideal, hecho ocasionado por la naturaleza discreta de las señales sintetizadas y la frecuencia fundamental considerada, lo que indica que se deben extremar cuidados al generar señales con estos valores.

Los resultados obtenidos sugieren que el modelo desarrollado es útil para generar voces artificiales, tanto sanas como patológicas, para un amplio rango de *Shimmer* y *Jitter*. Futuros trabajos de este grupo en esta línea incluirán la mejora en la calidad con que se perciben las señales sintetizadas, la incorporación de otros parámetros de interés clínico en el modelo, profundizar en la aplicabilidad a la síntesis de voces patológicas y el uso de técnicas de procesamiento avanzado de señales para el estudio de las señales generadas.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos"
Memorias del XVIII Congreso Argentino de Bioingeniería (SABI 2011), set, 2011.

Agradecimientos

Este trabajo fue realizado con el auspicio de la Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), el Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), y la Universidad Nacional de Entre Ríos (UNER). Los autores agradecen a la Dra. María C. Jackson Menaldi del Lakeshore Ear, Nose and Throat Center, St. Clair Shores (USA), y de Wayne State University, Detroit (USA) por sus valiosos comentarios.

Referencias

[1] Proakis J G, Hansen J H L and Deller J R 1993 *Discrete-Time Processing of Speech Signals* (New York: Macmillan Publishing Company)
[2] Rufiner H L 2009 *Análisis y modelado digital de la voz. Técnicas recientes y aplicaciones* 1st ed (Santa Fe: Ediciones UNL)
[3] Reynolds D A 2002 *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on* vol 4 pp IV4072–IV4075
[4] Rafieee M S and Khazaei A A 2010 *Computational Intelligence, Communication Systems and Networks, International Conference on* vol 0 (Los Alamitos, CA, USA: IEEE Computer Society) pp 250–255
[5] Xu N and Yang Z 2008 *Signal Processing, 2008. ICSP 2008. 9th International Conference on* pp 684–687
[6] Huang X, Acero A and Hon H 2001 *Spoken Language Processing: A Guide to Theory, Algorithm and System Development* (New Jersey: Prentice Hall PTR)
[7] McLoughlin I 2009 *Applied Speech and Audio Processing: With Matlab Examples* 1st ed (New York: Cambridge University Press)
[8] Story B H and Titze I R 1995 *The Journal of the Acoustical Society of America* **97** 1249–1260
[9] Adachi S and Yu J 2005 *The Journal of the Acoustical Society of America* **117** 3213
[10] Zhang Z, Neubauer J and Berry D A 2006 *The Journal of the Acoustical Society of America* **120** 1558
[11] Berry D A 2001 *Journal of Phonetics* **29** 431–450
[12] Makhoul J 1975 *Proceedings of the IEEE* **63** 561–580
[13] Torres M E, Schlotthauer G, Rufiner H L and Jackson-Menaldi M C 2009 *4th European Conference of the International Federation for Medical and Biological Engineering* (*IFMBE Proceedings* vol 22) ed Magjarevic R, Sloten J, Verdonck P, Nyssen M and Haueisen J (Springer Berlin Heidelberg) pp 252–255
[14] Schlotthauer G 2010 *Análisis de señales con descomposición empírica en modos y aplicaciones a la señal de voz* Tesis de doctorado en ingeniería, Universidad Nacional del Litoral, Santa Fe
[15] Baken R J and Orlikoff R F 2000 *Clinical measurement of speech and voice* (San Diego: Singular Thomson Learning)
[16] García M J V, Cobeta I, Martín G, Alonso-Navarro H and Jimenez-Jimenez F J 2011 *Journal of Voice* **25** 208–217
[17] Brockmann M, Drinnan M J, Storck C and Carding P N 2011 *Journal of Voice* **25** 44–53
[18] Deliyski D D 1993 *3rd Conference on Speech Communication and Technology EUROSPEECH 93* (Berlin, Alemania) pp 969—1972
[19] Murphy P 2008 *Journal of Voice* **22** 125 – 137 ISSN 0892-1997
[20] Ruinskiy D and Lavner Y 2008 *Electrical and Electronics Engineers in Israel, 2008. IEEEI 2008. IEEE 25th Convention of* (IEEE) pp 489–493
[21] Titze I R 1995 Workshop on acoustic voice analysis: summary statement Tech. rep. National Center for Voice and Speech Denver, USA

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
G. A. Alzamendi, G. Schlotthauer, H. L. Rufiner & M. E. Torres; "Desarrollo de un modelo para la síntesis de voz irregular basado en parámetros acústicos"
Memorias del XVIII Congreso Argentino de Bioingeniería (SABI 2011), set, 2011.