Evaluación de un nuevo modelo de síntesis de vocales con perturbaciones en los parámetros acústicos

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

† *Laboratorio de Señales y Dinámicas no Lineales (FI-UNER) - CONICET*
‡ *Centro de I+D en Señales, Sistemas e Inteligencia Computacional (FICH-UNL)*
♮ *metorres@santafe-conicet.gov.ar*

***Abstract*— La señal de voz presenta irregularidades intrínsecas que en presencia de patologías se hacen más evidentes. Los parámetros acústicos son útiles en la práctica médica para caracterizar la voz y detectar patologías. Aquí se propone un modelo para la síntesis de voz irregular a partir de los parámetros acústicos Shimmer y Jitter. Se generó la señal glótica artificial a partir de un tren de pulsos, perturbando la amplitud y periodo de cada pulso y aplicando a la señal resultante un filtro lineal autorregresivo equivalente al del tracto vocal. Se desarrollaron modelos para la perturbación de la amplitud y del periodo a partir de métodos estadísticos sencillos. Se generó un conjunto de señales y se analizó el desempeño del modelo utilizando una medida de calidad objetiva (PESQ). Los parámetros Shimmer y Jitter obtenidos coincidieron en su mayoria con los valores teóricos. Los resultados sugieren que el modelo desarrollado es útil para generar voces artificiales para un amplio rango de valores de Shimmer y Jitter y con una buena calidad.**

***Keywords*— Síntesis de voz, modelo de la voz, perturbación de parámetros acústicos.**

1. INTRODUCCIÓN

El estudio y modelado de los mecanismos de generación de la voz abarca diversas áreas de las ciencias y puntos de vistas inter-disciplinarios, dada la complejidad y diversidad de los elementos involucrados. Son sus ejes principales, el análisis de estructuras anatómicas y de los fenómenos involucrados en el proceso del habla, considerando su comportamiento dinámico y relaciones estructurales. Entre los diversos abordajes se puede mencionar: métodos para el reconocimiento de hablantes, estrategias para mejorar la calidad de las voces artificiales y su uso como interfaz hombre-máquina y diversas técnicas destinadas al modelado, acondicionamiento, síntesis, compresión y transmisión de señales de voz. Los modelos propuestos para analizar e imitar el proceso de generación del habla difieren en cuanto a las estrategias y métodos seleccionados y en relación con las aplicaciones consideradas. Se ha demostrado que incluso las voces sanas presentan irregularidades y que éstas son las responsables del grado de naturalidad con que se perciben [1, 11]. Recientemente los modelos de la voz han sido considerados para el estudio y síntesis de voces patológicas, permitiendo desarrollar un mayor entendimiento sobre las etiologías y alteraciones presentes en los diferentes trastornos [11, 13].

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 [11]. Los parámetros *Shimmer* y el *Jitter* son los parámetros más empleados para cuantificar las alteraciones instantáneas en la amplitud y la frecuencia, respectivamente, reportándose su utilidad para caracterizar diferentes tipos de voz y su sensibilidad a los diversos trastornos [1, 2, 3].

El presente trabajo tiene como finalidad proponer, desarrollar y evaluar 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 posee un marco teórico sencillo y ha demostrado ser útil en una gran variedad de aplicaciones [8]. 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), dando como resultado la señal de voz propiamente dicha [9]. A continuación, se estudiarán cada uno de los componentes del modelo.

2.1. Fuente glótica

La morfología de la fuente glótica (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:

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

donde A_i y P_j son la amplitud y periodo de cada uno de los pulsos [8]. El valor $1/P_j$ determina la frecuencia instantánea (F_0) del pulso. Las 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 [10] se utilizó este tipo de perturbación como ruido aleatorio adicionado a la FG para la síntesis de voz. 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. Para ello es necesario obtener una relación entre estas medidas y los parámetros de la FG.

Se define como *Shimmer* [1] 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* ($Shimmer_{\%}$)[1]:

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

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 [1]. De las diversas medidas para cuantificarlo, la más utilizada es la denominada *razón de Jitter porcentual* ($Jitter_{\%}$), dada por [1]:

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

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 estadísticamente independientes entre sí, lo que permite hacer uso de la Ec. (1). Además, suponemos que las series A_i y P_j presentan comportamiento Gaussiano siendo sus distribuciones $\mathcal{N}(A_0, \sigma_A^2)$ y $\mathcal{N}(P_0, \sigma_P^2)$ 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 estándar respectivos. Estas hipótesis han sido empleadas anteriormente con resultados muy satisfactorios, tanto en el análisis de la dinámica de la señal de voz [12] 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 poseen distribuciones dadas por $\mathcal{N}(0, 2\,\sigma_A^2)$ y $\mathcal{N}(0, 2\,\sigma_P^2)$ 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:

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

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

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

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 Ec. (2) se obtiene:

$$\sigma_A = \frac{\sqrt{\pi}\, A_0\, Shimmer_{\%}}{200}. \qquad (6)$$

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

$$\sigma_P = \frac{\sqrt{\pi}\, P_0\, Jitter_{\%}}{200}. \qquad (7)$$

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 [8, 9]. Se lo representa mediante una ecuación en diferencias de la forma:

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

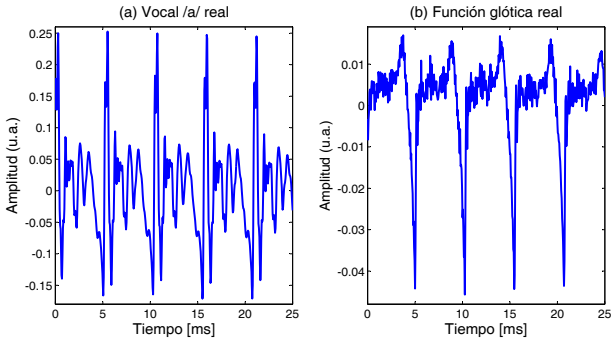

Figura 1: Señal de voz sana correspondiente a un individuo de sexo masculino ($F_0 = 189,295$ Hz, $Jitter_{\%} = 0,269\,\%$ y $Shimmer_{\%} = 1,826\,\%$). a) Vocal */a/* sostenida, b) Función glótica estimada (residuo) correspondiente.

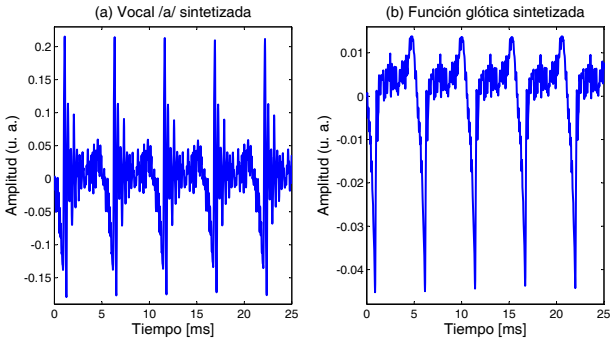

Figura 2: Señal de voz sintetizada considerando una voz sana, correspondiente a un individuo de sexo masculino ($F_0 = 189,295$ Hz, $Jitter_{\%} = 0,269\,\%$ y $Shimmer_{\%} = 1,826\,\%$). a) Vocal */a/* sostenida, b) Función glótica estimada (residuo) correspondiente.

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).

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

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

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.

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 la FG real y su aplicación en la síntesis de voz permite mejorar las características acústicas y perceptuales de la señal generada [8].

2.3. Señales reales

La base de datos utilizada [6] consta de grabaciones de vocal */a/* sostenida de 53 individuos con voces sanas y 654 que presentan voces alteradas debido a una variedad de patologías. 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.

En esta aplicación se trabajó con 22 coeficientes LPC. Esto se debe a que la frecuencia de muestreo de las señales empleadas en este trabajo es superior a la encontrada habitualmente en las aplicaciones y, como es sabido, se necesita una cantidad mayor de coeficientes LPC para representar el contenido frecuencial de la señal. En la Fig. 1 se puede apreciar 25 ms de una vocal (izquierda) y su correspondiente residuo (derecha) para el caso de una voz sana perteneciente a un individuo de sexo masculino.

2.4. Señales sintetizadas

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 descripto 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 señal de voz conforme a lo explicado en las Secciones 2.2 y 2.3. En la Fig. 2 se muestran 25 ms de una vocal sintetizada (izquierda) aplicando el procedimiento y con una *frecuencia de muestreo* (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 Fig. 1.

Se sintetizó un conjunto de señales tomando diferentes valores de $Jitter_{\%}$ y $Shimmer_{\%}$. Los rangos entre los que se trabajó fueron $0,00 \leq Jitter_{\%} \leq 3,00$ y $0,00 \leq Shimmer_{\%} \leq 5,00$, siendo el paso de 0,05. Los extremos se tomaron en función de los valores mínimos y máximos de cada parámetro acústico correspondientes a voces sanas, tomados de la base de datos (ver Tabla 1).

2.5. PESQ

A los efectos de evaluar la calidad perceptual de las voces sintetizadas con el modelo aquí propuesto, se utilizó una medida objetiva denominada *evaluación perceptual de la calidad de voz* (del inglés "perceptu-

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 patologías poseen valores superiores en comparación con las de los individuos sanos (DE indica el Desvío Estándar).

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

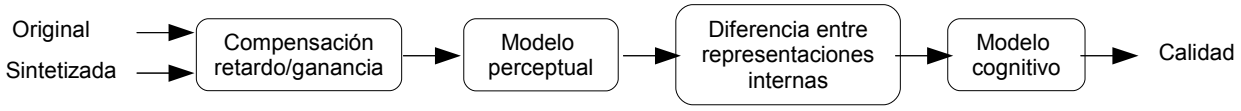

Figura 3: Diagrama de las etapas del método PESQ.

al evaluation of speech quality" o PESQ), definida en ITU P.862 como el estándar para la evaluación de la calidad de la voz transmitida por un canal de comunicación [7]. Ha sido ampliamente estudiada y se ha demostrado que tiene una correlación muy alta con las mediciones de calidad subjetiva en una amplia variedad de situaciones [5].

Esta medida utiliza varios niveles de análisis en un intento de imitar la percepción humana. La primera etapa consiste en una compensación de ganancia/retardo. Luego se realiza una transformación a un dominio perceptual y allí se calcula una densidad de distorsión a partir de la diferencia entre la señal analizada y la de referencia. Para la etapa final, se aplican algunos modelos cognitivos (ver Fig. 3).

En nuestro caso se utiliza una versión no lineal de la PESQ que toma valores en el rango $[0, 5]$ a partir de una regresión no lineal con pruebas subjetivas [4]. Se sintetizaron 53 vocales sostenidas correspondientes a las voces normales de la base de datos y se calculó el valor de PESQ en comparación con las reales. Para cada señal, se emplearon valores de F_0, $Jitter_{\%}$ y $Shimmer_{\%}$ extraídos de la información clínica. El algoritmo empleado se descargó de `http://www.utdallas.edu/~loizou/speech/software.htm`.

3. RESULTADOS

Analizamos aquí los resultados obtenidos a partir de la síntesis de las señales, considerando los parámetros acústicos, y evaluando su calidad perceptual.

Las Figs. 1 y 2 permiten comparar las señales sintetizadas mediante el modelo propuesto con las reales. En el caso de los residuos (Figs. 1.b y 2.b), se muestra que el comportamiento de éstos resultaron muy similares entre sí. Se aprecia que el residuo real presenta pequeñas oscilaciones en su amplitud, generadas por las irregularidades propias del aparato fonador, las cuales no se observan en el residuo sintetizado. Por el contrario, la morfología de la vocal sintetizada difiere considerablemente de la real (ver Figs. 1.a y 2.a), no así su regularidad. Esto último se debe a que el filtro TV únicamente simula aproximadamente el espectro de magnitud del tracto vocal, no garantizando la obtención de una réplica exacta de la señal original [8]. El filtro de pre-énfasis y la convolución del tren de pulsos con el residuo mejoran la calidad de la señal sintetizada.

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 Ecs. (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 Fig. 4 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 [1].

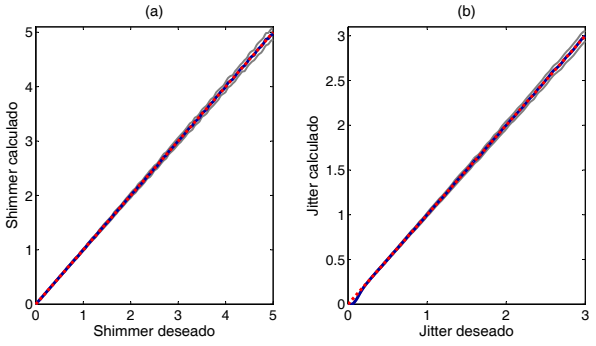

Figura 4: Parámetros acústicos calculados en función de los valores teóricos. En línea azul continua se representa el valor medio, en líneas grises continuas el desvío estándar y en línea punteada roja el valor teórico. a) $Shimmer_{\%}$, b) $Jitter_{\%}$.

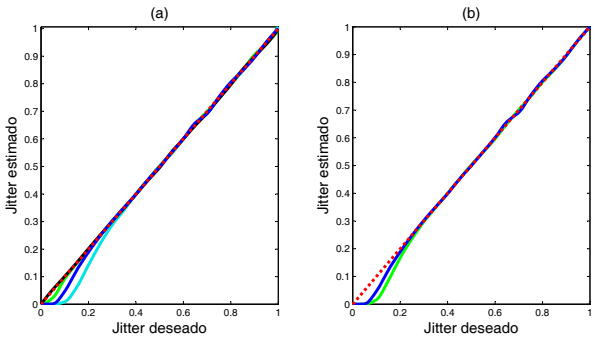

Figura 5: (a) Valor de $Jitter_{\%}$ para señales con diferentes F_m. Se representa en celeste $F_m = 35$ kHz, en azul $F_m = 50$ kHz, en verde $F_m = 75$ kHz y en negro $F_m = 100$ kHz. (b) Valor de $Jitter_{\%}$ para señales con diferentes F_0, con $F_m = 50$ kHz. En azul se presenta la voz de un hombre con $F_0 = 189,295$ Hz y en verde la voz de una mujer con $F_0 = 230,323$ Hz.

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 Fig. 4 y se debe principalmente a la naturaleza discreta de la señal sintetizada. Considerando la Ec. (3) para valores pequeños de $Jitter_{\%}$, se observa que la dificultad radica en el cálculo de $|P_{j+1} - P_j|$ dado 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.

Para corroborar esta hipótesis, se repitió el experimento variando la F_m del conjunto de señales. En la Fig. 5.a 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 Ec. (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 Fig. 5.b 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). Ambas señales fueron generadas con $F_m = 50$ kHz.

Si bien este comportamiento parecería ser un falla del modelo, al estudiar los valores de la Tabla 1 se observa que el menor valor encontrado es 0,175. Sin embargo, el análisis de la base de datos indica que la 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 debería seleccionar una F_m apropiada.

Para llevar a cabo el análisis de la calidad perceptual se obtuvo el valor PESQ tomando ventanas de 2500 muestras, tanto en las voces reales como en las sintetizadas. Se empleó esta cantidad de elementos teniendo en cuenta que la PESQ ha sido diseñada para el estudio de la voz hablada continua y que en este caso la estabilidad de las vocales se mantiene por periodos de $20 - 30[ms]$. Por otra parte, en las vocales sostenidas de larga duración la amplitud presenta oscilaciones que enmascararían la comparación. En la Fig. 6 se muestra un histograma de los valores PESQ obtenidos. Se puede observar que la mayoría de las señales sintetizadas lograron un valor de PESQ superior a 3,0 alcanzando un valor medio de 3,9 y un desvío estándar de 0,4, lo que permite inferir que la calidad perceptual de las señales sintetizadas es muy buena.

4. CONCLUSIÓN Y TRABAJOS FUTUROS

En este artículo se ha propuesto un modelo para la generación de voces artificiales con perturbaciones controladas, considerando los parámetros acústicos *Shimmer* y *Jitter*. En función de ellos se desarrollaron reglas para la modificación de la amplitud y del periodo de la FG. A partir de señales de voz reales, este modelo se aplicó a la síntesis de vocales sostenidas, para

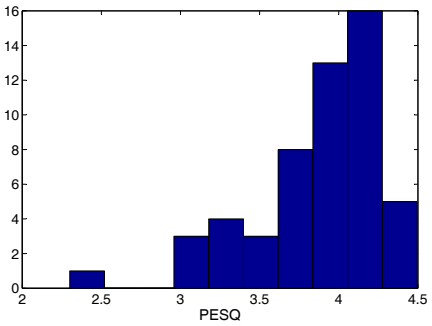

Figura 6: Histograma de los valores de PESQ obtenidos para las vocales sintetizadas en comparación con la correspondiente voz real.

un amplio rango de *Shimmer* y *Jitter*. Se mostró que la vocales sintetizadas poseen un comportamiento similar a las reales y que los parámetros obtenidos se correspondieron con los valores teóricos, para casi la totalidad del rango considerado. La bondad del modelo propuesto en cuanto a la calidad perceptual de las voces normales sintetizadas fue confirmada por los valores de PESQ obtenidos.

Como trabajos futuros en esta línea de investigación, se prevé avanzar en la aplicación del modelo a la síntesis de voces patológicas mediante la incorporación de otros parámetros de interés clínico en el modelo y el uso de técnicas de procesamiento avanzado de señales para el estudio de las señales generadas.

AGRADECIMIENTO

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] R. J. Baken and R. F. Orlikoff. *Clinical measurement of speech and voice*. Singular Thomson Learning, San Diego, 2000.

[2] Meike Brockmann, Michael J. Drinnan, Claudio Storck, and Paul N. Carding. Reliable jitter and shimmer measurements in voice clinics: The relevance of vowel, gender, vocal intensity, and fundamental frequency effects in a typical clinical task. *J Voice*, 25(1):44–53, 2011.

[3] María Jesús Velasco García, Ignacio Cobeta, Gonzalo Martín, Hortensia Alonso-Navarro, and Félix Javier Jimenez-Jimenez. Acoustic analysis of voice in Huntington's disease patients. *J Voice*, 25(2):208–217, 2011.

[4] Y. Hu and P.C. Loizou. Evaluation of objective quality measures for speech enhancement. *Audio, Speech, and Language Processing, IEEE Transactions on*, 16(1):229–238, 2008.

[5] K Kokkinakis and P C Loizou. Evaluation of objective measures for quality assessment of reverberant speech. In *Proc. ICASSP 2011, Prague*, pages 2420–2423, 2011.

[6] Massachusetts Eye and Ear Infirmary MEEI Voice and Speech Lab. Disordered voice database, model 4337, 2009.

[7] ITU-T P.862. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs, 2001.

[8] J G Proakis, J H L Hansen, and J R Deller. *Discrete-Time Processing of Speech Signals*. macmill, NY, 1993.

[9] Hugo L. Rufiner. *Análisis y modelado digital de la voz. Técnicas recientes y aplicaciones*. Ediciones UNL, Santa Fe, 1 edition, 2009.

[10] D. Ruinskiy and Y. Lavner. Stochastic models of pitch jitter and amplitude shimmer for voice modification. In *Proc IEEEI 2008*, pages 489–493, 2008.

[11] Gastón Schlotthauer. *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, 2010.

[12] I. R. Titze. Workshop on acoustic voice analysis: summary statement. Technical report, National Center for Voice and Speech, Denver, USA, 1995.

[13] María E. Torres, G. Schlotthauer, H. L. Rufiner, and M. C. Jackson-Menaldi. Empirical mode decomposition. spectral properties in normal and pathological voices. In *Proc IFMBE 2009*, volume 22, pages 252–255. Springer Berlin, 2009.