SAV: un sistema de análisis acústico para la evaluación de la voz

P.D. Agüero, J.C. Tulli, E.L. Gonzalez, A.J. Uriz y F. De la Cruz Arbizu
Laboratorio de Comunicaciones-Universidad Nacional de Mar del Plata
Juan B. Justo 4302, Mar del Plata, Argentina, Tel.: 481-6600, pdaguero@fi.mdp.edu.ar

Resumen. SAV (Sistema de Análisis de la Voz) es un software de análisis de audio que calcula una gran cantidad de parámetros de la voz, los cuales permiten hacer un seguimiento de los pacientes durante su tratamiento con medidas objetivas. Las principales motivaciones en el desarrollo de SAV son el logro de un software gratuito que pueda ser utilizado por los especialistas de la voz, proporcionando mecanismos de comunicación con los usuarios para incorporar en el software distintas funcionalidades sugeridas y corregir posibles deficiencias, y además ofrecer estudios de casos reales y análisis del rendimiento de los distintos parámetros usando voz sintética. Resultados experimentales en la estimación de jitter local promedio para un conjunto de voces sintéticas de jitter conocido demuestran que la exactitud de SAV es comparable a Praat y MDVP.

Palabras clave: análisis de la voz, jitter, shimmer, Praat, MDVP.

1. Introducción.

El análisis acústico de la voz mediante ordenadores ha alcanzado un importante desarrollo en los últimos tiempos gracias al progreso y difusión experimentados por los medios informáticos que lo hacen posible. Entre sus ventajas se destaca el ser un método no invasivo de evaluación de la voz, ofreciendo la oportunidad de objetivizar la evaluación en unos parámetros numéricos.

Uno de los problemas principales en el diagnóstico perceptivo de la voz por el oído del clínico es que el sistema auditivo humano está preparado fundamentalmente para percibir la voz o el habla como un todo integrado, lo cual es altamente beneficioso desde el punto de vista de la comunicación lingüística. Sin embargo, esta capacidad se ve limitada cuando se trata de individualizar componentes acústicos que son relevantes desde una perspectiva clínica.

En muchas ocasiones existe dificultad en determinar por un procedimiento exclusivamente perceptivo el origen de ciertas anomalías o particularidades de la voz. Por ejemplo, algunos rasgos del tono o pitch son más el producto de las resonancias del tracto vocal que de la frecuencia de vibración de las cuerdas vocales [1]. La hipernasalidad percibida en una voz puede ser la consecuencia de una desincronización en los tiempos de oclusión velar antes que una oclusión incompleta. Es decir, un mismo atributo o alteración de la calidad vocal puede tener su origen en subsistemas distintos difícilmente aislables por la mera audición.

En otras ocasiones, una adecuada percepción no puede matizarse con el grado de precisión que ofrece una medida numérica. Así, en una voz percibida como soplada puede establecerse el grado de aspiración, o "breathiness", a través del parámetro pertinente (índice de turbulencia de voz). En este sentido, junto a la evaluación subjetiva por parte del clínico experimentado, el diagnóstico se enriquece y gana precisión cuando se complementa con la medida objetiva de parámetros relevantes de la voz. Las ventajas de ello se

traducen en una mayor objetividad en el informe o la comunicación de los datos y en una mayor exactitud en la evaluación del progreso terapéutico, especialmente cuando éste es lento.

Para poder evaluar la voz disfónica a través de parámetros numéricos, es importante definir previamente la voz normal o no disfónica y disponer de valores normativos de comparación. Como algunos autores señalan, en nuestro ámbito geográfico son muy escasos los estudios llevados a cabo en este sentido [2]. Por otra parte, algunos valores pueden ser dependientes de los algoritmos usados por el software específico que los calcula, lo que hace más necesario, si cabe, el disponer de normas específicas de los principales programas empleados en la clínica [3][4].

La validez y confiabilidad del análisis acústico se ve afectado por numerosos factores, tales como el tipo de micrófono, los niveles de ruido, el sistema de adquisición digital, la frecuencia de muestreo y el software usado [5][6]. Por ejemplo, medidas de perturbación de amplitud y frecuencia no deberían depender del software utilizado. El Jitter y el shimmer están definidos por fórmulas simples y estándar [1]. Las diferencias observadas entre los diferentes valores numéricos provienen de los valores estimados de frecuencia fundamental en que se basan estas fórmulas. No existe un algoritmo estándar para el cálculo de la frecuencia fundamental que sea adoptado por todos los programas.

Junto a esta necesidad, una propiedad que debe disponer todo instrumento de medida es el de una fiabilidad adecuada que permita confiar en la estabilidad de los valores obtenidos. En la extracción de parámetros de la voz no podemos esperar una coincidencia absoluta entre dos medidas sucesivas del mismo individuo y, dada la enorme variabilidad de la voz humana, tanto entre individuos como en el individuo mismo, es aceptable cierta variación siempre que ésta se mantenga dentro de ciertos límites. No obstante, la robustez y validez clínica de los parámetros descansa necesariamente, como condición previa, en el grado de consistencia de sus valores. Dos medidas repetidas de la voz del mismo individuo en las mismas condiciones de registro deberían ser lo suficientemente semejantes para que podamos confiar en ellas [7].

En este artículo describimos SAV, un software de análisis de la voz programado en lenguaje JAVA©. Los motivos para el uso de este lenguaje de programación fueron la posibilidad de generar un software que pueda ser ejecutado en múltiples sistemas operativos, tales como Windows, Linux y Macintosh. De esta manera, los usuarios no deberán utilizar complejos mecanismos de instalación para usarlo.

Otra de las razones por las cuales hemos utilizado JAVA es la posibilidad de contar con herramientas de desarrollo gratuitas, como es el caso de Netbeans, una interfaz visual de programación de aplicaciones en JAVA muy versátil. Finalmente, la principal motivación ha sido la utilización de JAVA para ofrecer un software gratuito, y de este modo, llegar a la mayor cantidad de usuarios con independencia de sus posibilidades económicas.

A pesar de ser un software gratuito, SAV posee siempre un canal abierto para críticas y sugerencias de los usuarios, y también incorporará los avances desarrollados tanto en nuestro laboratorio como en publicaciones relacionadas con el tema.

Con el objeto de conocer la precisión de SAV en el análisis de la frecuencia fundamental, se incluye en este artículo un estudio comparativo en el cálculo del jitter local promedio con respecto a dos softwares de análisis clínico: Praat [8] y MDVP [9].

En la Sección 2 se incluye una descripción de las distintas funcionalidades de SAV. Luego, los resultados experimentales en la estimación del jitter local promedio se presentan en la Sección 3. Finalmente, en la Sección 4 se detallan las conclusiones.

2. Metodología.

El funcionamiento de SAV es sencillo. En la pantalla se encuentra toda la información disponible para el usuario de manera simultánea, con el objetivo de evitar la utilización de un menú de opciones que impida observar toda la capacidad del programa con un simple vistazo.

En la pantalla inicial de SAV de la Figura 1 se puede observar la presencia de cuatro solapas: *análisis gráfico*, *análisis estadístico*, *gráfico radial* y *seleccionar archivo*.

2.1. Solapa Seleccionar Archivo.

Si seleccionamos la solapa *seleccionar archivo* podremos elegir el archivo que se desea analizar con SAV. Por ejemplo, si seleccionamos el archivo test.wav, SAV procederá a hacer el análisis del archivo de audio y pasará automáticamente a visualizarlo en la solapa de *análisis gráfico*, tal como se muestra en la Figura 1.

El formato del archivo de sonido debe ser WAV, grabado con una frecuencia de muestreo de 44100Hz y 16 bits. En caso de no respetar estos requisitos, el software le indicará que el formato del archivo seleccionado es incorrecto.

2.2. Solapa Análisis Gráfico.

La pantalla de *análisis gráfico* presenta en orden descendente:

- *Señal de audio.* Es la representación visual en un gráfico XY de la señal de audio, donde el eje X corresponde al tiempo y el eje Y es la amplitud de la señal.

- *Espectrograma.* Es la representación visual de la evolución espectral de la señal de audio, donde el eje X corresponde al tiempo y el eje Y es la frecuencia. La mayor intensidad de potencia en un determinado rango de frecuencia se representa con tonos oscuros.

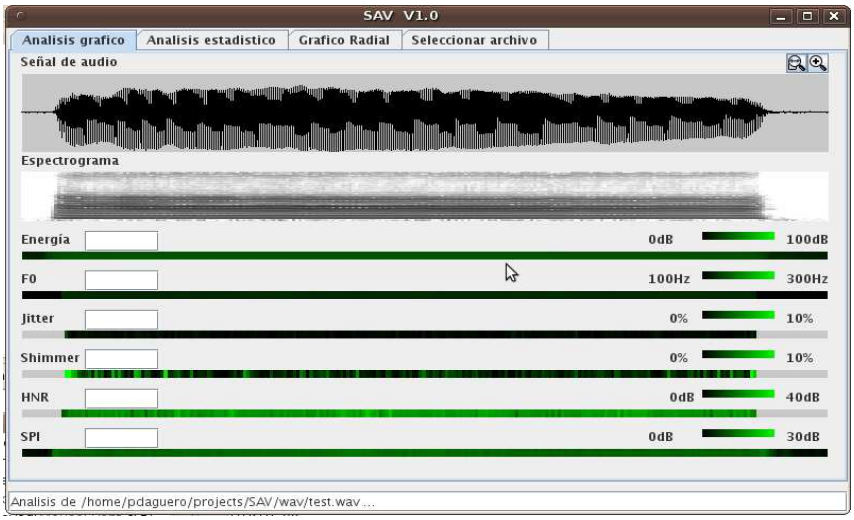


Figura 1. Pantalla principal de SAV.

- *Energía.* La energía es un parámetro que permite estimar la intensidad de la fonación. Su valor está calculado en decibeles (dB), ya que esta escala permite representar un gran rango de valores con números pequeños.

- *Frecuencia fundamental.* Este parámetro estima la frecuencia instantánea de vibración de la glotis en los fonemas sonoros. Su valor está directamente vinculado con el jitter y el shimmer.

- *Jitter.* Este parámetro estima la variación del período de vibración de la glotis entre dos períodos consecutivos.

- *Shimmer.* Este parámetro estima la variación de la máxima amplitud de la fonación entre dos períodos consecutivos de vibración de la glotis

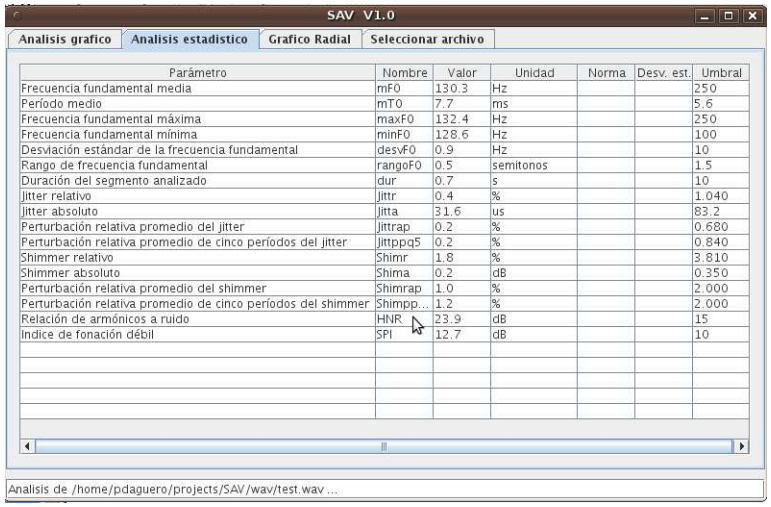
.

- *HNR.* La relación de armónicos a ruido (o por sus siglas en inglés HNR: Harmonics-to-Noise Ratio) estima el grado de periodicidad y armonicidad de la fonación de un fonema sonoro para un período de vibración de la glotis [8].

- *SPI.* El índice de fonación débil (o por sus siglas en inglés SPI: Soft Phonation Index) es un indicador de aducción de las cuerdas vocales y cierre glotal durante la fonación.

La información está representada usando una escala de intensidad de colores, donde el color verde más intenso indica un valor mayor del parámetro. Esta representación visual permite observar a intervalos regulares de 5,8ms el valor de los distintos parámetros, con lo cual se puede analizar distintos segmentos de la fonación.

Es posible ampliar una región del audio seleccionando una porción del mismo pulsando el botón izquierdo ratón en la señal de audio para indicar el comienzo de la selección y arrastrando el ratón hasta llegar al final de la misma. Luego, liberando el botón del ratón, la porción seleccionada se resaltará con color amarillo.



Parámetro	Nombre	Valor	Unidad	Norma	Desv. est.	Umbral
Frecuencia fundamental media	mF0	130.3	Hz			250
Período medio	mT0	7.7	ms			5.6
Frecuencia fundamental máxima	maxF0	132.4	Hz			250
Frecuencia fundamental mínima	minF0	128.6	Hz			100
Desviación estándar de la frecuencia fundamental	desvF0	0.9	Hz			10
Rango de frecuencia fundamental	rangoF0	0.5	semitonos			1.5
Duración del segmento analizado	dur	0.7	s			10
Jitter relativo	Jittr	0.4	%			1.040
Jitter absoluto	Jitta	31.6	us			83.2
Perturbación relativa promedio del jitter	Jittrap	0.2	%			0.680
Perturbación relativa promedio de cinco períodos del jitter	Jittppq5	0.2	%			0.840
Shimmer relativo	Shimr	1.8	%			3.810
Shimmer absoluto	Shima	0.2	dB			0.350
Perturbación relativa promedio del shimmer	Shimrap	1.0	%			2.000
Perturbación relativa promedio de cinco períodos del shimmer	Shimpp...	1.2	%			2.000
Relación de armónicos a ruido	HNR	23.9	dB			15
Indice de fonación débil	SPI	12.7	dB			10

Figura 2. Pantalla de análisis estadístico de SAV.

2.3. Solapa Análisis Estadístico.

La porción de audio visualizado en la pantalla de análisis gráfico puede ser analizada estadísticamente, para lograr un conjunto de valores compacto que permita comparar distintos audios. Esa es la función de la solapa de análisis estadístico.

En esta pantalla (Figura 2) se presenta un gran número de parámetros, tales como:

- *Frecuencia fundamental media.* El valor promedio de la frecuencia fundamental.

- *Período medio.* El valor promedio del período de vibración de la glotis.

- *Frecuencia fundamental máxima.* El valor máximo de la frecuencia fundamental.

- *Frecuencia fundamental mínima.* El valor mínimo de la frecuencia fundamental.

- *Desviación estándar de la frecuencia fundamental.* El valor de dispersión de la frecuencia fundamental.

- *Rango de frecuencia fundamental.* La amplitud de variación de la frecuencia fundamental en semitonos.

- *Duración del segmento analizado.* La duración de la porción de voz analizada en segundos.

- *Jitter y sus diferentes variantes.* Jitter relativo (jittr), absoluto (jitta) y perturbación relativa promedio (jittrap y Jittppq5).

- *Shimmer y sus diferentes variantes.* Shimmer relativo (shimr), absoluto (shima) y perturbación relativa promedio (shimrap y shimppq5).

- *Relación de armónicos a ruido.*

- *Indice de fonación débil.*

2.4. Solapa de Gráfico Radial.

Con el objeto de una visualización simple del valor de los parámetros indicando si son normales o anormales, se incorporó a SAV un gráfico radial (Figura 3).

En la solapa de Gráfico Radial se pueden observar los distintos parámetros del análisis estadístico, y dependiendo de su valor con respecto al umbral se ubicará en la región verde (normal) o roja (anormal).

3. Resultados y discusión.

Con el objeto de conocer la precisión de SAV en el análisis de la frecuencia fundamental, se desarrolló un estudio comparativo en el cálculo del jitter local promedio con respecto a dos softwares de análisis clínico: Praat y MDVP. Para ello se utilizaron voces con una frecuencia de muestreo de 44100Hz y 16 bits de resolución. Las señales contienen solamente voz, sin ninguna otra información adicional, como sería el caso de una señal del electroglotógrafo.

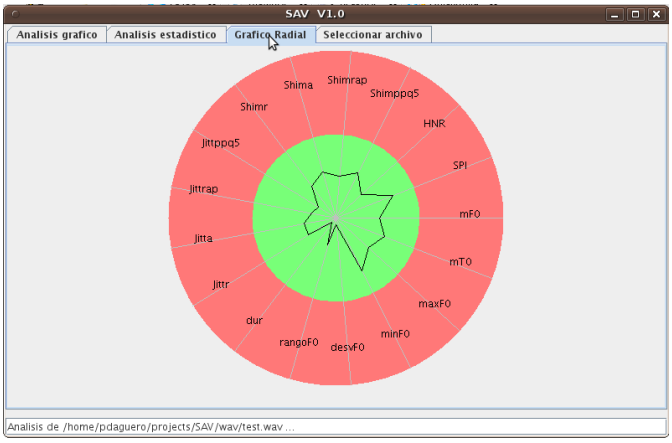


Figura 3. Pantalla de gráfico radial de SAV.

Con el objeto de obtener resultados experimentales en condiciones controladas, se generaron voces sintéticas con diferentes grados de jitter local promedio. La fuente glotal usada en los experimentos corresponde al modelo glotal de Liljencrants-Fant [10].

La fuente glotal sintética con el jitter deseado se filtra usando un conjunto de coeficientes de predicción lineal (LPC) estimados de una voz real sin ninguna patología. La forma de onda resultante contendrá un jitter local promedio conocido que podrá ser usado como referencia en los experimentos. Los valores de jitter local promedio simulados cubren un amplio rango de valores desde una voz normal hasta una voz patológica: 0.01% hasta 20%.

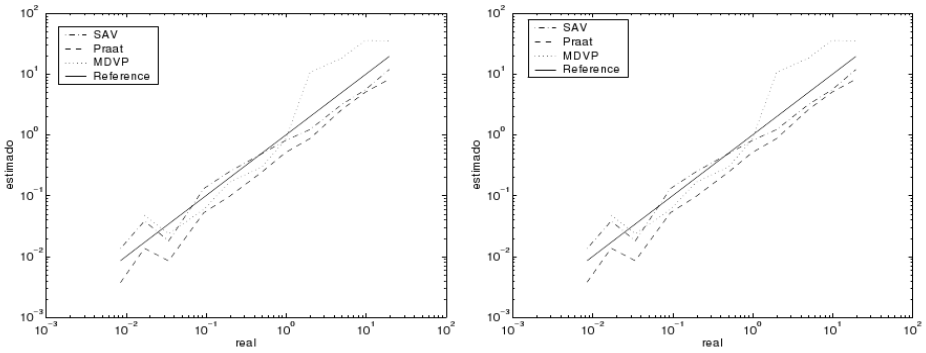


Figura 4. Resultados experimentales.

Los resultados del experimento realizado con las voces sintéticas con diferentes valores de jitter local promedio y relación señala ruido (SNR=40dB y SNR=20dB) se muestran en la Figura 4. Los valores de jitter local promedio estimados con SAV tienen una diferencia más pequeña con respecto a los valores de referencia que aquellos valores obtenidos con Praat y MDVP, tanto para una SNR de 40dB como para una SNR de 20dB.

Valores pequeños de jitter son detectados en forma más precisa por Praat, principalmente para valores inferiores a 0.1%. Para valores de jitter más grandes, SAV ofrece una mejor aproximación al valor real.

4. Conclusiones

En este artículo se presenta a SAV, un software de análisis acústico de la voz con propósitos clínicos. El software es de distribución libre y gratuita, y está realizado en lenguaje JAVA. De esta manera, es posible utilizar SAV en múltiples plataformas: Windows, Linux y Macintosh. Es importante observar que el jitter detectado con SAV es más preciso que Praat y MDVP en el rango de valores cercanos al umbral para patologías fijado para MDVP: 1.040%. Este es un resultado remarcable, ya que es importante que un software de análisis de la voz tenga precisión en el rango de valores necesario para el seguimiento del tratamiento de pacientes: 0.1% a 20%.

Referencias

[1] R. Baken, R. Orliko. Clinical measurement of speech and voice. Second Edition. San Diego, CA: Singular Publishing Group, 2000.

[2] R. Fernandez, D. Damborenea, P. Rueda, E. Garca, J. Leache, M.A. Campos, E. Llorente, M.J. Naya. Análisis acústico de la voz normal en adultos no fumadores. Acta Otorrinolaringologa, Vol. 50, pp. 134-141, 1999.

[3] V. Parsa, D. Jamieson. A comparison of high precision F0 extraction algorithms for sustained vowels. Journal of Speech and Hearing, Vol. 42, pp. 112-126, 1999.

[4] Ch. Read, E. Buder, R. Kent. Speech analysis systems: An evaluation. Journal of Speech and Hearing, vol. 35, pp. 314-332, 1992.

[5] D. Deliyski, H. Shaw, M. Evans. Influence of sampling rate on accuracy and reliability of acoustic voice analysis. Logopedics, Phoniatrics, Vocology. Vol. 30, pp. 55–62, 2005.

[6] D. Deliyski, H. Shaw, M. Evans. Regression tree approach to studying factors influencing acoustic voice analysis. Folia Phoniatrica et Logopaedica. Vol. 58, pp. 274–288, 2006.

[7] J. Gonzalez, T. Cervera, J.L. Miralles. Análisis Acústico de la voz: fiabilidad de un conjunto de parámetros multidimensionales. Acta Otorrinolaringol Esp, no. 53, pp. 256-268, 2002.

[8] P. Boersma. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. Proceeding of Institute of Phonetic Sciences, Vol. 17, pp. 97-110, 1993.

[9] D. Deliyski: Acoustic model and evaluation of pathological voice production. Proceedings of Eurospeech ’93, pp. 1969-1972, 1993.

[10] G. Fant, J. Liljencrants, Q. Lin. A four parameter model of glottal flow. STL-QPSR. Vol. 26, pp. 1–13, 1985.