Base de datos de Voces Venezolanas para pruebas de calidad de voz

Jesús J. Jiménez G.[*,a], José. A. Díaz[a], José Pacheco[b]

[a]*Centro de análisis y tratamiento de señales, Facultad de Ingeniería, Universidad de Carabobo. Valencia, Venezuela.*
[b]*Centro de procesamiento de imágenes, Facultad de Ingeniería, Universidad de Carabobo. Valencia, Venezuela.*

Resumen.-

El enfoque principal de este trabajo es crear una base de datos de voces venezolanas en edades comprendidas entre 16 y 26 años, la grabación de la voz de cada uno de los locutores se hizó en un ambiente aislado de ruido, utilizando técnicas y equipos de grabación profesionales, recogiendo información, tales como edad, sexo, si es fumador, entre otras, que pudiera servir de base a investigaciones futuras. Este trabajo se propone dar información a investigadores en el área de voz, e iniciar el soporte a la posible creación de una base de datos más amplia, en donde se pueda contar con el desempeño de profesionales como otorrinolaringólogos, logopedias, etc.

Palabras clave: voces venezolanas, base de datos de voces, muestras de voces

Venezuelan voice database for voice quality testing

Abstract.-

The main focus of this work is to create a database of Venezuelan voices aged 16 to 26 years, the voice recording of each of the speakers was done in an isolated environment noise, using techniques and equipment professional recording, collecting information, such as age, sex, whether it is smoking, among others, that could be the basis for future research. This paper aims to provide information to researchers in the area of voice, and start supporting the possible creation of a comprehensive database, where they can count on the performance of professionals and otolaryngologists, logopedias, etc.

Keywords: Venezuelan voices, Database voice, voice samples

Recibido: octubre 2012
Aceptado: febrero 2013.

1. Introducción

En las investigaciones se reportan la utilización de bases de datos de voz, realizadas por centros médicos, en donde la evaluación de la calidad de voz es realizada por expertos, dígase evaluación perceptiva, ya que en la actualidad no se cuenta con una evaluación objetiva de la misma.

Las bases de datos de voz para investigación de la calidad de voz son escasas y no se conoce ninguna que sea de libre obtección.

No se conocen criterios que establezcan la cantidad de muestras que puede arrojar resultados confíables en una base de datos de voz, y las cantidades de muestras en las investigaciones conocidas son muy variadas. En algunos casos se desconoce la integridad y/o la precisión en las medidas de la señal de voz, necesaria para iniciar cualquier proceso de investigación y análisis al respecto.

No existe un procedimiento estándar para la obtención de muestras de señales de voz. Los algoritmos de análisis de señales de voz que calculan los mismos parámetros en muchos casos arrojan resultados disimiles para la misma

[*]Autor para correspondencia
Correo-e: jjjimenezgriman@gmail.com (Jesús J. Jiménez G.)

muestra. En algunas de las bases de datos que se utilizan para la investigación son preprocesadas y queda la incertidumbre de poder establecer si eso influyo negativamente o no en la investigación.

2. Calidad de voz.

Aronson 1990, citado por [1] afirma que "hay alteración de la voz cuando esta difiere de las voces de otras personas del mismo sexo, similar edad y grupo cultural, en timbre, tono, volumen, flexibilidad y en dicción" (p.62). De la anterior definición se desprende que no existen criterios objetivos para la determinación de la voz normal y por ende la patológica, y es el aspecto cultural muy importante ya que una voz considerada normal en una región puede ser considerada patológica en otra. O como lo indica Kreiman [2] en su conferencia "la calidad general de un sonido se define formalmente como el atributo de la sensación auditiva en términos de que el oyente puede juzgar que dos sonidos presentadas en forma análoga y con la misma intensidad y el tono son diferentes (ANSI, 1960)" .

3. Creación de bases de datos de voz

3.1. La muestra de voz.

Casado [1] nos indica que el fonema más usado es |*ae*| , pero en español no resulta sencillo, por lo que se ha utilizado la |*a*| en el tono habitual del paciente y a una intensidad confortable. En lengua inglesa se ha utilizado el fonema |*ah*| [3] y en otras lenguas se utiliza un fonema de vocal específico [4] [5] [6].

El uso de vocales aisladas para esta evaluación constituye una situación comunicativa irreal, por lo que algunos autores prefieren considerar los resultados en voz continua, utilizan refranes, lecturas de libros, e incluso oraciones religiosas en pacientes donde hay problemas visuales [4].

No existe una metodología para la obtención de la muestra de voz, ni siquiera una en lo que refiere al tipo de equipo a utilizar, ya sea micrófono, distancia del micrófono a la boca, ángulo del micrófono, equipo de grabación, ni en la frecuencia de muestreo, etc., factores estos que influyen en la toma de la muestra [7].

Al revisar los diferentes trabajos se orientan a las particularidades de la investigación y con una gama de diferencias importantes, que van desde el fonema a utilizar para el especifico idioma nativo en la investigación, la duración de la grabación propuesta, si es con el tono habitual del paciente y a una intensidad confortable, si es monólogo de situaciones del día día, si es la realización de una lectura o lecturas con ritmo y emoción.

Adicionalmente no existe norma que establezca el número de muestras a considerar, para obtener respaldo estadístico confiable, el control en el tipo de sexo del individuo y la correspondiente edad como factores influyentes en la determinación de la calidad de voz; así como tampoco el control para la obtención de muestras con aspectos culturales semejantes.

3.2. Muestras de voz utilizadas en investigaciones.

Como se indicó anteriormente existen diferentes criterios para la conformación de una base de datos de voz, en esta sección se muestra las diferencias de cuatro trabajos donde se considera el idioma, el tipo de muestra, los equipos de grabación, la edad de los individuos que suministran la muestra y la presencia o no de patologías.

Como primera comparación se toma los individuos que participaron en la investigación como locutores y los equipos utilizados vea la Tabla 1. De la Tabla 1 se observa que los rangos de edades son bastante grande, 72 años la diferencia mayor y la voz en los seres humanos cambia por la edad. Otra información de esa tabla es la diversidad de equipos para la obtención de las grabaciones, lo que evidencia que no existe un estándar para la grabación a pesar de que todos los equipos son profesionales.

En la Tabla 2 se observa la diversidad ahora de la duración de la muestra, el tipo de muestra y su escogencia dependiendo de la investigación que se realice es amplia.

3.3. Equipos para la obtención de muestras utilizados en investigaciones.

Se han realizado trabajos con múltiples tipos de equipo, tratando de obtener métodos rápidos, precisos y rentables para el registro de datos, como se observa en los trabajos descritos anteriormente (Tabla 1) y como lo reseña el trabajo Comparison of Voice Acquisition Methodologies in Speech Research [8], en donde además se compara la obtención simultanea con tres métodos distintos y se analiza mediante técnicas de análisis idénticos. En esta investigación fueron utilizadas técnicas estándar de la industria para la adquisición de señales acústicas de alta calidad, como grabadora con unidad de disco duro incorporado y grabadora de estado sólido, y métodos de adquisición de datos basados en computadoras portátiles estándar. Los métodos utilizados son los siguientes:

1. El primer método utilizó un grabador portátil de estado sólido Marantz PMD671 y un micrófono de condensador ***cardioide*** AKG C 420 fijado en la cabeza a una distancia de 8 cm de la boca del locutor.
2. El segundo método utiliza una grabadora de disco duro Alesis Masterlink ML-9600 junto con un mezclador Behringer EURORACK UB802, con un micrófono de condensador cardioide modelo AKD C 420 fijado en la cabeza a una distancia de 8 cm de la boca del locutor.
3. El tercer método implicaba el uso de un PC portátil con una configuración básica de fábrica (Sony) y audífonos con conexión usb y micrófono unidireccional Logitech A-0374A en un ángulo de 45 ° y a 8 cm de la boca.
4. En todos los métodos se tomaron muestras con resolución a 16 bit y a 44,1 kHz.

Tabla 1: Comparacion de muestras

Trabajo:	**Locutores:**	**Equipos:**
Rusz [4]	46 pacientes, hombres y mujeres, con y sin enfermedad de Parkinson y en edades entre 34 y 83 años	Se utilizó para la grabación un micrófono condensador colocado a 15 cm de la boca aproximadamente, acoplado a una cámara de vídeo Panasonic NVGS 180 y se muestreo a 48 kHz con resolución de 16 bit
Hillman [5]	24 hombres y 16 mujeres con trastornos de voz con edades 19-85 años. 4 hombres y 3 mujeres sin trastornos de voz con edades entre 20-52 años	Se utilizó un micrófono condensador (MKE104, Sennheiser electronic GmbH) colocado en la cabeza a 4 cm de los labios y a 45º. Se utilizó un preamplificador (Symetrix 302 Dual Microphone Preamplifier) y se muestreo la señal de voz a 100 kHz con resolución de 16 bit

Continúa en la próxima página

Tabla 1 – *Continuación comparación de muestras*

Trabajo:	**Locutores:**	**Equipos:**
Michaelis [6]	447 diferentes individuos (hombres y mujeres) de edades comprendidas entre 10 y 80 años (media: 48 años) con una gran variedad de trastornos de la voz orgánica y funcional 88 personas sin antecedentes de problemas de la voz de 18 a 90 años (media: 47 años)	Se utilizó el equipo Computer Speech Lab (CSL 4300) a una frecuencia de muestreo de 50 kHz
Crevier-Buchman [9]	Participaron 12 pacientes de sexo masculino sometidos a laringectomía parcial supracricoidea con cricohioidoepiglotopexia	Audio Tape recorder Sony 60ES y un micrófono condensador colocado a 20 cm de los labios

Tabla 2: Comparacion de tipos de muestras

Trabajo:	**Idioma:**	**Tipo de muestra:**
Rusz [4]	Checo	Vocal sostenida de |*i*| en un tono y volumen cómodo, constante y mayor tiempo posible (por lo menos 5 s). Repetición de las silabas |*pa*|, |*ta*| |*ka*|, tan constantes y tantas veces como sea posible (por lo menos 5 veces). Vocales sostenidas de |*a*|, |*i*| y |*u*| en una respiración con un tono y volumen cómodo (en aproximadamente 5 s). Leer un mismo texto de 136 palabras fonéticamente no balanceado. Monólogo de aproximadamente 90 s donde el locutor expresa situaciones de su trabajo, etc,
Hillman [5]	Inglés	Vocal sostenida de |*i*| en un tono y volumen cómodo, constante y por lo menos 4 s
Michaelis [6]	Aleman	La vocal en alemán $\|\varepsilon :\|$ sostenido en un tono y volumen cómodo durante varios segundos.
Crevier-Buchman [9]	Inglés	Vocal sostenida de |*a*| en un tono y volumen cómodo, constante y por lo menos 3 s

4. Metodología

Se invitó a participar en la investigación a individuos aparentemente sanos y en forma voluntaria, que se encontraban en la Facultades de Ingeniería y Medicina de la Universidad de Carabobo.

Para esta investigación se obtuvo la colaboración del Especialista en Otorrinolaringología Prof. Reinaldo Sanchez de la Escuela de Medicina de la Universidad de Carabobo, en la evaluación médica de algunos de los individuos que participaron en esta investigación. Hay que resaltar que la

aparición de patologías en la evaluación médica no indica afección en la voz.

Para la elaboración de la parte de la historia clínica que contiene los datos culturales y personales de los individuos, sometidos a la investigación, se colocó un cuestionario electrónico en una computadora portátil, previo a esto fueron seleccionadas las preguntas de tal manera que pudiesen dar indicios del funcionamiento de su oído y su voz.

En esta investigación se utilizaron los fonemas |*a*| , |*e*|, |*i*|, |*o*| y |*u*| en español en un tono uniforme, cómodo y de volumen constante. La grabación de estos fonemas se realizó tres veces cada uno, a satisfacción del locutor y del técnico que realiza la grabación, y con períodos de descanso entre una grabación y otra. La grabación se realizó en una habitación insonorizada en el Centro de Análisis y Tratamiento de Señales de la Facultad de Ingeniería.

Se utilizó una parte de la oración padre nuestro, conocida por los individuos a grabar y se le indicó que deberían decir en un tono de conversación y volumen normal. La parte de la oración es la siguiente:

"Padre nuestro, que estás en los cielos, santificado sea tu nombre, venga a nosotros tu reino, hágase tu voluntad, así en la tierra como en el cielo."

La grabación de la parte de la oración se realizó también tres veces con períodos cortos de descanso entre oración y oración.

4.1. Equipos utilizados

Los equipos utilizados son los siguientes:

1. Micrófono Samson Q8 : patrón de captación súper cardioide, respuesta de frecuencia de 50 Hz a 16 kHz , sensibilidad -52 dBV/pa (2.5mV/Pa) e impedancia 300Ω
2. Mezclador profesional behringer xenyx 1204 USB: salida digital vía puerto USB, relación señal/ruido 110 dB, distorsión armónica total 0,005 y respuesta en frecuencia de 10Hz - 150KHz (-1 dB)
3. Computadoras portátiles: Lenovo Ideapad Y450 y Hp Pavilion dv4.

4.2. Procedimientos

A los individuos que participaron en la investigación se le solicitó la información personal y cultural con ayuda de un cuestionario electrónico, en una computadora portátil y se registró la información en una base de datos. La información solicitada fue la siguiente: sexo, profesión, fecha de nacimiento, país de nacimiento, lengua nativa, estado de nacimiento, ciudad de nacimiento, raza, fecha de grabación, ¿canta?, ¿fuma?, el consumo de cajas de cigarrillo si fuma, ¿está conforme con su voz?, ¿alguna persona le ha manifestado disconformidad con su voz?, ¿escucha bien?, ¿con que oído atiende el teléfono?, ¿alguna persona le ha indicado que escucha musica con volumen alto?, ¿reconoce su voz grabada? y ¿usa audífono para escuchar musica?

Como segunda actividad pasa el individuo a la sala insonorizada y el técnico le ajusta el micrófono a una distancia entre 15 cm a 20 cm de su boca. El técnico le solicita el número de cédula al individuo para que el sistema constate que sus datos están registrados en la base de datos. Si los datos no están registrados se le indica que conteste el cuestionario y regrese posteriormente a la sala insonorizada.

El individuo también es instruido en las acciones que podrían dañar la grabación a fin de evitarlo y como se desarrolla el proceso de grabación.

El técnico procede a realizar las grabaciones, apoyado por un programa de computación desarrollado en Python que evita la posibilidad de error en la secuencia y etiquetado de la grabación. Las grabaciones son realizadas en formato WAV con una frecuencia de muestreo de 48 kHz.

Al finalizar la grabación el técnico reproduce la grabación de la oración del individuo y le pregunta si reconoce su voz en la grabación. Si la respuesta a la pregunta ¿reconoce su voz en la grabación? no concuerda con lo que contesto en el cuestionario se modifica su respuesta en la base de datos.

Los individuos que fueron evaluados por el especialista en Otorrinolaringología de acuerdo a la disponibilidad del servicio, fueron auscultados y se completó su historia médica. La evaluación constó de los siguientes exámenes: faringoscopia,

rinoscopia anterior, rinoscopia posterior, laringoscopia y otoscopia.

Posteriormente el especialista en Otorrinolaringología vacía en la base de datos la información de la evaluación médica en la historia del paciente. Esto lo hace a través de un programa desarrollado en Python que disminuye la posibilidad de error de transcripción a la base de datos.

5. Resultados

5.1. Grabaciones

Las cantidades de grabaciones obtenidas se encuentran en la Tabla 3, discriminadas en el intento por ejemplo **a1** el fonema |*a*| en primer intento, la abreviatura **fr2** corresponde a la oración en su segundo intento y las cantidades en **Examinados** representan las cantidades de la grabación que corresponden a evaluación por otorrino del individuo que aporto su voz.

Tabla 3: Cantidades de grabaciones

Sexo:	**Grabación:**	**Total:**	**Examinados:**	**No_examinados:**
F	a1	52	33	19
F	a2	53	34	19
F	a3	53	34	19
F	e1	53	34	19
F	e2	53	34	19
F	e3	53	34	19
F	fr1	53	34	19
F	fr2	53	34	19
F	fr3	52	33	19
F	i1	53	34	19
F	i2	53	34	19
F	i3	47	28	19
F	o1	53	34	19
F	o2	53	34	19
F	o3	53	34	19
F	u1	53	34	19
F	u2	53	34	19
F	u3	53	34	19
M	a1	79	46	33
M	a2	79	46	33
M	a3	79	46	33
M	e1	78	46	32
M	e2	78	46	32
M	e3	78	46	32
M	fr1	80	47	33
M	fr2	79	47	32
M	fr3	78	45	33
M	i1	76	46	30
M	i2	76	46	30
M	i3	70	40	30
M	o1	79	46	33

Continúa en la próxima página

Tabla 3 – *Continuación cantidades de grabaciones*

Sexo:	**Grabación:**	**Total:**	**Examinados:**	**No_examinados:**
M	o2	78	46	32
M	o3	79	46	33
M	u1	78	46	32
M	u2	78	46	32
M	u3	78	46	32

Cada una de las grabaciones fue etiquetada con la información de edad, sexo, índice del individuo y de visita de acuerdo a base de datos y tipo de grabación. Las edades correspondiente a los individuos masculinos es de 16 a 25 años y la edad de las femeninas es de 15 a 26 años. Los individuos que prestaron su voz a esta investigación son venezolanos y su lengua nativa es el español.

5.2. Evaluación médica

Los resultados de la evaluación médica de los individuos sometidos a está investigación presentaron las siguientes afecciones: amigdalitis, sindrome temporomandibular, desviación del septum nasal, rinitis, hipertrofia de cornetes, mala oclusión dentaria, asma, voz nasal, tapón de cerumen en oído, sindrome vertiginoso, otomicosis, cofosis, entre otras. Hay que resaltar que individuos sanos después del examen del otorrino fueron muy escasos (11 casos) y que muchas personas tenían más de una afección.

6. Conclusión

Del análisis de resultados se observa que aunque cuando se realizo la invitación a participar como sujetos de estudio a estudiantes con la presunción de que se trataba de individuos sanos, en la evaluación realizada por el especialista en otorrinolaringología el individuo diagnosticado **Normal** solo se obtuvo 11 veces lo que llama poderosamente la atención, ya que se auscultaron 81 individuos femeninos y masculinos que se presumían sanos. Es de hacer notar que el que aparezcan diagnósticos de patologías no necesariamente estos influyen en la calidad de la voz por que pueden ser asintomáticos.

El número de grabaciones obtenidas permiten comenzar a realizar pruebas objetivas y perceptivas sobre las grabaciones obtenidas a fin de asegurar la calidad de las mismas.

El poder contar con grabaciones en el rango de edades de 15 a 26 años discriminadas en sexo (femenino y masculino) servirá de base a investigaciones futuras.

Considerando la influencia del aspecto cultural del español hablado por venezolanos en la selección de fonemas, luego del estudio de lo fonemas grabados 'a', 'e', 'i', 'o' y 'u' en un tono uniforme, cómodo y de volumen constante, se podría establecer la mejor selección de muestra o muestras para un procedimiento de evaluación objetiva de la calidad de señales de voz del español hablado por venezolanos.

La comparación de los fonemas grabados con la grabación realizada de la oración, en un tono de conversación y volumen normal, nos da la oportunidad de poder valorar su aporte en la clasificación objetiva de la voz.

La información del diagnóstico del especialista en otorrinolaringología puede ser un factor importante en la obtención de patrones para establecer la calidad objetiva de la voz.

Por último la presencia de tres intentos en cada una de las grabaciones nos permite la comparación entre las grabaciones del mismo individuo y/o la de otros individuos.

7. Recomendaciones

Se requiere seguir obteniendo grabaciones para aumentar la cantidad de las mismas.

Se requiere que se incluya la participación de especialistas en el área de voz como foniatras,

entre otros, para facilitar la obtención de patrones de calidad objetiva de la voz. Además hay que lograr incrementar la cantidad de personal especializado que trabaja en el Centro de Análisis y Tratamiento de Señales a fin de alcanzar resultados más rápidamente.

Es factible utilizar las instalaciones médicas con que consta la Universidad de Carabobo para enfocar la búsqueda en individuos con patologías para realizar las grabaciones de los mismos, ya que el énfasis en este trabajo se enfocó en individuos sanos.

Agradecimientos

Hay que realizar un enorme agradecimiento al Prof. Reinaldo Sanchez por su invalorable colaboración en esta investigación.

Y no menos importante el agradecimiento al C.D.C.H. por el apoyo financiero de este trabajo.

Referencias

[1] J. C. Casado M. and José A. Adrián T. *La evaluación clínica de la voz: fundamentos médicos y logopédicos*. Ediciones Aljibe, 2002.

[2] J. Kreiman, D. Vanlancker-Sidtis, and B.R. Gerratt. Defining and measuring voice quality. In *ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis*. Citeseer, 2003.

[3] A.A. Dibazar, S.Ñarayanan, and TW Berger. Feature analysis for automatic detection of pathological speech. *Engineering Medicine and Biology Symposium02*, 1:182–183, 2002.

[4] J. Rusz, R. Cmejla, H. Ruzickova, and E. Ruzicka. Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated parkinson's disease. *The Journal of the Acoustical Society of America*, 129(1):350–367, 2011.

[5] R.E. Hillman, T.F. Quatieri, D.D.D. Mehta, et al. *Impact of human vocal fold vibratory asymmetries on acoustic characteristics of sustained vowel phonation*. PhD thesis, Massachusetts Institute of Technology, 2010.

[6] D. Michaelis, M. Fröhlich, and H.W. Strube. Selection and combination of acoustic features for the description of pathologic voices. *The Journal of the Acoustical Society of America*, 103:1628, 1998.

[7] I. R. Titze and W.S. Winholtz. Effect of microphone type and placement on voice perturbation measurements. *Journal of Speech and Hearing Research*, 36(6):1177, 1993.

[8] A.P. Vogel and P. Maruff. Comparison of voice acquisition methodologies in speech research. *Behavior research methods*, 40(4):982–987, 2008.

[9] L. Crevier-Buchman, O. Laccourreye, F.L. Wuyts, M.C. Monfrais-Pfauwadel, C. Pillot, and D. Brasnu. Comparison and evolution of perceptual and acoustic characteristics of voice after supracricoid partial laryngectomy with cricohyoidoepiglottopexy. *Acta oto-laryngologica*, 118(4):594–599, 1998.