Detección automática de patologías laríngeas mediante análisis de la voz

Rufiner Hugo L., Martínez César E.

Laboratorio de Cibernética - Facultad de Ingeniería, Bioingeniería
Universidad Nacional de Entre Ríos

Resumen

En la valoración de la calidad de la voz, el análisis acústico está siendo aplicado cada vez más como herramienta de apoyo diagnóstico. Se describe la implementación de un sistema para detección de patologías de la voz utilizando el análisis de las señales en el dominio frecuencial. Los datos obtenidos sirven de alimentación a redes neuronales que clasifican la voz en tres grupos: normal y dos tipos de patologías. Se obtienen altos grados de reconocimiento, lo que indica que este tipo de análisis provee una caracterización de la voz en condición patológica de manera directa y no invasiva.

Análisis de la voz • Voces patológicas • Redes neuronales

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner & C. E. Martínez; "Detección automática de patologías laríngeas mediante análisis de la voz"
Anales del XII Congreso Argentino de Bioingeniería, Vol. 1, pp. 55-58, jun, 1999.

1. Introducción

Se sabe que la presencia de patologías en las cuerdas vocales causa cambios significativos en los patrones vibratorios normales de las mismas, lo cual impacta en la calidad resultante de la producción de voz.

Los problemas en la producción de la voz pueden surgir a partir de **[1,2]**: 1) desordenes funcionales (debido al abuso o mal uso del sistema vocal anatómica y fisiológicamente intacto), los cuales son corregidos mediante técnicas de terapia de voz; ó 2) patologías laríngeas (nódulos de cuerdas vocales, pólipos, úlceras, carcinomas y parálisis del nervio laríngeo), las cuales pueden ser corregidas mediante terapia de voz, cirugía y, en algunos casos, radioterapia.

Existen diversos procedimientos de rutina para examinación de la laringe con propósitos clínicos o de investigación, los cuales incluyen laringoscopía fibroscópica rígida y flexible (entubación del paciente con un instrumento de fibra óptica), videoestroboscopía (iluminación estroboscópica de la laringe, útil para la visualización de movimientos), electromiografía (observación indirecta del estado funcional de la laringe) y videofluoroscopía (técnica radiográfica en la cual se hacen ingerir al paciente distintos bolos de bario, como sustancia de contraste).

Sin embargo, existen también limitaciones para llevar a cabo estos estudios, como ser: 1) la laringe está situada fuera de la vista, en un lugar profundo del cuello; 2) el interior de la laringe es obscuro y debe ser adecuadamente iluminado para la examinación; 3) los movimientos de las cuerdas vocales durante la fonación son demasiados rápidos para ser capturados por un sistema óptico convencional.

En los últimos años ha crecido el interés por el análisis acústico de voces normales y patológicas como método alternativo de diagnóstico. En este tipo de estudio se aplican diferentes técnicas de procesamiento de señal (procesamientos temporales y frecuenciales para la obtención de parámetros acústicos) **[3]**, y extracción de la señal glótica de presión sonora mediante modelos de la dinámica del aire en la laringe (a partir de filtros inversos, análisis de predicción lineal, etc.) **[4,5]**.

El análisis acústico demuestra ventajas sobre los estudios anteriormente citados debido a su naturaleza no invasiva y a su potencial para proveer datos cuantitativos acerca del estado clínico de las funciones de la laringe y del tracto vocal, con adecuados tiempos de análisis.

En el campo del *Reconocimiento Automático de Patologías del Aparato Fonador* (RAPAF) se han utilizado diversas arquitecturas de *redes neuronales artificiales* (RNA), como así también modelos matemáticos del tracto vocal y la laringe **[3,5]**.

Las RNA son excelentes sistemas de clasificación y se especializan en trabajar con datos ruidosos, incompletos, solapados, etc. El reconocimiento de patrones de voces patológicas es una tarea de clasificación de datos que tiene todas estas características, haciendo a las RNA una alternativa atractiva a la aproximación descripta **[6]**.

2. Material y Métodos

2.1 Acerca de los datos

Los patrones para entrenamiento de las redes fueron obtenidos a partir de grabaciones de voces de personas con fonación normal y pacientes con patologías del aparato fonador. Cada registro es una grabación de la fonación sostenida de una vocal o un fonema vocálico. El uso de un estímulo de tipo vocal tiene ciertas ventajas. Primero, las vocales aisladas son usadas en la rutina de la práctica clínica para evaluación de la calidad de voces patológicas. Segundo, las medidas objetivas son relativamente directas, comparadas con el discurso continuo. Además, permiten una separación fácil y efectiva entre voces normales y patológicas **[3]**. El estudio del habla continua es un objetivo superior y un paso próximo evidente. Sin embargo, se requieren primero

resultados válidos basados en estímulos de menor complejidad.

2.2 Organización de los datos

Se crearon diferentes juegos de patrones para realizar los experimentos con redes neuronales. Cada juego de patrones estaba compuesto por 8 archivos: 4 archivos de patrones de voz normal y 4 archivos de patrones de voz patológica.

Cada uno de los conjuntos mencionados de 4 archivos se componen de la siguiente manera:

- 2 archivos para entrenamiento, con aproximadamente 500 patrones cada uno (un archivo de voz masculina y un archivo de voz femenina).
- 2 archivos para prueba, con aproximadamente 150 patrones cada uno (un archivo de voz masculina y un archivo de voz femenina).

2.3 Datos de voz normal

Las señales de voz normal fueron obtenidas del corpus de voz continua TIMIT **[7]**. Esta base de datos ha sido confeccionada en forma conjunta por Texas Instruments (TI) y el Massachusetts Institute of Technology (MIT). Es una de las bases multi-hablante más empleadas en el ámbito del Reconocimiento Automático del Habla (RAH) del discurso continuo por ser la más grande, completa y mejor documentada de su tipo. Esta base o corpus posee una gran cantidad de fonemas en diversos ambientes y pronunciados por más de 600 hablantes diferentes. El corpus TIMIT incluye la señal de voz correspondiente a cada oración hablada, así como también las transcripciones ortográficas, fonéticas y las palabras alineadas temporalmente.

De la oración SA1.WAV, del conjunto original de oraciones de entrenamiento, se extrajeron las porciones de señal que contienen el fonema /aa/. Al crear los distintos juegos de patrones, los hablantes se seleccionaron al azar entre las regiones dialécticas DR1 a DR8, de tal manera de tener representada una amplia variedad de dialectos y no repetir los patrones para el entrenamiento de distintas redes.

2.4 Datos de voz patológica

Se obtuvieron señales de una biblioteca de grabaciones de voces tomadas en VA Hospital (West L.A.) por el equipo de investigación del *Speech Processing and Auditory Perception Laboratory*, UCLA. Las señales fueron grabadas con un micrófono miniatura montado sobre la cabeza AKG C410 colocado a 4 cm de los labios del paciente. Las señales fueron pasadas por un filtro pasabajos de 8 Khz, digitalizadas directamente a 20 Khz y muestreadas a 10 Khz. Un segmento de 1 segundo fue extraído de la porción media de cada grabación **[8]**.

Para el propósito de este trabajo, las señales fueron remuestreadas a 16 Khz para obtener la misma referencia temporal que las señales de voz de TIMIT.

La clasificación de las señales fue realizada por el equipo de investigación mencionado, siendo agrupadas en las siguientes categorías: voz ronca y aspirada-ronca: 11 archivos, voz bicíclica (también conocida como diplofonía o fonación bifurcada): 8, voz bicíclica ronca: 1, voz aspirada-forzada: 2 y voz ronca-forzada: 2.

En la figura 1 se muestra un segmento de señal de voz normal y voz patológica, en donde pueden apreciarse las diferencias temporales de ambas ondas, mientras que en la figura 2 se observan los espectros de las mismas. Una diferencia que se aprecia a simple vista es la aparición de componentes de alta frecuencia en el caso patológico.

2.5 Procesamiento de señales

Para la extracción de patrones se utilizó una ventana móvil de 256 muestras, con solapamiento de 128 muestras. En cada segmento se aplicó una ventana de Hamming y luego se obtuvieron los patrones extrayendo los primeros 16 coeficientes cepstrales **[4]**. Cada patrón se completó con la información de 1 y 0 según fuera la activación en las salidas deseadas de la red.

Otros procesamientos empleados fueron el Cepstra en escala Mel (16 coeficientes) y la transformada rápida de Fourier (FFT) de 128 coeficientes.

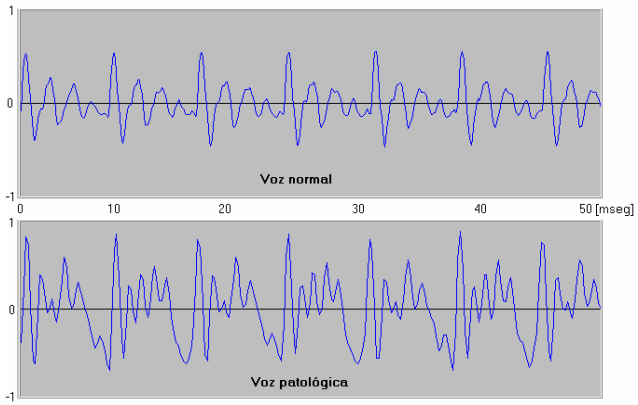

Figura 1: señales temporales correspondientes a fonema vocálico /aa/, de magnitud normalizada. Arriba: voz normal. Abajo: voz patológica.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner & C. E. Martínez; "Detección automática de patologías laríngeas mediante análisis de la voz"
Anales del XII Congreso Argentino de Bioingeniería, Vol. 1, pp. 55-58, jun, 1999.

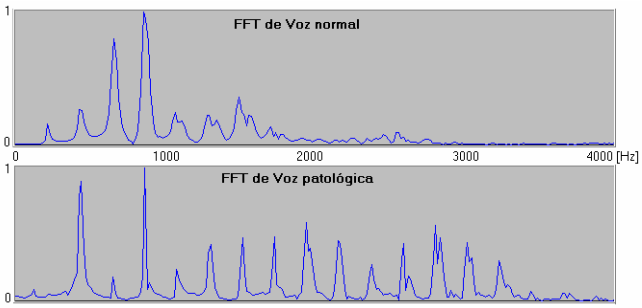

Figura 2: espectros de magnitud correspondientes a los segmentos temporales de la Figura 1.

2.6 Redes neuronales artificiales

Las RNA intentan simular, al menos parcialmente, la estructura y funciones del cerebro y sistema nervioso de los seres vivos. Una RNA es un sistema de procesamiento de información o señales compuesto por un gran número de elementos simples de procesamiento, llamados *neuronas artificiales* o simplemente *nodos*. Dichos nodos están interconectados por uniones directas llamadas *conexiones* y cooperan para realizar procesamiento en paralelo con el objetivo de resolver una tarea computacional determinada **[9]**.

Si bien los patrones a clasificar son dinámicos, estos tienen naturaleza estacionaria debido a que las muestras se toman de fonemas vocálicos pronunciados en forma sostenida, como se mencionó anteriormente. Esto hace innecesario el uso de *redes neuronales con retardos temporales* (RNRT) **[10,11]**, por lo que en nuestro trabajo se utiliza un *Perceptrón multicapa* (PMC) de una capa oculta. No existe un límite para fijar la cantidad de capas de un PMC, pero se ha demostrado que un PMC con una capa oculta y con un número suficiente de nodos es capaz de solucionar casi cualquier problema.

Para entrenar el PMC se utilizó el algoritmo de retropropagación **[9]**. Las entradas se normalizaron para cada dimensión del patrón en forma independiente. El entrenamiento se detuvo en el pico de generalización medido con respecto al archivo de prueba. Cada red se entrenó tres veces con los mismos patrones pero cambiando la semilla de inicialización aleatoria, reportándose el mejor de los resultados.

Para los experimentos con Cepstra y Mel Cepstra, el número de neuronas en la capa oculta se fijó en 50 luego de realizar una serie de experimentos cuyos resultados se muestran en la figura 3. Se debe destacar que debido a que los patrones obtenidos mediante FFT poseen dimensión 128, las redes utilizadas para su clasificación poseían diferente cantidad de neuronas en la capa oculta, que se fijó en 100 por un método similar.

Se trabajó con dos tipos de redes diferentes: una entrenada para distinguir entre voz normal y patológica (sin importar la patología); y otra para distinguir entre voz normal, bicíclica y ronca.

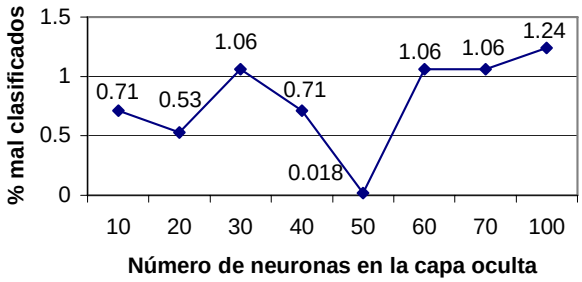

Figura 3: Porcentaje mínimo de patrones mal clasificados durante el entrenamiento.

3. Resultados

En las tablas 1 y 2 se presentan los resultados de los experimentos realizados para las redes de dos y tres clases respectivamente. Se muestra el porcentaje de frames clasificados correctamente para el archivo de entrenamiento (TRN) y prueba (TST).

TABLA 1: RESULTADOS PARA LAS REDES DE DOS CLASES

	% FRAMES BIEN CLASIFICADOS	
	TRN	TST
CEPSTRA	97.14	86.58
MEL CEPSTRA	91.30	85.50
FOURIER	61.00	63.33

TABLA 2: RESULTADOS PARA LAS REDES DE TRES CLASES

	% FRAMES BIEN CLASIFICADOS	
	TRN	TST
CEPSTRA	91.42	81.90
MEL CEPSTRA	81.05	77.70
FOURIER	48.37	46.86

4. Conclusiones

En este trabajo se presenta una alternativa para el diagnóstico automático de patologías laríngeas, basada en la extracción de características acústicas de la señal de voz y la clasificación de estos patrones mediante redes neuronales estáticas. Como se puede apreciar de las tablas 1 y 2, el análisis cepstral es el preprocesamiento que logra los desempeños más altos. Esto puede deberse a que la información relevante que permite realizar la distinción entre patologías se halla en la envolvente del espectro de magnitud de la señal, la cual se encuentra contenida en los primeros coeficientes cepstrales. En el caso de Mel

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner & C. E. Martínez; "Detección automática de patologías laríngeas mediante análisis de la voz"
Anales del XII Congreso Argentino de Bioingeniería, Vol. 1, pp. 55-58, jun, 1999.

Cepstra, la integración por bandas puede afectar esta información, mientras que en el caso de Fourier el aumento en la cantidad de dimensiones de los patrones hace más difícil la tarea de entrenamiento de la red y posiblemente más propensa a caer en mínimos locales.

La tarea de separación en tres clases obtuvo un desempeño menor debido a la mayor dificultad de la misma. Sin embargo, existe la posibilidad de entrenar diferentes redes para cada patología, lo que permitiría aumentar el número total de clases sin perder demasiada precisión, e inclusive agregar patologías sin necesidad de reentrenar totalmente el sistema.

5. Discusión

El desempeño obtenido en la clasificación hace promisoria la aplicación de esta alternativa como herramienta de apoyo para el diagnóstico de patologías del aparato fonador, siendo posible inclusive que cada profesional médico cree una base de datos propia con las patologías que sean de su interés o cuya incidencia sea mayor en su zona de influencia.

Entre las aplicaciones a explorar se encuentra la posibilidad de realizar mediante esta técnica no solamente un diagnóstico o análisis cualitativo, sino también un análisis cuantitativo en base a la cantidad de frames correctamente clasificados para un archivo de patrones de clase única. Esto permitiría, por ejemplo, seguir la evolución de alguna terapia de rehabilitación o medicación.

6. Referencias

[1] Paul W. Flint, Charles W. Cummings, *"The John Hopkins Center for Laryngeal and Voice Disorders"*. [http://www.med.jhu.edu/voice/index.html]. Department of Otolaryngology-Head & Neck Surgery, John Hopkins University. Baltimore, Maryland, August 1997.

[2] James A. Koufman, Gregory N. Postma, *"Center for Voice Disorders"*. [http://www.bgsm.edu/voice]. Wake Forest University, November 13, 1998.

[3] B. Boyanov, S. Hadjitodorov, "Acoustic analysis of pathological voices", *IEEE Engineering in Medicine and Biology*, pp. 74-82, july/august 1997..

[4] J. R. Deller, J. G. Proakis, J. H. L. Hansen, *Discrete-Time Processing of Speech Signals*, Macmillan Series. Englewood Cliffs, NJ: Prentice-Hall, 1993.

[5] John H. L. Hansen, "A nonlinear operator-based speech feature analysis method with application to vocal fold pathology assessment", *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 3, pp. 300-312, 1998.

[6] Judith A. Markowitz, *Using speech recognition*, Prentice-Hall, NJ, 1996.

[7] Garofolo, Lamel, Fisher, Fiscus, Pallett, Dahlgren, *DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus Documentation*, National Institute of Standards and Technology, February 1993.

[8] A. Alwan, P. Bangayan, J. Kreiman, and C. Long, "Time and Frequency Synthesis Parameters for Severe Pathological Voice Qualities", *Proc. of ICPhS*, Stockholm, Sweden, Vol. 2, 250-253, August 1995

[9] Mohamad H. Hassoun, *Fundamentals of Artificial Neural Networks*, The MIT Press, 1995.

[10] J.L. Elman, "Finding structure in time", *Cognitive Science* 14 (1990) 179-211.

[11] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, K. Lang; "Phoneme Recognition Using Time-Delay Neural Networks". *IEEE Trans. ASSP* Vol. 37, No 3 (1989).

Dirección para Correspondencia: Laboratorio de Cibernética – Facultad de Ingeniería (UNER). Ruta 11 Km.10 – Oro Verde (Paraná), Entre Ríos.

Correo electrónico:

Hugo L. Rufiner: lrufiner@arcride.edu.ar

César E. Martínez: cesarmart@arnet.com.ar

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
H. L. Rufiner & C. E. Martínez; "Detección automática de patologías laríngeas mediante análisis de la voz"
Anales del XII Congreso Argentino de Bioingeniería, Vol. 1, pp. 55-58, jun, 1999.