DETECCIÓN AUTOMÁTICA DE PATOLOGÍA POR ABUSO VOCAL MEDIANTE MODELOS ESTADÍSTICOS DE MEZCLAS DE GAUSIANAS

Juan I. Godino Llorente

Departamento de Ingeniería de Circuitos y Sistemas
EUITT. Universidad Politécnica de Madrid
igodino@ics.upm.es

Santiago Aguilera Navarro

Departamento de Ingeniería Electrónica
ETSIT. Universidad Politécnica de Madrid
aguilera@die.upm.es

Pedro Gómez Vilda

Departamento de Arquitectura y Tecnología de Sistemas Informáticos
FI. Universidad Politécnica de Madrid
pedro@pino.datsi.fi.upm.es

ABSTRACT

There is an increasing risk of vocal and voice diseases due to the modern way of life. It is well known that most of the vocal and voice diseases cause changes in the acoustic voice signal. These diseases have to be diagnosed and treated at an early stage. Acoustic analysis could be a useful tool to detect this kind of diseases. In this paper, we apply GMMs to the automatic detection of voice disorders. Former and actual works demonstrate that impaired voice detection can be carried out by means of supervised neural nets [5]: MLP (Multilayer perceptron). We have focused our task in detection of impaired voices by means of gaussian mixture models (GMMs) and parameters such MFFC and LPCC extracted from the voice signal. Results are comparable to those obtained with MLP.

1. INTRODUCCIÓN

La realidad presente del análisis acústico permite elaborar múltiples mediciones de parámetros vocales para cada uno de los pacientes a estudiar (jitter, shimmer, HNR, NNE...[3]). No existen estudios poblacionales rigurosos que hayan fijado los valores de normalidad para cada uno de estos parámetros en los distintos grupos de edad y sexo.

El objetivo final del trabajo es detectar de forma automática la presencia de patología vocal a partir del estudio de la señal de voz haciendo uso de técnicas de tratamiento digital de señal.

Por su naturaleza, nos centramos en las patologías de raíz orgánica que son las que afectan el registro de voz de manera directa. Este tipo de patologías se producen generalmente por "abuso vocal" y se manifiestan como una modificación en la morfología del elemento excitador, produciéndose un aumento de masa en las cuerdas que da lugar a una vibración menos regular (en los tramos sonoros).

2. MATERIAL DE PARTIDA

Utilizamos como material de partida los registros de la base de datos [1] recopilada por el Massachusetts Eye and Ear Infirmary Voice and Speech Lab (MEEI). Las muestras han sido grabadas en un entorno controlado con frecuencias de muestreo de 25 KHz y una resolución de 16 bits.

Dado el carácter estacionario de un segmento sonoro de habla, y dado que el conjunto de patologías en estudio afectan al elemento excitador, estudiamos la fonación sostenida de una vocal. Las alteraciones en la flexibilidad de la voz son más fácilmente mensurables en un registro de fonación sostenida.

El número de registros de voz utilizados para las pruebas fue de 135 (53 voces normales normal y 72 patológicas). La asimetría observada se debe a que los registros de voces normales son de 5 segundos de duración, mientras que los correspondientes a voces patológicas son de 3 segundos. El tipo de patología contemplada en la base de datos es de tipo orgánico: pólipos, nódulos, quistes, edemas, cáncer... Los datos se dividieron en dos subconjuntos: 70% para entrenamiento, y 30% para test.

3. PARAMETRIZACIÓN

El objetivo de la parametrización es el de reducir la dimensionalidad de los datos o vectores de entrada intentando garantizar la máxima separabilidad.

La aproximación paramétrica utilizada, es la que se conoce como parámetros mel-cesptrales (MFCC.-Mel Frecuency Cepstral Coefficients). Realiza el cálculo de la transformada del coseno del logaritmo de la energía, calculada sobre ventanas frecuenciales con ancho de banda dependiente de la frecuencia central del filtro. Conforme aumenta la frecuencia en estudio se aumenta el ancho de banda de la ventana frecuencial. Este método se basa en el sistema de percepción humana estableciendo una relación logarítmica entre la escala de frecuencia real (*Hz*) y la escala de frecuencia perceptual (*mels*).

Podemos obtener una representación mejorada extendiendo el análisis para incluir información sobre la derivada temporal de los parámetros calculados. Usamos la primera y la segunda derivada temporal. La inclusión en el vector de rasgos de la derivada, nos da cierta idea de le evolución en el tiempo y de la inercia de los parámetros en estudio. Con esto deslocalizamos temporalmente el estudio.

La descripción completa de los métodos de cálculo de los parámetros enumerados, está recogida en [2].

4. GMMs

La motivación esencial para utilizar GMMs es la capacidad del modelo para representar una muy amplia gama de funciones distribución. Cada clase queda modelada mediante una v.a. cuya función densidad de probabilidad está caracterizada por una mezcla de *Q* gausianas, de la forma [4]:

$$p(x/\lambda) = \sum_{i=1}^{Q} c_i \cdot p_i(x), \quad \sum_{i=1}^{Q} c_i = 1, \quad c_i \geq 0 \qquad (1)$$

donde $p_i(x)$, $i=1,...,Q$ son las densidades componentes, y c_i, $i=1,...,Q$ son los pesos. Cada densidad componente es una función gausiana de dimensión n.

La mezcla queda caracterizada por los vectores media, las matrices de covarianza, y los pesos de las mezclas, calculados de forma iterativa mediante el algoritmo de maximización de la media (algortimo EM.-Expectation Maximization) [4].

5. METODOLOGÍA

La Figura 1 muestrea un diagrama de bloques que describe gráficamente el procesado anterior a la clasificación. La señal analógica es filtrada y digitalizada, después de lo cual es enventanada utilizado ventanas tipo Hanning de 1024 muestras. El siguiente módulo es un detector de silencios y/o bordes con objeto de evitar modelar segmentos sin voz. Por ultimo tenemos el bloque de extracción paramétrica MFCC) y el clasificador basado en GMM. El esquema ha sido probado con y sin preénfasis de las señal ($1-0.95z^{-1}$).

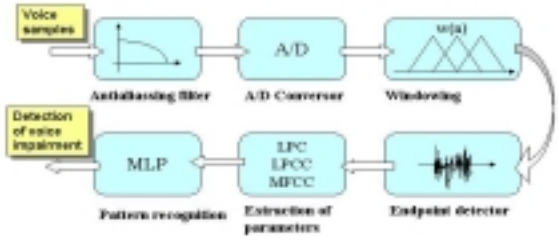

Figura 1: *Diagramas de bloques del módulo de procesado y detección.*

El clasificador está basado en GMMs, que caracterizaremos con un número de 128 gausianas. El modelo se inicializa mediante el algoritmo de cuantificación vectorial de las k-medias. El entrenamiento se lleva a cabo a lo largo de 20 iteraciones del algortimo, momento a partir del cual el error cometido con la aproximación de mezclas de gausianas varía por debajo del 1% en cada iteración. Las matrices de covarianza utilizadas para el modelo son de tipo diagonal.

6. EVALUACIÓN Y RESULTADOS

Con el conjunto de muestras se entrenan los modelos λ_c o $\lambda_{\bar{c}}$. Una vez entrenado el modelo se puede calcular el ratio de verosimilitudes para la secuencia de vectores de características de prueba X, que en el dominio logarítmico queda [4]:

$$\Lambda(X) = \log\left[p(X/\lambda_C)\right] - \log\left[p(X/\lambda_{\bar{C}})\right] \qquad (2)$$

El ratio de verosimilitudes se compara con un umbral θ. La voz será patológica si $\Lambda(X)>\theta$ y normal si $\Lambda(X)<\theta$. La Figura 2 muestra las curvas de falsa aceptación y falso rechazo vs. umbral del cociente de verosimilitudes. Ambas curvas cruzan en el punto de igual error (ERR.- Equal Error Rate).

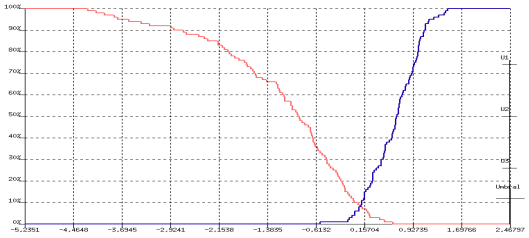

Figura 2: *Falsa aceptación (derecha) y falso rechazo (izquierda).*

La decisión final se toma en función del porcentaje de vectores que pertenecen a una clase u otra. La tabla muestra los resultados de EER para distinto número de parámetros utilizados. En la Tabla 1 se observa que los mejores resultados se obtienen a partir de la energía y 16 parámetros MFCC, junto con sus derivadas primera y segunda. En este caso el ERR es de 0% para el conjunto de datos de test utilizado. Aunque no se muestra en la tabla, utilizando preénfasis puede también conseguirse la detección, si bien el número de parámetros ha de aumentarse considerablemente.

	12	13	14	15	16	17	18
MFCC	46,3	48,1	30,8	47,5	25,7	0	0.6
MFCCp	46,6	49,6	50	42,8	33,9	43,42	32,8

Tabla 1: *Matriz de resultados usando parametrización MFCC y LPCC, con y sin preénfasis. Se muestra el EER.*

7. CONCLUSIONES Y TRABAJOS FUTUROS

Parametrización MFCC parece ser prometedora para la detección automática de patología laríngea. No obstante, hay que ser prudente con los resultados y evaluar la bondad del detector con una base de datos de mayor tamaño.

Se ha de buscar el óptimo de funcionamiento del detector, en cuanto a número de parámetros y número de mezclas.

Los resultados con GMM, son similares a los obtenidos con redes neuronales de tipo MLP [5].

Detectada la patología laríngea, el siguiente paso será tratar de distinguir entre un conjunto acotado de patologías (siempre dentro las pertenecientes al conjunto de origen orgánico).

8. REFERENCIAS

[1] "Disordered Voice Database", Version 1.03, Kay Elemetrics Corp, 1994

[2] "Fundamentals of speech recognition" L. Rabiner. Prentice Hall. 1993.

[3] "Clinical measurement of speech and voice" R.J. Baken. Taylor & Francis. 1993

[4] "Speaker identification and verification using Gaussian mixture seaker models" D. A. Reynolds. Speech Communication, 1995, pp 91-108

[5] JI. Godino-Llorente et al. "LPC, LPCC and MFCC parameterization applied to the detection of voice impairment". Proceedings of ICSLP'00, Beijing, China, 2000.