RECONOCIMIENTO DE EMOCIONES EMPLEANDO PROCESAMIENTO DIGITAL DE LA SEÑAL DE VOZ

Emotional recognition applying speech signal processing

RESUMEN
Se presenta en este trabajo una metodología para la caracterización de la señal de voz aplicada en el reconocimiento de estados emocionales. Los diferentes estados emocionales de un hablante producen cambios fisiológicos en el aparato fonador, lo que se ve reflejado en la variación de algunos parámetros de la voz. Las técnicas de procesamiento empleadas son: transformadas tiempo-frecuencia, análisis de predicción lineal y *raw data*.

PALABRAS CLAVES: Caracterización del la señal de voz, reconocimiento de emociones.

ABSTRACT
In this paper a methodology for extraction of features in emotional speech recognition is presented. Different emotional states of a speaker produce physiological changes in the human speech system, which is reflected in the variation of voice parameters. Processing techniques applied in this work are: time-frequency representations, linear prediction and raw data.

KEYWORDS: *Characterization of signal voice, emotional recognition.*

MAURICIO MORALES PÉREZ
Ingeniero Electricista
Estudiante de Maestría en Ingeniería Eléctrica
Universidad Tecnológica de Pereira.
mua2rmo@gmail.com

JULIÁN DAVID ECHEVERRY
Ingeniero Electrónico, M.Sc.
Profesor Auxiliar
Universidad Tecnológica de Pereira
jde@utp.edu.co

ÁLVARO ÁNGEL OROZCO
Ingeniero Electricista, M.Sc.
Profesor Titular
Universidad Tecnológica de Pereira.
aaog@utp.edu.co

1. INTRODUCCIÓN

Múltiples investigaciones han abordado el tema del procesamiento de la señal de voz con el fin de encontrar patrones que permitan la identificación del locutor, el reconocimiento de palabras o la generación de voz sintetizada. Pero sólo algunas investigaciones se han encaminado al reconocimiento de emociones a partir de la señal de voz [1].

La voz está directamente relacionada con los diferentes estados emocionales, ya que estos producen cambios a nivel fisiológico en el aparato fonador. Por ejemplo una palabra expresada con estado emocional triste presenta baja intensidad y corta duración en comparación con la misma palabra expresada con un estado emocional alegre o neutral.

Se plantea el desarrollo de una metodología que permita la extracción de características de la señal de voz a partir de las cuales se pueda determinar el estado emocional en el que se encuentra un hablante. Dicha metodología pretende ser implementada en dispositivos programables para trabajo en tiempo real, de modo que pueda identificar, reconocer y seguir los cambios emocionales. Entre otras aplicaciones serviría como una herramienta de ayuda en el diagnóstico y tratamiento de enfermedades psicológicas como los trastornos de ansiedad.
La señal de voz es un proceso aleatorio de carácter no estacionario, motivo por el cual las transformadas tiempo-frecuencia (TTF) se presentan como una buena alternativa de análisis, permitiendo una representación conjunta en el dominio del tiempo y la frecuencia [2].

2. CONTENIDO

2.1 Transformadas tiempo-frecuencia

A diferencia de los métodos convencionales las transformadas tiempo-frecuencia permiten una representación conjunta en los espacios del tiempo y la frecuencia, es decir, podemos conocer la dinámica de los contenidos espectrales a lo largo del tiempo [3], [4].
Para este trabajo se emplearon las transformadas Gabor, Wavelet discreta (DWT) y la distribución de *Wigner Ville* (WVD).

2.1.1 Transformada *Gabor*

La transformada Gabor consiste en la aplicación de una ventana $g(t)$ sobre una porción de la señal $s(t)$ permitiendo una aplicación local de la transformada de Fourier. De este modo, se releva la información en frecuencia localizada temporalmente en el dominio efectivo de la ventana. Desplazando temporalmente la ventana un intervalo τ se cubre el dominio de la señal obteniéndose una representación tiempo-frecuencia de la misma.

Fecha de Recepción: 07 Septiembre de 2007
Fecha de Aceptación: 30 Noviembre de 2007

$$s_g(\tau,\omega) = \int_{-\infty}^{\infty} s(t)g(t-\tau)e^{-j\omega t}dt \qquad (1)$$

2.1.2 Transformada Wavelet discreta

La idea básica de la transformada wavelet discreta es representar la señal por medio de los coeficientes de aproximación y detalle $\{a_j, d_j\}$ los cuales se obtienen al pasar la señal a través de filtros pasabanda [5],[6]. La descomposición por medio de la transformada Wavelet discreta se presenta en la figura 1.

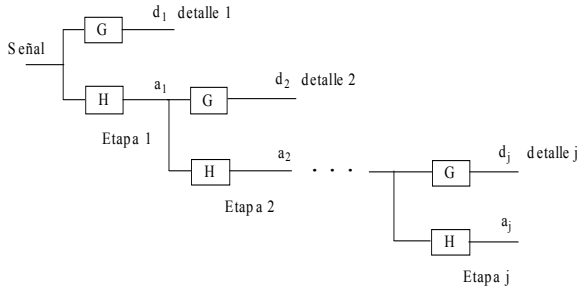

Fig 1. Descomposición multi-resolución por medio de la transformada wavelet discreta.

2.1.3 Distribución de *Wigner Ville*

La distribución de *Wigner Ville* es una transformación no lineal que permite determinar la energía de la señal en un intervalo de tiempo, se define como:

$$WV_S(t,f) = \int_{-\infty}^{\infty} s(t+\frac{\tau}{2})s^*(t-\frac{\tau}{2})e^{-j2\pi f\tau}d\tau \qquad (2)$$

Aunque la distribución de *Wigner Ville* presenta ventajas en cuanto a su buena resolución temporal involucra el inconveniente de la apariencia de términos cruzados $f(s_1(t).s_2(t))$. Esto debido a que la WVD es una distribución de energía, es decir, está relacionada con el cuadrado de la señal o lo que es lo mismo, el producto de la señal consigo misma. Aparecen dos tipos de componentes: autocomponentes y términos cruzados, pudiendo en algunos casos este último alcanzar valores superiores al de las autocomponentes haciendo el espectro prácticamente indescifrable [2],[7].

2.1.4 Análisis de predicción lineal (LPC)

Una de las técnicas más utilizada para la caracterización de la voz con base en un modelo predeterminado es el Análisis de Predicción Lineal. Esta metodología se basa en un modelo general de producción de voz de la forma *entrada - filtro - salida*, en el cual el tracto vocal es el filtro con función de transferencia *H(z)* cuyos parámetros varían en el tiempo en función de la acción que se realiza al pronunciar una palabra. Existen dos posibles señales de entrada para el filtro: sonora (tren de pulsos) o no sonora (ruido blanco), un esquema básico del modelo se presenta en la figura 2

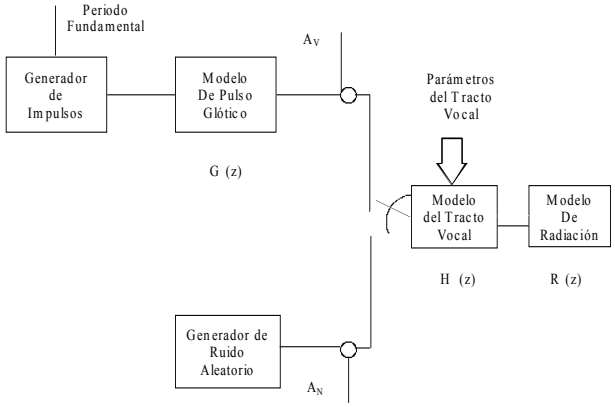

Fig 2. Diagrama básico modelo general de producción de voz.

Para el modelado del tracto vocal *H(z)* se emplea una función todo polos de la forma:

$$H(z) = \frac{G}{1-\sum_{i=1}^{N} a_i z^{-i}} \qquad (3)$$

Finalmente la señal queda representada por la estimación de los coeficientes a_i y la ganancia G [1].

2.1.5 *Raw Data*

Como su nombre lo indica (datos crudos), este tipo de análisis consiste en trabajar directamente sobre todo el universo de datos, es decir, la señal no es llevada a algún tipo de representación en la que se reduzca o transforme la información contenida en ella.

2.2 Extracción de características

La caracterización consiste en la obtención de parámetros, que de acuerdo a su relevancia (es decir, su importancia dentro de la señal) permitan de forma completa o parcial la descripción de la misma. Los principales objetivos de la caracterización son obtener una reducción en la dimensionalidad y realzar aspectos de la señal que contribuyan significativamente a realizar procesos posteriores (reconocimiento, segmentación o clasificación). En el análisis de la voz es común el empleo de dos tipos de características: las características acústicas, las cuales poseen un sentido físico determinado; y las características de representación, que corresponden a valores calculados a partir de alguna forma de representación de la voz, y a los cuales, en general, no les corresponde algún sentido físico [6]. En el caso de la voz, la extracción de características se realiza en intervalos de tiempo que comprenden entre 20 y 40ms, tramas de señal donde ésta se puede considerar como cuasi-estacionaria, es decir sus parámetros estadísticos permanecen invariantes dentro de la trama de observación [8].

2.2.1 Características acústicas

Como se mencionó anteriormente las características o parámetros acústicos son aquellos que poseen un significado físico, por lo que permiten una calificación de las cualidades vocales [9].

Los parámetros acústicos se pueden clasificar de la siguiente forma:

[1] *Parámetros cuasiperiódicos*: que reflejan las variadas formas de periodicidad presentes en la señal de voz: Frecuencia fundamental (F_0).

[2] *Parámetros de perturbación*: que reflejan una variación relativa de un parámetro determinado: Jitter, Shimmer, HNR.

Los parámetros de perturbación se calculan empleando la siguiente expresión:

$$V_P = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}|p(i+1)-p(i)|}{\frac{1}{N}\sum_{i=1}^{N-1}|p(i)|} \qquad (4)$$

Donde P_i es el valor del parámetro para la trama i, N es el número de tramas y V_P es el valor de la perturbación del parámetro.

2.2.2 Características de representación

Las características de representación se encargan de describir el comportamiento dinámico de la señal, son calculadas a partir de algún método de representación de señales (análisis de predicción lineal, transformadas tiempo- frecuencia), y generalmente no se les asocia algún sentido físico desde el punto de vista acústico.

En la figura 3 se presenta el espectrograma para la palabra /*experiencia*/ en estado emocional triste, calculado a partir de la transformada *Gabor*. A cada punto del gráfico se le asocia un valor, donde los colores más oscuros presentan un mayor contenido energético.

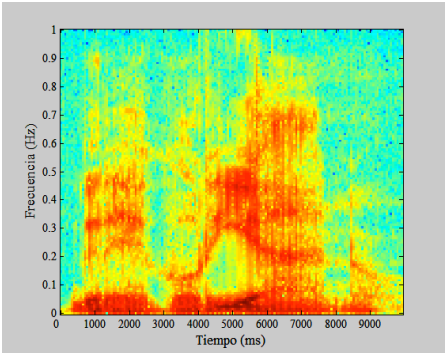

Fig. 3. Espectrograma de la palabra /*experiencia*/ en estado emocional triste calculado a partir de la transformada *Gabor*.

Generalmente se obtienen momentos estadísticos a partir de los datos de representación, estos momentos son utilizados como características de representación.

2.3 Marco experimental y resultados

En la figura 4 se presenta un diagrama reducido de la metodología implementada para el reconocimiento de emociones.

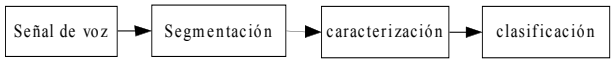

Fig. 4. Diagrama de flujo metodología implementada en el reconocimiento de emociones.

2.3.1 Descripción de la base de datos

Se trabajó sobre la Base de Datos SES[1], la cual es una base de datos monolocutor de habla emocional en español en la que el locutor, un actor profesional, simula habla ***triste, alegre, sorprendido, enfadado*** y ***neutro***. Consta de diversas sesiones de grabación, donde cada sesión contiene diversas palabras, frases o párrafos. Los ficheros de voz son archivos *.PCM, grabados a 16 kHz, 16 bits, sin cabecera y en formato Intel (little endian). En total consta de 30 palabras, 14 frases y 4 párrafos con cada uno de los estados emocionales mencionados. Este trabajo se realizó sobre las 30 palabras aisladas en cada uno de los diferentes estados emocionales.

En la tabla 1 se muestra la matriz de confusión de la identificación de emociones realizada por receptores humanos sobre la base de datos SES [9].

	Salida				
Entrada	**Alegre**	**Enojado**	**Neutro**	**Sorprendido**	**Triste**
Alegre	61.9%	7.9%	3.2%	11.1%	9.5%
Enojado		95.2%			
Neutro	3.2%	6.3%	76.2%	1.6%	7.9%
Sorprendido			3.2%	90.6%	1.6%
Triste	7.9%			4.8%	81.0%
Precisión	84.8%	87.0%	92.3%	83.8%	81.0%

Tabla 1. Matríz de confusiónidentificación de emociones realizada por receptores humanos sobre la base de datos SES

Se obtiene un acierto global de identificación de 80.89%

2.3.2 Caracterización de la señal de voz

Como no existe un conocimiento a priori de las características que van a proporcionar un mejor resultado en el reconocimiento de emociones, se consideró

[1] Esta base de datos es propiedad de la Universidad Politécnica de Madrid, Departamento de Ingeniería Electrónica, Grupo de Tecnología del Habla, ETSI Telecomunicación, Ciudad Universitaria, 28040 Madrid España.

adecuado contar un gran número de parámetros para posteriormente descartar aquellos que resultaran redundantes.

Como se dijo anteriormente, se plantean dos tipos de características a extraer en señales de voz: de representación y acústicas [10]. En primer lugar, las características de representación se extraen de las transformaciones que se aplican a la señal. Pueden ser vistas como parámetros estadísticos que aportan información de la naturaleza de la densidad espectral de energía de la señal. Las características acústicas pueden ser vistas como parámetros que aportan información sobre cualidades físicas de la voz (evolución de la tonalidad o variación de la amplitud).

En la figura 5 se presentan contornos de la frecuencia fundamental, calculada a partir de la función de autocorrelación [11] para la palabra /*coche*/ en diferentes estados emocionales, para este caso se pueden observar altos contenidos frecuenciales para los estados sorprendido y alegre y bajos para neutro y triste.

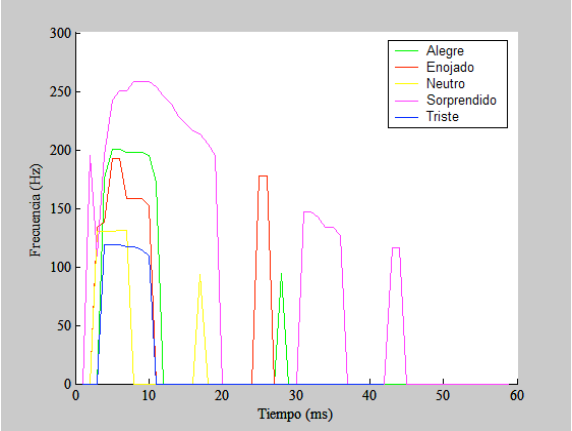

Fig 5. Contornos del pitch para la palabra /*coche*/ en diferentes estados emocionales.

Para todos los casos (excepto en Gabor que la ventana de análisis realiza una segmentación) la señal fue previamente segmentada con ventanas de 30ms de duración y un traslape del 50%, se promediaron los valores de cada uno de los parámetros extraídos por cada trama con lo que se obtuvo un vector de características promedio para cada señal. Se obtuvieron un total de 104 características agrupadas de acuerdo a la metodología usada en su extracción, un esquema general de la caracterización se muestra en la figura 6.

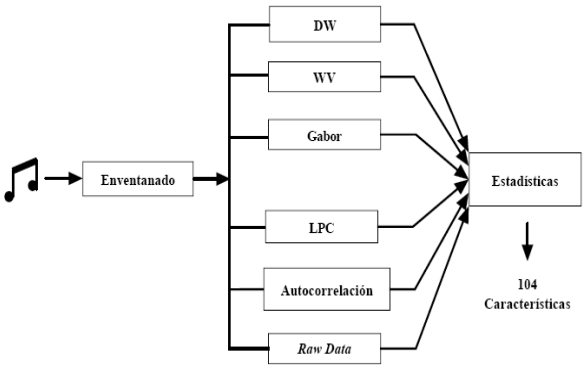

Fig 6. Diagrama general de caracterización

2.3.3 Validación de resultados

Los resultados de clasificación y reconocimiento pueden verse afectados por la falta de cuidado en el diseño de los experimentos. Con el fin de garantizar la validez estadística de los resultados, se emplea el método de validación cruzada *leave−one−out*, en el cual se emplea una sola muestra de todo el conjunto muestral como dato de validación y las muestras restantes como datos de entrenamiento. Este procedimiento se repite de tal manera que todas las muestras del espacio inicial sean empleadas como datos de validación una única vez. Los datos de validación y entrenamiento son llevados a un clasificador Bayesiano lineal donde se obtiene un porcentaje de acierto.

A modo de descartar las características que presenten redundancia se llevan al clasificador todas las posibles combinaciones del grupo de características, escogiendo el sub-grupo de éstas que presente mayor porcentaje de acierto.

2.3.4 Resultados

Se presentan las matrices de confusión y los respectivos porcentajes de acierto del reconocimiento de emociones para cada una de las metodologías utilizadas:

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	**Acierto (%)**
Alegre	15	3	2	8	2	50.00
Enojado	1	21	5	3	0	70.00
Neutro	1	1	25	0	3	**83.33**
Sorprendido	1	2	0	24	3	80.00
Triste	5	6	1	6	12	40.00
Acierto global						**64.66**

Tabla 2. Matríz de confusión empleando LPC

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	**Acierto (%)**
Alegre	14	6	7	3	0	46.66
Enojado	4	21	4	1	0	70.00
Neutro	4	0	26	0	0	**86.66**
Sorprendido	3	0	0	26	1	**86.66**
Triste	4	1	3	8	14	46.66
Acierto global						**67.33**

Tabla 3. Matríz de confusión empleando WVD

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	**Acierto (%)**
Alegre	22	2	1	2	3	**73.33**
Enojado	2	18	6	3	1	60.00
Neutro	2	7	20	1	0	66.66
Sorprendido	5	2	1	21	1	70.00
Triste	4	3	1	1	21	70.00
Acierto global						**68.00**

Tabla 4. Matriz de confusión empleando DWT

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	**Acierto (%)**
Alegre	22	3	2	2	1	73.33
Enojado	6	21	1	2	0	70.00
Neutro	0	2	25	1	2	**83.33**
Sorprendido	5	1	0	24	0	80.00
Triste	1	1	4	0	24	80.00
Acierto global						**77.33**

Tabla 5. Matriz de confusión empleando transformada Gabor

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	**Acierto (%)**
Alegre	25	2	0	1	2	83.33
Enojado	2	21	0	4	3	70.00
Neutro	3	2	23	0	2	76.66
Sorprendido	1	1	0	28	0	**93.33**
Triste	3	1	0	2	24	80.00
Acierto global						**80.66**

Tabla 6. Matriz de confusión empleando RAW DATA

A modo de obtener un mejor resultado se compone una nueva matriz de características a partir de las características discriminantes obtenidas con las técnicas anteriores, de nuevo se realizan todas las posibles combinaciones para descartar las características que presenten mayor redundancia. El proceso de discriminación se realiza nuevamente en base al mayor porcentaje de acierto en la clasificación. El resultado obtenido se presenta en la tabla 7.

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	**Acierto (%)**
Alegre	28	1	0	1	0	93.33
Enojado	1	27	2	0	0	90.00
Neutro	0	0	30	0	0	**100.00**
Sorprendido	0	0	0	30	0	**100.00**
Triste	2	0	0	1	27	90.00
Acierto global						**94.66**

Tabla 7. Matriz de confusión empleando método combinado

Las características a partir de las cuales se obtiene el resultado anterior se presentan en la tabla 8.

Características
Media de la curtosis de los coeficientes de la Wavelet Discreta
Shimmer del *Raw Data*
Mediana de la perturbación de la amplitud del *Raw Data*
Media de la mediana del *Raw Data*
Media de la desviación estándar del *Raw Data*
Media del mínimo del *Raw Data*
Varianza de la función de sonoridad
Media de la varianza de los LPC
Media de la curtosis de los LPC

Tabla 8. Características que presentaron mayor discriminancia

3. DISCUSIÓN

Los resultados de los trabajos de reconocimiento de emociones son difíciles de comparar, pues se utilizan bases de datos muy distintas. Algunos utilizan bases de datos multilocutor y otros monolocutor y el conjunto de emociones básicas consideradas no es el mismo en todos los casos.

Un trabajo destacable es el de Luengo y Navas (2005), en el cual se trabajó sobre una base de datos monolocutor de 97 grabaciones de habla emocional en euskara, con un conjunto de emociones básicas similares a las utilizadas en este trabajo. Se extrajeron características espectrales calculadas por medio de MFCC y parámetros prosódicos (con ayuda de un laringógrafo), las características fueron validadas con dos tipos de clasificadores *GMM* (*Gaussian Mixture Models*) y (*SVM*) máquinas de soporte vectorial. Se obtuvo un porcentaje de acierto global de 98.4% utilizando 512 componentes gaussianas y un porcentaje de 92.3% con la aplicación de un discriminador de características [12].

El trabajo más próximo al nuestro es el de J.M. Montero y J. Macías (2006) los cuales utilizaron la base de datos SES sobre la cual se calcularon características espectrales por medio de MFCC y parámetros prosódicos (con ayuda del software Praat), dichas características

fueron validadas por medio de un clasificador bayesiano lineal, obteniendo un porcentaje de acierto global de 94.66%.

4. CONCLUSIONES

Se desarrolló una técnica de caracterización de la señal de voz, basada en la utilización de características de representación y acústicas, la cual presentó buenos resultados en el reconocimiento de estados emocionales. Dicha técnica se validó por medio de un clasificador bayesiano lineal, dada la simplicidad del clasificador y su poca influencia dentro del resultado se garantizó la validez de la metodología propuesta para la extracción de características.

Al comparar la tablas I y VII se comprueba la efectividad de la metodología empleada superando con un 13.68% la identificación realizada por un grupo de oyentes humanos. Seria entonces factible una implementación de un sistema en tiempo, real que sirva como herramienta de apoyo a un especialista en tratamientos de enfermedades psicológicas como el estrés post-traumático.

Es importante destacar que la metodología de caracterización que presento mejores resultados individualmente fue el *Raw Data*. Los datos crudos contienen información sobre la intensidad, duración y pausas de la voz. Aunque pareciera poco óptimo trabajar sobre todo el espacio de datos, la capacidad y velocidad de los sistemas de cómputo actuales lo permiten, sin tener costos adicionales respecto a otras metodologías.

5. AGRADECIMIENTO

Este trabajo se desarrollo en el marco del siguiente proyecto de investigación: "IMPLEMENTACIÓN Y EFECTIVIDAD DE UN SISTEMA BASADO EN INTELIGENCIA ARTIFICIAL COMO HERRAMIENTA PARA EL TRATAMIENTO PSICOLÓGICO DE PERSONAS CON TRASTORNO DE ESTRÉS POSTRAUMÁTICO", financiado por Colciencias. Código 111037019600 y la Universidad Tecnológica de Pereira. Código 511-3-243-08.

6. BIBLIOGRAFÍA

[1] X. Huang, Spoken language processing. Prentice-Hall Inc, 2001.

[2] J. D. Echeverry, "Caracterización acústica de bioseñales empleando transformadas tiempo-frecuencia y modelado paramétrico," Master's thesis, Universidad Tecnológica de Pereira - U.T.P, 2006.

[3] L. Cohen, Time –Frequency Análisis. Prentince Hall Signal Processing Series, 1995.

[4] F. Hlawatsch and G. Bourdreaux-Bartels, "Linear and quadratic time-frequency signal representations," in signal processing Magazine, April 1992, pp.21-67.

[5] S. Mallat, Wavelet Tour of signal Processing. Boston: American Press, 1998.

[6] F. Vargas, "Selección de características en el análisis de voces," Master's thesis, Universidad Nacional de Manizales, 2003.

[7] H. I. Choi and W. J. Williams, "Improved time-frequency representation of multicomponent signals using exponential kernels," in Speech and Signal Processing, vol. 35, 1989, pp. 217-250,276-300 y 372-389.

[8] J. L. Mesa and P. Q. Morales, Codificación, Síntesis y Reconocimiento de Voz. España: Universidad de las Palmas de Gran Canaria, 1994.

[9] J. M. Montero and J. Macias, "Prosodic and segmental rubrics in emotion identification," Acoustic , Speech and Signal processinf, ICASSP 2006 Proceeding. IEEE Internacional Conference, 2006.

[10] P. Cortes y C. Suárez, "Indice de incapacidad vocal: Factores predictivos," Acta Otorrinolaringiol, vol. 57, pp. 101-108, 2006.

[11] F. Watkins, and P. López, "Implementación de un modelo digital de análisis y síntesis de voz empleando técnicas lpc y de autocorrelación," Universidad de Santiago de Chile, 2002.

[12] I. Lugeno y E. Navas, "Reconocimiento automático de emociones utilizando parámetros prosódicos," en Procesamiento del lenguaje natural, vol 35, pp. 13-20, 2005.