Reconocimiento de emociones en señales de voz

Leonardo N. Arato y Nahuel Boutet
Trabajo práctico final de "Procesamiento Digital de Señales", II-FICH-UNL.

***Resumen*— Con el progreso de la tecnología es cada vez más frecuente la interacción entre hombres y máquinas, a razón de esto, resulta de interés desarrollar mecanismos que sean capaces de simplificar ésta interacción. Una posibilidad de llevar adelante esta simplificación es implementar la capacidad de generar y reconocer el habla.**

Dado que la voz es el medio de comunicación más natural para los humanos, es necesario proporcionar a las computadoras la capacidad de reconocer las emociones presentes en el habla.

A partir de estas interpretaciones se pueden generar nuevas herramientas que sean útiles en diversas situaciones tales como llamadas de emergencias o de auxilio, detección de enfermedades psicológicas o comunicaciones con Call-Centers.

Los diferentes estados emocionales afectan el sistema del aparato fonador del ser humano y esto puede verse reflejado mediante la variación de algunos de sus parámetros. Teniendo en cuenta esas variaciones hemos identificado en la voz las emociones de felicidad, enojo, tristeza y neutralidad. A partir de estas variaciones hemos creado un procedimiento que dada de una señal de voz de entrada intenta identificar el estado emocional del hablante.

***Palabras clave*— Voz, Emoción, Vector Característico, Distancia de Mahalanobis.**

I. Introducción

En el análisis de la voz tenemos en cuenta su naturaleza aleatoria de carácter no estacionario. Para el tratamiento de este tipo de señales debemos tener conocimiento del comportamiento del sistema que las genera. Sabemos que el sistema del aparato fonador en el hombre tarda aproximadamente entre 10 y 30 milisegundos en cambiar los parámetros característicos que le dan forma al mismo, este es el tiempo promedio en el que podemos considerar a la señal como estacionaria. Es aquí donde el uso técnicas convencionales [1] son las herramientas adecuadas para extraer la información de la señal.

El enfoque utilizado para este desarrollo está orientado a la identificación de las emociones en la voz dependiente del hablante, lo cual reduce la complejidad del problema.

El uso de la Transformada de Fourier nos permite representar la señal de voz en otro dominio donde discriminamos los componentes que forman la señal. La utilización de ventanas son las herramientas que permiten extraer la información de los componentes en periodos donde la hipótesis de estacionariedad se cumple [2].

Entre los métodos convencionales, el método de Predicción Lineal resulta apropiado para obtener la información necesaria para modelar al aparato fonador como un sistema auto regresivo.

El análisis en el dominio frecuencial del sistema auto regresivo permite identificar las frecuencias formantes y la frecuencia fundamental de la señal de la voz. El comportamiento de estas componentes frecuenciales será lo que caracterice a cada emoción.

De cada señal de prueba se extraen sus componentes característicos que serán comparados con las características de las señales de entrenamiento que fueron agrupadas para cada emoción. Este agrupamiento de los datos hace oportuno el uso de la distancia de Mahalanobis [3] para la comparación y posterior clasificación.

II. Metodología

A. Tratamiento de la base de datos.

Los datos utilizados han sido extraídos de una base de datos en idioma Hindi, la cual cuenta con cuatro emociones diferentes, divididas en treinta y tres frases distintas para siete hombres y cuatro mujeres, donde cada señal fue muestreada a 8000Hz.

Se procedió separando veinticinco señales de cada emoción (enojo, felicidad, neutral y tristeza) por persona para su procesamiento como parámetros de entrenamiento del algoritmo, las ocho restantes se han utilizado para realizar las pruebas. Cada señal se corresponde con una palabra que estaba contenida en una frase.

B. Tratamiento de señales

Cada señal ha sido ventaneada con una ventana de Hamming de ciento sesenta muestras que se corresponden con una duración de veinte milisegundos. El solapamiento entre ventanas consecutivas se realiza en saltos de ochenta muestras (media ventana). Para cada ventana se calcula la energía y la cantidad de cruces por cero y se comparan con un umbral para descartar aquellas que no contengan fonemas sonoros, Fig. 1.

En las ventanas con fonemas sonoros se calcula la energía, la frecuencia fundamental (F0) y las tres primeras frecuencias formantes (F1, F2, F3), debido a la limitación de la frecuencia de muestreo.

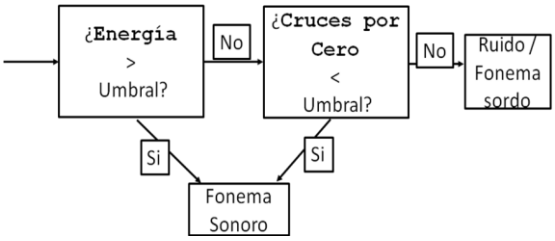

Fig. 1: Modelo de detector de ventanas con fonemas sonoros.

C. Vector de características

Las frecuencias formantes de cada ventana se obtienen mediante el cálculo de los tres primeros máximos locales de

la respuesta en frecuencia del tracto vocal modelado a partir de los coeficientes obtenidos con el método de predicción lineal Fig. 2.

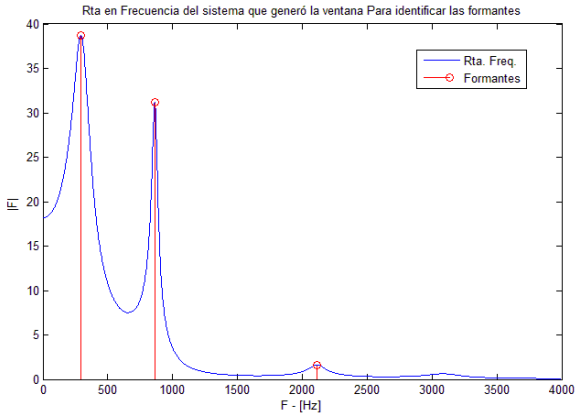

Fig. 2: Respuesta en frecuencia y frecuencias formantes.

La frecuencia fundamental se calcula mediante el método de Auto-correlación Fig. 3.

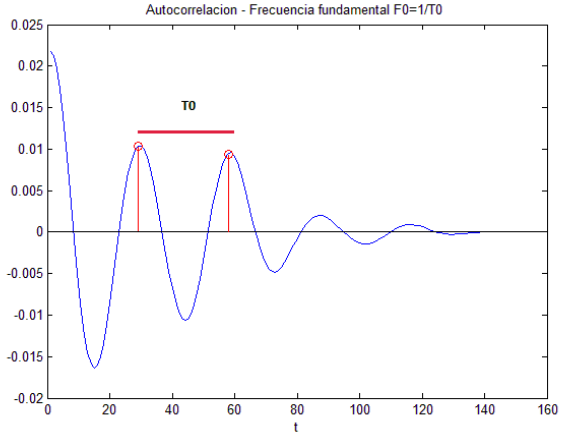

Fig. 3: Calculo de F0.

Al finalizar el ventaneo de cada señal construimos un vector de características que contiene la varianza y el promedio de energía, la varianza y el promedio de la frecuencia fundamental y las frecuencias formantes promedio de las ventanas que contienen fonemas sonoros Fig. 4.

Var(E)	Ẽ	Var(F0)	F0	F1	F2	F3

Fig. 4: Vector Característico.

Se tomó la energía como un componente característico debido a que la energía de una señal de voz varía para cada emoción [4] Fig. 5.

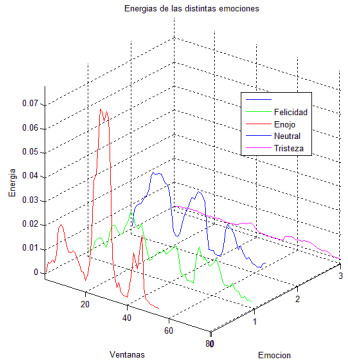

Fig. 5: Variación de energía según emoción.

D. Criterio de comparación

Para cada señal de prueba obtenemos su vector de características que usaremos para poder clasificarla según su tipo de emoción. Por cada persona tenemos veinticinco vectores de características correspondientes a veinticinco señales de voz distintas clasificados según su tipo de emoción. Considerando a cada uno como una observación de un proceso aleatorio, calculamos la distancia de Mahalanobis entre el vector de características de la señal de prueba y cada una de las distribuciones correspondiente a cada tipo de emoción Fig. 6.

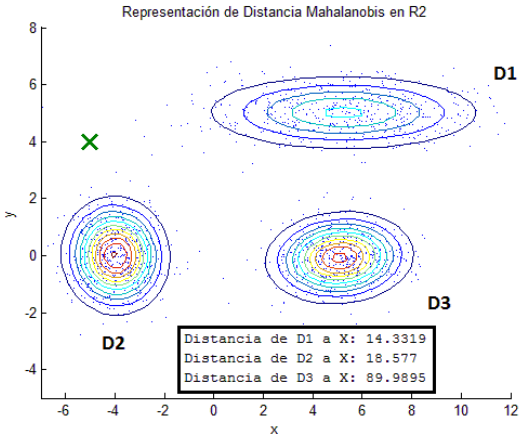

Fig. 6: Distribuciones Representativas en R2.

La utilidad de la distancia de Mahalanobis radica en que permite determinar la similitud entre dos variables aleatorias multidimensionales independizándose de la variabilidad de las componentes del vector característico. La distancia de Mahalanobis se calcula de la siguiente manera:

$$d_m(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T \Sigma^{-1} (\vec{x}-\vec{y})}.$$

E. Proceso de clasificación

Para realizar la clasificación de una señal de prueba correspondiente a una emoción X, se calcula la distancia de Mahalanobis entre el vector característico de la señal y cada una de las cuatro distribuciones correspondientes a cada emoción.

Si la menor distancia obtenida, de entre las cuatro posibles, corresponde a la distancia entre el vector característico de la señal de prueba y la distribución correspondiente a la emoción X, se considera una clasificación exitosa, de caso contrario es una falla Fig. 7.

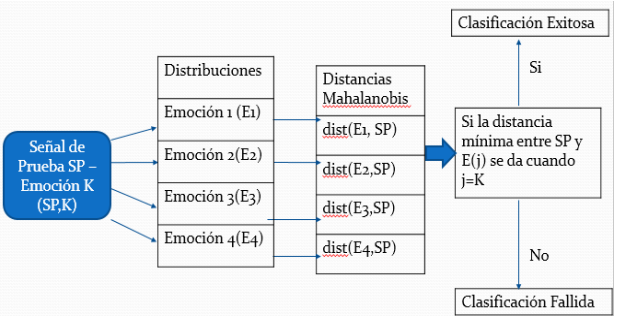

Fig. 7: Proceso de Clasificación.

III. Resultados

Para analizar los resultados hemos tomado como primer caso el tamaño de ventana de ciento sesenta muestras. Con ello hemos obtenido un promedio de 81.69 % de aciertos en los hombres donde la emoción con mayor porcentaje de acierto fue Enojo con un 98.21%. Para las pruebas en mujeres te obtuvo un promedio de 84.37% de aciertos donde la emoción con mayor porcentaje de acierto fue la Felicidad Tabla I.

TABLA I

PORCENTAJES ACIERTOS CON VENTANAS DE 160 MUESTRAS

(8 SEÑALES DE PRUEBA POR EMOCIÓN).

Emoción	Hombres	Mujeres
Enojo	98.21	84.37
Felicidad	73.21	93.75
Neutral	76.78	84.37
Tristeza	78.57	75.00

En el segundo caso tomamos ventanas de doscientos cincuenta y seis muestras, Con la cual se obtuvo un promedio de aciertos de 82.53% en los hombres, donde la emoción con mayor porcentaje de acierto fue Enojo con un 100% de aciertos. Para las pruebas en mujeres se obtuvo un promedio de 86.71% de aciertos, observando que la emoción con mayor porcentaje de acierto fue 93.75% para la emoción de Felicidad Tabla II.

TABLA II

PORCENTAJES ACIERTOS CON VENTANAS DE 256 MUESTRAS

(8 SEÑALES DE PRUEBA POR EMOCIÓN).

Emoción	Hombres	Mujeres
Enojo	100	84.37
Felicidad	73.21	93.75
Neutral	76.57	84.37
Tristeza	80.35	84.37

Para el tercer caso se ha agregado ruido únicamente a las señales de prueba de manera que se obtenga una SNR de 20dB y 0dB.

En el caso de la SNR=20dB se obtuvo una porcentaje promedio de acierto de 74.10% en hombres, donde la emoción con mayor porcentaje de acierto fue el Enojo con 98.21% y notando una disminución hasta el 53.37% en las pruebas de Felicidad. Para las pruebas en mujeres se obtuvo un promedio de 74.21% de aciertos, observando que la emoción con mayor acierto fue Felicidad con 87.5% y notando una disminución hasta el 50.0% para la emoción Neutral Tabla III.

TABLA III

PORCENTAJES ACIERTOS CON SNR=20DB

(8 SEÑALES DE PRUEBA POR EMOCIÓN).

Emoción	Hombres	Mujeres
Enojo	98.21	81.25
Felicidad	53.57	87.5
Neutral	73.21	50.0
Tristeza	71.42	78.12

Cuando la SNR=0dB se obtuvo porcentajes de aciertos de 0.0% para las emociones de Tristeza y Neutralidad tanto en hombres como mujeres, y el porcentaje más alto de aciertos fue para la emoción de Enojo, no alcanzando el 80.0% en ningún caso Tabla IV.

TABLA IV

PORCENTAJES ACIERTOS CON VENTANAS DE 160 MUESTRAS

(8 SEÑALES DE PRUEBA POR EMOCIÓN).

Emoción	Hombres	Mujeres
Enojo	77.5	74.37
Felicidad	26.78	13.75
Neutral	0.0	0.0
Tristeza	0.0	0.0

IV. Conclusiones

En primer lugar, podemos decir que el porcentaje de acierto ha sido elevado superando ampliamente nuestras expectativas iniciales.

Los mayores porcentajes de acierto se han dado bajo los sentimientos de Enojo para el hombre y Felicidad para la mujer.

Al aumentar el tamaño de las ventanas de 160 a 256 muestras el algoritmo ha arrojado los resultados de clasificación han mejorado levemente.

Ante la presencia de señales de prueba con bajos niveles de ruido el algoritmo se ha comportado satisfactoriamente.

Cuando la SNR se acerca a cero, el algoritmo empieza a mostrar sus falencias.

De todas formas los resultados obtenidos no son para nada concluyentes ya que la base de datos utilizada es pequeña, contiene frases con emociones sobre actuadas y presenta una alta relación señal-ruido.

Referencias

[1] H. Milone, L. Rufiner, C. Acevedo, L. Di Persia, M. Torres "Introducción a las Señales y los Sistemas Discretos", 2009.

[2] L. Vignolo http://pdsfich.wdfiles.com/local--files/clases-de-teoria/teoria09speech2014.pdf, "Procesamiento de la señal de voz" 2014

[3] http://www.mathworks.com/help/stats/mahal.html

[4] S. Ramamohan, S. Dandapat "Sinusoidal Model-Based Analysis and Classification of Stressed Speech", *IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING, VOL. 14, NO. 3, MAY 2006.*