Clasificación de emociones a través de la señal de voz parametrizada

BELLO - Vicente †*, MARTÍNEZ - Miriam, MONTERO - José Antonio, DE LA CRUZ - Eduardo.

Tecnológico de Acapulco

Recibido: septiembre, 22, 2018; Aceptado Febrero 9, 2019.

Resumen

El reconocimiento automático de las emociones humanas a través de la señal de voz parametrizada, es un área de investigación activa debido a la amplia variedad de aplicaciones que puede tener, entre otras: telecomunicaciones, aprendizaje, interfaz humano-computadora y entretenimiento. En este trabajo se muestra una metodología para el reconocimiento de emociones analizando segmentos de voz. La metodología se basa principalmente en la transformada rápida de Fourier (FFT) y coefícientes de correlación de Pearson. El Tono (Pitch), frecuencia fundamental (F_0), la intensidad de la señal de voz (energía) y la tasa de habla se han identificado como parámetros importantes para identificar emciones. El sistema tiene una interfaz gráfica que permite la interacción del usuario por medio de un micrófono integrado en la computadora, la cual procesa automáticamente los datos adquiridos. El corpus contiene 16 frases por emoción creada por 11 usuarios (9 mujeres y 2 hombres) con un total de 880 muestras de audio. Se consideran las siguientes emociones básicas: disgusto, ira, felicidad, miedo y neutral. El algoritmo de reconocimiento de emociones ofrece un 80% de efectividad en los resultados obtenidos.

Palabras clave: Señal de voz parametrizada, interfaz humano-computadora, corpus.

Abstract

The automatic recognition of human emotions through the parametrized voice signal, is an area of active research due to the wide variety of applications that may have, among others: telecommunications, learning, human-computer interface and entertainment. In this work, a methodology for the recognition of emotions analyzing voice segments is shown. The methodology is based mainly on the fast Fourier transform (FFT) and Pearson correlation coefficients. The pitch (Pitch), fundamental frequency (F0), the intensity of the voice signal (energy) and the speech rate have been identified as important parameters to identify emulations. The system has a graphical interface that allows user interaction through a microphone integrated in the computer, which automatically processes the acquired data. The corpus contains 16 phrases per emotion created by 11 users (9 women and 2 men) with a total of 880 audio samples. The following basic emotions are considered: disgust, anger, happiness, fear and neutrality. The emotion recognition algorithm offers 80% effectiveness in the results obtained.

Keywords: Parameterized voice signal, human-computer interface, corpus.

Citación: BELLO - Vicente †, MARTÍNEZ – Miriam, MONTERO - José Antonio, DE LA CRUZ - Eduardo. Clasificación de emociones a través de la señal de voz parametrizada. Foro de Estudios sobre Guerrero. 2018, Mayo 2019- Abril 2020 Vol.5 No.6 247-254.

*Correspondencia al Autor (luanberry@hotmail.com)
† Investigador contribuyendo como primer autor.

Introducción

La voz es sonido y como tal, se caracteriza por una serie de elementos. La información es acústica cuando la extracción se hace únicamente sobre la señal de voz, la cual describe los sonidos básicos del lenguaje y trata de explicar cómo se realizan acústicamente en una expresión hablada.

Las técnicas empleadas en el análisis de la señal de voz se pueden dividir en dos categorías: Transformadas Tiempo – Frecuencia y Análisis Paramétrico. La primera de estas categorías hace referencia a la representación de la señal en espacios conjuntos del tiempo y la frecuencia, permitiendo conocer la ubicación temporal del contenido espectral, esta técnica es efectiva en el tratamiento de señales no estacionarias como es la señal de voz.

El análisis paramétrico busca estimar un modelo matemático que de forma aproximada represente el sistema de producción vocal.

En la actualidad en México, existen pocos repositorios de datos por lo tanto es probable que los sistemas actuales tarden algún tiempo en madurar lo suficiente como para presentarse como una alternativa de solución para el reconocimiento de estados de ánimo por medio del análisis de la voz.

Se abarcan los estudios relacionados con tecnologías del habla utilizando reconocimiento de emociones en la voz y el desarrollo de un prototipo para reconocer el estado anímico de jóvenes de nivel superior de edades entre 17 y 25 años.

Se creo un corpus de emociones y se implementaron las fases de extracción de parámetros acústicos y clasisicación de la emoción que dan como resultado el reconocimiento de las emociones primarias. Se probo con mas de 800 muestras en la clasificación.

Emoción y estado emocional son conceptos diferentes: mientras que las emociones surgen repentinamente en respuesta a un determinado estímulo y duran unos segundos o minutos, los estados de ánimo son más ambiguos en su naturaleza, perdurando durante horas o días. Las emociones pueden ser consideradas más claramente como algo cambiante y los estados de ánimo son más estables.

Objetivos

Diseñar un sistema de reconocimiento de emociones a través de la señal de voz parametrizada.

Objetivos Especificos

- Diseñar una interfaz gráfica de usuario para realizar el proceso de grabado de voz.
- Crear una base de datos de emociones.
- Diseñar el algoritmo de reconocimiento de emociones (Preprocesamiento, extracción de características y clasificación).
- Evaluar los resultados obtenidos.

Metodología de Desarrollo

La metodología se divide en el estudio de parámetros acústicos y lingüísticos que contienen características de los estados emocionales, el diseño del modulo de grabación, la captura de frases para tener el corpus emocional y las pruebas con el algoritmo de reconocimiento de estados emocionales primarios. La identificación de un conjunto de características acústicas y modelos estadísticos, permitirá clasisicar emociones en la voz con un porcentaje de más del 80% de eficiencia.

ISSN: 2007-882X

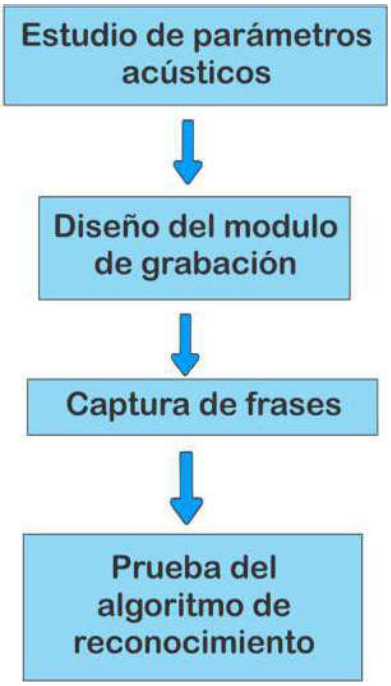

Figura 1: *Etapas del proyecto.*

Estudio de parámetros acústicos

Los parámetros acústicos son medidas que se emplean para el análisis acústico de la voz que deben observarse en toda exploración acústica, e incluyen la frecuencia fundamental (F_0), la intensidad, las perturbaciones de amplitud (shimmer), perturbaciones de frecuencia (jitter) y la expresión del ruido espectral (calculada mediante la relación armónico/ ruido), de modo que es posible evaluar hasta los más pequeños cambios en la masa y tensión, así como el carácter bioquímico de las cuerdas vocales (Adrián Torres and Casado Morente, 2002).

La voz no es otra cosa que un sonido y como tal, se caracteriza por una serie de elementos. Los rasgos que han sido mas recurrentes en la literatura son el pitch, duración, calidad de voz y forma del pulso glotal y tracto vocal. Hasrul, por ejemplo, agrupa en su trabajo las características que han sido utilizadas para la detección de emociones en la voz. (Hasrul, 2012) (Palacios, 2017). Estos parámetros se muestran en la tabla 1.

Características Utilizadas	Descripción
Ancho de banda	Este rango se mide en Hercios *(Hz).*
Áreas del tracto vocal	Numero de armónicos ocasionados por el flujo de aire no lineal en el tracto vocal que produce la señal de voz.
Características espectrales	Contenido energético de bandas de frecuencia divididas por la longitud de muestra.
Detección de la Actividad del Habla	Esta propiedad se define como el perfil rítmico del habla.
Duración	Diferencia entre el instante de inicio y final de una secuencia hablada obteniendo una taza de duración de sentencias de tipo emocional y neutras.
Energía	Es el valor de la magnitud física que expresa la mayor o menor amplitud de las ondas sonoras.
Formantes	Son frecuencias reforzadas por la resonancia.
Intensidad	Se mide en Decibelios *(dB).*
Pitch	Se representa como F_0 *(Frecuencia Fundamental).*
Velocidad del habla *(speaking rate)*	La proporción de unidades segméntales, silabas y pausas por unidad de tiempo producidas por un locutor.

Tabla 1: *Características usadas en el reconocimiento de emociones en el Habla.*

ISSN: 2007-882X

Diseño del modulo de grabación.

La interfaz gráfica de usuario (GUI) para el sistema de grabación tiene las siguientes características:

- Captura de señal de audio: El sistema debe permitir la captura de audio a una frecuencia de 44100 Hz, con una tasa de bits de 16 kbps (kilobits por segundo), un canal mono y en formato WAV.

- Capacidad para guardar archivos de audio: El sistema permite guardar la voz del locutor en tiempo real en una carpeta llamada corpus.

- Capacidad detener la grabación de audio: El sistema debe tener la opción de detener una grabación de voz en tiempo real.

- Capacidad de eliminar: El sistema debe eliminar archivos de audio.

- Capacidad de reproducir: El sistema debe permitir la reproducción de formatos de audio.

- Capacidad de detener reproducción: El sistema debe permitir detener una reproducción en curso.

- Capacidad de cambiar la ruta: El sistema debe permitir cambiar la ruta para guardar el formato de audio en otra dirección.

Captura de frases.

Para el desarrollo del corpus emocionalse se consideran la ira, felicidad, neutral, miedo, tristeza y disgusto. Los textos de estímulo para las frases fueron concebidos en el contexto de situaciones de la vida cotidiana. Se diseñaron 16 enunciados para cada emoción.

Estos enunciados fueron producidos por veinte hablantes: 2 hombres y 9 mujeres. Cada uno de los veinte hablantes produjo el enunciado con cada una de las emociones indicadas. Se han escogido frases cuyo contenido semántico no implique ninguna emoción en concreto de forma que la clasificación se pueda realizar con base a detalles prosódicos.

Acontinuacion se muestran las instrucciones para llenar el corpus emocional con fraces emottivas utilizando la interfaz grafica.

1.- Nombre de los archivos de audio: Para crear un archivo de audio se debe llenar de la siguiente manera:

- Sus iniciales; por ejemplo: **V**icente **B**ello **A**mbario **(VBA**)-
- Género (**M, F**).
- Emoción:
    - **D: Disgusto (parecido al Asco)**
    - **I: Ira**
    - **F: Felicidad**
    - **M: Miedo**
    - **N: Neutral**
- El número de la frase.

Ejemplo de un nombre de archivo para la primera frase con la emoción de disgusto: **VBAMD1**

2. Presione la tecla **Enter** para agregar el nombre del archivo a la ruta.

3.- A continuación, presione el Botón **“Grabar”** y diga la frase, al terminar presione **“Detener G”**.

4. Repita el paso 1 para cambiar de emoción.

ISSN: 2007-882X

Frases
1.- Los Tiempos ya no son como antes
2.- De que estas hablando pues
3.- ¿Quieres un consejo?
4.- La tarea es para mañana
5.- Él es el jefe de grupo
6.- Si, es verdad
7.- No lo creo, no seas chismoso
8.- Siempre llegas tarde
9.- ¿Puedes guardar silencio por favor?
10.- Si no te gusta, hazlo tu
11.- La computadora de mi mama está descompuesta
12.- La escuela está pintada de rosa
13.- Vivirás conmigo
14.- Mi punto de Vista es otro
15.- Esa actividad no me corresponde
16.- Ahí está un loco

Tabla 2: *Frases de Estimulo Diseñadas para cada Emoción.*

Prueba del algoritmo de reconocimiento.

Hay dos factores importantes durante este proceso. Para desarrollo del código se deben de cambiar los parámetros para ver lo que mejor funciona en el algoritmo. Haciendo uso de un programa de escritorio, se graban audios con una frecuencia de muestreo de 44100 Hz y una tasa de audio de 16 bits.

Se usa un canal (Mono) que da como resultado un vector de miles de datos, de los que se discriminarán los datos significativos.

Normalización

En general se entiende que la normalización es la operación mediante el cual un conjunto de valores de una determinada magnitud es transformado en otros de tal manera que estos últimos pertenezcan a una escala predeterminada.

Es posible normalizar un conjunto de valores en el intervalo [0,1] aplicado para cada valor la transformación mostrada en la ecuación 1.

$$v_i = \frac{a_i - min}{max - min} \qquad (1)$$

Donde a_i es el valor a transformar, min y max son el mínimo y el máximo del conjunto de valores y v_i es el valor normalizado.

La normalización consiste dar un tratamiento a la señal acústica y encontrar el conjunto óptimo de características que permitan realizar la clasificación de emociones.

El algoritmo de función que normalice los datos de un vector numérico que recibe como parámetro es el siguiente:

- Devuelve el valor absoluto máximo del vector a transformar.
- Devuelve el número de elementos del vector a transformar (Tamaño del vector = *n*)
- Devuelve un vector de ceros de *n* filas y 1 columna.
- Se hace un bucle donde el valor inicial de i es 1 y se va incrementando en 1 hasta que llegue a ser el valor de *n.*
- Se divide el vector en la posición *i* entre su valor máximo absoluto.

ISSN: 2007-882X

Extracción de características

Este módulo consiste en agrupar las características acústicas espectrales, ya que estas describen las propiedades de una señal en dominio de la frecuencia mediante armónicos y formantes, también se extrae información prosódica (volumen, velocidad, duración). El algoritmo para extraer características es la transformada rápida de furrier *(FFT*) el cual realiza lo siguiente:

- Se cortan los 60000 primeros valores del vector
- Se obtiene el valor absoluto de la transformada de Fourier de la grabación
- Se multiplica el resultado por el conjugado del vector original
- Solo acepta las Frecuencias arriba de 150 *Hz*
- Se normaliza el vector utilizando la norma euclidiana

La norma euclidiana (también llamada magnitud del vector, longitud euclidiana, *o 2-Norm*) de un vector *v* con los elementos de *N* es definido por la ecuación 2.

$$||v|| = \sqrt{\sum_{k=1}^{N} |v_k|^2} \qquad (2)$$

FFT es la abreviatura usual (de sus siglas en inglés Fast Fourier Transform), y es un eficiente algoritmo que permite calcular la transformada discreta de Fourier y su inversa dados vectores de longitud *N* por la ecuación 3.

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi \frac{n}{N}} \qquad (3)$$

Esta es la fórmula para la transformada discreta de Fourier, lo que convierte las señales (como una grabación de sonido digital) muestreadas a el dominio de la frecuencia. Siendo este el motor matemático detrás de una gran parte de la tecnología que utiliza hoy en día.

Se Obtienen las *FFT* de cada tramo, teniendo 5 vectores por cada emoción con el objetivo de generar una superficie en la que se pueda observar las frecuencias y su variación en el tiempo. Se promedian las *FFT* de cada tramo, para obtener un patrón de la frase pronunciada.

Clasificación.

Se define el coeficiente de correlación de Pearson como un índice que puede utilizarse para medir el grado de relación de dos variables siempre y cuando ambas sean cuantitativas y continuas. El coeficiente de correlación de Pearson es un índice de fácil ejecución. En primera instancia, sus valores absolutos oscilan entre 0 y 1. Si tenemos dos variables *X* e *Y*, entonces se define coeficiente de correlación de Pearson entre estas dos variables como $r_{x,\ y}$ como se muestra en la ecuación 4.

$$px, y = \frac{\sigma xy}{\sigma x \sigma y} = \frac{E[(X-\mu x)(Y-\mu y)]}{\sigma x \sigma y} \qquad (4)$$

Resultados

Una vez que se conocen los objetos y los eventos, se puede diseñar la interfaz para la aplicación. la figura 2 muestra la interfaz grafica de usuario. La entrada de voz se maneja con los botones mediante un clic sobre él. A dicha acción se le denomina Evento Clic.

ISSN: 2007-882X

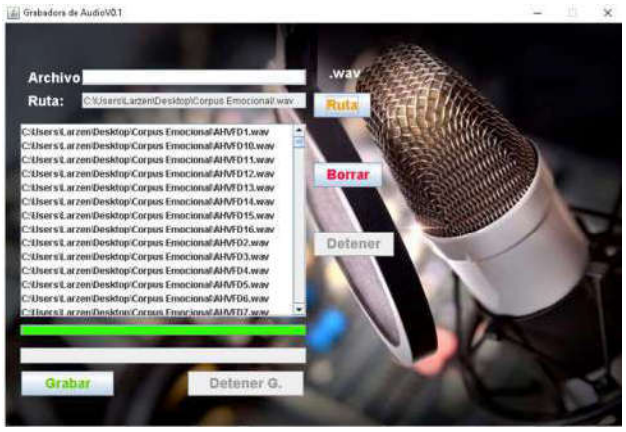

Figura 2: *Modulo de Grabación.*

Con la intención de determinar si los parámetros acústicos y la velocidad de habla funcionan como elementos caracterizadores de los distintos tipos de emociones, se construyo el corpus emocional recogido por locutores.

El corpus esta constituido por una serie de grabaciones en las cuales se recogen emociones simuladas por los locutores.

La figura 3 muestra la asesorías y uso adecuado del software a los participantes previo a la grabación. La Figura 4 muestra el proceso de grabación de audio realizado en un aula cerrada, con el proposito de reducir ruidos y distractores.

Figura 3: *Asesorías para los discursos emotivos.*

Fifura 4. *Proceso de grabación.*

La Tabla 3 muestra la matriz confusión del algoritmo utilizado en este trabajo donde se pueden observar que la emoción neutral tiene mayor confusión a diferencia de las demás emociones, también cabe mencionar que el disgusto, la Ira y Felicidad son emociones son claramente identificadas con mayor exactitud por este clasificador.

		Predicion					
		Disgusto	Ira	Felicidad	Miedo	Neutral	Totales
Observaciones	Disgusto	143	8	9	5	11	176
	Ira	23	134	12	2	5	176
	Felicidad	25	13	116	4	18	176
	Miedo	18	14	9	115	20	176
	Neutral	39	34	21	2	80	176
	Totales	248	203	167	128	134	880

Tabla 3: *Matriz de confusión.*

Conclusiones

Se diseño un algoritmo capaz de identificar los parámetros acústicos para el reconocimiento de estados emocionales en la voz, se obtuvo de acuerdo a los resultados un algoritmo capaz de reconocer un 80% de las frases con emoción actuada.

ISSN: 2007-882X

Como trabajo futuro se tiene previsto evaluar el desempeño en otros contextos tales como: llevar a cabo evaluaciones sobre diferentes bases de datos tanto de emociones, reales, como actuadas con el fin de evaluar el alcance del sistema, hacer una evaluación subjetiva con personas no especializadas o no entrenadas, realizar un análisis estadístico del sistema en general para demostrar su confiabilidad, realizar pruebas del sistema con diferentes locutores, utilizar medidas de tendencia central (Media, mediana, desviación, varianza, moda), calcular el coeficiente de reproductibilidad para tener la certeza si el número de errores es tolerable y finalmente integrar el sistema de reconocimiento de emociones a un sistema.

Referencias

Adrián Torres, J. A. and Casado Morente, J. C. (2002). La evaluación clínica de la voz: fundamentos médicos y logopédicos. Ediciones Aljibe.

Hasrul, M. N., Hariharan, M., & Yaacob, S. (2012, February). Human Affective (Emotion) behaviour analysis using speech signals: A review. In Biomedical Engineering (ICoBE), 2012 International Conference on (pp. 217-222). IEEE.

Palacios Alonso, D. (2017). Contribución al estudio de selección de parámetros para identificación de estrés en la voz (Doctoral dissertation, ETSI_Informatica).

ISSN: 2007-882X