IDENTIFICACIÓN DE ESTADOS EMOCIONALES A TRAVÉS DEL ANÁLISIS ACÚSTICO

Vicente Bello Ambario [1], Miriam Martínez Arroyo [2],
José Antonio Montero Valverde [3] y Cristina Barrera De Jesús [4]

***Resumen*—El propósito de esta investigación es conocer el estado emocional de un alumno (nivel superior) por medio de la voz para conocer si es apto o no para tomar una clase. Determinar el estado emocional por medio de la voz es el objetivo principal de este estudio que se realizara en el Instituto Tecnológico de San Marcos, Guerrero, México. La metodología se basa principalmente en el uso de parámetros para la extracción de características y modelos estadísticos para el clasificador. El estado emocional del alumno puede ser detectado de forma confiable a través del análisis de la voz. En este trabajo se presenta una propuesta para determinar el estado emocional de alumnos y saber si están en condiciones para involucrarse en el proceso enseñanza-aprendizaje.**

***Palabras clave*—Estado emocional, Parametrización, Modelos estadísticos, reconocimiento de patrones.**

Introducción

Las emociones humanas expresan el estado emocional o sentimientos debidos a diferentes factores que experimenta el ser humano y el entorno que lo rodea. Estas expresiones son útiles para establecer canales de comunicación e interactuar con otros humanos.

La primera sección del presente trabajo pretende estudiar los fundamentos del reconocimiento de los estados emocionales a partir de una serie de parámetros calculados en la señal de voz. Se presenta una comparativa para observar características y parámetros para que sirven para el reconocimiento emocional humano.

A continuación, se describe el corpus de voz sometido al preprocesamiento de las grabaciones de audio y el algoritmo empleado para identificar emociones. El tercer apartado detalla los experimentos realizados las métricas empleadas.

Fundamentos

La necesidad de dotar a las máquinas de inteligencia emocional ha provocado que en las últimas décadas la atención de los investigadores se haya centrado en crear sistemas de reconocimiento de voz sensibles a las emociones, con el fin de mejorar dicha comunicación, dando lugar al campo del reconocimiento de emociones por voz (Speech Emotion Recognition, SER) *(Pérez Pascual, F., 2017).*

Las emociones son físicamente "reacciones que representan modos de adaptación a ciertos estímulos del individuo y que tienen afectación en las respuestas de distintos sistemas biológicos, entre los que se incluye la voz" *(Levenson, 1994).* Las emociones ejercen una fuerza muy poderosa en el comportamiento humano, además de tener una gran influencia en la salud, tanto física como mental, de las personas. Un ejemplo claro de este hecho es que una persona que se encuentra en un estado de excitación es capaz de hacer alguna acción que nunca hubiera imaginado, y viceversa, que se vea paralizada por el miedo o la tristeza. Sin embargo, uno de los problemas que se encuentran los investigadores a la hora de trabajar con emociones es que éstas presentan un alto grado de subjetividad. Esto es debido a que la manera cómo se expresan depende en gran medida del hablante, de su cultura de procedencia y del entorno que frecuenta.

Emociones Primarias

Las emociones primarias son: enojo, miedo, tristeza, alegría disgusto y sorpresa. La voz neutral *(Kim et al., 2007)* se puede percibir de una forma uniforme, calmada, con un tono más o menos idéntico, sin alteraciones o

[1] El Ing. Vicente Bello Ambario es alumno de la Maestría en Sistemas Computacionales (MSC) en el Instituto Tecnológico de Acapulco (ITA) perteneciente al Tecnológico nacional de México (TecNM), en Guerrero, México. luanberry@hotmail.com

[2] La Dra. Miriam Martínez Arroyo es Profesora de la MSC en el ITA (TecNM), Gro., México. miriamma_ds@hotmail.com

[3] El Dr. José Antonio Montero Valverde es Profesor la MSC en el ITA (TecNM), Gro., Méx. jamontero1@infinitummail.com

[4]La MCC. Cristina Barrera De Jesús es Profesora en el Instituto Tecnológico de San Marcos (ITSM) (TecNM), Gro., Méx. 01cristinabarrera@gmail.com

interrupciones, posteriormente la emoción de enojado se puede apreciar una voz determinante, fuerte, irritable, agresiva y severa. Para el estado de la felicidad, se le puede considerar como una voz cantada, llena de alegría, de alguna forma como si el locutor tuviera una sonrisa en la cara; la forma de expresarse con la emoción del miedo denota una voz cambiante, interrumpida, un tono casi chillón, voz ansiosa, con susurros. Por último, el estado emocional de tristeza puede ser percibido como monótono, depresivo, lento, melancólico y lento *(Solís, 2011)*. El habla neutra suele caracterizarse por un tono con un rango de variación estrecho y unas transiciones de frecuencia fundamental suaves, además de una velocidad de locución alta. A continuación, plantearemos una de las clasificaciones de las emociones primarias:

- *Ira*: Se define como "la impresión desagradable y molesta que se produce en el ánimo". El enfado se caracteriza por un tono medio alto (229 Hz), un amplio rango de tono y una velocidad de locución rápida (190 palabras por minuto), con un 32% de pausas.
- *Felicidad*: Se manifiesta en un incremento en el tono medio y en su rango, así como un incremento en la velocidad de locución y en la intensidad.
- *Tristeza*: El habla triste exhibe un tono medio más bajo que el normal, un estrecho rango y una velocidad de locución lenta.
- *Miedo*: Comparando el tono medio con los otros cuatros emociones primarias estudiadas, se observó el tono medio más elevado (254 Hz), el rango mayor, un gran número de cambios en la curva del tono y una velocidad de locución rápida (202 palabras por minuto).
- *Disgusto*: Se caracteriza por un tono medio bajo, un rango amplio y la velocidad de locución más baja, con grandes pausas.

Análisis de los Parámetros de Voz

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 (2012)*, por ejemplo, agrupa su trabajo en 13 características que han sido utilizadas para la detección de emociones en la voz. Estos parámetros se muestran en el cuadro 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)*
LPCs *(Linear Prediction Coefficients)*	Conjunto de formulaciones esenciales equivalentes para modelar una forma de onda dada
MFCCs *(Mel Frequency Cepstrum Coefficients)*	Técnica de fraccionar la señal inicial en un conjunto discreto de bandas espectrales que contiene información analoga
Pitch	Se representa como F_0 *(Frecuencia Fundamental)*
Tasa de cruce por ceros	Representa cuantas veces la señal cambia de signo pasando por el eje de las abscisas
Velocidad del habla *(speaking rate)*	La proporción de unidades segméntales, silabas y pausas por unidad de tiempo producidas por un locutor

Cuadro 1. Características usadas en el reconocimiento de emociones en el Habla *(Hasrul , 2012) (Palacios, 2017)*.

El cuadro 2 presenta un resumen de las relaciones entre las emociones y los parámetros del discurso. Como se puede observar, únicamente aparecen cinco emociones. Estas corresponden con las emociones primarias o básicas.

Es conocido que existe una relación entre la información prosódica y la expresión de emociones en el habla; rasgos como la intensidad, la curvatura de frecuencia fundamental y la velocidad de locución son características importantes den la discriminación de emociones en la voz *(Nwe et al., 2003) (Montero Martínez, 2003).*

	Felicidad	Ira	Disgusto	Miedo	Tristeza
Velocidad del habla	Ligeramente acelerada con Incremento	Ligeramente acelerada	Lenta	Muy Acelerada	Pausada
F_0	Incremento de la media, variabilidad	Incremento de la media, mediana, variabilidad	----------	Incremento en la F0 media, perturbación, variabilidad del movimiento de F_0	Debajo de la F0 media normal
Articulación	Normal	Tensa	Normal	Precisa	Arrastrada
Intensidad	Alta	Alta	Baja	Normal	Baja
F0 promedia	Alta	Alta	Baja	Alta	Baja
Espectro	Incremento de la energía de alta frecuencia	Elevado en el punto medio	----------	Aumento de la energía de alta frecuencia	Disminución de la energía de alta frecuencia
Otros	Distribución irregular de acentos	Habla cortada	----------	Irregularidad en la sonorización	Ritmo con pausas irregulares

Cuadro 2. Comparativo de emociones *(Ortego Resa et al., 2009) (Cowie et al., 2001).*

Descripción del Sistema

El desarrollo del proyecto se realizó en un lapso de 13 semanas en el periodo de 4 meses. La figura 1 muestra la estructura general del sistema propuesto iniciando por la captura de voces para tener el corpus emocional. Para el diseño del corpus emocional se convocaron a alumnos del Instituto Tecnológico de San Marcos en el estado de Guerrero, México.

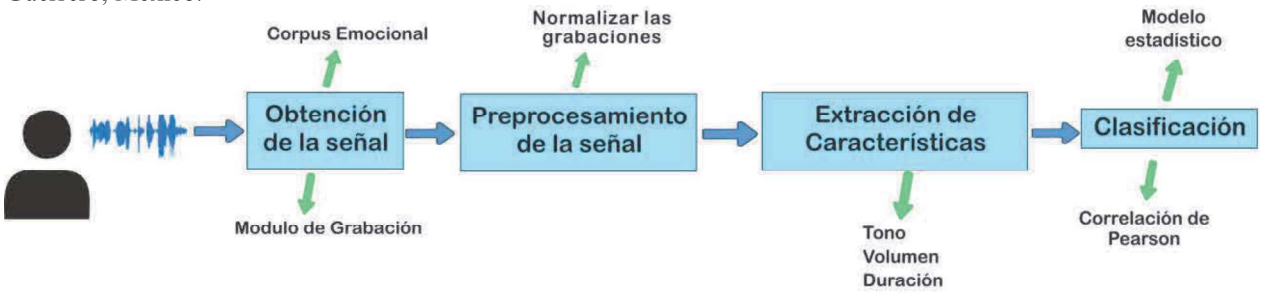

Figura 1. Diagrama de bloques de la estructura general del sistema propuesto.

Corpus Emocional

El corpus de voz consta de 880 frases en español grabadas por alumnos del *ITSM* con edades entre 18 y 26 años. estas frases expresan 5 estados emocionales diferentes: disgusto, ira, felicidad, miedo y neutral con un total de 16 frases por cada uno de ellos. 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. La figura 2 muestra la interfaz gráfica que se utilizó para crear el corpus emocional.

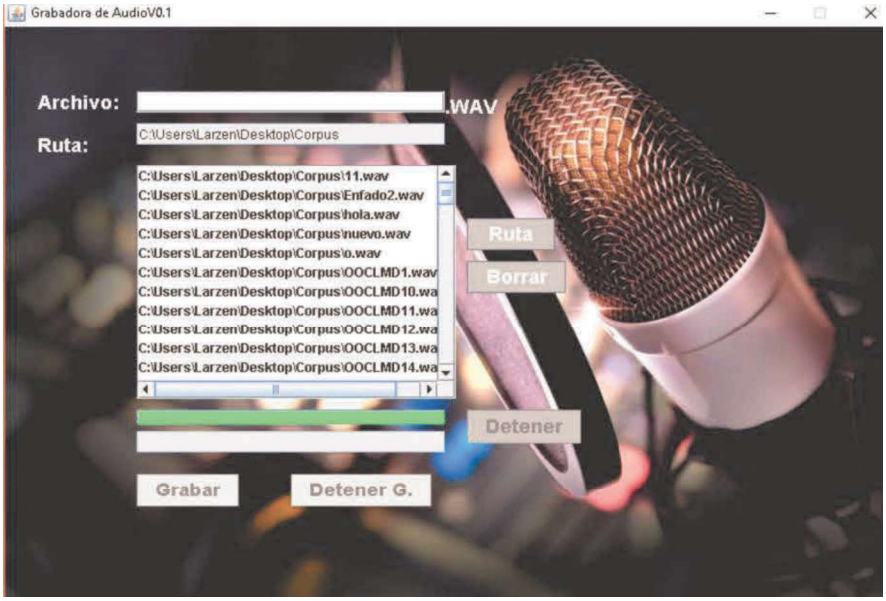

Figura 2. Interfaz principal de la grabadora de voz.

Como se muestra en la figura 3 la grabación de audio se realizó en un aula cerrada, ubicada en el laboratorio de computo del Tecnológico de San Marcos *(ITSM),* con el fin de reducir ruidos y distractores.

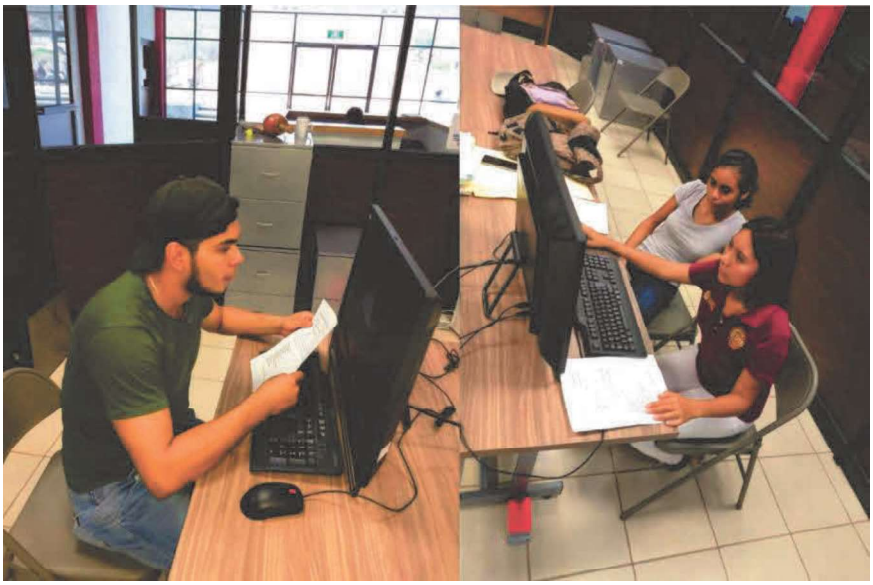
Figura 3. Alumnos del ITSM en el proceso de grabación

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.

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

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

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.

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

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.

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

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.

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

Comentarios Finales

Se ha creado corpus emocional mexicano para la prueba del algoritmo de reconocimiento de emociones en la voz utilizando un método estadístico como clasificador. Cabe mencionar que se esperan agregar más características al vector para aumentar la eficiencia del reconocedor utilizando técnicas de *Machine Learning*.

Resumen de resultados

En la etapa de procesamiento se logró procesar la señal de audio como se muestra en la figura 3.

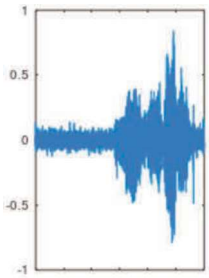

Figura 4. La Frase: "Vivirás conmigo" grabada por alumnos del ITSM.

En la etapa extracción de características se lograron obtener el espectro de frecuencia que contiene un vector con patrones necesarios para detectar las 5 emociones que se muestran en la figura 5.

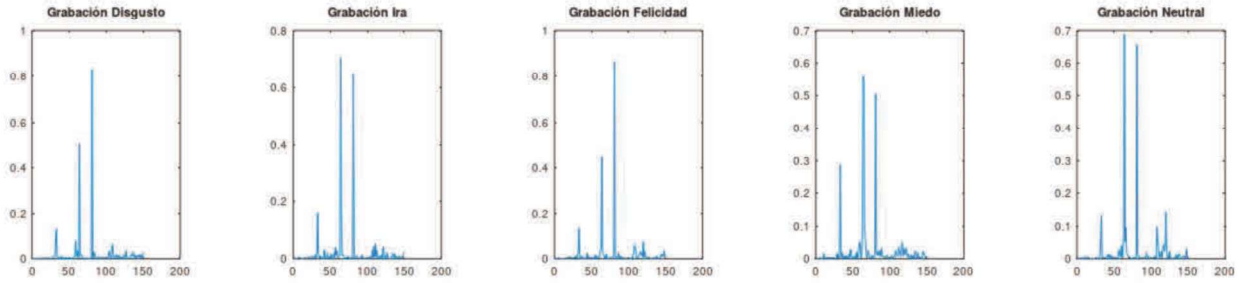

Figura 5. patrón de la frase pronunciada en cada emoción

En la etapa de clasificación que se utilizaron las diferencias entre el vector a clasificar y los vectores de características almacenados en la base de datos mediante la correlación de Pearson detectando las diferencias por medio del coeficiente de error. En el cuadro 3 se muestra la el éxito en la detección de la emoción "Disgusto" detectando el coeficiente de error que es el más cercano a 0 y así señalando la semejanza mas significativa en el vector de características con emoción a reconocer.

CORRELACIÓN DE PEARSON	0.15327
COEFICIENTE DE ERROR DISGUSTO:	0.018317
COEFICIENTE DE ERROR IRA:	0.021492
COEFICIENTE DE ERROR FELICIDAD:	0.022185
COEFICIENTE DE ERROR MIEDO:	0.020861
COEFICIENTE DE ERROR NEUTRAL:	0.052955
EMOCIÓN IDENTIFICADA:	DISGUSTO

Cuadro 3. Reconocimiento del "Disgusto" mediante el método de correlación muestral

El cuadro 4 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 y la Ira son emociones son claramente identificadas con mayor exactitud por este clasificador

		Prediccion					
		Disgusto	Ira	Felicidad	Miedo	Neutral	Totales
Observaciones	Disgusto	92	5	4	4	17	122
	Ira	10	96	5	1	14	126
	Felicidad	13	10	87	3	15	128
	Miedo	8	13	2	88	12	123
	Neutral	24	32	10	2	59	127
	Totales	147	156	108	98	117	

Cuadro 4. Matriz de confusión para el algoritmo de clasificación Utilizando correlación de Pearson

Conclusiones

Se realizo una investigación científica de los parámetros acústicos para el reconocimiento de estados emocionales en la voz en el área de Sistemas Inteligentes, se obtuvo de acuerdo a los resultados obtenidos un algoritmo capaz de reconocer más de un 80% de las frases con emoción actuada por los alumnos del ITSM. Los resultados demuestran la necesidad de más parámetros en la etapa de extracción de características. Fue necesario generar un corpus debido a la falta de estandarización en la obtención de emociones y la inexistencia de normas que den garantía en la reproductibilidad. Es indispensable utilizar más métodos de clasificación y técnicas de aprendizaje artificial para tener una mayor eficiencia en la clasificación.

Referencias

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., and Taylor, J. G. (2001). Emotion recognition in human computer interaction. IEEE Signal processing magazine, 18(1):32-80. Hasrul, M. N., Hariharan, M., & Yaacob, S. (2012, February). Human Affective (Emotion) Behaviour Analysis using Speech Signals: A Review. In International Conference on Biomedical Engineering (ICoBE) (Vol. 27, p. 28).

Kim, E. H., Hyun, K. H., Kim, S. H., and Kwak, Y. K. (2007). Speech emotion recognition using eigen-fft in clean and noisy environments. In Robot and Human interactive Communication, 2007. RO-MAN 2007. The 16th IEEE International Symposium on, pages 689-694. IEEE.

Levenson, R.W. (1994). Human emotion. A functional view. In P. Ekman & R.J. Davidson (Eds). The nature of Emotions: Fundamental Questions (pp. 123-126). New York: Oxford University Press.

Montero Martínez, J. M. (2003). Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano. PhD thesis, Telecomunicacion

Nwe, T. L., Foo, S. W., and De Silva, L. C. (2003). Speech emotion recognition using hidden markov models. Speech communication, 41(4):603-623.

Ortego Resa, C. et al. (2009). Detección de emociones en voz espontánea. B.S. thesis.

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

Pérez Pascual, F. (2017). Speech emotion recognition: Un sistema de reconocimiento de emociones por voz basado en Ivectors (Bachelor's thesis, Universitat Politècnica de Catalunya).

Solís, V. J. F. (2011). Modelo de procesamiento de voz para la clasificación de estados. PhD thesis, Instituto Politécnico Nacional. Centro de Investigación en Computación.

Puebla Romero, T., C. Dominguini y T. T. Micrognelli. "Situaciones inesperadas por el uso de las ecuaciones libres en la industria cocotera," *Congreso Anual de Ingeniería Mecánica*, Instituto Tecnológico y Científico Gatuno, 17 de Abril de 2005.

Washington, W. y F. Frank. "Six things you can do with a bad simulation model," *Transactions of ESMA*, Vol. 15, No. 30, 2007.

Wiley J. y K. Miura Cabrera. "The use of the XZY method in the Atlanta Hospital System," *Interfaces*, Vol. 5, No. 3, 2003.