Reconocimiento de emociones a través del análisis de la voz

Ing. Vicente Bello Ambario [1], Dra. Miriam Martínez Arroyo [2],
Dr. José Antonio Montero Valverde [3] y MTI. Juan Miguel Hernández Bravo [4]

***Resumen*— El reconocimiento automático de las emociones humanas mediante el análisis de la voz, 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 coeficientes de correlación de Pearson. Por el momento se muestran resultados parciales obtenidos en las fases iniciales de este proceso, para ello se utiliza la base de datos de Berlín la cuál es referencia en estos trabajos.**

Palabras clave*— *REV, FFT, Corpus Oral, coeficientes de correlación de Pearson, Reconocimiento de patrones.

Introducción

En un entorno cotidiano las personas expresan y comunican sus emociones y estados afectivos mediante información procedente del rostro (expresiones faciales), del habla tanto con información explícita o lingüística (el mensaje), como implícita o paralingüística (características prosódicas como el tono de la voz, la intensidad, la velocidad o el ritmo) y del cuerpo (gestos de las manos y posturas o movimientos del cuerpo) *(Marco G., 2017)*. El Reconocimiento de emociones de la voz (*REV*) es un sistema de identificación de emociones a través de un locutor humano. Este proceso permite reconocer el impulso emocional causado por un estímulo temporal llamado emoción interacción persona-computadora, a diferencia del estado emocional, la voz emotiva suele durar pocos minutos. Los diferentes estados emocionales de un hablante producen cambios fisiológicos en el aparato fonador, lo que se ve reflejado en la variación de dichas características. 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. *(Duque & Morales, 2007)*.

Se han presentado muchos enfoques para reconocer estados afectivos basados en características específicas del habla. Para este propósito se han utilizado características a corto plazo (formantes, ancho de banda de formantes, frecuencia de tono / fundamental y energía de registro) y características a largo plazo (media de tono, desviaciones estándar de tono, envolventes temporales de tono y energía). Las características a corto plazo reflejan las características del habla local en una ventana de corto tiempo, mientras que las características a largo plazo reflejan las características de la voz sobre un enunciado completo *(Li & Zhao, 1998)*. El Tono (*Pitch)*, frecuencia fundamental (F_0), la intensidad de la señal de voz (*energía*) y la tasa de habla se han identificado como importantes indicadores de la emoción en la voz (Ververidis & Kotropoulos, 2006) *(Duque &, 2007*)

Análisis de las características de la voz emotiva

La información acústica describe sonidos, lenguaje y la expresión emotiva. Estos elementos incluyen fonemas, la forma de articulación y en que estado de ánimo se pronuncie. 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. De acuerdo al tipo de información las características acústicas suelen agruparse en:

- Espectrales: Describen las propiedades de una señal en el dominio de la frecuencia mediante harmónicos y formantes
- Calidad de Voz: Definen estilos al hablar como neutral, susurrante, jadeante, estrepitoso resonante, sonoro, ruidoso

[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] 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] El MTI. Juan Miguel Hernández Bravo es Profesor la MSC del ITA (TecNM), Gro., Méx. jmhernan@yahoo.com

- Prosódicas: Describen fenómenos suprasegmentales como entonación, volumen, velocidad, duración, pausas y ritmo

Análisis de emociones.

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. Aunque el principio de una emoción puede ser fácilmente distinguible de un estado de ánimo, es imposible definir cuando una emoción se convierte en un estado de ánimo; posiblemente por esta razón, el concepto de emoción es usado como un término general que incluye al del estado de ánimo (Ortego Resa, 2009). Las emociones pueden ser vistas por su valor adaptativo con las tareas fundamentales de la vida. Cada emoción tiene características únicas y otras que son comunes por ser producto de nuestra evolución (Ekman, 1992). Las emociones básicas 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 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 Villarreal, 2011).

Análisis de señales.

La capacidad auditiva del ser humano varía en un rango de frecuencias de 20 Hz a 20,000 Hz (Herrera, 2006). Los sonidos emitidos al hablar se encuentran de 100 Hz a 15,000Hz en mujeres y en hombres de 400Hz a 15,000 Hz. (Hernández, 2016). 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. La alegría 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. El habla triste exhibe un tono medio más bajo que el normal, un estrecho rango y una velocidad de locución lenta. El miedo se distingue comparando el tono medio con los otros cuatros emociones primarias estudiadas, se observa 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). En la figura 1. Se pueden percibir, en las gráficas, las señales de voz que expresa en la palabra en serbio "da", que en castellano se puede traducir como "si"; dichas señales fueron expresadas en 5 diferentes emociones y cabe hacer notar las diferencias en duraciones de tiempo, así como las diferencias en amplitud (*Kim, et al.,2007)*.

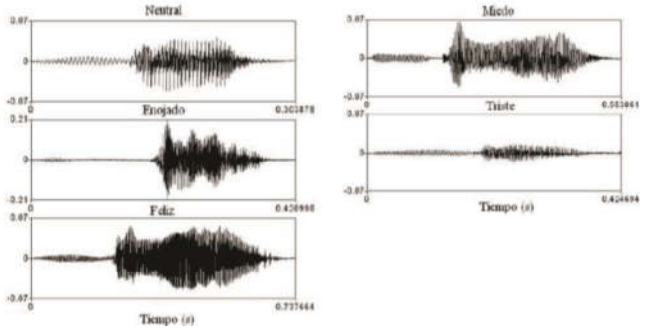

Figura 1. Palabra "da" en serbio, se traduce como "si" en castellano (Solís Villarreal, 2011).

Parámetros del habla y la transmisión de emociones

Los efectos fisiológicos en el habla (acústicos, prosódicos y léxicos), se utilizan para expresar emociones, dentro de los cuales se consideran los más importantes: *pitch*, duración, calidad de voz y forma del pulso glotal y tracto vocal.

Tono.

El tono *(pitch en inglés),* se podría definir como la impresión perceptiva que nos produce la frecuencia fundamental (F_0) de la onda sonora. es, por tanto, una cualidad subjetiva dependiente de una propiedad física (Monzo, 2010). El pitch también conocido como melodía (Garrido, 1991) tiene las siguientes propiedades:

- *Frecuencia fundamental (F_0).* Se define como el ciclo periódico de la señal de voz, siendo el resultado de la vibración de los pliegues vocales. Su medida habitual es el hercio *(Hz),* que da una medida de los ciclos por segundo.

- *Curva de F_0 o melódica*. Se trata de la secuencia de valores de F_0 para una elocución, y se relaciona con la percepción de la entonación del habla.
- *Jitter*. Parámetro que caracteriza la perturbación de F_0 debida a fluctuaciones en los tiempos de apertura y de cierre de los pliegues vocales de un ciclo al siguiente.

Volumen.

El volumen es aire que al salir de los pulmones golpea la glotis y produce vibraciones. Tiene efectos en el oyente porque transmite emociones. Un volumen de voz alto se asocia a la agresividad, nerviosismo, tensión y lejanía. Al contrario, un volumen bajo puede sugerir depresión, cansancio y proximidad. Las propiedades relacionadas con el volumen son los siguientes:

- *Intensidad*. Medida de la energía de la onda acústica. Habitualmente se utiliza una transformación logarítmica de la amplitud de la señal, llamada decibelio (*dB*), que representa mejor la percepción humana del sonido.
- *Shimmer*. Parámetro que caracteriza la perturbación en la intensidad debida a fluctuaciones en la amplitud de un ciclo al siguiente.

Duración

La duración es la componente de la prosodia descrita por la velocidad del habla y la situación de los acentos, y cuyos efectos son el ritmo y la velocidad. El ritmo en el habla deriva de la situación de los acentos y de la combinación de las duraciones de las pausas y de los fonemas. Las propiedades relacionadas con los aspectos temporales del habla son:

- *Velocidad del habla*. Se mide a partir de la duración de los segmentos del habla o como el número de unidades lingüísticas por unidad temporal (p.ej. sílabas por segundo).
- *Pausas*. El número y la duración de los silencios en la señal de voz es un parámetro del que habitualmente se realiza su medida

Efectos de las emociones en el habla

La tabla 1 presenta un resumen de las relaciones entre las emociones y los parámetros del discurso. Como se puede observar en la tabla únicamente aparecen cinco emociones. Estas corresponden con las emociones primarias o básicas.

	Felicidad	**Enfado**	**Asco**	**Miedo**	**Tristeza**
Velocidad del habla	Acelerada o retardada	Ligeramente acelerada	Mucho más acelerada	Muy acelerada	Pausada
F_0	Incremento de la media, variabilidad	Incremento de la media, mediana, variabilidad	----------	Incremento en la F_0 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
F_0 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

Tabla 1. Emociones y características del habla (Cowie et al. (2001) (Ortego Resa, 2009).

Para el desarrollo de un sistema de reconocimiento de emociones es importante contar precisamente con una base de datos apropiada para el entrenamiento (modelado) del mismo *(Pérez Gaspar, et al., 2015). Berlín Emotional Speech* es una base de datos alemana, cuenta con 7 emociones, 10 actores profesionales (5 hombres y 5 mujeres) expresan 10 diferentes emociones en idioma alemán. Este corpus fue grabado mediante frecuencia de muestreo de 16,000 Hz, con 16 bits de precisión en formato WAV (*Kadiri & Yegnanarayana, 2017)*. La base de datos consta de 535 instancias, de las cuales 127 corresponden al estado de enojado, 81 a aburrido, 46 para disgustado, 69 para miedo, 71 para feliz, 62 a triste y 79 para neutral (Solís Villarreal, 2011).

Metodología

La Figura 2 presenta el diagrama a bloques de la estructura general del sistema propuesto por cuatro bloques principales: El primer bloque define las características de los dispositivos para capturar la señal de la voz, el segundo bloque se encarga del preprocesamiento de la señal, en el cual se realizan tareas de normalización y segmentación de la señal de entrada. El tercer módulo se encarga de la extracción de características de la Señal de voz, el método de obtención de patrones presentado en este artículo es la Transformada rápida de Fourier (*FFT*), aplicada a la señal de voz. El cuarto módulo es el encargado de la clasificación. El reconocedor propuesto es la correlación entre señales, el cual arroja datos, dependiendo de qué tan parecida es una señal con la otra.

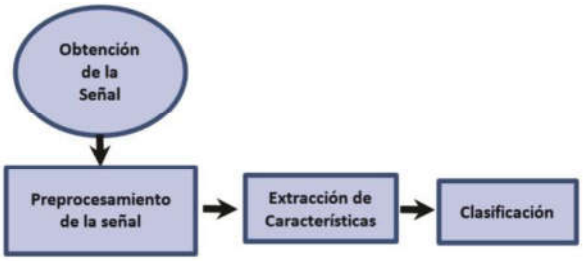

Figura 2. Proceso de reconocimiento de emociones en la voz

Obtención de la señal.

Hay dos factores importantes durante este proceso. Primero está la taza de muestreo o que tan seguido los valores de voltaje son grabados (Fs = 44100 Hz). Segundo, son los bits por segundo, o que tan exactamente los valores son grabados (Tasa de bits = 24). Un tercero podría ser el número de 32 canales (mono o estéreo), pero para las aplicaciones de reconocimiento de voz un canal mono es suficiente. La mayoría de aplicaciones vienen con valores pre-determinados, para desarrollo del código se debería de cambiar los parámetros para ver lo que mejor funciona en el algoritmo. En la figura 3 se muestran dos señales grabadas de dos actores con la frase en alemán "*Der Lappen liegt auf dem Eisschrank*" (en inglés: "*The cloth is lying on the fridge*." Y en español:" El paño esta tirado en la nevera"), expresando el enojo en sus palabras. Haciendo uso de un programa implementado en MATLAB: se graban dos segundos de audio con una frecuencia de muestreo de 44100 Hz y una tasa de audio de 24 bits. La grabación da como resultado un vector de 88 mil datos, de los que se discriminarán los datos significativos mediante un umbral de 0.1.

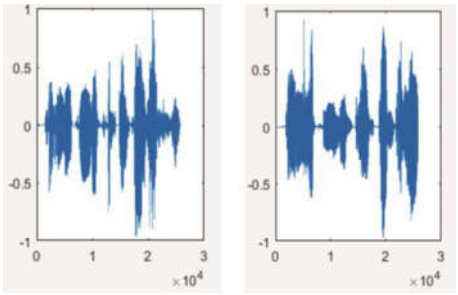

Figura 3. La Frase: *"Der Lappen liegt auf dem Eisschrank"* por dos actores alemanes.

Preprocesamiento de la señal

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

- Guardar los dos audios en variables para su tratamiento
- Obtener los parámetros acústicos como el pitch o la altura
- Normalizar las grabaciones
- Se preparan los audios almacenados en la base de datos con la misma frase a evaluar y cortando los primeros 25782 primeros valores

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. A continuación, se presenta el proceso para extracción características:

- Se obtiene la transformada de Fourier de grabación previamente normalizada

- Se obtiene el conjugado
- Guarda las frecuencias arriba de 100 Hz
- Se normaliza el Vector

El resultado de la extracción de la *FFT* de cada tramo de las grabaciones se muestra en la figura 4. El objetivo de generar un módulo de graficas con dominios de la frecuencia y tiempo es de observar las frecuencias y su variación en el tiempo. Se promedian las *FFT* de cada tramo, para obtener un patrón de la frase pronunciada. Se obtienen el espectrograma del audio para obtener los valores de tono e intensidad del audio.

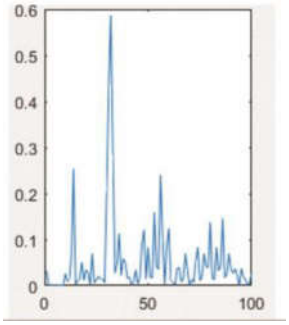

Figura 4. Espectro de grabación aplicando FFT

Clasificación

Una vez que se obtuvo el patrón de la frase "*Der Lappen liegt auf dem Eisschrank*" por los dos actores alemanes, se extraen las mismas características a la misma frase expresada con diferentes emociones que se encuentran en el repositorio de datos como se muestra en la figura 5. Para la clasificación de emociones se utilizó el coeficiente de correlación de Pearson, pensado para variables cuantitativas (escala mínima de intervalo), es un índice que mide el grado de covariación entre distintas variables relacionadas linealmente. Adviértase que decimos "variables relacionadas linealmente". Esto significa que puede haber variables fuertemente relacionadas, pero no de forma lineal, en cuyo caso no proceder a aplicarse la correlación de Pearson. El coeficiente de correlación de Pearson es un índice de fácil ejecución e, igualmente, de fácil interpretación. Digamos, en primera instancia, que sus valores absolutos oscilan entre 0 y 1. Esto es, si tenemos dos variables X e Y, y definimos el coeficiente de correlación de Pearson entre estas dos variables como r_{xy} entonces: $\leq \boldsymbol{r_{xy}} \leq \mathbf{1}$ **.** Se especifican "valores absolutos" ya que en realidad si se contempla el signo el coeficiente de correlación de Pearson oscila entre −1 y +1. No obstante ha de indicarse que las magnitudes de la relación vienen especificadas por el valor numérico del coeficiente, reflejando el signo la dirección de tal valor. En este sentido, tan fuerte es una relación de +1 como de -1. En el primer caso la relación es perfecta positiva y en el segundo perfecta negativa. Posteriormente se calcula el coeficiente de error de cada una de las emociones grabadas y la grabación que tenga el error mínimo será la emoción que evalúa el sistema.

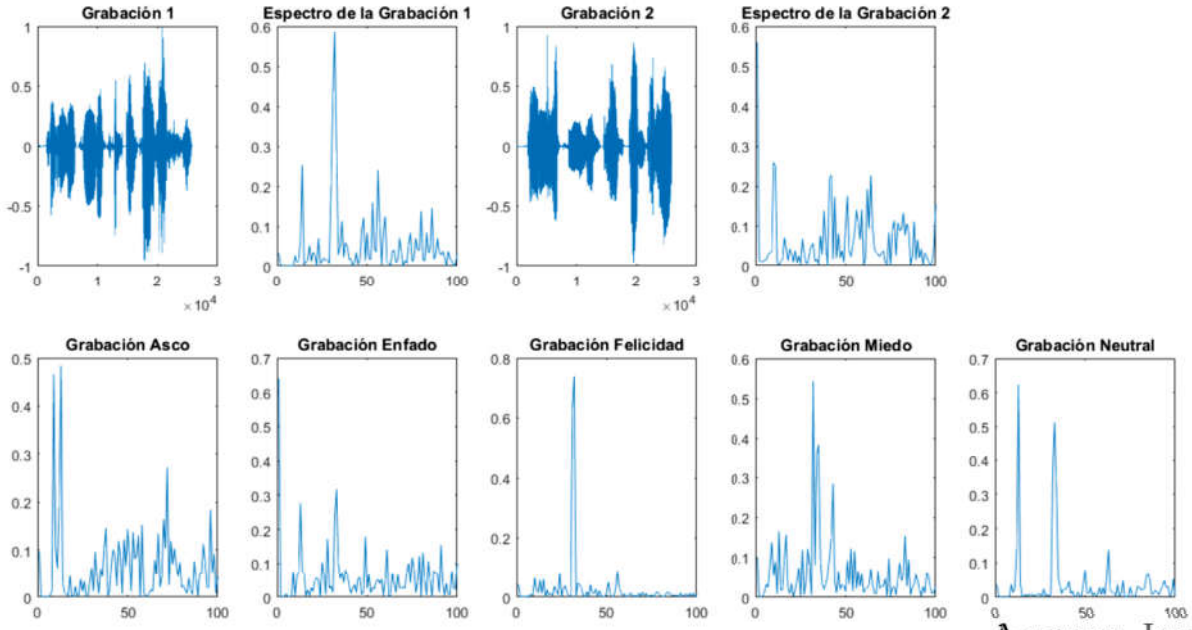

Figura 5. Espectro de las grabaciones la base de datos de la frase "Der Lappen liegt auf dem Eisschrank" en diferentes emociones

Pruebas y Resultados

Aplicando el método de correlación de Pearson la fase de clasificación, se compara la palabra a reconocer con los valores absolutos de los coeficientes de error de cada emoción se obtiene el reconocimiento de la emoción. Los resultados se muestran en la tabla 2.

Correlación de Pearson	**-0.0286**
Correlación de Error ASCO:	0.0678
Correlación de Error ENFADO:	0.0595
Correlación de Error FELICIDAD:	0.0661
Correlación de Error MIEDO:	0.0624
Correlación de Error NEUTRAL:	0.0702
Emoción Detectada:	ENFADO

Tabla 2. Reconocimiento del "Enfado" mediante el método de correlación muestral.

El coeficiente de correlación es un indicador que nos permite establecer la covariación conjunta de las dos señales de voz a reconocer y así tener la universalidad suficiente para poder establecer la comparación entre distintos casos de emociones (lineal, de Pearson). La correlación negativa se usa para detectar el coeficiente menor (0.0595), por lo tanto, esta se convierte en la emoción a reconocer por el sistema. Con la base de datos de Berlín, se ha detectado la emoción de Enfado únicamente con este proceso. El sistema está basado en el software de Matlab para el procesado de la señal y como interfaz gráfica, debido a que es muy precisó al trabajar con variables vectoriales y matriciales. Es apropiado para el caso de muchas señales de interés, donde la frecuencia de muestreo sea menor que 44.1 KHz

Comentarios Finales

El *REV* es un área de investigación que tiene diversas aplicaciones, sin embargo tiene mayor importancia en sistemas de interacción humano-computadora, como puede ser un Sistemas Tutor Inteligente (STI), ya que permite mejorar la calidad del aprendizaje basándose en el estado emocional del alumno. En este trabajo se muestra la comprensión de los elementos del habla que ayudan a reconocer las emociones y una metodología para el reconocimiento de emociones analizando segmentos de voz. La metodología se basa principalmente en la transformada rápida de Fourier (FFT) y coeficientes de correlación de Pearson. Por el momento se muestran resultados parciales obtenidos en las fases iniciales de este proceso, para ello se utiliza la base de datos de Berlín la cuál es referencia en estos trabajos. Gracias al tratamiento de señales acústicas se ha encontrado correlaciones significativas a partir de una base de datos de emociones. Se ha detectado exitosamente la emoción del enfado que junto con la alegría presentan una similitud en términos de valencia o de fuerza. Como trabajo futuro se pretende crear un corpus de emociones espontaneas con estudiantes, probar algoritmos de Clasificación que darán como resultado el reconocimiento de todas las emociones primarias y determinar el clasificador con mejores resultados.

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 y tener la certeza si el número de errores es tolerable y finalmente integrar este reconocimiento de emociones a un STI.

Referencias

Cowie, R. D.-C. (2001). Emotion recognition in human-computer interaction. IEEE Signal processing magazine, 18(1), 32-80.

Duque Sánchez, C., & Morales Pérez, M. (2007). Caracterización de voz empleando análisis tiempo-frecuencia aplicada al reconocimiento de emociones (Tesis de pregrado). Pereira: Universidad Tecnológica de Pereira.

Ekman, P. (1992). An argument for basic emotions. Cognition & emotion, 6(3-4), 169-200.

Garrido, J. M. (1991). Modelización de patrones melódicos del español para la síntesis y el reconocimiento. Barcelona, España: Depto. de Filologia Espanola, Universidad Autonoma de Barcelona.

Hernández, R. (2016). Sistema de control activado por voz para uso en domótica (Tesis de maestría). Xalapa: UNIVERSIDAD VERACRUZANA.

Herrera, A. L. (2006). Identificación automática del lenguaje hablado sin reconocimiento fonético de la señal de voz.

Kadiri, S. R. (2017). Epoch extraction from emotional speech using single frequency filtering approach. Speech Communication, 86, 52-63.

Kim, E. H. (2007). Speech emotion recognition using eigen-fft in clean and noisy environments. In Robot and Human interactive Communication. The 16th IEEE International Symposium on, 689-694.
Li, Y., & Zhao, Y. (1998). Recognizing emotions in speech using short-term and long-term features. In Fifth International Conference on Spoken Language Processing.
Marco Giménez, L. (2017). Evaluación y uso del estado emocional en entornos educativos interactivos (Tesis doctoral). Valencia: Departamento de Informática de la universidad de Valencia).
Monzo, C. (2010). Modelado de la cualidad de la voz para la síntesis del habla expresiva (Tesis Doctoral). Escola Técnica Superior d'Enginyeria Electronica i Informatica La Salle-Ramon Llull.
Ortego Resa, C. (2009). Detección de emociones en voz espontánea (Tesis de pregrado). Madrid: Universidad Autónoma de Madrid.
Pérez-Gaspar, L. A.-R. (2015). Integración de optimización evolutiva para el reconocimiento de emociones en voz. Research in Computing Science, 93, 9-21.
Solís Villarreal, J. F. (2011). Modelo de procesamiento de voz para la clasificación de estados. Instituto Politécnico Nacional.
Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. Speech communication, 48(9), 1162-1181.