Reconocimiento de emociones en el habla

Julián David Echeverry Correa[1]
Mauricio Morales Pérez[2]

Resumen

Se presenta en este trabajo una metodología para la caracterización de la señal de voz aplicada al reconocimiento de estados emocionales. Son estudiadas cuatro emociones primarias (alegría, enojo, sorpresa y tristeza) y un estado neutral. Se realizó un análisis en el dominio temporal y un análisis acústico empleando los *MFCC* (*Mel Frequency Cepstral Coefficients*). Las pruebas comprueban la efectividad de la metodología en el reconocimiento de las emociones superando el reconocimiento realizado por un grupo de personas. Se obtiene un porcentaje de 94.00% de acierto en el reconocimiento de emociones trabajando sobre la base de *SES* (*Spanish emotional speech*).

Palabras clave

Reconocimiento de emociones, procesamiento señal de voz, *MFCC*.

Abstract

A methodology of feature extraction in emotional speech for emotion recognition is proposed. Four primary human emotions, including happiness, anger, surprise and sadness are investigated.

---

[1] Ingeniero Electrónico, M.Sc. en Ingeniería Eléctrica. Profesor del Programa de Ingeniería Eléctrica. Universidad Tecnológica de Pereira.

[2] Ingeniero Electricista. Estudiante Maestría en Ingeniería Eléctrica. Universidad Tecnológica de Pereira.

Fecha de recepción: 7 de noviembre de 2008
Fecha de aceptación: 11 de diciembre de 2008

In order to recognize emotional states, acoustic MFCC (Mel frequency cepstral coefficients) and time representation features are extracted from voice recordings. Experiments indicate that emotion recognition effectiveness comparable to human listeners can be achieved. Recognition accuracy of 94.00% for emotion detection was obtained from database SES (Spanish emotional speech).

Key words

Emotion recognition, signal speech processing, MFCC.

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.

Los primeros estudios sobre la influencia de las emociones en el comportamiento de los seres fueron realizados por Darwin (Darwin, 1873), quien vinculó la expresión sonora de las emociones al instinto, o cuanto menos, a conductas heredadas y por tanto, que pueden aparecer en el individuo de modo totalmente involuntario. Desde ese momento, la expresión emocional ha tendido a ser considerada como el carácter del habla humana más claramente universal y transcultural.

La voz está directamente relacionada con los diferentes estados emocionales, ya que éstos 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 sorprendido.

Diferentes investigaciones han encontrado algunos de los componentes del habla que se emplean para la expresión de emociones: el *pitch*, la calidad de voz y la duración del habla (Scherer, 1979). Se plantea, entonces, el desarrollo de una metodología que permita la extracción de parámetros representativos del habla, empleando para dicha extracción un análisis temporal y acústico de la señal de voz, con el fin de 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.

2. Metodología

Con el fin de realizar el procesamiento de la señal, se emplea la metodología presentada en la figura 1.

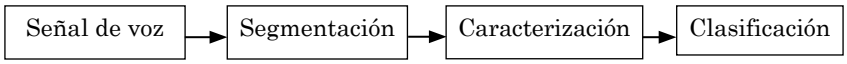

Figura 1. Metodología empleada para el procesamiento de la señal de voz

2.1 Segmentación

La voz es una señal biológica resultado de un proceso aleatorio con carácter estacionario o no, según la longitud del intervalo de observación. Generalmente se examina en intervalos de tiempo suficientemente cortos (entre 20 y 60 ms) donde sus características estadísticas permanecen invariantes (Ávila & Quintana, 1994).

Para definir el intervalo de estacionariedad, la señal fue dividida en intervalos de longitud L sobre los cuales se aplicó el test estadístico *run test* (Shamugan, 1998) obteniéndose una duración de 30 ms como intervalo óptimo de estacionariedad. La señal fue segmentada por medio de ventanas rectangulares de 30 ms y un traslape del 50% con el fin de reducir pérdida de información por los bordes de las ventanas.

2.2 Caracterización

Para la extracción de características se emplearon técnicas que representaran el comportamiento temporal y espectral de la señal, obteniendo parámetros referentes a las componentes del habla relacionadas con la expresión emocional. Las técnicas de extracción empleadas son *raw data* y *MFCC*.

2.2.1 *Raw Data*

Raw data, como su nombre lo indica (datos crudos). Este tipo de análisis consiste en trabajar directamente sobre todo el universo de datos en el dominio temporal, es decir, la señal no es llevada a

ningún tipo de representación o reducción previa de parámetros. En la figura 2 se muestra la evolución temporal de la palabra /*vivirás*/ en diferentes estados emocionales.

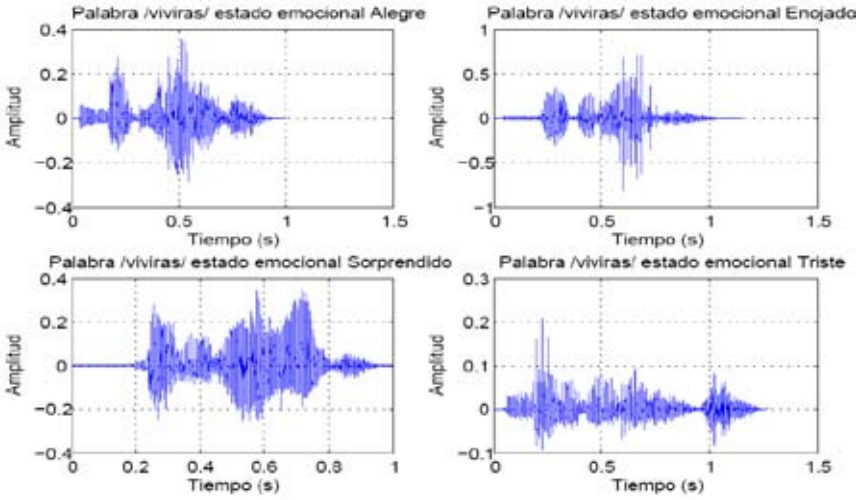

Figura 2. Evolución temporal palabra /vivirás/ en diferentes estados emocionales

Luego de la segmentación, sobre cada una de las tramas de la señal se calcularon los siguientes momentos estadísticos: media, mediana, máximo, mínimo, desviación estándar, varianza, asimetría y curtosis.

También fueron extraídos los parámetros conocidos como: **perturbación de la amplitud, perturbación de la amplitud máxima o *Shimmer* y coeficiente de amplitud.**

Para el cálculo de la perturbación de la amplitud, que es simplemente una medida de la variación de la amplitud de la señal para las distintas tramas, se empleó la ecuación 1.

$$V(p)=\frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A(i+1)-A(i)\right|}{\frac{1}{N}\sum_{i=1}^{N-1}\left|A(i)\right|} \qquad (1)$$

Donde *A(i)* es el valor de la amplitud para la muestra *i*, *N* es el número de muestras de la trama y *V(p)* es el valor de la perturbación de la amplitud para la trama *p*.

Para el cálculo del *Shimmer,* que es una medida de la variación de la amplitud máxima (pico a pico) de todas las tramas, se empleó la ecuación 2.

$$Sh = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1}\left|A\max(i+1) - A\max(i)\right|}{\frac{1}{N}\sum_{i=1}^{N-1}\left|A\max(i)\right|} \qquad (2)$$

Donde $A_{max}(i)$ es el valor de la amplitud máxima de la trama *i*, *N* es el número de tramas y *Sh* es el valor del *Shimmer* de la señal.

El coeficiente de amplitud es una medida de la desviación estándar relativa de la amplitud, calculado como la razón entre la desviación estándar de la amplitud y el valor promedio de la amplitud como se muestra en la ecuación 3.

$$Ca(p) = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N-1}\left(A(i+1) - A(i)\right)^2}}{\frac{1}{N}\sum_{i=1}^{N-1}\left|A(i)\right|} \qquad (3)$$

Donde *A* es el valor de la amplitud para la muestra *i*, *N* es el número de muestras de la trama y *Ca(p)* es el valor del coeficiente de amplitud para la trama *p*.

Se obtienen un total de diez parámetros entre los momentos estadísticos, la perturbación de amplitud y el coeficiente de amplitud, estas características son dinámicas pues se calcula un valor para cada una de las tramas de la señal y toman el nombre de contornos dinámicos, ya que el parámetro estudiado varía respecto al tiempo. El *Shimmer* se considera una característica estática pues se calcula un valor puntual para toda la señal.

Con el fin de trabajar sólo sobre características estáticas se obtiene un vector de características estadísticas compuesto por la media de cada uno de los contornos de los ocho momentos estadísticos. Un esquema que representa esta metodología se muestra en la figura 3.

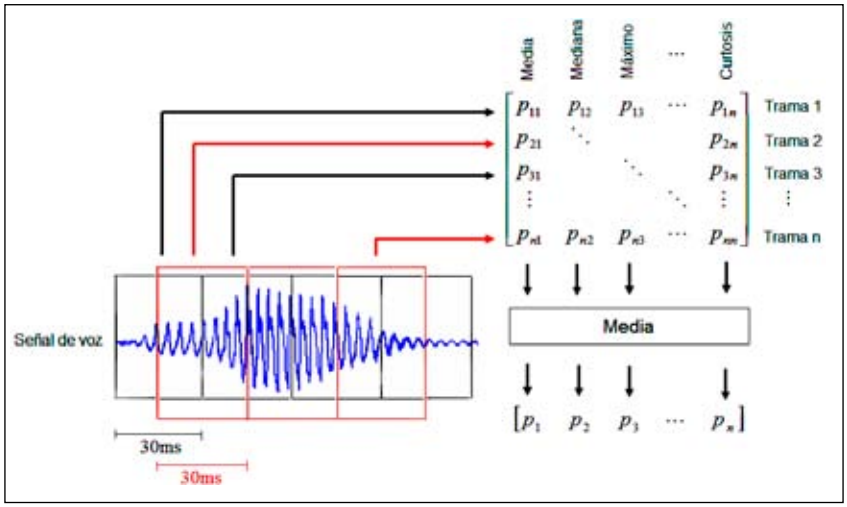

Figura 3. Extracción de características estadísticas

Sobre los contornos de perturbación de amplitud y coeficiente de amplitud se extraen como características los momentos estadísticos de media, mediana, máximo, mínimo, desviación estándar, asimetría y curtosis.

En total se obtuvo un grupo de 25 características que representan el comportamiento temporal de la señal. En el anexo se encuentra la totalidad de características y su descripción.

2.2.2 *MFCC (Mel frequency cepstral coefficients)*

MFCC está basado en el modelo del sistema auditivo periférico (Saha & Yadhunandan, 2004) en el cual la percepción de los contenidos de frecuencia no sigue un comportamiento lineal. Por esto se

hace necesario realizar una medida del contorno de la frecuencia fundamental sobre la escala Mel, que es una escala logarítmica para valores por encima de 1kHz y lineal por debajo de este valor.

Para el cálculo de los *MFCC* fue empleada la función mfcc.m del *Auditorytoolbox* del software Matlab cuyo algoritmo se describe en la figura 4.

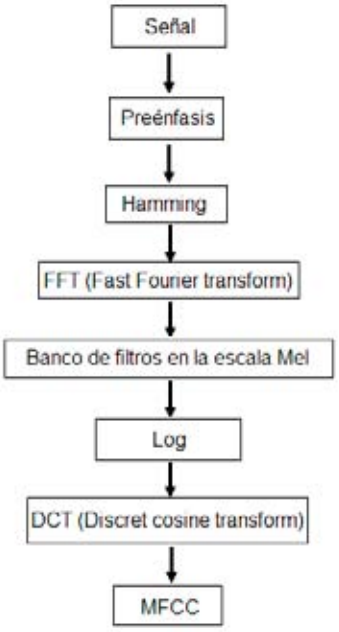

Figura 4. Algoritmo para el cálculo de los *MFCC*

En el banco de filtros utilizado, cada uno de los filtros se encuentra distribuido conforme a la escala *Mel* (ecuación 4), obteniéndose un banco de filtros como el presentado en la figura 5.

$$f_{mel} = 1127.01048 \log_e\left(1 + \frac{f_{Hz}}{700}\right) \qquad (4)$$

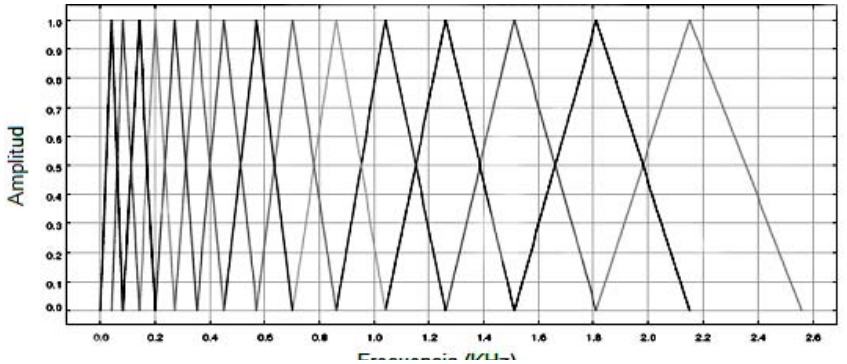

Figura 5. Banco de filtros pasa banda en la escala *Mel*

Para este trabajo se obtuvieron un total de veinte coeficientes *MFCC* por señal, para lo cual se empleó un banco de veinte filtros pasabanda, además se calculó la primera y segunda derivada de dichos coeficientes con el fin de obtener vectores de características que representen la dinámica acústica de la señal de voz.

Sobre los vectores de datos correspondientes a los coeficientes *MFCC*, su primera y segunda derivada se calcularon los siguientes momentos estadísticos: media, mediana, máximo, mínimo, desviación estándar, asimetría y curtosis, con el fin de emplearlos como características estáticas.

En el anexo se presentan las 24 características y su respectiva descripción extraídas a partir de los *MFCC*.

2.3 Clasificación

A partir de las características escogidas, se procede a la etapa de clasificación con el objetivo de validar la metodología de caracterización y entrenar un sistema capaz de realizar el reconocimiento automático de emociones. Para este caso se empleo un clasificador basado en la teoría de *Bayes*, debido a su fácil implementación y poca influencia dentro del proceso, lo que lo hace adecuado para validar la metodología de caracterización.

El teorema de *Bayes* establece que la probabilidad de la clase W_i, dado el vector de características *X,* es igual a la probabilidad *a priori* de la clase por la función de densidad de probabilidad $p(X / W_i)$ sobre la probabilidad total de las muestras, como se observa en la ecuación 5.

$$P\left(W_i / X\right) = \frac{p\left(x / w_i\right) P(w_i)}{p(x)} \qquad (5)$$

Se considera que al tener igual número de señales por clase las probabilidades *a priori* de cada clase deben ser iguales y con un valor de (0.2). Para la función de densidad de probabilidad se emplea una sola Gaussiana, donde la varianza y la media se calculan de acuerdo con las ecuaciones 6 y 7.

$$\overline{X} = \frac{1}{n} \sum_{i=0}^{N-1} X_i \qquad (6)$$

$$\sigma^2 = \frac{1}{n} \sum_{i=0}^{N-1} X_i^2 - \overline{X}^2 \qquad (7)$$

Con el fin de validar el sistema de identificación, la base de datos es dividida en dos grupos: uno de entrenamiento y otro de prueba. Para este caso, se utilizó el método de validación cruzada *leave-one-out,* en el cual una sola señal de la base de datos se emplea como prueba y las restantes como entrenamiento; este proceso se repite sucesivamente hasta que todas las señales hayan sido utilizadas para prueba.

Las pruebas de validación se realizaron para los dos grupos de características obtenidos mediante *raw data* y *MFCC*.

En la validación se tuvieron en cuenta todas las posibles combinaciones de cada uno de los grupos de características, con el fin de determinar cuales características presentaban el mejor acierto global de identificación (es decir el acierto promedio de todos los estados emocionales), para el caso de las 25 características obtenidas empleando el *raw data* se realizó la validación cruzada

de las 2^{25} posibles combinaciones, así mismo se realizó la validación cruzada de las 2^{24} posibles combinaciones de las características obtenidas con lo *MFCC*.

Por último, luego de obtener las características más discriminantes para el reconocimiento de emociones a partir del análisis temporal y acústico, se realiza una nueva validación utilizando ambos grupos de características discriminantes y todas las posibles combinaciones.

3. Materiales

Se trabajó sobre la base de datos SES[3], la cual es una base de datos monolocutor de habla emocional en español en la que el locutor, un actor profesional, simula habla en cuatro estados emocionales **(triste, alegre, sorprendido, enfadado)** y un estado **neutro** (Montero, Gutiérrez, Palazuelos, Aguilera, & Prado, 1998). 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; dicho estudio fue realizado por los creadores y propietarios de la base de datos SES.

[3] 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.

Tabla 1. Matriz de confusión reconocimiento de emociones realizada por receptores humanos sobre la base de datos SES

	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%

Todos los algoritmos de procesamiento fueron desarrollados sobre el paquete de software Matlab versión 7.0.

4. Resultados

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

La combinación de características que presentó mejor resultado empleando *raw data* se presenta en la tabla 2.

Tabla 2. Grupo de características discriminantes para el reconocimiento de emociones empleando *raw data*

Característica	Definición
Maca	Máximo del cociente de amplitud
Mev	Mediana de la perturbación de la amplitud
Mav	Máximo de la perturbación de la amplitud
Asv	Asimetría de la perturbación de la amplitud
Kuv	Curtosis de la perturbación de la amplitud
Mac	Máximo de la señal
Mic	Mínimo de la señal
Sdc	Desviación estándar de la señal
Sh	*Shimmer*

Los resultados de identificación con las características discriminantes empleando *raw data* se presentan en la tabla 3.

Tabla 3. Matriz de confusión reconocimiento de emociones empleando *raw data*

	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
						80.66

La combinación de características que presentó mejor resultado empleando *MFCC* se presenta en la tabla 4.

Tabla 4. Grupo de características discriminantes para el reconocimiento de emociones empleando *MFCC*

Característica	Definición
Mider	Mínimo de la primera derivada de los MFCC
Sdder	Desviación estándar de la primera derivada de los MFCC
Vader	Varianza de la primera derivada de los MFCC
Asder	Asimetría de la primera derivada de los MFCC
Kuder	Curtosis de la primera derivada de los MFCC
Mder2	Media de la segunda derivada de los MFCC
Mader2	Máximo de la segunda derivada de los MFCC
Mider2	Mínimo de la segunda derivada de los MFCC
Vader2	varianza de la segunda derivada de los MFCC
Asder2	Asimetría de la segunda derivada de los MFCC

Los resultados de identificación con las características discriminantes empleando *MFCC* se presentan en la tabla 5.

Tabla 5. Matriz de confusión reconocimiento de emociones empleando *MFCC*

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	26	2	0	1	1	86.67
Enojado	0	26	0	0	4	86.67
Neutro	0	1	29	0	0	**96.67**
Sorprendido	1	2	0	26	1	86.67
Triste	0	4	0	1	25	83.33
Acierto global						88.00

La combinación de características que presentó mejor resultado empleando *raw data* y *MFCC* se presenta en la tabla 6.

Tabla 6. Grupo de características discriminantes para el reconocimiento de emociones empleando *raw data* y *MFCC*

Característica	Definición
Mic	Mínimo de la señal
Sdc	Desviación estándar de la señal
Sh	*Shimmer*
Vader	Varianza de la primera derivada de los MFCC
Asder	Asimetría de la primera derivada de los MFCC
Mder2	Media de la segunda derivada de los MFCC
Mider2	Mínimo de la segunda derivada de los MFCC
Vader2	Varianza de la segunda derivada de los MFCC
Asder2	Asimetría de la segunda derivada de los MFCC

Los resultados de identificación con las características discriminantes empleando *raw* data y *MFCC* se presentan en la tabla 7.

Tabla 7. Matriz de confusión reconocimiento de emociones empleando *raw data* y MFCC

	Salida					
Entrada	Alegre	Enojado	Neutro	Sorprendido	Triste	Acierto (%)
Alegre	28	0	0	2	1	93.33
Enojado	0	28	0	1	1	93.33
Neutro	0	1	29	0	0	**96.67**
Sorprendido	1	0	0	28	1	93.33
Triste	1	1	0	0	28	93.33
Acierto global						94.00

5. Conclusiones

- Se desarrolló una metodología para la identificación de emociones a partir de la señal de voz empleando características acústicas y temporales, la cual presentó altos porcentajes de reconocimiento de los diferentes estados emocionales, la metodología se validó por medio de un clasificador bayesiano, que por su simplicidad y poca influencia dentro del resultado garantiza la robustez de las características y su metodología de extracción.
- Al comparar la tablas 1 y 7 se comprueba la efectividad de la metodología empleada superando con un 13.02% la identificación realizada por un grupo de escuchas humanos. Sería entonces factible la implementación de un sistema en tiempo real que sirva como herramienta de apoyo a un especialista en tratamientos de enfermedades relacionadas con desórdenes del comportamiento.
- Es importante destacar los altos porcentajes de identificación obtenidos, empleando la técnica de datos crudos; las características a partir de este método contienen información sobre la intensidad, duración, acentos, pausas y calidad de voz,

los cuales son componentes importantes en la expresión de emociones.

6. Agradecimientos

Este trabajo fue financiado por Colciencias y la Universidad Tecnológica de Pereira, bajo el contrato 1110-370-19600.

Los autores agradecen al Grupo de Tecnología del Habla, Departamento de Ingeniería Electrónica, Universidad Politécnica de Madrid, especialmente a Juan Manuel Montero el préstamo de la base de datos SES para el desarrollo de este proyecto.

7. Bibliografía

Ávila, E., & Quintana, P. J. (1994). Codificación, sintesís y reconocimiento de voz. Universidad de Gran Canaria.

Darwin, C. (1873). La expresion de las emociones en los animales y el hombre. Madrid: Alianza.

Montero, J. M., Gutiérrez, J., Palazuelos, S., Aguilera, S., & Prado, J. M. (1998). Emotional Speech Synthesis: From Speech Database to TTS . 5th International Conference on Spoken Language Processing.

Saha, G., & Yadhunandan, U. (2004). Modified Mel-Frequency Cepstral Coefficient. Proceedings of the IASTED. IEEE.

Scherer, K. (1979). Personality makers in speech. Cambridge University Press, pág. 147.

Shamugan, K. (1998). Random Signals: Detection Estimation and Data Analysis. Wiley.

8. Anexo

Características extraídas empleando *raw data*

	Característica	Definición
Datos crudos	Mc	Media de la señal
	Mec	Mediana de la señal
	Mac	Máximo de la señal
	Mic	Mínimo de la señal
	Sdc	Desviación estándar de la señal
	Vac	Varianza de la señal
	Asc	Asimetría de la señal
	Kuc	Curtosis de la señal
Perturbación de la amplitud	Mv	Media de la perturbación de la amplitud
	Mev	Mediana de la perturbación de la amplitud
	Mav	Máximo de la perturbación de la amplitud
	Miv	Mínimo de la perturbación de la amplitud
	Sdv	Desviación estándar de la perturbación de la amplitud
	Vav	Varianza de la perturbación de la amplitud
	Asv	Asimetría de la perturbación de la amplitud
	Kuv	Curtosis de la perturbación de la amplitud
	Sh	*Shimmer*
Coeficiente de amplitud	Mca	Media del coeficiente de amplitud
	Meca	Mediana del coeficiente de amplitud
	Maca	Máximo del coeficiente de amplitud
	Mica	Mínimo del coeficiente de amplitud
	Sdca	Desviación estándar del coeficiente de amplitud
	Vaca	varianza del coeficiente de amplitud
	Asca	Asimetría del coeficiente de amplitud
	Kuca	Curtosis del coeficiente de amplitud

Características extraídas empleando MFCC

	Característica	Definición
Coeficientes MFCC	Mmfcc	Media de los MFCC
	Memfcc	Mediana de los MFCC
	Mamfcc	Máximo de los MFCC
	Mimfcc	Mínimo de los MFCC
	Sdmfcc	Desviación estándar de los MFCC
	Vamfcc	Varianza de los MFCC
	Asmfcc	Asimetría de los MFCC
	Kumfcc	Curtosis de los MFCC
Primera derivada de los MFCC	Mder	Media de la primera derivada de los MFCC
	Meder	Mediana de la primera derivada de los MFCC
	Mader	Máximo de la primera derivada de los MFCC
	Mider	Mínimo de la primera derivada de los MFCC
	Sdder	Desviación estándar de la primera derivada de los MFCC
	Vader	Varianza de la primera derivada de los MFCC
	Asder	Asimetría de la primera derivada de los MFCC
	Kuder	Curtosis de la primera derivada de los MFCC
Segunda derivada de los MFCC	Mder2	Media de la segunda derivada de los MFCC
	Meder2	Mediana de la segunda derivada de los MFCC
	Mader2	Máximo de la segunda derivada de los MFCC
	Mider2	Mínimo de la segunda derivada de los MFCC
	Sdder2	Desviación estándar de la segunda derivada de los MFCC
	Vader2	varianza de la segunda derivada de los MFCC
	Asder2	Asimetría de la segunda derivada de los MFCC
	Kuder2	Curtosis de la segunda derivada de los MFCC