Automatic Emotion Detection in Speech Using Mel frequency Cesptral Coefficients

S.Bedoya-Jaramillo[1*], E.Belalcazar-Bolaños[1], T.Villa-Cañas[1], J.R.Orozco-Arroyave[1], J.D. Arias-Londoño[2], J.F.Vargas-Bonilla[1]
[1]Departamento de Ingeniería Electrónica y Telecomunicaciones - Universidad de Antioquia
[2]Departamento de Ingeniería de Sistemas - Universidad de Antioquia

***Abstract*—(Categoría4) Emotional states produce physiological alterations in the vocal tract introducing variability in the acoustic parameters of speech. Emotion recognition in speech can be used in human-machine interaction applications, speaker verification, analysis of neurological disorders and psychological diagnostic tools. This paper proposes the use of Mel Frequency Cesptral Coefficients (MFCC) for automatic detection of emotions in running speech. Experiments were conducted on the Berlin emotional speech database for a three- class problem (anger, boredom and neutral emotional states). In order to evaluate the discrimination ability of the features three different classifiers were implemented: k-nearest neighbor, Bayesian Linear and quadratic. The highest accuracy results are obtained when neutral and anger emotions are evaluated.**

***Keywords*— MFCC, speech emotion recognition, feature selection, running speech.**

I. INTRODUCCIÓN

Durante muchos años, la voz, como principal forma de comunicación entre humanos ha motivado la investigación en áreas de interacción hombre-máquina. La voz además de contener información lingüística, es una señal biológica que contiene información sobre características físicas, estados fisiológicos y emocionales que intervienen en el proceso de comunicación. Consecuentemente la detección de contenido emocional en la señal de voz permite que en la interacción hombre-máquina, se produzca una respuesta en función del estado emocional de una persona [1-2]. Otro tipo de aplicación es usado para el reconocimiento de emociones negativas en centros de atención de llamadas [4], también es útil en herramientas de diagnostico y tratamiento psicológico como trastornos de ansiedad [3].

Los estados emocionales son estados que todos los seres humanos reconocemos, sin embargo, no existe una definición teórica de ellos. Existen investigaciones dirigidas a buscar aspectos contenidos en los estados emocionales y que permitan describirlos [4]. En general se define la activación y el balance como dos dimensiones para la caracterización de la voz [5]. El aspecto evaluado en ambas dimensiones corresponde a la energía requerida para expresar cierta emoción. Activación corresponde a un alto requerimiento de energía, contrario al balance. Estados emocionales como la alegría, la ira o el miedo, causan un aumento en el ritmo cardiaco, cambios en los movimientos respiratorios, mayor presión sub-glótica y temblor muscular, produciendo variaciones en las características de la señal de voz [6]. La señal resultante tiene entonces un nivel alto de energía y pitch promedio. Por el contrario estados emocionales como el aburrimiento causan una disminución en el ritmo cardiaco y la presión arterial, generando un aumento en la producción de saliva. La señal resultante tiene un nivel de energía y un pitch promedio en un nivel bajo.

En la tarea de reconocimiento de emociones no es claro cuáles son las mejores características que permiten realizar una distinción entre los diferentes estados emocionales. Existen estudios enfocados en encontrar cuáles son las características más apropiadas y que al mismo tiempo no dependan del hablante o del contenido léxico [4]. De acuerdo con la literatura, las características de la voz pueden ser agrupadas en diferentes categorías: espectrales, cepstrales y de perturbación. Las características espectrales incluyen los LPC (por las siglas en Inglés de Linear Predictive Coding) y los LFPC (por las siglas en Inglés de Log – Frequency – Power – Coefficients) [7-8] y algunas modificaciones matemáticas aplicadas sobre el espectro de voz, tales como el operador de energía de Teager (TEO, por las siglas en Inglés de Teager Energy Operator) [10]. Las características cepstrales son descritas clásicamente mediante los MFCC (por las siglas en inglés de Mel Frequency Cesptral Coefficients), y las características de perturbación incluyen el jitter y el shimmer [9].

En la literatura se encuentran diferentes trabajados desarrollados en el tema de detección de emociones, sin embargo las diferentes metodologías y bases de datos utilizadas pueden dificultar la comparación de resultados. Se presentan algunos trabajos implementados sobre la base de datos *Berlin emotional Database:* En [7] y [11] proponen el reconocimiento de 7 estados emocionales usando diferentes características, en el primero proponen el uso de características espectrales y características prosódicas obteniendo un rendimiento del 88.6%, mientras que en el segundo usando características de modulación espectral y características prosódicas obtienen un rendimiento del 91.6%. El trabajo propuesto en [12] utiliza características no lineales para la clasificación de tres estados emocionales obteniendo un rendimiento del 93.78%.

En este artículo se estudia la capacidad discriminante de los MFCC para la caracterización de diferentes estados emocionales.

* Autor de correspondencia: stefany.bedoya@udea.edu.co

978-1-4673-2761-9/12/$31.00 ©2012 IEEE

El artículo esta organizado de la siguiente manera: la sección II presenta una descripción de la metodología que se ha implementado para abordar el problema. En la sección III se describen los experimentos realizados. La sección IV muestra los resultados obtenidos y finalmente en la sección V se presentan las conclusiones.

II. METODOLOGÍA

El objetivo de este artículo es evaluar la capacidad discriminativa de los MFCC para la clasificación de tres tipos de estados emocionales: ira, aburrimiento y neutral. La Fig.1 resume los componentes más importantes de la metodología aplicada.

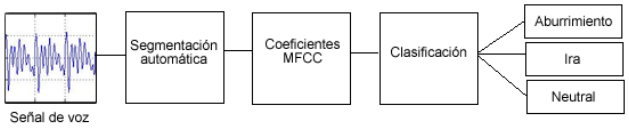

Fig. 1 Metodología propuesta

A. *Cálculo de Coeficientes Ceptrales en la escala de Mel*

Los MFCC han sido ampliamente utilizados en aplicaciones de procesamiento de voz debido a la relación física que tienen con los mecanismos de la producción de voz y su capacidad para caracterizar la energía de la señal en bandas de frecuencia de acuerdo con la escala auditiva humana.

En este artículo, los MFCC son calculados de forma no paramétrica, por lo tanto se derivan de la Transformada de Fourier, la diferencia básica se encuentra en la distribución de las bandas de frecuencias del primero, las cuales se encuentran espaciadas según la escala de Mel para modelar la respuesta auditiva humana. La Escala de Mel tiene una mayor resolución en regiones de baja frecuencia tal y como se muestra en la Fig. 2, donde el ancho de banda de los filtros es constante para frecuencias hasta 1KHz, y crece exponencialmente hasta 4KHz.

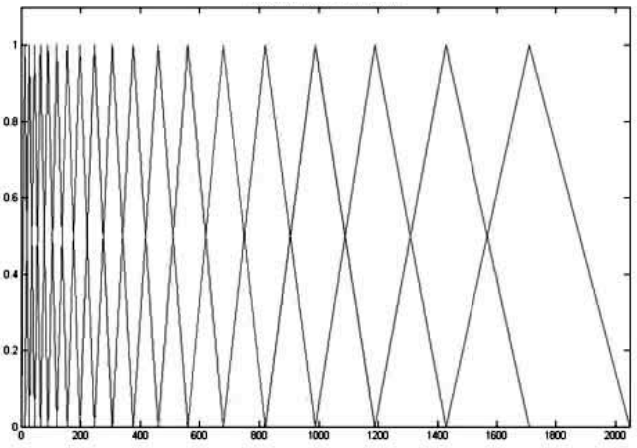

Fig. 2 Banco de filtros según la escala de Mel

El espectro de frecuencias de Mel en un tiempo de análisis n, se describe en la ecuación (1) [13].

$$E_{mel}(n,l) = \frac{1}{A_l}\sum_{k=L_l}^{U_l}|V_L(\omega_k)X_{\hat{n}}(n,\omega_k)|^2 \quad (1)$$

Donde $V_L(\omega_k)$ es la función de ponderación triangular para el filtro de orden *l-ésimo* de la DTF (por las siglas en Inglés de Discrete Fourier Transform) definida entre L_l y U_l, y donde la expresión dada en (2) es un factor para normalizar el filtro de orden *r-ésimo.*

$$A_l[\omega_k] = \sum_{k=L_l}^{U_l}|V_L(\omega_k)|^2 \ (2)$$

Es posible calcular los MFCC de acuerdo con la siguiente ecuación.

$$mfcc[n,l] = \frac{1}{R}\sum_{l=1}^{R}\log(E_{mel}[n,l])\cos\left[n\left(l-\frac{1}{2}\right)\frac{\pi}{l}\right] \ (3)$$

La Fig.3 describe el procedimiento para la extracción de los 12 coeficientes Cepstrales en la escala de Mel, incluyendo el coeficiente C_0, que representa el logaritmo de la energía de la señal.

B. *Clasificación*

1. *Vecinos más cercanos (K-nn)*

El método K-nn (por las siglas en Inglés de k-nearest neighbor) es un método de clasificación no paramétrico que permite estimar la función de densidad de probabilidad de que un elemento x pertenezca a la clase c.

Si consideramos que la probabilidad de que $\theta' = \omega_i$; siendo θ' la etiqueta asociada con el vecino más cercano; es igual a la probabilidad a posteriori $P(\omega_i|X)$ y $P(\omega_i|X') \approx P(\omega_i|X)$, el vector de características X pertenece a la clase c, entonces podemos decir que nuestro vector de características tiene una tienen una probabilidad a posteriori similar con un prototipo de esta clase.

1. *Bayesiano Lineal y Cuadrático*

Los clasificadores Bayesiano Lineal y Bayesiano Cuadrático son de tipo paramétrico, basados en la covarianza de las características. Se define un conjunto de funciones discriminantes dadas en la ecuación (4) donde para cada uno de los datos se toman distribuciones normales multivariadas con el fin de obtener un conjunto de funciones discriminantes para los datos con densidad normal.

$$g_i(x) = P(\omega_i|X) + \ln\big(P(\omega_i)\big) \ (4)$$

Es posible llevar la expresión dada en la ecuación (4) a (5), con $(X-\mu_i)^T\sum^{-1}(X-\mu_i)$ denominada como la distancia de Mahalanobis.

$$g_i(X) = -\frac{1}{2}(X-\mu)^T \sum^{-1}(X-\mu) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\sum| + lnP(w\_i) \quad (5)$$

Para un clasificador cuadrático, Un vector de características X pertenece a la clase c cuando la distancia de Mahalanobis entre el vector y la media de c es la menor.

Para el clasificador lineal las fronteras de decisión serán hiperplanos y no hiperelipsoides como en la cuadrática, es necesario entonces expandir la expresión para la distancia de Mahalanobis para obtener la ecuación dada en (6)

$$g_i(X) = W_i^T X + w_{i0} \quad (6)$$

Para el clasificador lineal las fronteras de decisión serán hiperplanos y no hiperelipsoides como en la cuadrática, es necesario entonces expandir la expresión para la distancia de Mahalanobis para obtener la ecuación dada en (6)

$$g_i(X) = W_i^T X + w_{i0} \quad (6)$$

Donde W_i dado en (7)

$$W_i = \sum^{-1}\mu_i \text{ y } w_{i0} = -\frac{1}{2}\mu_i^T \sum^{-1}\mu_i + lnP(\omega_i) \quad (7)$$

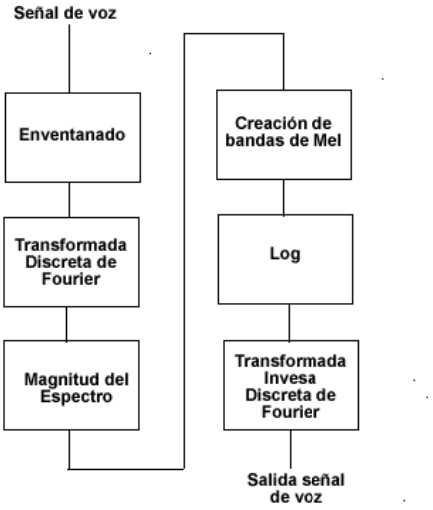

Fig. 3 Diagrama de bloques para obtención de MFCC

III. EXPERIMENTACIÓN

A continuación se explican todos los detalles relacionados con los experimentos mediante los cuales se validó la metodología planteada en el artículo.

A. Base de datos

En este trabajo, son analizados registros de audio de la base de datos *Berlin emotional speech database* [14]. Particularmente son incluidos registros correspondientes a tres tipos de emociones: aburrimiento, ira y neutral. La base de datos fue producida por un conjunto de 10 actores, 5 mujeres y 5 hombres. Cada actor pronunció 10 frases en alemán. El número de registros de voz usados para cada emoción es: 127 para neutral, 81 para aburrimiento, y 79 para ira. Todos los registros de audio fueron muestreados a 16 kHz con 16 bits de resolución.

B. Pre-procesamiento de los datos

Teniendo en cuenta que los registros corresponden a habla continua, las muestras de audio fueron segmentadas usando un detector sonoro-sordo y considerando el contenido de energía y la tasa de cruces por cero, tal y como en [12].

C. Análisis de término corto

Las señales de voz fueron procesadas siguiendo un esquema de tiempo corto, tomando ventanas de 40 ms traslapadas al 50%, por cada ventana se calculan los coeficientes MFCC siguiendo el procedimiento descrito en la sección II. A continuación se calculan la media sobre cada vector de características, formando un vector con 12 características por cada registro de audio (12 valores promedio).

D. Clasificación y validación

El proceso de clasificación es llevado a cabo mediante los clasificadores descritos en la sección II.*B*. Los datos para el proceso de clasificación son divididos aleatoriamente en dos conjuntos: 70% de los datos para entrenamiento y el 30% restante para validación.

El entrenamiento del sistema se realizó empleando una metodología de validación cruzada, repitiendo el proceso diez veces con el fin de garantizar que los parámetros de entrenamiento son los óptimos. El subconjunto de entrenamiento es divido en 10 partes y para cada una de las diez iteraciones, nueve de las partes son tomadas para ajustar los parámetros de los clasificadores y la parte restante es utilizada para el test. En cada repetición se garantiza que el conjunto de test es diferente. Una vez son aplicadas las diez iteraciones, se eligen los parámetros óptimos para los clasificadores fijando como criterio de selección el menor error medio y desviación. Finalmente, utilizando los parámetros seleccionados, el proceso de clasificación es aplicado sobre el conjunto de validación, el cuál no participa nunca dentro del proceso de entrenamiento del sistema.

El experimento completo es repetido diez veces, de tal forma que el error de validación global es el error promedio obtenido.

IV. RESULTADOS

Las tablas I, II y III presentan los resultados en forma de matrices de confusión para cada uno de los clasificadores. La primera columna de cada matriz indica la clase real de los datos y las demás columnas indican el porcentaje de datos que fueron asignados por el clasificador a cada clase.

La diagonal de las matrices representa el porcentaje de

acierto obtenido para cada una de las clases (emociones), y los demás valores representan el error cometido para cada clase.

De acuerdo con los resultados obtenidos, los tres clasificadores implementados identifican, con una precisión similar, los estados emocionales de aburrimiento y neutral. Adicionalmente, es posible observar que ninguno de los clasificadores comete error al clasificar los registros de voz correspondientes a ira.

Tabla I. Matriz de confusión: Clasificador Knn

Estado emocional	%Decisión del clasificador		
	Aburrimiento	Ira	Neutral
Aburrimiento	79.17±8.33	0±0	20.83±8.33
Ira	0±0	100±0	0±0
Neutral	20.83±8.33	0±0	79.17±8.33

Tabla II. Matriz de confusión: Clasificador Bayesiano Lineal

Estado emocional	%Decisión del clasificador		
	Aburrimiento	Ira	Neutral
Aburrimiento	70.83±8.33	0±0	29.17±8.33
Ira	0±0	100±0	0±0
Neutral	25±8.33	0±0	75±8.33

Tabla III. Matriz de confusión: Clasificador Bayesiano Cuadrático

Estado emocional	%Decisión del clasificador		
	Aburrimiento	Ira	Neutral
Aburrimiento	83.33±4.17	0±0	16.67±4.17
Ira	0±0	100±0	0±0
Neutral	16.67±4.17	0±0	83.33±4.17

V. CONCLUSIONES

Se presenta una metodología para la detección automática de tres tipos de estados emocionales a partir del habla, basada en la implementación de 12 Coeficientes Cepstrales en la escala de Frecuencias de Mel, incluyendo el coeficiente C_0, que representa el logaritmo de la energía de la señal.

Mediante la técnica de caracterización presentada se obtienen altas tasas de acierto cuando se busca caracterizar emociones con grados de activación diferentes. La ira y la neutralidad son estados emocionales que permiten un alto grado de discriminación entre ellas a causa de que los requerimientos de energía de uno con respecto al otro son muy diferentes. Por el contrario cuando se analiza el aburrimiento con respecto a la ira o la neutralidad, no es posible lograr altas tasas de acierto debido a que éste puede ser considerado como una mezcla entre activación y balance.

Cuando se requiere discriminar entre emociones que tengan un rango de activación similar, es decir, ira y alegría, es necesario implementar características que aporten mayor información y adicionalmente es necesario aplicar técnicas de clasificación más robustas que permitan incrementar las tasas de acierto.

REFERENCIAS

[1] B. Schuller, G. Rigoll,M.Lang, Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture, in: Proceedings of the ICASSP 2004, pp. 577–580.

[2] Yildirim, S., Narayanan, S., Potamianos, A.: Detecting emotional state of a child in a con- versational computer game. Computer Speech and Language 25, 29–44 (2011)

[3] Burkhardt, F., Polzehl, T., Stegmann, J., Metze, F., Huber, R.: Detecting real life anger. In: IEEE Int. Conf. on Acoustics, Speech and Signal Processing, pp. 4761–4764. IEEE Press, New York (2009)

[4] Moataz El Ayadi , Mohamed S. Kamel, Fakhri Karray, Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 2010.

[5] R. Fernandez, A computational model for the automatic recognition of affect in speech, Ph.D. Thesis, Massachusetts Institute of Technology, February 2004.

[6] Williams, K. Stevens, "Vocal correlates of emotional states, Speech Evaluation in Psychiatry", Grune and Stratton, 1981, pp. 189–220

[7] Wu, S., Falk, T.H., Wai-Yip, C.: Automatic recognition of speech emotion using long- term spectro-temporal features. In: 16th Int. Conf. on Digital Signal Proc., July 5-7, pp. 1– 6 (2009)

[8] Giannakopoulos, T., Pikrakis A., Theodoridis, S.A.: Dimensional Approach to Emotion Recognition of Speech from Movies. In: IEEE Int. Conf. on Acoustic, Speech and Signal Proc., pp. 65–68 (2009)

[9] Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Comm. (2011), doi:10.1016/j.specom.2011.01.011.

[10] S. Bou-Ghazale, J. Hansen, A comparative study of traditional and newly proposed features for recognition of speech under stress, IEEE Trans. Speech Audio Process. 8 (4) (2000) 429–442.

[11] Wu, S., Falk, T.H., Wai-Yip, C.: Automatic speech emotion recognition using modulation spectral features. Speech Comm. 53, 768–785 (2011)

[12] Patricia Henríquez, Jesús B. Alonso, Miguel A. Ferrer, Carlos M. Travieso, and Juan R. Orozco-Arroyave, Nonlinear Dynamics Characterization of Emotional Speech, Neurocomputing (IN PRESS), 2012.

[13] Lawrence R. Rabiner. Ronald W. Schafer. Introduction to Digital Speech Processing. Foundations and Trends Signal Processing 1:1-2 (2007).

[14] Database of German Emocional Speech, http://pascal.kgw.tu-berlin.de/emodb/