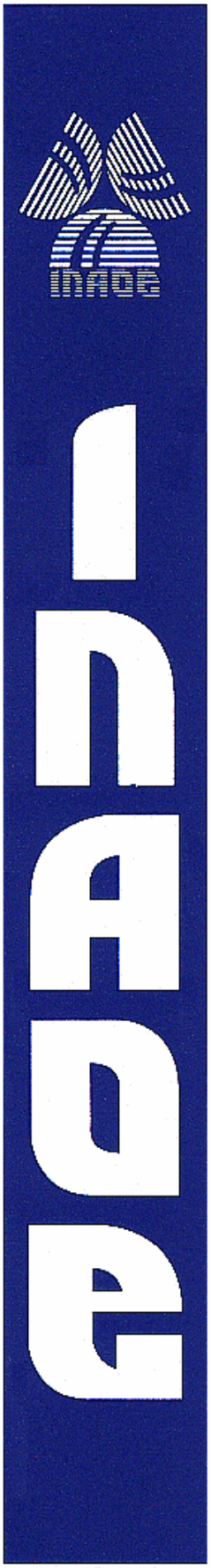


Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo

Por

Humberto Pérez Espinosa

Tesis sometida como requisito parcial para obtener el grado de

DOCTOR EN CIENCIAS EN LA ESPECIALIDAD DE CIENCIAS COMPUTACIONALES

En el

Instituto Nacional de Astrofísica, Óptica y Electrónica

Tonantzintla, Puebla, 2013

Supervisada por:

Dr. Carlos Alberto Reyes García y

Dr. Luis Villaseñor Pineda

Investigadores titulares de INAOE

Resumen

La presente investigación doctoral aporta en la comprensión de los elementos del habla que ayudan a determinar las emociones y en la creación de un método de reconocimiento de patrones basado en un modelo emocional continuo apropiado para emociones espontáneas. Para abordar este problema se adoptó un modelo psicológico emocional continuo el cual permitió tratar las emociones en un sentido más amplio de lo que se ha hecho tradicionalmente. El trabajo también exploró a fondo el espacio de características acústicas en datos realistas para identificar aquellas con mayor aporte de información en esta tarea. Se experimentó con varios tipos de características acústicas incluyendo nuevas características utilizadas en otros campos y se recurrió a técnicas de selección de atributos para encontrar las más importantes. Se estudió la trascendencia de dichos atributos en datos multilingües. El modelo final está basado en la representación de emociones basado en agrupamiento difuso y estudiamos la predicción de emociones apoyado en contexto. Los resultados obtenidos en la estimación de primitivas emocionales y en la clasificación de emociones discretas usando nuestro conjunto de características y métodos son comparables con los mejores resultados en el estado del arte.

Abstract

In this thesis, we worked on emotion recognition from the speech signal. To address this problem we have adopted a psychological continuous emotions model. With this emotion model, emotions are studied in a broader sense of what has been done traditionally. It has been thoroughly explored the acoustic feature space on realistic scenarios. We applied soft computing and probabilistic techniques for estimating emotional states. This work contributes to the understanding of speech elements that help to determine the emotions and to create a pattern recognition method based on a continuous emotional model suitable for spontaneous emotions detection. We experimented with various types of acoustic and linguistic features, including new features used in other fields. We had used feature selection techniques to find the most important ones. In addition, we have studied the importance of these attributes on multilingual data, we propose a method for representing emotional states based on fuzzy clustering and studied the prediction of emotional states based on context. The results obtained in the estimation of emotion primitives and classification of basic emotions using our feature set and methods are comparable with the best results in the state of the art.

Dedicatoria

A mí esposa e hija por ser mi inspiración y motivación.

A mis padres y hermanas por su apoyo y cariño

Agradecimientos

A mis asesores de tesis Carlos Alberto Reyes García y Luis Villaseñor Pineda por su guía, apoyo y tiempo dedicado a este proyecto.

A mis revisores los doctores Angélica Muñoz Meléndez, José Francisco Martínez Trinidad, Aurelio López López, Manuel Montes y Gómez y Luis Alberto Pineda Cortés por darme valiosos comentarios para mejorar el trabajo.

Al INAOE por brindarme todos los medios necesarios para llevar a cabo el trabajo presentado en este documento.

A Conacyt por la beca para estudios doctorales otorgada.

Índice General

Índice de Tablas

Índice de Figuras

Índice de Formulas

Índice de Algoritmos

Notación

C	Número de *clusters* a formar o número de emociones en la base de datos
D	Número de emociones a generar a partir de los datos
A	Conjunto de muestras etiquetadas
E	Conjunto de muestras a clasificar
UE	Matriz de membresías de muestras a clasificar
UA	Matriz de membresías de muestras etiquetadas
U	Matriz de membresías de muestras de referencia
D	Número de emociones conocidas
T	Temperatura del sistema
m	Grado de difusión

Capítulo 1: Introducción

Las emociones son inherentes a los seres humanos. El afecto y la emoción juegan un papel importante en nuestras vidas y están presentes en todo lo que hacemos. De ahí que las emociones conforman un aspecto natural y social de la comunicación humana (Averill, 1990). Mediante la expresión de emociones durante la comunicación oral se transmite información implícita importante sobre el hablante, que complementa la información explícita contenida en el intercambio de mensajes en una conversación.

Inicialmente filósofos y psicólogos se interesaron en el estudio del efecto que tienen las emociones sobre la voz y expresiones faciales de los individuos. Más recientemente, los científicos en computación también se han involucrado en el estudio de las emociones, en cómo reconocerlas automáticamente y han intentado incorporar esta tecnología en aplicaciones del mundo real (Vidrascu & Devillers, 2005) (González, 1999). Las primeras preguntas que surgen al involucrarse en el reconocimiento de emociones a partir de la voz son: ¿Qué evidencias existen de que en realidad las emociones de las personas se reflejan en sus voces? ¿Las emociones se reflejan de manera semejante en todas las personas? ¿De qué depende la manera en que expresamos emociones con nuestra voz? Estás preguntas se han tratado de responder desde el enfoque de diferentes disciplinas como la filosofía (Gómez, 1971) (James W. , 1884), la biología (Darwin, 1872), la química (Liebowitz, 1983), la psicología (Ekman, 1972) y la antropología (Lutz & Miles White, 2001).

Como antecedentes históricos se puede mencionar al filósofo Platón que formuló la doctrina del alma tripartita la cual sugería que el alma tiene una estructura compuesta por tres áreas: cognición, emoción y motivación (Gómez, 1971). Charles Darwin estableció que las emociones son patrones relacionados con la supervivencia que han evolucionado para resolver ciertos problemas que una especie ha enfrentado a través de su evolución. La postura de Darwin es que las emociones son, más o menos, las mismas en todos los humanos y en particular independientes de la cultura (Darwin, 1872). Aún cuando los antropólogos afirman que las emociones son productos socioculturales, varios autores han trabajado en demostrar la hipótesis de Darwin. Esto da pie a un debate que tocaremos en la sección *5.5 Análisis multilingüe de características.*

Uno de los primeros trabajos que buscó definir que es una emoción fue el presentado por William James, psicólogo y filósofo estadounidense, quien publicó en 1884 un artículo titulado *"What is an emotion?"* (James W. , 1884). En este trabajo se establece que ciertos cambios físicos suceden directamente a la percepción de un hecho excitante y que nuestro sentimiento o percepción de esos cambios es lo que conocemos como emoción. Además, estos cambios físicos se transmiten a través de diversos canales, de esta manera el evento acústico asociado al habla es afectado por el estado del sistema nervioso central, y por lo tanto el habla conlleva información sobre el estado emocional de un individuo.

El desarrollo del reconocimiento automático de emociones se basa en estos avances teóricos y conclusiones alcanzadas en las disciplinas que estudian el fenómeno emocional humano y permiten hacer suposiciones razonables en el modelado computacional. La definición del término *emoción* es la base para cualquier tipo de investigación en esta área. Al tener una definición común, entre los investigadores en el área, es posible comparar los resultados de la investigación realizada por dichos investigadores, ya que se tiene la certeza de estar analizando el mismo fenómeno. Schuller y Steidl, importantes investigadores en el área de reconocimiento automático de emociones, han propuesto (Steidl, 2009) usar la definición de emociones hecha por Scherer, quien es un reconocido especialista en la psicología de las emociones:

> *Las emociones son episodios de cambios coordinados en varios componentes (incluyendo al menos activación neuropsicológica, expresión motriz y sentimientos subjetivos pero también posiblemente tendencias a la acción y procesos cognitivos) en respuesta a eventos externos o internos de mayor significancia para el organismo. Los eventos disparadores externos pueden ser, por ejemplo, el comportamiento de otros, un cambio en la situación actual, o un nuevo estímulo. Los eventos internos son, por ejemplo, pensamientos, recuerdos y sensaciones.* (Scherer K. R., 2000, p. 137)

Esta definición menciona diferentes características de las emociones para las cuales, de acuerdo a Scherer (Scherer K. R., 2000), se ha encontrado un creciente consenso entre los psicólogos. Dichas puntos de acuerdo son:

- Las emociones son de naturaleza episódica, los episodios emocionales duran cierto tiempo y normalmente no se detienen abruptamente, sino se desvanecen disminuyendo su intensidad haciendo la detección del final más difícil que del comienzo.
- Las emociones son perceptibles, es decir, es notorio el cambio en el funcionamiento del organismo causado por algún evento.
- Las emociones se componen de una *triada de reacción* que incluye:

    1. Excitación Psicológica
    2. Expresión Motriz
    3. Sentimiento Subjetivo

    Algunos psicólogos han añadido a estos tres elementos la motivación generada por una evaluación cognitiva el evento o estímulo.
- Las emociones son provocadas por eventos o estímulos importantes o relevantes para el individuo

No obstante, muchas veces es difícil diferenciar entre fenómenos afectivos que cumplen con algunos de los componentes mencionados arriba. Sin embargo, las emociones se diferencian de otros fenómenos afectivos como humor, posturas interpersonales, actitudes y rasgos de personalidad por las siguientes características:

- **Intensidad:** Las emociones son el fenómeno afectivo con mayor intensidad.
- **Duración:** Las emociones son el fenómeno afectivo con menor duración.
- **Sincronización:** Cuando se experimenta una emoción existe un grado muy alto de coordinación de diferentes sistemas orgánicos
- **Focalizado en evento:** Las emociones están fuertemente ligadas a un evento o estimulo particular que las provoca.
- **Apreciación:** La naturaleza de la emoción experimentada está fuertemente relacionada con el resultado de un proceso de evaluación que antecede a dicha reacción emocional.
- **Rapidez:** El cambio de una emoción a otra es relativamente rápido.
- **Impacto:** Las emociones afectan fuertemente al comportamiento del individuo.

Es importante notar que a pesar de que la distinción es muy fina entre distintos tipos de fenómenos afectivos hay características particulares de las emociones que se pueden identificar para uso práctico y llevar su reconocimiento automático a aplicaciones del mundo real.

Una de las primeras aplicaciones identificadas del reconocimiento automático de emociones es en la Interacción Humano Computadora (IHC). La necesidad por el reconocimiento automático de emociones ha surgido a partir de la tendencia hacia una interacción más natural entre humanos y computadoras que la que existe actualmente. Computación Afectiva es un tópico dentro de la IHC que incluye esta tendencia de investigación tratando de dotar a las computadoras con la habilidad de detectar, reconocer, modelar y tomar en cuenta el estado emocional de los usuarios.

Los sistemas de IHC incorporan sistemas de habla y visión debido a que éstos son los canales más naturales de comunicación humana. Uno de los objetivos que persiguen los sistemas de IHC es que la interacción sea bidireccional, para lo cual una máquina debe escuchar el mensaje del usuario y responder de manera natural. Para alcanzar esta forma de interacción, la expresión emocional debe ser reconocida y sintetizada. De esta manera, los sistemas de IHC podrán adaptarse al estado emocional del usuario, como lo hacemos los humanos al conversar, alcanzando una interacción más natural, eficiente y amigable que la interacción actual entre humanos y computadoras. Las emociones son esenciales para el proceso de pensamiento humano e influencian las interacciones con personas y sistemas inteligentes. El reconocer el estado de ánimo de los usuarios en un sistema de IHC le brinda información relevante al sistema, retroalimentándolo y haciéndolo capaz de reaccionar y adaptarse.

Las siguientes aplicaciones son un ejemplo de cómo se puede aprovechar el conocimiento del estado emocional de los usuarios para tomar decisiones en sistemas de IHC. Un tutorial interactivo (Hernández, Sucar, & Conati, 2008) en el que se podría adaptar la carga emocional de la respuesta del sistema buscando motivar y captar el interés dependiendo del estado emocional del alumno. Un sistema telefónico de atención automática a clientes que provee asistencia médica a usuarios que llaman pidiendo ayuda (Vidrascu & Devillers, 2005). Dichos usuarios podrían presentar diferentes emociones como tensión, miedo, dolor o pánico dependiendo de la enfermedad o de la emergencia que están experimentando. El manejo de una llamada será diferente dependiendo de la clasificación del estado emocional del usuario, dando prioridad a las llamadas más urgentes, dirigiéndolas a la persona indicada.

Otra aplicación es un Sistema de Respuesta Interactiva por Voz (IVR) que atiende pacientes con problemas psicológicos (González, 1999). El sistema detecta si hay algún grado de depresión basándose principalmente en características articulatorias de la calidad de voz del paciente. El sistema alerta a un experto humano cuando detecta en el paciente un grado de depresión alarmante.

Las aplicaciones del reconocimiento automático de carga emocional en la voz no se limitan únicamente a la IHC. En la interacción humano – humano también puede usarse, por ejemplo, para monitorizar conversaciones entre agentes y clientes en *call centers* y detectar emociones no deseadas (Devillers & Vidrascu, 2006). Por ejemplo, un cliente enojado o frustrado o un agente inseguro o nervioso. De esta manera un inspector de calidad puede tomar decisiones sobre la administración y mejora del personal y de los servicios.

En áreas médicas el reconocimiento automático de emociones podría ser usado en el soporte médico remoto. Este tipo de ambiente permite la comunicación de médicos y pacientes para casos de monitorización regular y situaciones de emergencia. En este escenario un sistema de reconocimiento de emociones puede estimar las emociones del paciente y transmitir datos indicando si el paciente está experimentando tristeza o depresión. Las instituciones de salud que monitorizan a estos pacientes estarían mejor preparadas para responder. Dicho sistema tendría el potencial de mejorar la satisfacción y salud del paciente (González, 1999) (Nasoz, Alvarez, & Lisetti, 2004) (Vidrascu & Devillers, 2005).

Como muestran estos ejemplos de aplicación, mediante el reconocimiento automático de emociones se puede incrementar el desempeño, la usabilidad y en general la calidad de sistemas de interacción humano computadora, sistemas de atención a clientes y otros tipos de aplicaciones. Sin embargo, el reconocimiento automático de emociones es un problema complejo, por lo cual ha sido difícil de implementar en aplicaciones reales.

1.1 Problemática actual

El área de reconocimiento automático de emociones ha sido un área de investigación muy activa en los últimos años, no obstante, aún se está lejos de una solución clara para este problema. Diversos obstáculos han influido en la construcción de una solución apropiada. Por un lado, un factor que afecta el desempeño de los reconocedores de emociones en contextos reales es la dificultad de generar bases de datos con emociones espontáneas. Generalmente se ha trabajado con bases de datos actuadas las cuales proporcionan "retratos de emociones" representando expresiones prototípicas e intensas que facilitan la búsqueda de correlación acústica y la subsecuente clasificación automática. Este tipo de bases de datos suelen grabarse en un ambiente controlado lo cual elimina problemas en el procesamiento de la señal, por ejemplo, ruido o reverberación. Además, se puede garantizar una cantidad balanceada de muestras por clase. Como consecuencia, no se han tenido buenos resultados al trasladar el conocimiento extraído de éstas bases de datos a contextos reales (Steidl, 2009). En contraparte, las bases de datos con emociones espontáneas muestran elocuciones con contenido emocional no perteneciente a una sola clase, sino que son una mezcla de emociones. Además, en ciertos casos, existen muestras con una carga emocional muy ligera, cercana a un estado emocional neutro. Aunado a esto, las bases de datos con emociones espontáneas suelen grabarse en ambientes ruidosos como conversaciones telefónicas o programas de televisión lo que conlleva la inclusión de ruido. Finalmente, por la naturaleza misma del problema se trata de una situación con un gran desbalance entre los ejemplos por clase.

Otro reto a resolver es la identificación de un conjunto de características acústicas que permitan reconocer emociones en el habla espontánea. El trabajo hecho a la fecha se ha centrado principalmente en características relacionadas con aspectos prosódicos, como acento, entonación y ritmo; sin embargo, se ha descubierto que entre más nos alejamos de emociones actuadas y nos acercamos a un escenario realista, menos fiable es la prosodia como un indicador del estado emocional del hablante (Batliner, Fischer, Humber, Spliker, & Nöth, 2003), ya que, mientras el objetivo de los actores es mostrar cierto estado emocional, no es evidente que los hablantes en la vida real muestren del todo sus emociones y hagan uso de los mismos recursos lingüísticos (Selting, M., 1994). Por lo tanto, es necesario encontrar características que complementen la información que proporciona el aspecto prosódico del habla.

Otro problema a enfrentar es el modelo psicológico a utilizar. Dos enfoques psicológicos son de particular interés: el discreto y el continuo. Los modelos discretos se basan en el concepto de emociones básicas, como: enojo, alegría, tristeza, etc., que son la forma más intensa de las emociones, a partir de las cuales se generan todas las demás, mediante variaciones o combinaciones de ellas. De esta forma, los modelos discretos parten del supuesto de la existencia de emociones universales, que pueden ser distinguidas claramente una de otra. En contraste, los modelos continuos, conocidos también como dimensionales, representan las emociones mediante un espacio multidimensional continuo, en el cual cada eje corresponde a un atributo emocional llamado, por algunos autores, *primitiva emocional*. Las primitivas emocionales son propiedades o atributos mostrados por todas las emociones por lo tanto, virtualmente cualquier emoción puede ser definida en función de estas primitivas emocionales. Estas primitivas funcionan como dimensiones en un espacio multidimensional donde se pueden distinguir emociones a partir de sus componentes genéricos. Uno de estos modelos es el modelo tridimensional, cuyos ejes son: Valencia, Activación y Dominación. La primera, también llamada placer por algunos autores, describe qué tan negativa o positiva es una emoción. La Activación describe la excitación interna de un individuo y va desde el estar muy tranquilo hasta llegar a ser muy activo; y la Dominación describe el grado de control del individuo sobre la situación o, en otras palabras, qué tan fuerte o débil se muestra el individuo. En el trabajo hecho por Osgood (Osgood, May, & Miron, 1975) se muestra que prácticamente cualquier concepto relacionado con emociones puede ser localizado en este espacio tridimensional.

Hasta el momento la mayoría de los trabajos en reconocimiento automático de emociones han utilizado modelos emocionales discretos, donde las emociones a reconocer están claramente identificadas en el corpus de entrenamiento. Bajo este enfoque no existe una valoración de la emoción sino la búsqueda de una o varias reglas que permitan la discriminación de las emociones en cuestión. De esta forma, es necesario repetir el proceso de entrenamiento de modelos si se desea agregar una nueva emoción o un nuevo conjunto de emociones. Otra complicación de los modelos discretos es la dificultad de trabajar con emociones espontáneas ya que no es posible representar apropiadamente el traslape e intensidad de emociones en el habla.

A pesar de que los avances en el área han sido importantes, se ha comprobado que en contextos realistas aún falta mucho por hacer. Por lo tanto es necesario proponer y explorar otros enfoques que permitan llegar a un buen desempeño del reconocimiento de emociones en aplicaciones del mundo real. Para ello, en esta tesis se propone trabajar con

características diversificadas que expandan el uso de características acústicas, emplear un modelo continuo más general nos permita acercarnos más a situaciones reales. Además se buscará aprovechar el contexto emocional para mejorar la predicción de las emociones. Es decir, se incluirá información de las emociones previas para predecir la emoción presente.

1.2 Preguntas de investigación

- ¿Qué características acústicas son útiles para reconocer emociones en el habla espontánea?
- ¿Cuáles de esas características son más útiles para un modelo emocional continuo?
- ¿De qué forma podemos emplear un modelo emocional continuo en el diseño de un método de reconocimiento de emociones aplicable a emociones espontáneas?
- ¿El uso de modelos continuos mejorará el reconocimiento de emociones con respecto al uso de modelos discretos?

1.3 Objetivos

1.3.1 Objetivo general

Desarrollar un método para el reconocimiento automático de emociones espontáneas basado en un modelo emocional continuo a partir de la información acústica extraída de la señal de voz, flexibilizando la transferencia de modelos entre aplicaciones y alcanzando un desempeño similar o mejor que los reconocedores de emociones actuales basados en modelos discretos.

1.3.2 Objetivos particulares

1. Identificar diferentes tipos de características relevantes en el reconocimiento de Primitivas Emocionales en bases de datos de emociones espontáneas.

2. Diseñar un esquema de reconocimiento de patrones basado en un modelo emocional continuo.

3. Estudiar la relación entre primitivas en nuestro modelo emocional y determinar la manera de interpretarlas para ubicar emociones discretas en aplicaciones específicas.

4. Evaluar nuestros resultados en emociones espontáneas con métricas que permitan la comparación con otros trabajos.

1.4 Organización de esta tesis

En el capítulo 1 se presenta la motivación del trabajo desarrollado y se plantean los objetivos de la tesis.

En el capítulo 2 se describen los conceptos y técnicas existentes usados en el desarrollo de esta tesis.

En el capítulo 3 se describen los diferentes enfoques que se han usado para resolver la problemática ligada al reconocimiento automático de emociones en voz. Se hace un análisis de los trabajos más importantes relacionados con nuestra propuesta. Se construye una taxonomía de acuerdo a tres criterios para agrupar las propuestas hechas por los investigadores en el área y se hace una diferenciación con nuestra propuesta.

En el capítulo 4 se describe le método propuesto y se hace una diferenciación con los métodos del estado de arte. También se describe el proceso de diseño y generación de nuestra propia base de datos de habla emocional. Basada en la aplicación de una prueba psicológica aplicada a niños.

En el capítulo 5 se hace un análisis de características acústicas. Se describen los métodos de extracción y selección usados. Se cuantifica el aporte y la importancia de diferentes tipos de descriptores acústicos. Se explora el aporte de información complementaria como información lingüística e información contextual, y se hace un estudio multilingüe de características acústicas.

En el capítulo 6 se describe a detalle el método de estimación de emocionales basado en un modelo psicológico tridimensional continuo y lógica difusa. Se describe un modo multinivel de representación de emociones.

En el capítulo 7 se describe un método para reclasificación de segmentos basado en el contexto temporal y en evidencia acústica

En el capítulo 8 se hace una evaluación del método propuesto. Se evalúan dos aspectos: 1) el modo tradicional de clasificación de emociones discretas, y 2) el aspecto difuso de nuestros métodos propuestos.

En el capítulo 9 se hace un resumen del trabajo y se presentan las conclusiones, las contribuciones y el trabajo propuesto de nuestro trabajo de investigación.

Capítulo 2: Marco Teórico

En este capítulo se presentan conceptos importantes para la comprensión de los capítulos subsecuentes. En primer lugar se explican los diferentes tipos de características acústicas de la voz con los que estuvimos trabajando. En segundo lugar se explican técnicas de aprendizaje automático y computación suave de las que hicimos uso para desarrollar el método propuesto en este trabajo.

2.1 Caracterización de voz

Las características acústicas de la voz estudiadas en esta tesis se pueden dividir en tres grandes grupos: Características prosódicas, características de calidad de voz y características espectrales.

Prosodia: Las características prosódicas describen fenómenos suprasegmentales, es decir, características detectables en unidades de voz mayores que fonemas, tales como: entonación, melodía, velocidad, volumen, duración, pausas y ritmo. La prosodia es una fuente rica de información en el procesamiento de voz, porque contiene información paralingüística importante, que complementa el mensaje con una intención que puede reflejar una actitud o un estado emocional (Kehrein, 2002). Este tipo de características son las más comúnmente usadas en reconocimiento de emociones en voz.

Calidad de Voz: Las características de calidad de Voz nos da la distinción primaria de la voz de una persona cuando los aspectos prosódicos son excluidos. Algunos calificativos de la voz con respecto a su calidad son: tensa, susurrada, ronca, débil, nasal etc. Algunos autores (Steidl, 2009) han estudiado la importancia de la calidad de voz, estableciendo que la clave de la diferenciación vocal de emociones discretas parece ser la calidad de voz.

Espectro: Las características espectrales describen las propiedades de una señal de voz en el dominio de la frecuencia; más allá de la frecuencia fundamental, nos dan información de armónicos y formantes (Steidl, 2009), es decir, de las concentraciones de energía acústica en torno a ciertas zonas de resonancia.

En esta tesis se incluye la caracterización espectral por medio de cocleograma los cuales no se habían probado antes en reconocimiento de emociones. Un cocleograma (Boersma, 2001) representa la excitación de los filamentos de los nervios auditivos de la membrana basilar, la cual está situada en la cóclea en el oído interno. Esta excitación es representada como una función sobre el tiempo (en segundos) y en la frecuencia de Bark que es una escala psico-acústica. Un cocleograma también modela el volumen del sonido y el enmascaramiento de frecuencias. El enmascaramiento de frecuencias en el oído sucede cuando escuchamos dos sonidos de diferente intensidad al mismo tiempo; el sonido más débil no es distinguido por el cerebro que sólo procesa el sonido más fuerte, es decir el sonido más débil es enmascarado por el sonido más fuerte. Las características basadas en cocleogramas han sido usadas para reconocimiento de voz con buenos resultados. En el trabajo de (Byrne, 1989) los cocleogramas fueron usados para reconocimiento de secuencias de fonemas, mejorando los resultados obtenidos por características LPC.

Los wavelets son una alternativa a la transformada de Fourier, esta caracterización también fue probada en esta tesis. La transformada wavelet permite una buena resolución en bajas frecuencias. También incluimos características ampliamente usadas en procesamiento de voz como MFCCs (Mel-frequency cepstral coefficients) (Zbynik, 1999) que representan la percepción de la voz basada en el oído humano y han sido usadas exitosamente para discriminar fonemas. Los MFCCs han mostrado que no sólo son útiles para determinar lo que se dice, sino también, cómo se dice.

Para la extracción de características en esta tesis se utilizaron dos herramientas de procesamiento de audio Praat (Boersma, 2001) y openSMILE (Eyben, Wöllmer, & Schuller, 2009).

2.2 Selección de características

Una parte muy importante de esta tesis es la búsqueda y análisis de las características más valiosas para discriminar emociones en la voz. A pesar de que ya hay mucho trabajo al respecto basado en emociones discretas, hay muy poco trabajo basado en primitivas emocionales. El principal problema en nuestro caso fue probamos muchas características, cerca de 7,000, con relativamente pocas muestras por cada base de datos, alrededor de 2,000.

Se probó la técnica *SubSetEval* (Witten & Frank, 2005) la cual es una implementación en Weka del algoritmo *Correlation-based Feature Selector* (Hall, 1998). Este algoritmo de selección cae en la categoría de *filtro*, ordena los subconjuntos de características de acuerdo a una función de evaluación heurística. Dicha función evalúa la valía de cada subconjunto de atributos considerando la habilidad predictiva individual de cada característica junto con el nivel de redundancia entre ellas, prefiriendo los subconjuntos de características que están altamente correlacionadas con una clase y bajamente correlacionadas entre ellas. La función de evaluación de subconjuntos se muestra a continuación.

$$Ms = \frac{k\, r_{cf}}{\sqrt{k + k\,(\,k - 1) r_{ff}}}$$

Formula 1 Formula para evaluar subconjuntos de acuerdo a *SubSetEval*

Donde *Ms* es el *mérito* heurístico de un subconjunto de características *S* conteniendo *k* características, r_{cf} es la correlación media entre la clase y las características ($f \in S$), y r_{ff} es el promedio de la inter-correlación entre características. El numerador se puede ver como un indicador de que tan predictivo es un conjunto de características para la clase; el denominador se puede ver como un indicador de que tanta redundancia existe entre las características.

También se probó *ReliefAttribute* (Witten & Frank, 2005), el cual es una implementación en Weka del algoritmo *Relief* (Kira & Rendell, 1992) cuya idea clave es estimar atributos de acuerdo a los valores que permiten distinguir las muestras más parecidas. Para una muestra dada, el algoritmo *Relief* busca los dos vecinos más cercanos: uno de la misma clase y otro de una clase diferente. Los buenos atributos deben de ser diferentes valores entre muestras de diferentes clases por un lado y los mismos valores para muestras de la misma clase por otro lado. Dado un conjunto de entrenamiento *S*, el tamaño del subconjunto *m*, y un umbral de relevancia *r*, el algoritmo *Relief* detecta aquellas características que son estadísticamente relevantes para el concepto objetivo. *T* codifica un umbral de relevancia. Se asume la escala de cada característica como nominal o numérica. Las diferencias de valores de características entre dos instancias *X* y *Y* son definidas por la siguiente función *diff*: Cuando Xk y Yk son numéricos, como en nuestro caso:

$$diff(x_k, y_k) = \frac{(X_k - Y_k)}{nu_k}$$

Formula 2 Formula para cálculo diferencia en el algoritmo Relief

Dónde nu_k hace una normalización de los valores de *diff* en el intervalo [0,1]. Relief toma un subconjunto compuesto de *m* tripletas de una muestra **X**, su muestra *Near-hit* y su muestra *Near-miss*. Relief usa la distancia euclidiana *p*-dimensional para seleccionar el *Near-hit* y *Near-miss*. Relief llama una rutina para actualizar el vector de pesos de características *W* para cada tripleta y determina el vector de relevancia de pesos de características promedio de todas las características para la clase objetico. Finalmente Relief selecciona aquellas características cuyo peso promedio, nivel de relevancia, esta sobre el umbral dado.

La principal técnica de selección de características usada en esta tesis es Selección hacia Adelante Flotante Lineal (*Linear Floating Forward Selection LFFS)* el cual hace una búsqueda *Hill Climbing*, empezando con un conjunto vacío de características o con un conjunto predefinido. Se evalúan todas las inclusiones posibles de un atributo al conjunto solución. En cada paso el atributo con la mejor evaluación es añadido. La búsqueda termina cuando ya no hay inclusiones que mejoren la evaluación. Adicionalmente, LFFS dinámicamente cambia el número de características incluido o eliminado en cada paso.

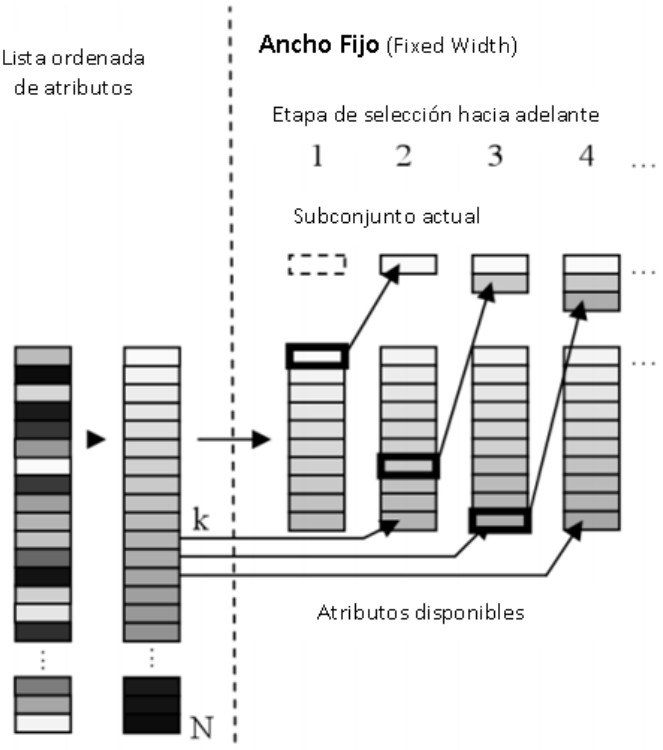


Figura 1 Selección hacia Adelante Flotante Lineal con Ancho Fijo (Basado en (Gutlein, Frank, Hall, & Karwath, 2009))

En nuestros experimentos usamos LFFS en una modalidad llamada Ancho Fijo (*Fixed Width)* que es ilustrada en la Figura 1 y funciona de la siguiente manera: Se parte de un conjunto de *N* atributos los cuales se evalúan individualmente de acuerdo a cierto criterio y son ordenados del mejor al peor según el resultado de dicha evaluación. Los *k* mejores atributos son seleccionados para formar el conjunto de atributos disponibles en la primera etapa del proceso de selección hacia adelante, el resto se mantiene en un conjunto de atributos eliminados. El primer paso en la selección hacia adelante consiste en añadir el mejor de los atributos del conjunto de atributos disponibles al conjunto actual de atributos. Para mantener un ancho fijo, es decir, un número constante en el conjunto de atributos disponibles se añade a éste el mejor elemento del conjunto de atributos eliminados. En cada iteración del proceso de selección hacia adelante los atributos añadidos al conjunto actual solución son reemplazados por el siguiente mejor atributo del conjunto de atributos eliminados.

En nuestros experimentos usamos como criterio de evaluación de características el resultado de un esquema envolvente en el cual se generan modelos de regresión usando máquinas de vectores de soporte. La métrica para calificar el desempeño individual de cada atributo es el coeficiente de correlación de Pearson entre los valores estimados por el modelo y los valores anotados manualmente.

2.2.1 Métricas de medición de calidad de características acústicas

Para medir el aporte de cada grupo de características acústicas usamos tres medidas Coeficiente de correlación de Pearson, Share y Portion.

El coeficiente de correlación es el parámetro más común para medir el desempeño de algoritmos de aprendizaje automático en tareas de regresión, como es nuestro caso. Usamos *Share* y *Portion* que son medidas propuestas en (Batliner A. , et al., 2011) para calcular el impacto de diferentes tipos de características en el desempeño del reconocimiento automático de emociones discretas. Este coeficiente indica la fuerza y dirección de la relación lineal entre la las primitivas anotadas y las primitivas estimadas por el modelo entrenado. Esta es nuestra principal métrica para medir los resultados de clasificación. El coeficiente de correlación de Pearson es usado para medir la calidad de la variable estimada determinando la fuerza y la dirección de una relación lineal entre el valor

estimado y el valor real de una variable. Entre más cercano esté el coeficiente de -1 o 1 es más fuerte la correlación entre las variables. A medida que dicho coeficiente se acerca a cero existe una relación menos fuerte. En la Formula 3 se muestra la fórmula para el cálculo del coeficiente de Pearson.

$$r = \frac{\sum(x - x')(y - y')}{\sqrt{\sum(x - x')^2 \ \sum(y - y')^2}}$$

Formula 3 Formula para cálculo de coeficiente de correlación de Pearson. x e y son las medias de muestra

Share: Muestra la contribución de cada grupo de características con relación al total de características seleccionadas.

$$\boldsymbol{Share} = \frac{<\text{número de características seleccionadas del grupo } X> * 100}{<\text{número de características seleccionadas de todos los grupos}>}$$

Formula 4 Formula para cálculo de *Share*

Por ejemplo, si se seleccionan 28 características del grupo Tiempos y en total se seleccionaron 150.

Share = (28 x 100) / 150 = 18.7

Portion: Muestra la contribución de los grupos de características pesados por el número de características por grupo.

$$\boldsymbol{Portion} = \frac{<\text{número de características seleccionadas del grupo } X> * 100}{<\text{número total de características del grupo } X>}$$

Formula 5 Formula para cálculo de *Portion*

Por ejemplo, si 28 características son seleccionadas de un total de 125 características del conjunto Tiempos entonces:

$$\textbf{\textit{Portion}} = (28 \times 100) / 125 = 22.4$$

2.3 Selección de muestras mediante auto-entrenamiento

El método de selección de muestras usado en esta tesis está inspirado en la técnica conocida como auto-entrenamiento o *self-training* ampliamente usado con aprendizaje semi-supervisado (Zhu, 2006), (Chapelle, Scholkopf, & Zien, 2006), (Zhou & Li, 2005). Dicho método se basa en la generación de un modelo de clasificación base generado a partir de datos etiquetados, con este modelo se clasifica un conjunto de muestras no etiquetadas. Se aplica cierto criterio para determinar un nivel de confianza sobre la clasificación de cada una de las muestras recién clasificadas. Se toma un cierto número de las más confiables y se vuelve a generar un nuevo modelo de clasificación base, generado con las muestras iniciales más las muestras recién clasificadas. Este ciclo se repite hasta alcanzar cierto criterio de paro.

2.4 Agrupamiento difuso mediante Fuzzy C-means

El agrupamiento difuso, mediante la técnica *Fuzzy C-means* (FCM), forma una parte muy importante de nuestro método ya que nos permite suavizar la clasificación de emociones análogamente a como suelen darse las expresiones emocionales en el mundo real. El resultado del agrupamiento de muestras por FCM es una matriz de partición difusa, la cual indica el grado de pertenencia de cada muestra a cada clase. La determinación de una matriz de partición difusa U (dividiendo n muestras en C clases) usando agrupamiento *FCM* supervisado es un procesamiento de optimización iterativo. El núcleo del *SFCM* (Kalyani S., 2010) es usar las muestras de datos etiquetadas para guiar la optimización iterativa. La función objetivo del problema de optimización *SFCM* es definida como indica la Formula 6.

$$J_m(U, v) = \sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik})^m\, d_{ik}^2 + a\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik} - f_{ik})^m d_{ik}^2$$

Formula 6 Función objetivo Supervised Fuzzy C-means

Donde

U	Matriz de partición difusa
v	Matriz de centros de cada grupo
u_{ik}	Grado de membresía del *k*-ésimo dato perteneciente al *i*-ésimo *cluster* (valor entre 0 y 1)
d_{ik}	Distancias del *k*-ésimo dato del *i*-esimo centro de *cluster*
f_{ik}	Grado de membresía de la muestra etiquetada perteneciente al *i*-ésimo *cluster* (valor es 0 o 1)

El coeficiente *a* denota el factor de escalamiento y *m* denota el coeficiente de difusión. El papel de *a* es mantener el balance entre el componente supervisado y no supervisado dentro del mecanismo de optimización. El parámetro *m,* controla el nivel de difusión en la clasificación, entre más grande es, la pertenencia de cada muestra a cada clase se diluye más. El valor típico de *m* es 2 y *a= L/n*, *L* denotando el tamaño de las muestras etiquetadas. La función J_m puede tomar números grandes de valores, el más pequeño es asociado con el mejor agrupamiento.

El método propuesto toma algunas ideas del Agrupamiento supervisado difuso y algunas de las operaciones realizadas en el algoritmo Fuzzy C-Means original como las siguientes el cálculo de la matriz de membresías mediante la siguiente formula:

$$U_{ik} = \left[\sum_{j=1}^{c}\left(\frac{d_{ik}}{d_{ij}}\right)^{\frac{2}{m-1}}\right]^{-1}$$

Formula 7 Formula para calcular matriz de membresías

Donde, d_{ik}, es la distancia euclidiana entre el *i*-esimo centro de cluster y la *k* esima muestra.

2.5 Campos Aleatorios de Markov

Está técnica probabilista nos permitió incorporar a nuestro método de clasificación información contextual, es decir incluir información de la clasificación de muestras anteriores cronológicamente al momento de clasificar una nueva muestra. A continuación se da una definición formal de esta técnica según (Chávez Garcia, 2010, p. 31). Un Campo Aleatorio de Markov (CAM) es un modelo gráfico probabilista que caracteriza relaciones contextuales dadas por observaciones con conocimiento obtenido de las interacciones con variables vecinas (Li Z. S., 1994). El concepto de Campo Aleatorio de Markov procede del intento de colocar dentro de un marco probabilista general un modelo físico específico. Caracteriza las relaciones contextuales locales de fenómenos físicos. Puede complementar FCM incorporando información contextual que ayude a la clasificación.

CAM se ha usado en procesamiento de voz, en específico para incluir información complementaria a la información comúnmente usada en el proceso de reconocimiento de voz (Gravier, Sigelle, & Cholle, 1998). También se han usado en Reconocimiento Automático de Voz como una extensión de HMM donde se incorpora información de interacción entre bandas (Gravier, Sigelle, & Cholle, 1998)

En (Wallach, 2004) se usaron para *POS tagging*[1] donde,

- CAM Define una probabilidad condicional sobre secuencias de etiquetas dada una secuencia de observación particular
- HMM Define una distribución de probabilidad conjunta sobre secuencias de etiquetas y observaciones

Para generar expresiones faciales emocionales (Ju & Lee, 2008) existen dos factores que determinan la probabilidad de una configuración de valores. La primera es la probabilidad *a priori* de cada estado, que se ejemplifica con un campo magnético externo.

La segunda es la probabilidad conjunta o condicional, representada por la intersección de los campos magnéticos de estados vecinos. Ambas funciones se combinan en la probabilidad máxima *a posteriori*. En análisis de imágenes también se ha usado esta

[1] *Part of Speech tagging* es el proceso de etiquetado de una palabra en un texto como correspondiente a una parte particular de habla, basándose tanto en su definición, así como su contexto, es decir en la relación con las palabras adyacentes y similares en una frase, oración o párrafo.

técnica (Dutta, 2009), donde se considera que la información contextual en el análisis de imágenes debería ser más completo y adquiere la habilidad de reducir la ambigüedad y recuperar información faltante.

Formalmente se puede definir un campo aleatorio de Markov como sigue: Sean $F = \{F_1, F_2, \ldots, F_n\}$ variables aleatorias dentro de un conjunto S, donde cada F_i puede tomar un valor f_i de un conjunto de valores L. A F se le conoce como campo aleatorio, y a la "instanciación" de cada una las variables F_i con un valor f_i, se le llama configuración de F, por lo tanto, la probabilidad de que una variable aleatoria Fi tome el valor fi se denota como $P(fi)$, y la probabilidad conjunta es denotada como $P(F_1 = f_1, F_2 = f_2 \ldots, F_n = f_n)$.

Se dice que un campo aleatorio es un campo aleatorio de Markov si éste tiene la propiedad de *localidad*, es decir que el campo satisfaga la siguiente propiedad:

$$P(f_i \,|f_{S-\{i\}}) = P(f_i|f_{Ni})$$

Donde $S - \{i\}$ representa el conjunto S sin el elemento i, $f_{Ni} = \{f_i' | i' \in N_i\}$ y Ni representan el conjunto de variables vecinas del nodo fi.

Un sistema de vecindad para S se define como:

$$V = \{V_i | \forall_i \in S\}$$

y cumple con las siguientes propiedades:

1.- Un sitio no es vecino de sí mismo.
2.- La relación de vecindad es mutua.

La probabilidad conjunta puede expresarse como:

$$P(f) = \frac{e^{-U_p(f)}}{Z}$$

Donde Z es conocida como la función de partición o constante de normalización, y $U_p(f)$es conocida como la función de energía. La función de energía $U_p(f)$ representa la

información externa e interna necesaria para cambiar o no el valor de una variable aleatoria. La configuración óptima es obtenida cuando se minimiza la función de energía $U_P(f)$, obteniendo un valor para cada una de las variables aleatorias en F. Obtener la configuración de menor energía (mayor probabilidad) es una operación muy costosa, por lo que se plantea como un problema de optimización. Es decir que se busca la configuración de mayor probabilidad, sin tener que calcular directamente las probabilidades de cada configuración.

Para plantear la obtención de la configuración más probable, como un problema de optimización, se necesitan definir tres componentes principales (Chellappa & Jain, 1993):

1. Representación del CAM. Se representan las variables aleatorias del CAM, así como sus valores, su sistema de vecindad y los potenciales asociados a los dos factores de probabilidad.

2. Función objetivo. Se define la función de energía, que incluya los potenciales definidos anteriormente, el objetivo es minimizar esta función de energía para encontrar el valor más probable para una variable.

3. Algoritmo de optimización. Se selecciona un algoritmo que permita seleccionar el valor más apropiado para la variable analizada, de acuerdo al valor obtenido por la función de energía.

En nuestro caso utilizamos el algoritmo de optimización conocido como simulado recocido el cual se describe en la sección *7.2 Algoritmo de optimización.*

Capítulo 3: Estado del Arte

Como se comentó anteriormente en la problemática del reconocimiento de emociones en voz, ésta puede visualizarse a través de tres aspectos: el modelado del fenómeno emocional, la caracterización de la información y la generación de bases de datos. En este capítulo usamos estos criterios para agrupar los enfoques empleados por los autores en esta área para estudiar el problema de la clasificación automática de emociones a partir de voz. Como se muestra en la Figura 2 el primer criterio es de acuerdo al tipo de modelo emocional adoptado, el segundo es de acuerdo al tipo de información usada y el tercero es de acuerdo a la manera de obtener los datos. En este capítulo se hace una revisión de los avances logrados hasta el momento en cada uno de los aspectos mencionados y se sitúa nuestra propuesta dentro de este marco.

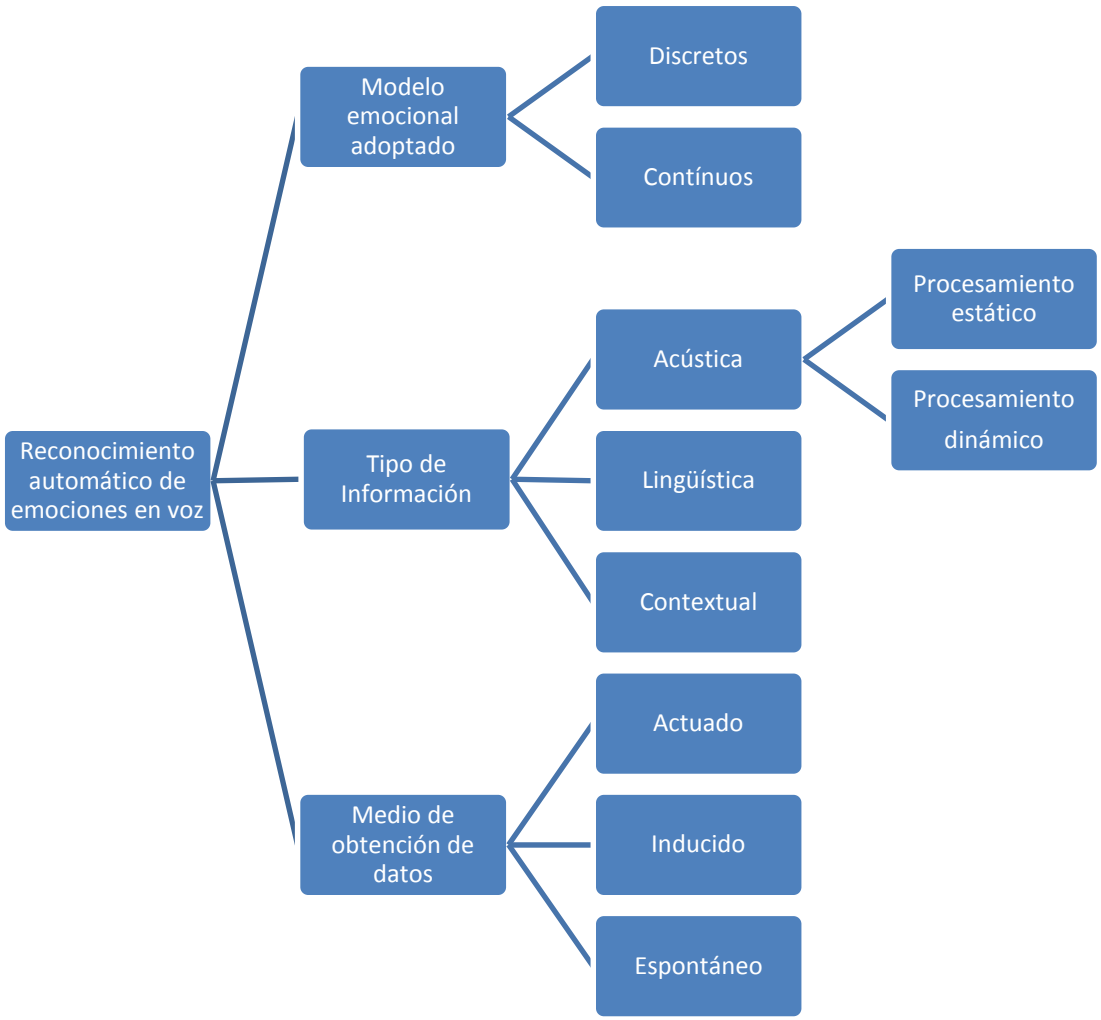


Figura 2 Taxonomía de reconocimiento automático de emociones

3.1 Modelo emocional

El primer gran reto en el área de reconocimiento automático de emociones consiste en determinar qué emociones se desean reconocer, para abordarlo se ha recurrido a los modelos psicológicos de emoción que tienen como objetivo explicar y representar la generación, composición y clasificación de las emociones humanas para comprender los mecanismos subyacentes a los procesos emocionales. Esta representación es usada en modelos computacionales para su análisis, categorización y organización. Para el reconocimiento automático de emociones se han empleado los enfoques discreto y continuo para capturar y describir las manifestaciones emocionales de los individuos.

3.1.1 Modelos discretos

Los Modelos Discretos se basan en el concepto de *Emociones Básicas*, como enojo, alegría, tristeza, etc., que son la forma más intensa de las emociones a partir de las cuales se generan todas las demás mediante variaciones o combinaciones de ellas. Suponen la existencia de emociones universales, al menos en esencia, que pueden ser distinguidas claramente una de otra por la mayoría de la gente, asociadas con funciones cerebrales que evolucionaron para lidiar con diferentes situaciones (Ekman, 1992). Las emociones básicas son experimentadas por los mamíferos sociales y tienen manifestaciones particulares asociadas con ellas tales como expresiones faciales, patrones fisiológicos y tendencias de comportamiento. En este trabajo usamos el término emociones básicas refiriendo al concepto de emociones a partir de las que se generan otras emociones. También usamos el término emociones discretas para referirnos a emociones categóricas, es decir emociones especificadas por una etiqueta emocional sin que necesariamente sea considerada una emoción básica.

El predominio de esta teoría en el reconocimiento automático de emociones puede explicarse por el hecho de que dada una aplicación la diferenciación entre emociones es relativamente clara, sin embargo, aún en esas situaciones existe la necesidad de definiciones más detalladas, por ejemplo, distinguiendo entre ira y cólera. En el enfoque discreto, las representaciones prototípicas son asimiladas más fácilmente y esto repercute en su generación y reconocimiento, y por lo tanto son más útiles para una construcción rápida de bases de datos y como un punto de partida para la investigación emergente en este campo de investigación. Sin embargo, bajo este enfoque no es claro cuáles son las emociones básicas. En la Tabla 1, tomada de (Ortony, Clore, & Collins, 1988), se muestran varios conjuntos de emociones básicas en inglés propuestos por distintos autores, como se

puede observar, no existe un criterio claro para definir qué emociones forman este conjunto.

Tabla 1 Conjuntos de emociones básicas propuestos por diferentes autores (Ortony, Clore, & Collins, 1988)

Autor	Emociones Básicas	Base de Inclusión
Plutchik	Acceptance, anger (enfado), anticipation, disgust (asco), joy (alegría), fear, sadness (tristeza), surprise (sorpresa)	Relacionado con el proceso biológico adaptativo
Ekman, Friesen, Ellsworth	Anger (enfado), disgust (asco), fear, joy (alegría), sadness (tristeza), surprise (sorpresa)	Expresiones faciales universales
Gray	Rage and terror, anxiety (ansiedad), joy (alegría)	Fijo
Izard	Anger (enfado), contempt (contento), disgust (asco), distress (aflicción), fear, guilt (culpabilidad), interest (interés), joy (alegría), shame , surprise (sorpresa)	Fijo
James	Fear, grief (aflicción), love, rage	Enredo corporal
Mowner	Pain, pleasure	Estados emocionales no aprendidos
Oatley and Johnson-Laird	Anger (enfado), disgust (asco), anxiety (ansiedad), happiness (felicidad), sadness (tristeza)	No requieren contenido proposicional
Paksepp	Expectancy , fear, rage, panic	Fijo
Tomkins	Anger (enfado), interest (interés), contempt (contento), disgust (asco), distress (aflicción), fear, joy (alegría), shame, surprise (sorpresa)	Densidad en actividad neuronal
Watson	Fear, love, rage	Fijo
Weiner and Graham	Happiness (felicidad), sadness (tristeza)	Reconocimiento independiente

No obstante, los modelos discretos permiten una representación más particularizada de las emociones en las aplicaciones donde solamente se requiere reconocer un conjunto predefinido de emociones y tiende a ignorar la mayor parte del espectro de expresiones emocionales. Si un conjunto reducido de emociones básicas es usado como un punto de partida para el reconocimiento de emociones, surge la pregunta de si las mismas

características y patrones de comportamiento son válidas tanto para emociones extremas como para emociones más sutiles (Sobol-Shikler, 2008).

Otro de los problemas de estos modelos es la investigación intercultural de emociones y la traducción correcta de términos emocionales o afectivos usados. Muchos de estos términos tienen significados connotativos y denotativos diferentes en distintos idiomas. Por el momento no hay una solución satisfactoria a estos problemas (Hillsdale & Erlbaum, 1998). Algunos autores han llegado a la conclusión que la representación del espectro emocional mediante emociones básicas es demasiado compleja para su utilización en aplicaciones prácticas (Iriondo, 2008). Hemos agregado a la Tabla 1, entre paréntesis**,** la traducción al español de acuerdo a la lista de descriptores afectivos en cinco lenguas indoeuropeas. No todas las emociones de la Tabla 1 están incluidas en la lista. Dicha lista es un subproducto de las actividades de investigación de un equipo de psicólogos de diferentes países que realizaron una serie de estudios mediante cuestionarios interculturales, involucrando respuestas libres acerca de experiencias emocionales (Scherer, Wallbott, & Summerfield, 1986). Los autores hacen la aclaración de que la lista no es exhaustiva y que no todos los términos son estrictamente emociones.

Tabla 2 Comparativa modelos discretos

Trabajo	Emociones	Desempeño	Datos	Idea Clave
(Tóth, Sztahó, & Vicsi, 2007)	Sorpresa, disgusto, nerviosismo, tristeza	70%	Actuados	Pre-procesamiento de la señal de voz y modelado mediante HMM
(Pittermann & Pittermann, 2006)	Enojo, aburrimiento, disgusto, miedo, felicidad, tristeza	72%	Actuados	Usar un modelo HMM y tratar emociones como palabras o fonemas como si se estuviera haciendo reconocimiento de voz.
(Sato & Obuchi, 2007)	Enojo, neutro, tristeza, felicidad	66.4%	Actuados	Algoritmo que emplea múltiples "machotes" de clasificación emocional. Los *Codebooks* están entrenados mediante agrupamiento
(Luengo, Navas, & Hernandez, 2005)	Sorpresa, tristeza, alegría, miedo, asco, ira	98.4%	Actuados	Construir tres clasificadores diferentes, uno para cada tipo de característica acústica, combinando GMM y SVM

Tradicionalmente, los trabajos en reconocimiento de emociones en voz se sitúan en este enfoque de modelado emocional ya que se basan en un conjunto de emociones discretas, que pueden ser emociones básicas como el *Big Six* de Ekman (alegría, enojo, tristeza, sorpresa, asco, miedo) (Tóth, Sztahó, & Vicsi, 2007) (Pittermann & Pittermann, 2006) (Sato & Obuchi, 2007) (Luengo, Navas, & Hernandez, 2005) u otro conjunto de emociones que pueden ser derivadas de emociones básicas.

En este enfoque de modelado también se sitúan trabajos que no usan emociones discretas reconocidas teóricamente como emociones básicas pero cuyo objetivo es detectar emociones específicas, como decepción, confianza, frustración, enojo, de acuerdo a un dominio de aplicación. Por ejemplo, identificación de tensión en un sistema de cobranza donde suelen surgir conflictos entre el agente que solicita un pago y el cliente. Otro ejemplo es la identificación de frustración en un sistema automatizado de información, donde los clientes frecuentemente no logran obtener la información que necesitan (Fell & MacAuslan, 2003). La categorización es principalmente hecha sobre bases subjetivas ya que es difícil estandarizar un conjunto de etiquetas emocionales. A pesar de los muchos intentos tratando de establecer una correspondencia entre emociones y voz no existe un conjunto definido de emociones universalmente aceptado.

En la Tabla 2 se hace una comparación de algunos trabajos que utilizan el enfoque discreto para hacer reconocimiento de emociones en voz. Todos ellos usan bases de datos de mociones actuadas en diferentes idiomas como euskara, inglés y alemán. Se puede observar que algunos de ellos obtienen muy buenos resultados debidos principalmente a que los datos son actuados y libres de ruido.

Uno de los trabajos más completos dentro de la clasificación de emociones discretas es la tesis doctoral de Stefan Steidl (Steidl, 2009). En este trabajo se construyó el corpus FAU Aibo que está diseñado para realizar investigación en reconocimiento de emociones en voz orientada a emociones que aparecen en escenarios realistas, donde las emociones son sutiles y además, existen mezclas de diferentes emociones. Dicho corpus de habla emocional espontánea está en alemán, las voces grabadas son de niños entre 10 y 13 años de edad interactuando con el robot *Aibo* de *Sony*. El habla emocional fue inducida mediante un experimento de *Mago de Oz*. Se les pidió a los niños que le dieran instrucciones al robot de cómo ir de un punto a otro como si estuvieran hablando con un amigo. El corpus generado muestra espontaneidad emocional ya que los niños, como adaptadores tempranos dentro de un escenario de escolar son destinatarios plausibles para el modelado automático de emociones (Batliner, Steidl, & Noeth, 2008). Etiquetaron once

emociones discretas a nivel de palabra y frase. Con los datos obtenidos llevaron a cabo experimentos de clasificación en tres niveles de segmentación: nivel de palabra, nivel de turno y nivel de bloque intermedio.

Steidl propuso un conjunto de características acústicas agrupándolas en prosódicas, espectrales, y de calidad de voz. Encontró que las características acústicas que tuvieron mejor desempeño fueron la Intensidad y las duraciones del grupo de prosódicas y los MFCCs del grupo de espectrales. Caracterizó el aspecto lingüístico de las interacciones usando técnicas conocidas de las cuales los modelos de unigramas y bolsa de palabras fueron los que mostraron mejor desempeño. Encontró que el desempeño de las características lingüísticas es ligeramente peor que el de las características acústicas. Además obtuvo una mejora mediante la combinación de ambas fuentes de información. Los mejores resultados se obtuvieron usando a segmentación a nivel de bloque intermedio donde alcanzaron una tasa promedio de reconocimiento de casi el 70% para 4 clases, Enojo, Enfático, Neutral y Maternal.

El trabajo hecho por Steidl sin duda es muy completo, aborda una gran parte de la problemática en el reconocimiento automático de emociones en voz, sin embargo, los experimentos de clasificación hechos en el citado trabajo se basan únicamente en modelos emocionales discretos dejando abierta la interrogante de cómo podrían mejorar los resultados obtenidos al aplicar el enfoque de los modelos emocionales continuos.

3.1.2 Modelos continuos

Los modelos continuos también conocidos como modelos dimensionales representan emociones usando un espacio multidimensional continuo, donde cada eje corresponde a un propiedad emocional llamada por algunos autores *primitiva emocional* (Grimm M. , Kroschel, Mower, & Narayanan, 2007). Las primitivas emocionales son propiedades presentes en todas las emociones, por lo tanto, cualquier emoción puede ser definida en función de primitivas emocionales. Cualquier emoción puede ser representada por un punto en un espacio de coordenadas multi-dimensional. Las primitivas emocionales pueden considerarse como los componentes genéricos de una emoción mediante los cuales se puede describir y diferenciar de otras emociones. Uno de estos modelos es el modelo tridimensional, cuyos ejes son: Valencia, Activación y Dominación. La primera, también llamada placer por algunos autores, describe qué tan negativa o positiva es una emoción. La Activación describe la excitación interna de un individuo y va desde el estar muy

tranquilo hasta llegar a ser muy activo; y la Dominación describe el grado de control del individuo sobre la situación o, en otras palabras, qué tan fuerte o débil se muestra el individuo. En el trabajo hecho por Osgood (Osgood, May, & Miron, 1975) se muestra que prácticamente cualquier concepto relacionado con emociones puede ser localizado en este espacio tridimensional.

Un ejemplo es el modelo bidimensional circunflejo (Russel, 1980) (Steidl, 2009) en el que se proponen las primitivas Valencia y Activación para la representación de emociones. Otro ejemplo es el modelo tridimensional Valencia – Activación – Dominación donde *Valencia,* también llamada placer por algunos autores, describe qué tan negativa o positiva es una emoción de acuerdo a si es una emoción agradable o desagradable para quien la experimenta. *Activación* describe la excitación interna de un individuo y va desde estar muy tranquilo hasta estar muy activo y *Dominación* que describe el grado de control del individuo sobre la situación o, en otras palabras, qué tan fuerte o débil se muestra el individuo.

El modelo tridimensional, ilustrado en la Figura 3, surge por la necesidad de distinguir entre emociones que se encuentran traslapadas en un espacio bidimensional. Añadir la tercera dimensión ayuda a distinguir entre emociones como miedo y enojo ya que ambas tienen Valencia y Activación similar pero difieren en el eje Dominación.

Los modelos continuos ofrecen mayor flexibilidad en la representación de emociones ya que no se limitan a un conjunto fijo de emociones, sino que pueden representar un amplio espectro de emociones en el espacio multidimensional y trasladarlo a un conjunto de emociones discretas si así se requiere (Grimm M. , Kroschel, Mower, & Narayanan, 2007). Este tipo de modelos tiene la capacidad de representar de mejor manera la forma en que suceden las emociones en el mundo real, ya que muchas veces las emociones no se generan de forma prototípica sino que pueden manifestarse como una mezcla de emociones o como ligeras expresiones emocionales difíciles de detectar. Al etiquetar bases de datos emocionales los modelos discretos son más adecuados que los continuos para asignar estados preseleccionados a patrones psicológicos, mientras el enfoque continuo es más adecuado que el discreto para evaluar la carga emocional (Beale & Peter, 2008).

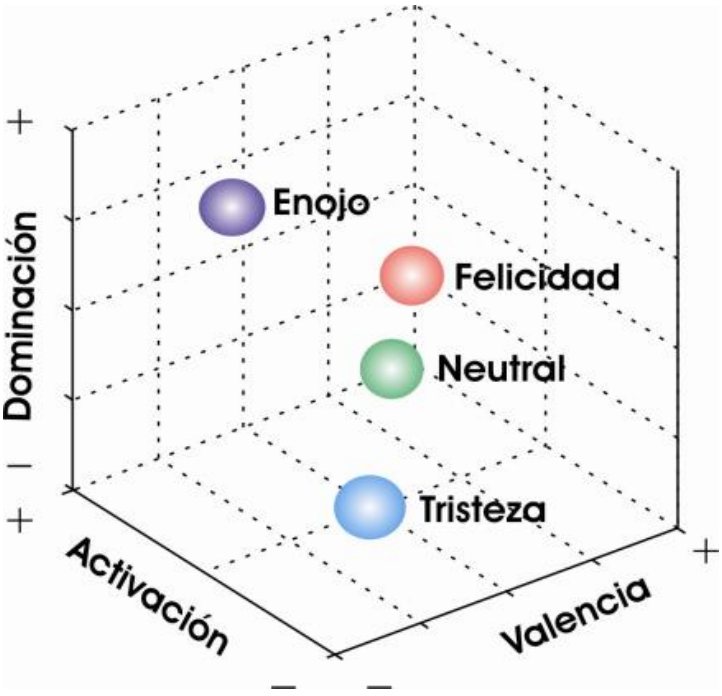


Figura 3 Modelo tridimensional continuo de las emociones

El enfoque continuo elimina algunas limitantes del enfoque discreto, como por ejemplo, el definir y nombrar las emociones discretas necesarias para representar un espectro suficientemente amplio de emociones de acuerdo a una aplicación específica.

Uno de los primeros y más importantes trabajos en predicción de primitivas emocionales fue realizado por (Grimm M. , Kroschel, Mower, & Narayanan, 2007) en el cual se emplea un modelo emocional tridimensional. Las ideas propuestas en este trabajo se prueban en una base de datos actuada y en una base de datos de emociones espontaneas etiquetadas con las primitivas emocionales Valencia, Activación y Dominación mediante la técnica denominada *Self Assessment Manikins (SAMs)* (Grimm & Kroschel, 2005). En dicho trabajo se extrajeron características acústicas tales como melodía, Intensidad, velocidad y características espectrales y calcularon la correlación que existe entre características acústicas y primitivas emocionales con el fin de establecer reglas que determinen qué combinaciones de valores en las características acústicas corresponden a cierto grado de una primitiva emocional.

El proceso de clasificación propuesto por Grimm consiste en extraer las características acústicas de las muestras de prueba las cuales se fuzifican. Estos valores fuzificados son dados como entrada al sistema de reglas generadas a partir de la

correlación existente entre características acústicas y primitivas emocionales. Después de esto, se realiza el proceso de implicación para obtener las conclusiones y determinar la salida. Finalmente se defuzifican los resultados obtenidos de la implicación. La salida final son tres valores entre uno y cinco que corresponden a cada una de las primitivas emocionales que a su vez son mapeadas hacia un conjunto de emociones discretas usando un clasificador *K Vecinos más cercanos (KNN).* Realizaron varios experimentos con subconjuntos de muestras de su corpus de entrenamiento. Alcanzaron un coeficiente de correlación promedio de 0.60 para las tres primitivas usando el corpus completo.

En este trabajo las reglas derivadas de los coeficientes de correlación para la representación de la relación entre las características acústicas y las primitivas emocionales parece ser una generalización burda. En el trabajo de Grimm no se ha explotado en su totalidad el potencial de técnicas de computación suave, como la lógica difusa, para abordar este tipo de problemas. La configuración del sistema de inferencia difusa parece muy básica ya que no se experimenta con otros tipos de funciones de membresía u operaciones difusas. El conjunto de características acústicas utilizadas no incluye información de calidad de voz la cual se ha demostrado que es importante para estimación de emociones (Lugger & Yang, 2008).

Por su parte Lugger realizó un experimento (Lugger & Yang, 2008) donde usó una base de datos en alemán llamada *Berlin Emotional* que consta de seis emociones discretas: tristeza, aburrimiento, neutral, ansiedad, felicidad e ira. Las características acústicas que probó fueron prosódicas y de calidad de voz. Partió de la suposición de que el conjunto óptimo de características acústicas depende fuertemente de las emociones a ser clasificadas y a partir de esto se hizo una clasificación en cascada de 3 fases basada en el modelo psicológico emocional continuo.

En la primera etapa de su cascada o árbol de clasificación, se clasifican dos diferentes niveles de Activación. Una clase incluye ira, felicidad, y ansiedad con un nivel de Activación alto mientras en la segunda clase se incluyen neutral, aburrimiento y tristeza con un nivel de Activación bajo. Para esta discriminación de Activación se alcanzó una buena taza de clasificación del 98.8% en promedio. En la segunda etapa se clasifican dos niveles de Dominación en cada clase de Activación. Esto significa que todos los patrones que fueron clasificados con una Activación alta en la primera etapa son clasificados en una clase conteniendo felicidad e ira o en una segunda clase solo conteniendo ansiedad. Similarmente, todos los patrones que fueron clasificados en Activación baja en la primera etapa son clasificados a una clase conteniendo neutral, aburrimiento o en una conteniendo

sólo tristeza. En la tercera etapa, se distingue entre emociones que difieran sólo en la dimensión Valencia: felicidad vs ira, así como neutral contra aburrimiento.

Lugger propone un enfoque interesante para sacar provecho de ambos tipos de modelado emocional, continuo y discreto, sin embargo, trabajó sobre una base de datos actuada y etiquetada con emociones discretas basándose en una categorización manual de los niveles correspondientes de cada emoción con las tres primitivas emocionales. Esto suscita la exploración de técnicas automáticas para la evaluación de primitivas emocionales y por supuesto la evaluación del método presentado en bases de datos de emociones espontáneas.

Otro trabajo relevante es el de Lichtenstein (Lichtenstein, Oehme, Kupschick, & Jürgensohn, 2008) quien realizó un experimento para comparar ambos enfoques, discreto y continuo, en cuanto a cuál es más adecuado para estimar emociones y cuál se debería adoptar como estándar para el estudio de emociones. Se observó que al etiquetar una base de datos, estimar en términos de Valencia y Activación en una escala continua resulta más fácil que hacerlo asignando una de las clases emocionales definidas. En la Tabla 3 se hace una comparación de algunos trabajos que utilizan el enfoque continuo para hacer reconocimiento de emociones en voz.

Tabla 3 Comparativa modelos continuos

Trabajo	Dimensiones	Desempeño	Idea Clave
(Grimm M. , Kroschel, Mower, & Narayanan, 2007)	Valencia, Activación, Dominación	0.60	Estimador lógico difuso y una base de reglas derivadas de características acústicas
(Grimm, Kroschel, & Narayanan, 2007)	Valencia, Activación, Dominación	0.69	Optimización de parámetros para un algoritmo de Regresión de Vectores de Soporte
(Lugger & Yang, 2008)	Activación, Potencia, Evaluación	-	Estimación en cascada de las tres dimensiones usadas. Después de la estimación se usa un clasificador de emociones discretas

3.2 Caracterización de la información

Las emociones provocan una serie de cambios fisiológicos (James W. , 1890) en las personas que pueden ser monitorizados y medidos para estimar el estado emocional actual del individuo. Estos eventos físicos se pueden clasificar en dos tipos, internos y externos. Los eventos internos son bioseñales emitidas por el sistema nervioso central. Algunas de las bioseñales que han sido usadas para la medición de emociones son: repuesta galvánica de la piel, electromiografía, ritmo cardiaco y señales cerebrales. Hoy en día para medir estas bioseñales son necesarios dispositivos especiales portados por el individuo lo que dificulta su uso en muchas aplicaciones. Los eventos externos donde se refleja el estado emocional son eventos audibles y visibles. Los eventos visibles son expresiones faciales, ademanes y movimientos corporales. La relación entre emociones y expresiones faciales ha sido estudiada ampliamente por el psicólogo Paul Ekman que ha definido seis emociones básicas usadas ampliamente tanto en la síntesis como en el reconocimiento de emociones. Los eventos audibles conllevan información emocional a través de mensajes explícitos, es decir, información lingüística que es lo que se está diciendo y mensajes implícitos, es decir, información acústica que es la manera en que se dicen las cosas. Debido a su fácil disponibilidad, la mayoría de los trabajos en reconocimiento de emociones en voz utilizan solamente información acústica (Sato & Obuchi, 2007) (Tóth, Sztahó, & Vicsi, 2007) (Schuller, Lang, & Rigoll, 2005) (Luengo, Navas, & Hernandez, 2005). Las características acústicas suelen agruparse en:

- ***Espectrales*** que describen las propiedades de una señal en el dominio de la frecuencia mediante armónicos y formantes.
- ***De Calidad de Voz*** que definen estilos al hablar como neutral, susurrante, jadeante, estrepitoso resonante, sonoro, ruidoso.
- ***Prosódicas*** que describen fenómenos suprasegmentales como entonación, volumen, velocidad, duración, pausas y ritmo.

En este trabajo nos enfocamos en el estudio de características acústicas para la estimación de emociones en voz desde el punto de vista de los modelos emocionales continuos. Además exploramos el aporte de información complementaria como información lingüística y de contexto.

3.2.1 Procesamiento dinámico y estático

En el reconocimiento automático de emociones en voz se ha detectado que es importante tomar en cuenta la evolución y el cambio o transformación que sufre el estado emocional de los individuos a través del tiempo. En este enfoque, denominado procesamiento dinámico, se captura información acerca de cómo evolucionan las características en el tiempo. Por el otro lado, en el procesamiento estático se evita el sobreajuste en el modelado fonético aplicando funciones estadísticas sobre descriptores de bajo nivel en periodos de tiempo. El procesamiento estático es más común en reconocimiento de emociones en voz; sin embargo, el procesamiento dinámico ha mostrado buenos resultados en publicaciones recientes (Dumouchel, Dehak, Attabi, Dehak, & Boufaden, 2009) (Bozkurt, Erzin, Erdem, & Erdem, 2009) (Vlasenko, Schuller, Wendemuth, & Rigoll, 2007). En el modelado estático se clasifica usando métodos estáticos como *Support Vector Machines* o Redes Neuronales. La clasificación se hace a nivel de la elocución completa por lo que los segmentos de análisis son de diferentes tamaños. Las características son obtenidas de la extracción de LLDs (*Low Level Descriptors*), por ejemplo entonación, intensidad, o coeficientes espectrales, y de la aplicación de funciones estadísticas, como media, desviación estándar, cuantiles, sobre las características, lo cual resulta en vectores de características del mismo tamaño para todas las muestras de voz (Vogt & André, 2009) (Planet, Socoró, Monzo, & Adell, 2009) (Lee, Mower, Busso, Lee, & Narayanan, 2009). En el modelado dinámico se emplean características como tono, intensidad, MFCCs y sus derivativas etc. con modelos de clasificación dinámicos como *Hidden Markov Models* (Pitterman & Schmitt, 2008) o *Gaussian Mixture Models*.

El análisis se hace a nivel de ventanas del mismo tamaño, por lo que para cada elocución se tienen vectores de características de diferentes tamaños dependiendo de su duración. Las características que usualmente se extraen son MFCCs y otros tipos de coeficientes, por ejemplo, coeficientes de intensidad, velocidad (Vlasenko, Schuller, Wendemuth, & Rigoll, 2007).

En (Vlasenko, Schuller, Wendemuth, & Rigoll, 2007) se realiza una comparación entre clasificar emociones mediante un procesamiento estático y uno dinámico en dos bases de datos. Se obtienen mejores resultados con el procesamiento estático en ambas bases de datos. Adicionalmente, se realiza una fusión de ambos procesamientos tomando como una característica más la estimación hecha por el clasificador dinámico y pasando

este nuevo vector de características al clasificador estático. Está fusión mejoró sustancialmente los resultados obtenidos por los dos clasificadores por separado.

En la Tabla 4 se hace una comparación de algunos trabajos que utilizan procesamiento dinámico de características para hacer reconocimiento de emociones en voz.

Tabla 4 Comparativa de procesamiento dinámico

Trabajo	Emociones	Desempeño	Idea Clave
(Pitterman & Schmitt, 2008)	Enojo, aburrimiento, disgusto, miedo, alegría, neutro, tristeza	76%	Usa un clasificador HMM basado en MFCCs y una base de datos de palabras clave. El reconocimiento se hace en dos etapas. La primera para palabras y la segunda para emociones.
(Vlasenko, Schuller, Wendemuth, & Rigoll, 2007)	Enojo, aburrimiento, disgusto, miedo, alegría, neutro, tristeza	83%	Estrategia de clasificación uno contra todos. HMM con uno y dos estados.
(Wollmer, et al., 2008)	4 niveles de Valencia, 7 niveles de Activación	48%	Modelado de la evolución temporal de las emociones basada en la estimación de niveles de Valencia y Activación

3.3 Corpora para reconocimiento de emociones

La obtención de habla natural y la riqueza de su anotación son muy importantes para encarar los retos del reconocimiento automático de emociones en voz. Dentro del área de computación afectiva, la generación adecuada de datos es un factor clave para trabajar en el desarrollo e investigación de nuevos modelos emocionales útiles en aplicaciones reales. Deben ser considerados varios puntos importantes para la creación de bases de datos emocionales. Primero, la naturaleza y origen del habla capturada. Hay diferentes tipos de fuentes de datos. Tradicionalmente, la fuente de datos emocionales más usada es la actuación de emociones. IEMOCAP (Busso, et al., 2008), Emo-DB (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005) y SSE (Montero, 2003) contienen habla emocional actuada. Otra forma de adquirir datos es a través de la inducción de emociones. La

inducción puede ser hecha presentando estímulos tales como imágenes, videos, audio, con el objetivo de generar reacciones emocionales. La inducción también puede ser hecha por interacción, donde los experimentadores intentan generar algunas reacciones al realizar cierta actividad. Por ejemplo, realizando un experimento de *Mago de Oz*, en el que los sujetos interactúan con un sistema computacional que creen ser autónomo, pero en realidad es controlado por un humano. EmoTaboo (Devillers & Martin, 2008), SAL 1 (Douglas-Cowie, et al., 2007), FAU Aibo (Steidl, 2009) son algunos ejemplos de bases de datos de emociones inducidas.

También hay datos espontáneos adquiridos en ambientes de interacción real, tales como programas de televisión o sistemas telefónicos de atención a clientes. Las bases de datos *VAM* (Narayanan, Grimm, & Kroschel, 2008), y *EMOTV1* (Abrilian, S.; Devillers, L.; Buisine, S.; Martin, 2005) fueron grabadas de programas de televisión. Las bases de datos *Genova Airport Lost Luggage Database* (Scherer & Ceschi, 1997) and *CEMO* (Devillers & Vidrascu, 2006) fueron grabadas de servicios telefónicos. El uso de habla emocional actuada simplifica el reconocimiento automático de emociones.

Es difícil usar reconocedores entrenados con datos actuados en el mundo real dado que en las emociones generadas en interacción entre personas no son expresadas tan intensa y prototípicamente como un actor usualmente lo interpreta (Steidl, 2009). Por otro lado, adquirir datos en ambientes reales es problemático. Es difícil obtener los derechos de autor para hacer los datos disponibles al público, la posición de micrófonos y cámaras no es ideal, hay ruido como fondos visuales y acústicos, y no hay control en el contenido emocional (Busso, et al., 2008). La tendencia actual es grabar habla espontánea natural y realista en ambientes con condiciones restringidas tales como una entrevista (Gunes, Schuller, Pantic, & Cowie, 2011).

Otro punto importante a considerarse es el esquema de anotación el cual debe ser fácilmente entendible por los evaluadores humanos con el propósito de alcanzar un acuerdo alto entre evaluadores. Los dos esquemas de anotación más usados son el categórico, donde se asignan emociones discretas a muestras de voz, y el continuo, donde cada muestra de voz se evalúa con valores numéricos correspondientes a los niveles de primitivas emocionales.

La anotación continua se puede hacer estáticamente, por ejemplo, usando *The Self Assessment Manikins* (SAM) (Lang, 1980) de esta manera los segmentos completos son evaluados con el mismo valor. La anotación de primitivas también puede ser hecha

dinámicamente, es decir siguiendo un estado emocional continuamente en el tiempo, como lo permite la herramienta de etiquetado Feeltrace (Cowie, et al., 2000). También es valioso anotar otro tipo de información como información lingüística, eventos, historia, contexto y otros tipos de descriptores lingüísticos y emocionales. Es deseable una alta diversidad de edad, género, idioma, contexto sociocultural y por supuesto diversidad emocional. Para el caso de anotación de primitivas emocionales, lo ideal es tener suficientes muestras distribuidas en el espacio tridimensional.

Tabla 5 Bases de datos que incluyen anotaciones de primitivas emocionales. V = Valencia, A = Activación, D = Dominación, E = Anticipación / Expectación, I = Intensidad Emocional

Base de datos	EmoWisconsin (Nuestro corpus)	IEMOCAP	VAM	SAL 1	SEMAINE
Horas Grabadas	11:38	12:00	12:00	4:11	6:30
Hablantes	28	10	20	4	20
Segmentación	Turnos	Turnos	Turnos	Sesión	Sesión
Muestras	2,040	10,039	947	4	25
Tipo de Datos	Inducidos	Actuados	Espontáneos	Inducido	Inducido
Idioma	Español	Inglés	Alemán	Inglés	Inglés
Primitivas	V,A,D	V,A,D	V,A,D	V,A	V,A,D,E,I
Evaluadores	11	3	17	4	4

La Tabla 5 muestra algunas de las bases de datos existentes, anotadas con primitivas emocionales y sus propiedades más importantes. Es necesario generar más bases de datos en idiomas diferentes, cubriendo un rango más amplio de fenómenos emocionales, tomando en cuenta puntos relevantes, como espontaneidad y riqueza de anotación para permitir estudios posteriores y comparación de resultados entre diferentes características acústicas y clasificadores. (Schuller, Steidl, & Batliner, 2009).

Para los propósitos de nuestra tesis es necesario tener al menos una base de datos etiquetada con las primitivas emocionales Valencia, Activación y Dominación. Adicionalmente, necesitamos anotaciones de emociones discretas para cada muestra, que nos ayuden a validar las estimaciones de primitivas emocionales, evaluando el mapeo hecho desde el enfoque continuo hacia el discreto. Hay varias bases de datos etiquetadas con emociones discretas como FAU Aibo (Steidl, 2009), *Berlin Database of Emotional Speech* (Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005), *Spanish Emotional Speech* (Montero, 2003) y unas cuantas etiquetadas con primitivas emocionales como

VAM Corpus (Narayanan, Grimm, & Kroschel, 2008) y IEMOCAP Database (Busso, et al., 2008). La base de datos que usamos en la mayor parte de nuestros experimentos está etiquetada con categorías discretas comunes y con primitivas emocionales. Esta base se llama IEMOCAP (Busso, et al., 2008) (*Interactive Emotional Dyadic Motion Capture Database*).

Este corpus fue colectado por *Speech Analysis and Interpretation Laboratory en University of Southern California*. Esta base de datos está en inglés fue grabada por diez actores en pares hombre - mujer. Incluye información del movimiento de manos, cara y cabeza, así como información detallada de expresiones faciales y ademanes en audio y video. Para generar un diálogo emocional los autores diseñaron dos escenarios; en el primero los actores siguieron un guion mientras que en el segundo, los actores improvisaron de acuerdo a una situación preestablecida. El corpus completo contiene alrededor de doce horas de grabación. Las muestras fueron etiquetadas con emociones discretas como felicidad, enojo, tristeza, frustración, miedo, sorpresa y neutral. Las categorías "otro" y "no identificado" también fueron incluidas. Para evaluar las primitivas emocionales se asignó un valor entero entre uno y cinco, como: Valencia (1-negativo, 5-positivo), Activación (1-calm, 5-excited), and Dominación (1-debil, 5-fuerte).

Las características de esta base de datos la hacen muy interesante para los propósitos de nuestro trabajo ya que su anotación incluye los dos enfoques más importantes. Se ha puesto atención a la interacción espontánea, a pesar de usar actores. Además ha sido capturada en condiciones ideales para la grabación. La base de datos muestra una diversidad significativa de emociones. Usamos el audio de la primera sesión segmentada en turnos. Para estos experimentos utilizamos únicamente las categorías: enojo, felicidad, neutral y sorpresa. El total de muestras incluidas es 1,820.

El Corpus FAU Aibo, descrito en (Steidl, 2009), es un corpus con grabaciones de niños interactuando con el robot mascota de Sony Aibo. El corpus consiste de habla con emociones espontáneas. Se hizo creer a los niños que el robot respondía a sus órdenes, mientras el robot estaba en realidad respondiendo a las órdenes de un operador humano. El operador hacía que el robot se comportara de acuerdo a una secuencia de acciones predeterminada; en algunas ocasiones el robot era desobediente, provocando reacciones emocionales. Los datos fueron recopilados en dos escuelas diferentes en Alemania. Los participantes fueron 51 niños en edades de 10 a 13 años, 21 niños y 30 niñas; alrededor de 9.2 horas de habla. La voz fue transmitida con una diadema inalámbrica de alta calidad.

Las grabaciones fueron segmentadas automáticamente en turnos. Cinco personas entrenadas escucharon cada grabación en orden secuencial para etiquetarlas con una de 10 clases. El corpus está etiquetado a nivel de frase. Para otorgar una clase a una palabra se hizo una votación entre las opiniones de los etiquetadores. Si tres o más coinciden se atribuye la etiqueta a la palabra.

El corpus VAM se describe en (Narayanan, Grimm, & Kroschel, 2008). Consta de 12 horas de grabaciones en audio y video del *Talk Show* alemán *"Vera am Mittag".* Este corpus tiene la particularidad de estar etiquetado con tres primitivas emocionales: Valencia, Activación y Dominación. Para etiquetar este corpus se usaron 17 evaluadores humanos. Cada evaluador etiquetó todas las muestras con la idea de calcular el grado de acuerdo entre etiquetadores. Se cuenta con 947 muestras emocionales con 47 hablantes (11 h / 36 f) con una duración promedio de 3.0 segundos por elocución. Se cuenta con el audio, así como con transcripciones.

3.4 Discusión del estado del arte

De acuerdo a la revisión del estado del arte presentado en este capítulo identificamos tres áreas de oportunidad:

1. Generación y diversificación de datos confiables: Existen muchas bases de datos de habla emocional, sin embargo, la gran mayoría están grabadas en escenarios demasiado controlados, lo que restringe su confiabilidad para el estudio del reconocimiento de emociones genuinas en contextos realistas. Además, no existe una base de datos de este tipo en español mexicano.

2. Modelado de las emociones: A pesar de que es posible alcanzar un buen desempeño en el reconocimiento de emociones mediante el modelado discreto de emociones en cierto tipo de aplicaciones, se ha detectado que este enfoque tiene varias limitaciones por lo que recientemente el modelado continuo ha captado el interés de los principales grupos de investigación en el área. Sin embargo la investigación con este enfoque aún es incipiente y no se ha mostrado cual puede ser el verdadero beneficio de usar este enfoque.

3. Procesamiento de la voz: Aún no es claro que características acústicas son las más relevantes para la discriminación automática de emociones en la voz sobre todo desde el enfoque de los modelos continuos.

Las principales aportaciones buscadas en esta tesis están dirigidas a la solución de estos tres puntos. En los siguientes capítulos se explicará el trabajo realizado al respecto

Capítulo 4: Método de Reconocimiento de Emociones Basado en un Modelo Continuo

El diagrama de la Figura 4 muestra el modelo de reconocimiento de emociones propuesto en esta tesis. En el lado izquierdo, en los recuadros; C1, C2 y C3, se muestra el flujo de los datos de entrenamiento para la creación de los modelos usados en el método de reconocimiento de emociones. Estos componentes toman como entrada archivos de audio de voz grabada y sus respectivas etiquetas de Valencia, Activación, y Dominación. La salida son tres modelos entrenados los cuales son usados en la etapa de prueba como se indica con líneas punteadas.

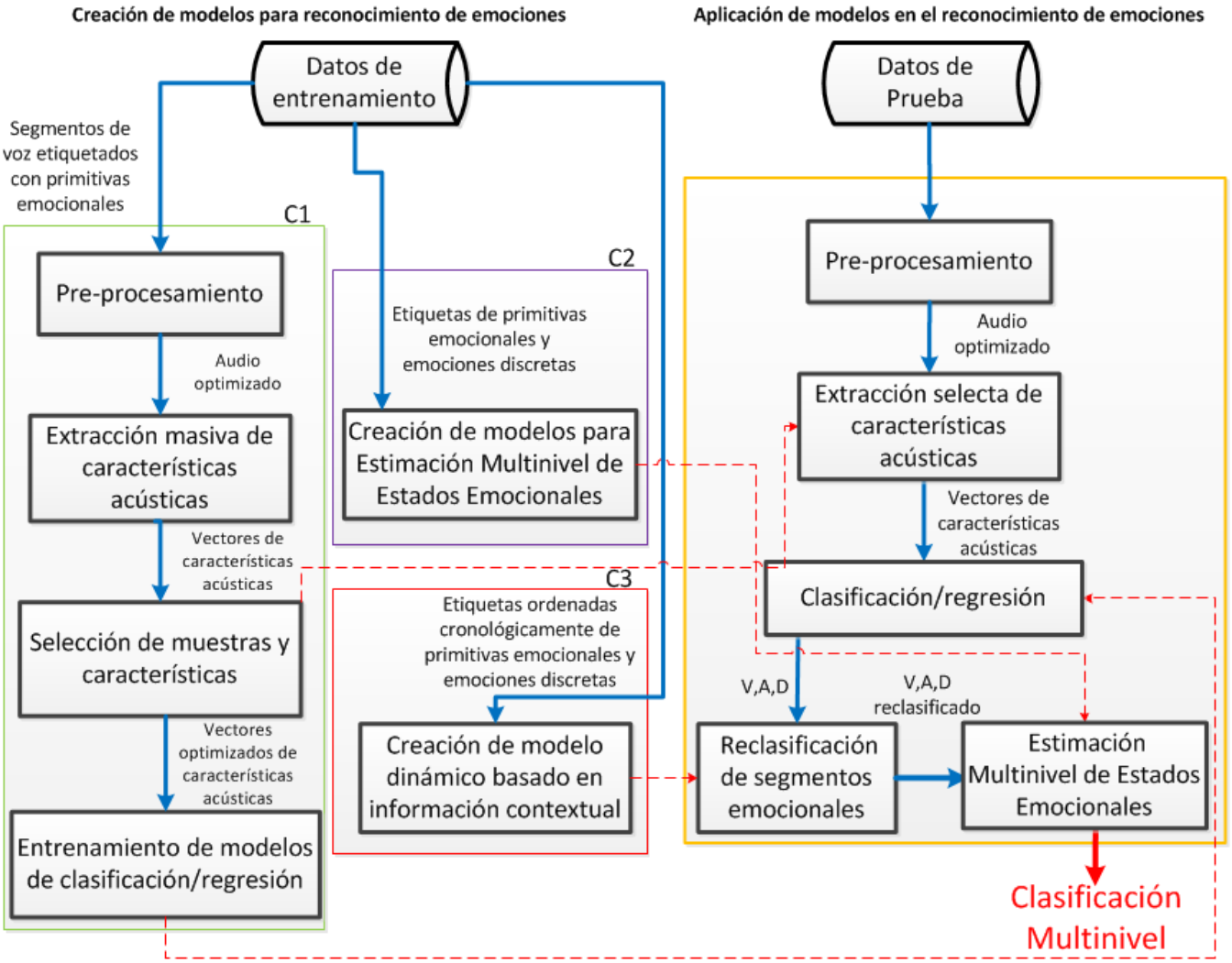


Figura 4 Modelo Propuesto

El lado derecho (recuadro amarillo) muestra el flujo de los datos en la etapa de aplicación de modelos para el reconocimiento de emociones, el cual toma como entrada archivos de audio de voz grabada y como salida genera una clasificación emocional de los segmentos dados. La etapa de creación de modelos para reconocimiento de emociones consiste de tres componentes complementarios que son: Creación de modelos para estimación multinivel de emociones, Creación de modelo dinámico basado en información contextual y Entrenamiento de modelos de clasificación y regresión, que a su vez toma como entrada el resultado de procesos previos.

En la Figura 4 se agrupan los componentes de esta etapa en recuadros de colores. En el componente C1, de color verde, se pre-procesa el audio y se extraen características acústicas. Se identifican los atributos acústicos más adecuados para caracterizar la señal acústica y de esa manera permitir el reconocimiento de emociones. Esto implica la selección de las mejores muestras y características. En el componente C2, de color morado, se crean modelos para la estimación multinivel de emociones, dichos modelos están basados en agrupamiento difuso. En el componente C3, de color rojo, se crea un modelo dinámico considerando la información contextual de los datos de entrenamiento; esta información se usa en la reclasificación de segmentos emocionales.

Antes de extraer características acústicas de la voz grabada se realiza un *pre-procesamiento* de la señal para mejorar la calidad de la grabación. Este pre-procesamiento no siempre es necesario, depende de la relación entre ruido y voz en la grabación. El audio pre-procesado pasa a la etapa *de extracción masiva de características acústicas* que consiste en generar vectores de atributos que describen acústicamente los segmentos de audio. Para determinar qué características acústicas usar, en el módulo de *selección de muestras y características*, se realizó un análisis monolingüe y multilingüe analizando una amplia variedad de propiedades acústicas relacionadas con la expresión de emociones en voz y se logró identificar los atributos acústicos más útiles para discriminar emociones. Las mejores características acústicas para cada primitiva emocional se usan en el *entrenamiento de modelos de clasificación y regresión* para estimar el nivel de primitivas emocionales.

El módulo de *creación de modelos para estimación multinivel de emociones* toma como entrada los datos de etiquetado emocional continuo del corpus de entrenamiento por lo cual, dichos datos, deben estar etiquetados con las primitivas emocionales Valencia, Activación y Dominación (VAD) opcionalmente; dichos datos pueden estar etiquetados con emociones discretas ya que las etiquetas de emociones discretas sirven para calcular

automáticamente la ubicación de ciertas emociones en el espacio tridimensional continuo definido por el modelo VAD o dicha ubicación puede ser indicada manualmente de acuerdo a las emociones que se deseen reconocer.

El módulo de *creación de modelo dinámico basado en información contextual* toma como entrada los datos etiquetados con primitivas emocionales ordenados cronológicamente. Realiza un proceso de ajuste de diversos parámetros y la salida es el modelo con una parametrización óptima.

En la etapa de aplicación de modelos en el reconocimiento de emociones se usan los modelos creados en la etapa anterior para reconocer emociones en datos de prueba. Esta etapa inicia con el módulo de pre-procesamiento que es idéntico al usado en la etapa de creación de modelos para reconocimiento de emociones. Los datos de prueba son grabaciones de conversaciones o monólogos que son pre-procesados con el objeto de normalizar el audio y dividir las grabaciones de habla continua en segmentos de audio equivalentes a un "turno". El audio pre-procesado pasa al módulo de extracción selecta de características dónde solo se extraen las mejores características identificadas en la etapa de entrenamiento. La salida de este módulo son vectores de características acústicas describiendo cada muestra de prueba. Dichos vectores pasan a los modelos de clasificación y regresión. La salida de este módulo son estimaciones de Valencia, Activación y Dominación para cada muestra de prueba. Dichas estimaciones pasan a la etapa de reclasificación de segmentos donde se aplica una reclasificación basada en el contexto emocional de cada muestra; de esta manera se corrigen posibles errores en la clasificación tomando en cuenta la clasificación de muestras anteriores y el grado de certeza de la clasificación de la muestra actual. La salida es nuevamente una estimación de Valencia, Activación y Dominación rectificada.

El último módulo es el de estimación multinivel de emociones, su entrada es la estimación de Valencia, Activación y Dominación y la salida es una interpretación en tres diferentes niveles de abstracción del contenido emocional de las muestras. Las emociones discretas son usadas en el módulo de creación de modelos para estimación multinivel de emociones. La utilidad de cada nivel de abstracción depende de la aplicación y hace que el modelo sea muy flexible para ser usado en cualquier escenario.

4.1 Innovación de la propuesta

La principal innovación de la propuesta radica en la forma en que se usa el modelo tridimensional continuo para hacer una estimación emocional flexible a diferentes aplicaciones. El objetivo del método propuesto es modelar fenómenos emocionales como intensidad y mezcla de emociones, que son fenómenos presentes en la expresión emocional cotidiana. Por supuesto, dependiendo de la aplicación particular se busca también mantener la referencia a la representación discreta de las emociones.

En trabajos relacionados ya se ha empezado a investigar sobre cómo sacar provecho de los modelos emocionales continuos para estimar de manera más adecuada la carga emocional en la voz (Grimm M. , Kroschel, Mower, & Narayanan, 2007)**,** (Lugger & Yang, 2008)**,** (Wöllmer M. , et al., 2009)**,** (Eyben, et al., 2010). Sin embargo, creemos que es posible obtener una representación más descriptiva y completa de lo que se ha propuesto hasta el momento.

Nuestra hipótesis es que mediante la interpretación de primitivas emocionales es posible obtener una representación más real del contenido emocional en la voz al manejar diferentes niveles de abstracción, ya que gracias al modelo continuo es posible obtener una interpretación a nivel de emoción discreta, de intensidad de emociones, de mezclas de emociones o de grupos de emociones.

De acuerdo al estudio del estado del arte y de su discusión en la sección *3.4 Discusión del estado del arte*, nuestra propuesta aporta principalmente en los siguientes aspectos:

Generación de datos y portabilidad a escenarios reales: Trabajamos con diferentes bases de datos, grabadas en diferentes contextos y diferentes idiomas etiquetadas con primitivas emocionales. Esto con el objetivo de probar la robustez del método a diferentes condiciones. Además, generamos nuestra propia base de datos haciendo hincapié en generar emociones genuinas.

Caracterización acústica de las emociones: Se exploró un espectro muy amplio de características acústicas desde el enfoque de los modelos continuos incluyendo

características prosódicas, de calidad de voz, y espectrales. Se incluyó tanto procesamiento estático como dinámico de características a nivel de segmentos de audio y a nivel de conversación mediante el uso del contexto emocional.

Modelado de fenómenos emocionales: Nuestro trabajo está basado en un modelo continuo tridimensional cuyas primitivas son Valencia, Activación y Dominación. Se propone un método basado en la estimación automática de primitivas emocionales, una estimación multinivel de emociones y conservando el mapeo hacia emociones discretas desde el modelo tridimensional continuo.

La Tabla 6 indica los aspectos en los que estamos interesados en abordar y como se compara la propuesta que hacemos con trabajos importantes en el área y que tomamos como punto de partida.

Tabla 6 Comparativo de los trabajos relacionados con la propuesta hecha para esta tesis

Autor	Tipo de Base de datos	Etiquetado de Base de datos	Modelo	Procesado	Características			
					Prosodia	Calidad	Espectro	Texto
Grimm 07	Espontánea	Continuo	Continuo	Estático	√		√	
Steidl 09	Espontánea	Discreto	Discreto	Estático	√	√	√	√
Lugger 08	Actuada	Discreto	Continuo	Estático	√	√	√	
Esta tesis	Espontánea	Continuo/Discreto	Continuo	Estático/ Dinámico	√	√	√	√

4.2 Creación de corpus de habla emocional

En esta sección se describe la creación de una base de datos de habla emocional en el español hablado en México. Se grabaron 28 niños y niñas de entre 7 y 13 años en un ambiente controlado, moderadamente ruidoso. Se grabaron dos sesiones de entre 10 y 15 minutos por niño. Para crear esta base de datos se adaptó la prueba neuropsicológica *Wisconsin Card Sorting Test (WCST)*, buscando provocar diferentes emociones en niños. La prueba se aplicó en dos partes. La primera sesión era fácil de realizar y motivante para el niño, con lo que se evocaban emociones positivas como alegría y excitación. La segunda sesión era difícil y estresante, provocando en el niño emociones negativas como frustración y nerviosismo. Se obtuvo una buena diversidad de emociones. Los datos fueron

etiquetados con los enfoques discreto y continuo a nivel de turno. En el resto de este documento llamamos a este corpus *EmoWisconsin*. En la Figura 5 se muestra la configuración del escenario para la realización de las sesiones de grabación.

Figura 5 Configuración de escenario para grabación de la base de datos EMOWisconsin

4.2.1 Diseño de la prueba

El experimento propuesto para construir nuestra base de datos es modificar el la prueba psicológica *Wisconsin Card Sorting Test (WCST)* (Grant & Berg, 1948) para inducir emociones en niños. Debido a que no existía una base de datos con habla emocional espontanea en español, se decidió crear esta base de datos, diseñada de acuerdo a las siguientes necesidades:

I. El idioma es el español hablado en México, debido a nuestro interés en estudiar las particularidades de nuestra lengua materna. Dado que actualmente no existe otra base de datos similar en español, la creación de esta nos permite extender el análisis multilingüe de primitivas emocionales.
II. El habla es espontánea e inducida. Es espontánea con el objetivo de estudiar los fenómenos que sólo se producen espontáneamente, como mezcla y variación en la intensidad de las emociones. Es inducida con el fin de tener un ambiente controlado y producir nuestros propios datos, lo cual nos da derecho a usarlo sin restricciones y ponerlo a disposición de la comunidad científica.
III. La anotación del contenido emocional se hizo usando primitivas emocionales y emociones discretas. Esto se hizo con el objetivo de comparar ambos enfoques, continuo y discreto, y aumentar su aprovechamiento por investigadores interesados en uno o entro enfoque.

IV. Se trabajó con una cantidad relativamente alta de participantes para permitir la creación de modelos independientes del hablante.

V. Se trabajó con un número relativamente elevado de evaluadores para hacer frente a la alta subjetividad en la anotación de emociones.

VI. Se grabaron varias horas de audio. Dichas grabaciones no están libres de ruido, con el objetivo de emular las condiciones de audio en aplicaciones reales.

VII. Se trató de obtener diversidad emocional induciendo emociones en las regiones altas y bajas de cada primitiva emocional.

4.2.2 WCST Emocional

La prueba de ordenamiento de cartas Wisconsin, también conocida como WCST, por sus siglas en inglés, fue diseñada para evaluar las funciones cognoscitivas abstractas de los individuos. Para aplicar esta prueba se usan 132 tarjetas con diferentes figuras. Ver Figura 6. La tarea requiere que los sujetos encuentren el criterio de ordenamiento de tarjetas que el examinador tiene en mente. El criterio de ordenamiento es una combinación de las dimensiones perceptuales: color, número y forma. La búsqueda del criterio de ordenamiento se hace mediante prueba y error, basada en la retroalimentación de examinador. El individuo siendo evaluado tiene las tarjetas en sus manos y va eligiendo tarjetas que pone sobre la mesa. El examinador no le dice cómo hacer que las tarjetas cumplan con el criterio de ordenamiento; solo le dice si la tarjeta elegida cumple o no con el criterio. Una vez que el sujeto cree haber encontrado el criterio de ordenamiento, debe seguir eligiendo cartas que confirmen que el criterio de ordenamiento se sigue cumpliendo al cambiar algunas dimensiones perceptuales y conservar otras. Después de seis aciertos consecutivos, el criterio de ordenamiento cambia sin aviso (Nyhus & Barcelo, 2009).

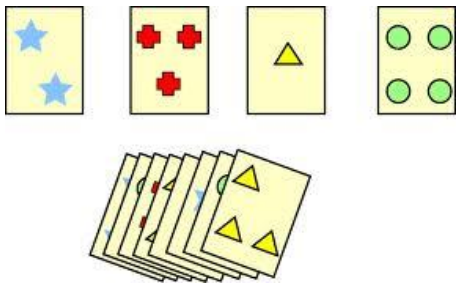


Figura 6 Tarjetas usadas en la prueba

EL WCST no tiene tiempo límite y el ordenamiento continúa hasta que todas las cartas están ordenadas o un máximo de seis criterios de ordenamiento son alcanzados. La prueba es llevada a cabo con el examinador de un lado de la mesa viendo de frente al individuo en el otro lado de la mesa. La prueba toma entre doce y veinte minutos en realizarse y genera indicadores psicométricos, incluyendo porcentajes y percentiles de: criterios de ordenamiento alcanzados, pruebas, y errores. Con estos resultados se calcula y diagnostica al individuo.

Se hicieron algunas variaciones del WCST original con el objetivo de estimular el diálogo e inducir diferentes emociones en los individuos. La prueba se le presenta al niño como un juego en el que tiene que estar muy concentrado. El desempeño de los niños y la interacción con los examinadores durante la prueba activa reacciones que generan cambios emocionales lo cual a su vez, se refleja en sus voces. Cada niño participó en dos sesiones que nombramos: sesión positiva y sesión negativa. A continuación se describe en que consiste cada una.

Sesión Positiva: Un examinador amable inspira confianza al niño antes y durante esta sesión. El examinador anima al niño a expresar verbalmente todas sus impresiones sobre el juego. Las instrucciones del juego le son claramente explicadas. El examinador diseña criterios de ordenamiento que el niño puede resolver causando satisfacción en él. En esta fase de la prueba, se espera que el niño inicie en un estado emocional neutro y se dirija hacia un estado emocional con Valencia, Activación, y dominio altos. Las emociones esperados en esta sesión fueron serenidad, seguridad, motivación, alegría.

Sesión Negativa: Cuando la sesión positiva finaliza, el examinador le dice al niño que va a participar en la segunda parte del juego y que espere un par de minutos. Al comienzo de la sesión negativa se presenta un examinador de muy mal humor. Durante el juego el examinador parece estar molesto e impaciente. En esta fase de la prueba, se espera que el niño inicie en estado emocional neutral y se dirija hacia emociones con Valencia y Dominación baja. La Activación puede variar de bajo a alto. Inicialmente, las emociones esperadas en esta sesión son nerviosismo, inseguridad, estrés y frustración.

Como en la prueba original, al participante se le entrega el conjunto de 132 tarjetas. El examinador decide sobre la marcha, la dificultad del criterio de ordenamiento de acuerdo con la edad del niño, su desempeño durante la prueba y, si se trata de la sesión positiva o negativa. Durante las dos sesiones el niño es incitado a conversar. Después de la

finalización de las sesiones, un formulario de identificación es llenado con la edad, el sexo, la duración de cada sesión y criterios de ordenamiento alcanzados.

Los dos elementos más recurridos para la inducción de emociones durante las sesiones son: En primer lugar, la postura o personalidad que el examinador demuestra mediante gesticulaciones, ademanes, tono de voz, palabras para animar o presionar a los niños. En segundo lugar, el desempeño del niño en el juego. Se espera que eventos como los siguientes desencadenen reacciones emocionales en los niños:

- Enfrentarse a criterios de ordenamiento muy difíciles de deducir puede causar estrés o ansiedad.
- Encontrar el criterio de ordenamiento puede conducir a satisfacción, confianza.
- Malos entendidos con el examinador podrían causar inseguridad, incertidumbre.
- Enfrentarse recurrentemente a criterios de ordenamiento demasiado fáciles eventualmente conducirá al aburrimiento, tedio.
- Se incapaz de encontrar el criterio de ordenamiento puede conducir a decepción, frustración
- Darse cuenta de que se ha desarrollado habilidad para el juego podría conducir a alegría, entusiasmo.
- Presión por parte de examinador para resolver el juego rápidamente podría conducir a estrés, nerviosismo.

Antes de la grabación de la base de datos final, se realizó una prueba piloto con nueve niños. Durante esta prueba, nos dimos cuenta que era importante conocer el estado emocional de los niños antes de las sesiones de grabación. Cuando estaban en una situación personal dolorosa o estresante, la sesión negativa llegó a ser demasiado abrumadora para ellos. Estos niños se encontraban demasiado estresados al final de la segunda sesión. La personalidad y la experiencia personal de cada niño juegan un papel importante, ya que algunos niños estaban más habituados a manejar situaciones de presión y hacer frente a adultos intimidantes. Para conocer estos importantes detalles acerca de los niños, se entrevistó brevemente a cada niño antes de las sesiones de grabación. En esta entrevista se indagó en su personalidad y estabilidad emocional actual para evitar abrumarlos durante las sesiones. Siempre se aplicó la primera sesión positiva y luego negativa con el objetivo de ganarse la confianza del niño durante la primera.

4.2.3 Adquisición y Segmentación del audio

Las sesiones fueron aplicadas por un grupo de cinco psicólogos quienes ayudaron a diseñar el protocolo final de la prueba. Estos examinadores estuvieron intercambiando los roles de examinador amable y enojón. Se trabajó con un grupo de 28 niños dentro del rango de edades que acepta la prueba. Once niños y diecisiete niñas entre siete y trece años de edad. La grabación se realizó en un ambiente con poco ruido, pero no completamente aislado. Fue grabado en dos computadoras con tarjeta de sonido *Sigmatel STAC 9200.* Las grabaciones fueron mono canal, con un tamaño de muestra de 16 bits, frecuencia de muestreo de 44.100 kHz y almacenadas en formato Windows WAV PCM. Se grabó 11:39 horas en 56 sesiones, dos sesiones por niño, durante siete días en el Instituto Nacional de Astrofísica Óptica y Electrónica.

Encontrar el tamaño óptimo de los segmentos de audio es un problema abierto en el reconocimiento de emociones en voz. Diferentes alternativas como palabras, turnos, frases, etc. han sido probadas. En general, es aceptado que la segmentación a nivel de turno es una buena opción (Steidl, 2009), (Busso, et al., 2008). La segmentación de nuestros datos se realizó manualmente a nivel de turnos. Después de la segmentación se obtuvo un total de 3.098 segmentos. 1.424 adquiridos en sesiones positivas y 1.674 adquiridos en sesiones negativas. Los criterios para la segmentación instruidos a las personas que lo hicieron son los siguientes: 1) Evitar los segmentos con más de una voz al mismo tiempo, es decir, que la voz del examinador y del niño se superponen 2) Dividir los turnos cuando contengan pausas largas. 3) Incluir expresiones no lingüísticas cuando se mezclan con palabras. 4) No incluir expresiones no lingüísticas aisladas.

4.2.4 Anotación emocional y acuerdo entre etiquetadores

Para describir el contenido emocional en nuestros datos usamos dos esquemas de anotación: el discreto y el continuo. Para la anotación continua usamos los métodos de etiquetado *SAMs* (Grimm M. , Kroschel, Mower, & Narayanan, 2007) y *Feeltrace* (Cowie, et al., 2000). Anotamos nuestros datos con seis emociones discretas. Además, todas las sesiones fueron transcritas. Para elegir las emociones discretas que serían usadas se realizó una prueba piloto. Se determinó que las seis emociones más recurrentes en el habla de los niños son: Inseguridad, Molestia, Motivación, Nerviosismo, Neutro y Seguridad. Los anotadores asistieron a una sesión de entrenamiento donde se dejó en claro que emoción representa cada una de las seis etiquetas y cada primitiva emocional. Para la anotación

discreta, además de anotación categórica los anotadores tenían a su disposición una etiqueta para segmentos que no coinciden con ninguna etiqueta y otra para segmentos que consideren mal segmentados. La Tabla 7 muestra el número de segmentos por categoría. Once etiquetadores participaron en este proceso. Usamos la plataforma de evaluación TRUE (Planet, Iriondo, Martinez, & Montero, 2008) para anotar nuestros datos, ver Figura 7

Tabla 7 Número de segmentos por emoción, SEG = Número de segmentos, ID = Indefinido, MS = Mal segmentado, INS = Inseguro, MOL = Molestia, MOT = Motivación, NER = Nerviosismo, NEU = Neutral, SEG = Seguridad

Sesión	SEG	ID	MS	INS	MOL	MOT	NER	NEU	SEG
Positiva	1,424	525	119	205	5	41	105	11	413
Negativa	1,674	533	118	310	12	31	162	10	498
Total	3,098	1,058	237	515	17	72	267	21	911

Para estimar el acuerdo entre evaluadores en la anotación de emociones discretas usamos la medida *Free-marginal multi rater Kappa* (Warrens, 2010). Para medir el acuerdo en la anotación de primitivas emocionales usamos el índice *Cronbach Alpha* (Cronbach, 1951). En los resultados mostrados en la Tabla 8 se observa que el acuerdo es bajo, principalmente para la anotación de emociones discretas. Sin embargo, el tener muchos anotadores nos permite incrementar el acuerdo eliminando los evaluadores con el menor acuerdo y de esta manera obtener una anotación más confiable.

Calculamos las etiquetas finales a partir de las anotaciones de todos los evaluadores usando el siguiente criterio: Para las anotaciones discretas contamos cuantas veces la muestra es anotada con cada emoción. La etiqueta final es la emoción que aparece más. Cuando hay empate la muestra se marca como indeterminada. Para la anotación continua, la etiqueta final es el promedio de todas las anotaciones.

Tabla 8 Acuerdo entre etiquetadores

Sesión	Kappa	Total	V	A	D
Positiva	0.2265	0.3443	0.6671	0.7045	0.6480
Negativa	0.2496	0.3493	0.5765	0.6667	0.6029
Promedio	0.2380	0.3468	0.6218	0.6856	0.6254

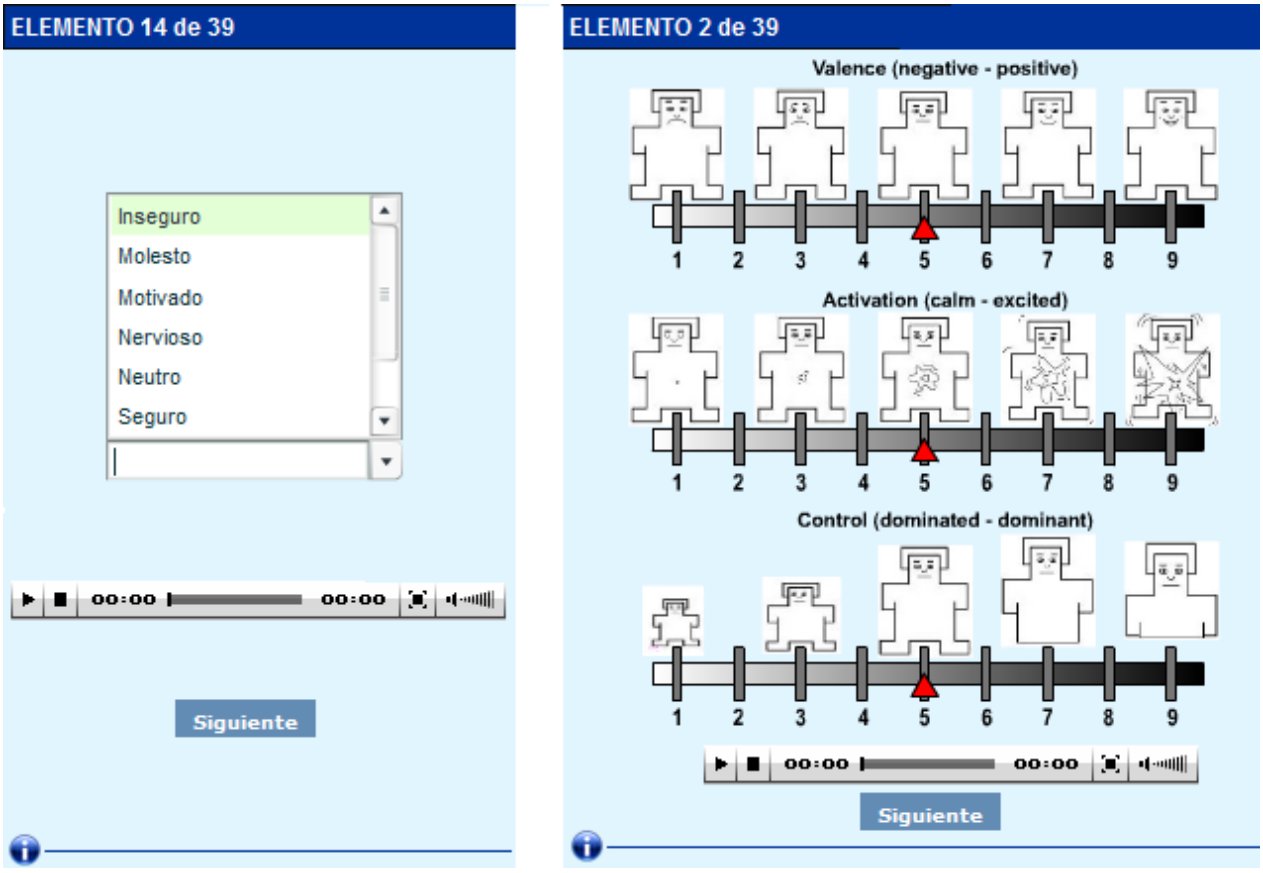


Figura 7 Interfaz de anotación TRUE

Capítulo 5: Caracterización de Voz y Selección de Datos

Las emociones son un fenómeno humano complejo. Un sinnúmero de investigadores han intentado una variedad de enfoques para modelar este fenómeno y encontrar el conjunto óptimo de descriptores. Varios autores (Xie, Chen, Chen, & Chen, 2005) (Lugger & Yang, 2007) (Batliner A. , et al., 2011) han trabajado en el análisis de las características acústicas más importantes desde el punto de vista de la categorización discreta. Sin embargo, no han estudiado con la misma profundidad la importancia de características acústicas desde el punto de vista de los modelos continuos. Creemos que el enfoque continuo tiene un gran potencial para modelar la ocurrencia de emociones en el mundo real. Este modelo continuo tridimensional es adoptado en esta tesis. Como un primer paso hacia la explotación del enfoque continuo, analizamos las características acústicas más importantes en la estimación automática de primitivas emocionales en voz. Posteriormente, usamos esta estimación para localizar el estado emocional de los individuos en el espacio multidimensional y, si es necesario, para mapearlo hacia una emoción discreta.

5.1 Extracción de características acústicas

La extracción de características acústicas se refiere al proceso mediante el cual se procesa la señal de voz y se obtiene una representación numérica describiendo propiedades acústicas de la voz. Como un paso previo a la extracción, se diseñó un módulo de pre-procesamiento de audio que mejora las grabaciones de voz y posibilita la segmentación automática a nivel de turnos, lo cual resulta útil para realizar pruebas con nuevos datos generados en ambientes con ruido. Con este proceso, mostrado en la Figura 8, se mejora el audio, se disminuye el ruido y se normaliza la magnitud acústica de la voz en las grabaciones. La segmentación automática está basada en la detección de pausas de acuerdo a un umbral establecido dada la relación de voz a ruido en la señal.

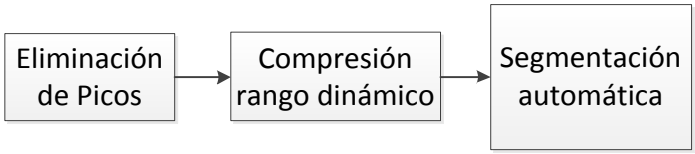


Figura 8 Pasos en el módulo de pre-procesamiento

Como se ilustra en la Figura 9, se extrajeron características acústicas de la señal de voz usando los programas para análisis acústico Praat (Boersma, 2001) y OpenSMILE (Eyben, Wöllmer, & Schuller, 2009). Se evaluaron dos conjuntos de características; uno diseñado mediante un enfoque selectivo, es decir, basado en un estudio tomando en cuenta las características que podrían ser útiles, características que han sido exitosas en trabajos relacionados y características usadas para tareas similares. El segundo conjunto de características fue obtenido aplicando un enfoque de fuerza bruta, es decir, generando una gran cantidad de ellas esperando que algunas sean de utilidad.

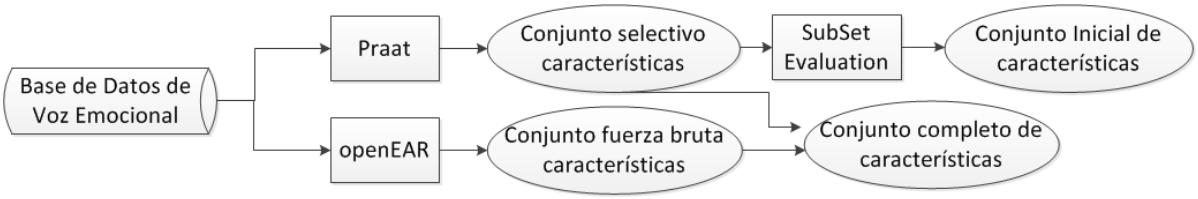


Figura 9 Proceso de extracción/selección de características

El conjunto de características selectivo es un conjunto de características que fue construido a través de nuestro trabajo de investigación con diferentes bases de datos (Pérez Espinosa & Reyes García, 2009) (Pérez Espinosa, Reyes García, & Villaseñor Pineda, 2010) y se obtuvo mediante el software Praat. Se diseñó este conjunto de características tratando de representar varios aspectos de la voz, incluyendo los atributos tradicionales asociados a la prosodia, es decir, duración, entonación e intensidad. También se incluyeron características que han mostrado buenos resultados en tareas similares como reconocimiento de voz, reconocimiento de hablante, clasificación de llanto de bebé (Santiago, Reyes G., & Gomez G., 2009), reconocimiento del idioma (Reyes, 2007), y detección de patologías en la voz (Dubuisson, Dutoit, Gosselin, & Remacle, 2009), (Ishi, Ishiguro, & Hagita, 2005).

Finalmente incluimos algunas características que por intuición y experiencia creímos podrían aportar información valiosa. La Tabla 9 muestra el número de características acústicas que hemos incluido en cada grupo. Dividimos los tres tipos de características en Prosódicas, Espectrales y Calidad de Voz.

Tabla 9 Conjunto de características - Enfoque Selectivo

Grupo	Tipo de Característica y descripción	Número de Características
	Prosódicas	
Tiempos	Estimación de la rapidez del habla basada en la detección de silabas por unidad de tiempo y en la duración de pausas y voz.	8
Entonación	Contorno Melódico (Pitch) basado en la periodicidad de la señal de voz	9
Intensidad	Contorno de intensidad de la señal	12
	Calidad de Voz	
Calidad de Voz	Diferentes medidas de aspectos presentes en voces patológicas como voz cortada, radio harmónicos a ruido y otras.	24
Calidad de Voz	Diferentes medidas de calidad de articulación basadas en la estimación de formantes, diferencias en bandas de energía y otras.	12
	Espectrales	
LPC	Transformada Rápida de Fourier	4
LPC	Promedio a Largo Plazo del Espectro (LTAS)	5
LPC	Ondeletas	6
MFCC	Coeficientes Cepstrales en las Frecuencias de Mel	96
Cocleogramas	Cocleogramas	96
LPC	Codificación Predictiva Lineal	96
TOTAL		**368**

Subdividimos las características prosódicas en tiempos de elocución, contorno melódico y contorno energético. Con respecto a las características de calidad de voz incluimos los dos descriptores más populares de este tipo de característica que son *Jitter* y *Shimmer* (Drioli, Tisato, Cosi, & Tesser, 2003). También incluimos otros descriptores de calidad de voz que han sido relacionados con la escala GRBAS[2] (*Grade, Roughness, Breathiness, Asthenia, Strain*) en trabajos afines (Dubuisson, Dutoit, Gosselin, & Remacle, 2009), (Ishi, Ishiguro, & Hagita, 2005), (Lugger & Yang, 2006), (Núñes B., Corte S., Suarez N., Señaris G., & Sequeiros, 2004).

Algunas de estas características nunca han sido usadas en reconocimiento de emociones en voz. Por ejemplo, las diferencias de energía entre bandas de frecuencia y el radio entre bandas de frecuencia, fueron usadas por (Dubuisson, Dutoit, Gosselin, & Remacle, 2009) para discriminar entre voces normales y patológicas. Para la detección

[2] Es una escala perceptual para analizar y medir la calidad vocal en pacientes con patologías en la voz

automática de *Voz Rota* (Vocal Fry) se han usado incrementos y decrementos en picos de energía (Ishi, Ishiguro, & Hagita, 2005). La articulación también es un parámetro importante para medir la calidad de voz. Incluimos algunas medidas estadísticas de los primeros cuatro formantes cómo descriptores articulatorios.

Incluimos varios tipos de representaciones espectrales. Algunas de estas representaciones nunca han sido usadas en reconocimiento de emociones como Cocleogramas, que han sido usados previamente para clasificación de llanto de bebé (Santiago, Reyes G., & Gomez G., 2009) y otras que han sido estudiadas muy poco para esta tarea como Wavelets (Kandali, Routray, & Basu, 2009).

El conjunto de características por fuerza bruta fue extraído usando el software OpenSMILE. Extrajimos un total de 6,552 características incluyendo funciones estadísticas de descriptores de bajo nivel tales como FFT-Spectrum, Mel-Spectrum, MFCC, Pitch (Frecuencia Fundamental F0 vía ACF), Intensidad, Espectro, LSP. El resultado del procesamiento de la señal con descriptores de bajo nivel es un arreglo de coeficientes, la longitud de dicho arreglo depende del tiempo de duración del segmento de audio. Para obtener vectores de características de la misma longitud para cada segmento, sin importar su duración, se calculan funciones estadísticas a los coeficientes de los descriptores de bajo nivel. Se calcularon 39 funcionales tales como: Extremos, Regresión, Momentos, Percentiles, Cruces, Picos, Promedios. La Tabla 10 muestra las características extraídas mediante el enfoque de fuerza bruta.

Tabla 10 Conjunto de características - fuerza bruta

Grupo	Tipo de Característica y descripción	Número de Características
	Prosódicas	
Intensidad	Energía LOG	117
Tiempos	Índice de cruces por cero	117
PoV	Probabilidad de voz	117
Entonación	F0 basado en el cálculo de frecuencia fundamental	234
	Espectrales	
MFCC	Coeficientes Cepstrales en las Frecuencias de Mel	1,521
MEL	Espectro de Mel	3,042
SEB	Energía espectral en bandas	469
SROP	Punto de partida spectral	468
sFlux	Flujo espectral	117
SC	Centroide espectral	117
SMM	Máximo y mínimo espectral	233
	TOTAL	**6,552**

Como se mencionó en el capítulo anterior, hay dos enfoques para el procesamiento de características en reconocimiento de emociones en voz: el enfoque estático y el enfoque dinámico. Nuestras características espectrales, por ejemplo los coeficientes MFCC son características estáticas ya que describen propiedades espectrales dentro de una ventana de muestreo donde la señal es aproximadamente estacionaria. El propósito de las características dinámicas es describir el cambio de las características a través del tiempo. Nuestro conjunto de características incluye características dinámicas generadas mediante el cálculo de la primera y segunda derivadas de las características estáticas del conjunto de características de fuerza bruta. El cálculo de la derivada se hace a partir de los vectores de coeficientes de características estáticas x^t y aplicando la Formula de regresión mostrada abajo donde W especifica la mitad del tamaño de la ventana a ser usada para calcular los coeficientes de regresión. El W usado fue 2. Para calcular la segunda derivada se aplica el mismo procedimiento sobre el vector de coeficientes de la primera derivada.

$$d^t = \frac{\sum_{i=1}^{W} i * (x^{t+i} - x^{t-i})}{2 \sum_{i=1}^{W} i^2}$$

Formula 8 Formula para cálculo de la derivada sobre una ventana de muestreo

Para extraer las características mencionadas en esta sección se implementaron scripts en Praat y openSMILE

5.2 Selección de características

A partir de experimentos de regresión usando el conjunto completo de características acústicas, es decir las características mostradas en la Tabla 9 más las mostradas en la Tabla 10, detectamos la necesidad de encontrar los mejores subconjuntos de características para construir modelos entrenados de estimación de primitivas emocionales en voz. En los experimentos mencionados, se usaron los datos del corpus VAM y se evaluó la precisión del reconocimiento automático mediante el cálculo del coeficiente de correlación de Pearson entre los valores estimados por el clasificador y los valores esperados dados por el etiquetado manual de las muestras del corpus VAM. Los resultados obtenidos mostraron una baja correlación en la estimación de primitivas emocionales cuando se entrenan modelos con el conjunto completo de 6,920 características acústicas, como se reporta a continuación. La Tabla 11 muestra el coeficiente de

correlación obtenido cuando se estima el valor de las primitivas emocionales con un modelo construido a partir de 942 muestras y 6,920 características.

Tabla 11 Índice de correlación obtenido usando todas las características acústicas

Primitiva Emocional	Índice de Correlación
Valencia	0.0151
Activación	0.0095
Dominación	-0.001

Como podemos ver, la correlación es muy baja, esto indica que no se encontraron patrones relevantes entre las variaciones de características acústicas y la emoción en la voz, por lo tanto, los modelos aprendidos de estos datos no son útiles. Incluir demasiadas características con relación al número de muestras complica la tarea del clasificador, SVM en este caso, impidiendo un modelo de predicción apropiado. Aun cuando tener muchos atributos podría mejorar el poder de discriminación, en la práctica, con una cantidad limitada de datos una cantidad excesiva de atributos retrasa significativamente el proceso de aprendizaje y frecuentemente resulta en un sobreajuste (Morales & González, 2009). Para resolver este problema, inicialmente, probamos diferentes selectores de atributos tales como *SubSetEval* y *ReliefAttribute*, cuyas descripciones formales pueden consultarse en la sección *2.2 Selección de características* de este documento. Usando dichos algoritmos de selección no fue posible mejorar los resultados de correlación en la estimación de primitivas emocionales. Debido a este problema, fue necesario idear una manera de seleccionar las mejores características entre un gran número de ellas, teniendo en mente que contamos con pocas clases. Propusimos dos esquemas para la selección de atributos trabajando con conjuntos pequeños de características a la vez, con la idea de evitar la búsqueda de las mejores características en el conjunto completo.

5.2.1 Selección no agrupada de características

El objetivo de este esquema de selección de características es hacer una búsqueda de las mejores características en un espacio de búsqueda muy grande pero, partiendo de un conjunto inicial de buenas características encontradas previamente. La Figura 10 muestra el proceso aplicado en este esquema donde el proceso inicia a partir de un conjunto base de características, obtenido de un proceso de selección de características aplicado a la base de datos VAM. Se decidió tomar como punto de partida la base de datos VAM dado que es la

base de datos más explorada para trabajar con el modelo emocional continuo de emociones. El conjunto inicial de características alcanzó buena correlación en la estimación de primitivas emocionales en la base de datos VAM. Este proceso de selección fue llevado a cabo con 252 características obtenidas mediante un enfoque selectivo y 949 muestras. Los detalles de dicho experimento se pueden consultar en (Pérez Espinosa, Reyes García, & Villaseñor Pineda, 2010). Una vez encontrado el conjunto inicial de características, se realiza el proceso de selección de muestras explicado en la sección *5.6 Selección de muestras* de esta tesis. Finalmente, se aplica el proceso de selección de características conocido como *Linear Floating Forward Selection (LFFS)* cuyo algoritmo puede consultarse en la sección *2.2 Selección de características* de este documento. Este proceso es repetido para cada primitiva emocional.

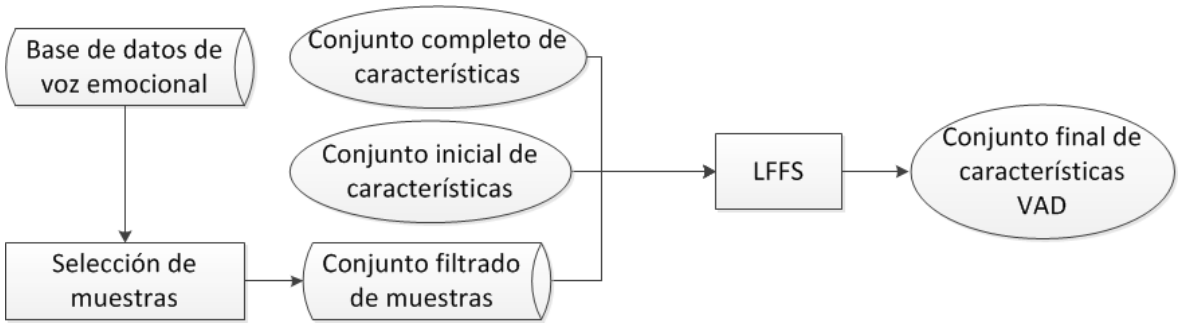


Figura 10 Selección no agrupada de características

5.2.2 Selección agrupada de características

En este esquema, la idea es dividir el conjunto de características completo en grupos más pequeños de acuerdo a las propiedades acústicas que representan. El objetivo de este esquema es en primer lugar, realizar la búsqueda de mejores características sobre un espacio de búsqueda más reducido y en segundo lugar, facilitar el análisis de que propiedades del habla aportan mayor información para discriminar emociones. Los grupos de características son mostrados en Tabla 9 y Tabla 10. La Figura 11 muestra los pasos seguidos en este esquema. Primero, aplicamos la selección de muestras explicado en la sección *5.6 Selección de muestras* de esta tesis. Segundo, dividimos el conjunto de datos completo en subconjuntos agrupando las características que modelan las mismas propiedades del habla por ejemplo, el grupo Tiempo contiene todas las características

orientadas a medir la velocidad en la voz, el grupo Entonación contiene todas las características orientadas a medir la entonación, etc. Una vez agrupados se aplica LFFS. A diferencia de la selección no agrupada de características, en esta ocasión el proceso de búsqueda mediante LFFS empieza de un conjunto vacio. Tercero, las características seleccionadas para cada grupo se unen en un conjunto final. Los tres pasos en este esquema de selección por grupos de características son repetidos para cada primitiva emocional.

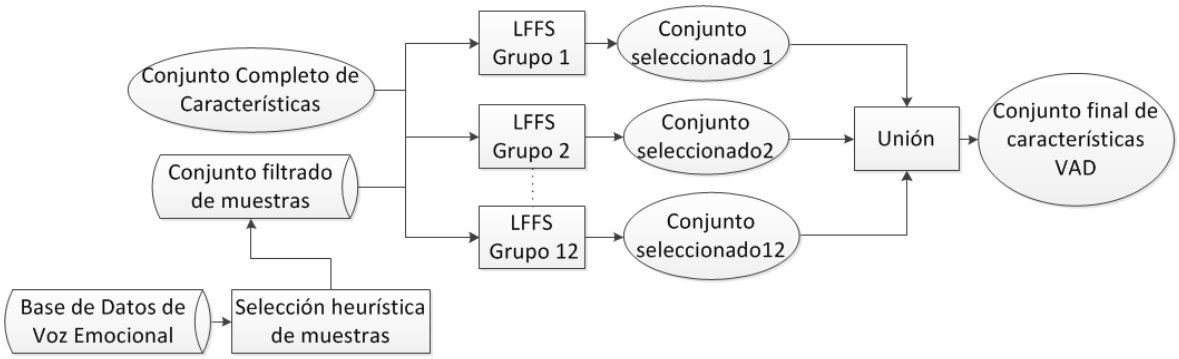


Figura 11 Esquema de selección por grupos de características

5.3 Resultados de selección de características

Todos los resultados de los experimentos de aprendizaje en este capítulo fueron obtenidos usando máquinas de vectores de soporte y validados mediante validación cruzada de diez pliegues. Se eligió este clasificador tras evaluar el desempeño de varios clasificadores. Las métricas usadas para medir la importancia de los grupos de características son coeficiente de correlación Pearson, Share y Portion explicados en la sección *2.2.1 Métricas de medición de calidad de características acústicas*. Una vez identificado los mejores conjuntos de características acústicas construimos clasificadores individuales para estimar cada primitiva emocional. La Tabla 12 muestra los resultados de evaluación del esquema de selección de muestras y atributos. Ilustrado en la Figura 10. La segunda columna muestra los resultados cuando se usan todos los atributos y todas las muestras en el proceso de aprendizaje. Como podemos ver el coeficiente de correlación es muy bajo. La tercera columna muestra los resultados cuando el proceso de aprendizaje es realizado usando las mejores características, propuestas en (Pérez Espinosa, Reyes García,

& Villaseñor Pineda, 2010) para cada primitiva. La cuarta columna muestra los resultados cuando el proceso de aprendizaje usa las mismas características que el experimento previo, pero con muestras filtradas por el proceso descrito en la sección 2.3 *Selección de muestras mediante auto-entrenamiento*. Finalmente, la quinta columna muestra los resultados después de filtrar las muestras y de aplicar el método de selección de características LFFS. Como se puede ver la mejora en los resultados fue gradual después de aplicar cada paso del proceso.

Tabla 12 Resultados para cada paso del esquema de selección no agrupada de características. Cada columna muestra el número de características / coeficiente de correlación

	Resultados de Referencia		Resultados con Esquema 1	
Primitiva Emocional	**Todos los Atributos**	**Selección Inicial**	**Selección Muestras**	**Selección LFFS**
Valencia	6920 / 0.0151	56 / 0.5188	56 / 0.5597	62 / 0.6189
Activación	6920 / 0.0095	67 / 0.7463	67 / 0.7861	60 / 0.7969
Dominación	6920 / -0.001	23 / 0.6536	23 / 0.7117	31 / 0.7437

Los dos esquemas de selección de características fueron aplicados para cada una de las tres primitivas emocionales, esto es, se ejecutó seis veces el proceso de selección de características, obteniendo seis diferentes conjuntos de características. Esta sección muestra los resultados de los mejores subconjuntos para cada primitiva. La Figura 12, Figura 13, y Figura 14 reflejan la efectividad de cada conjunto de características y las diferencias entre grupos. Los grupos POV y SC no se muestran en estas gráficas porque ninguna característica fue seleccionada de estos grupos. Es importante notar que la Correlación, Share y Portion mostrados en estas gráficas son obtenidos descomponiendo en grupos las características del conjunto solución encontrado por los esquemas de selección uno y dos y evaluándolos separadamente con estas métricas.

Mientras el Esquema 2: selección agrupada de características, asegura que por lo menos una característica de cada grupo será incluida, el Esquema 1: selección no agrupada de características puede no incluir ningún elemento de ciertos grupos en el conjunto solución.

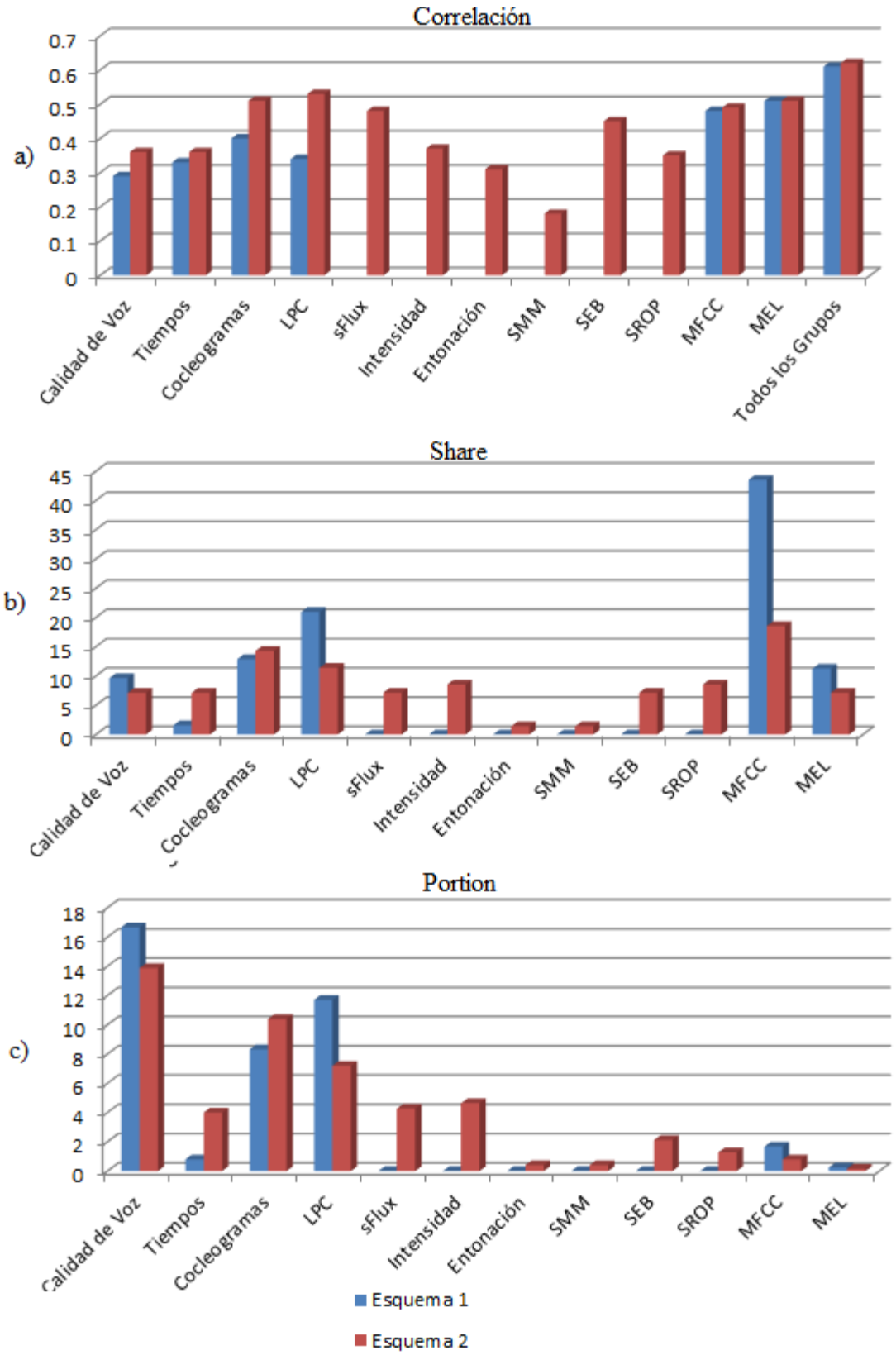


Figura 12 Esquema 1 / Esquema 2 resultados de selección de atributos para Valencia

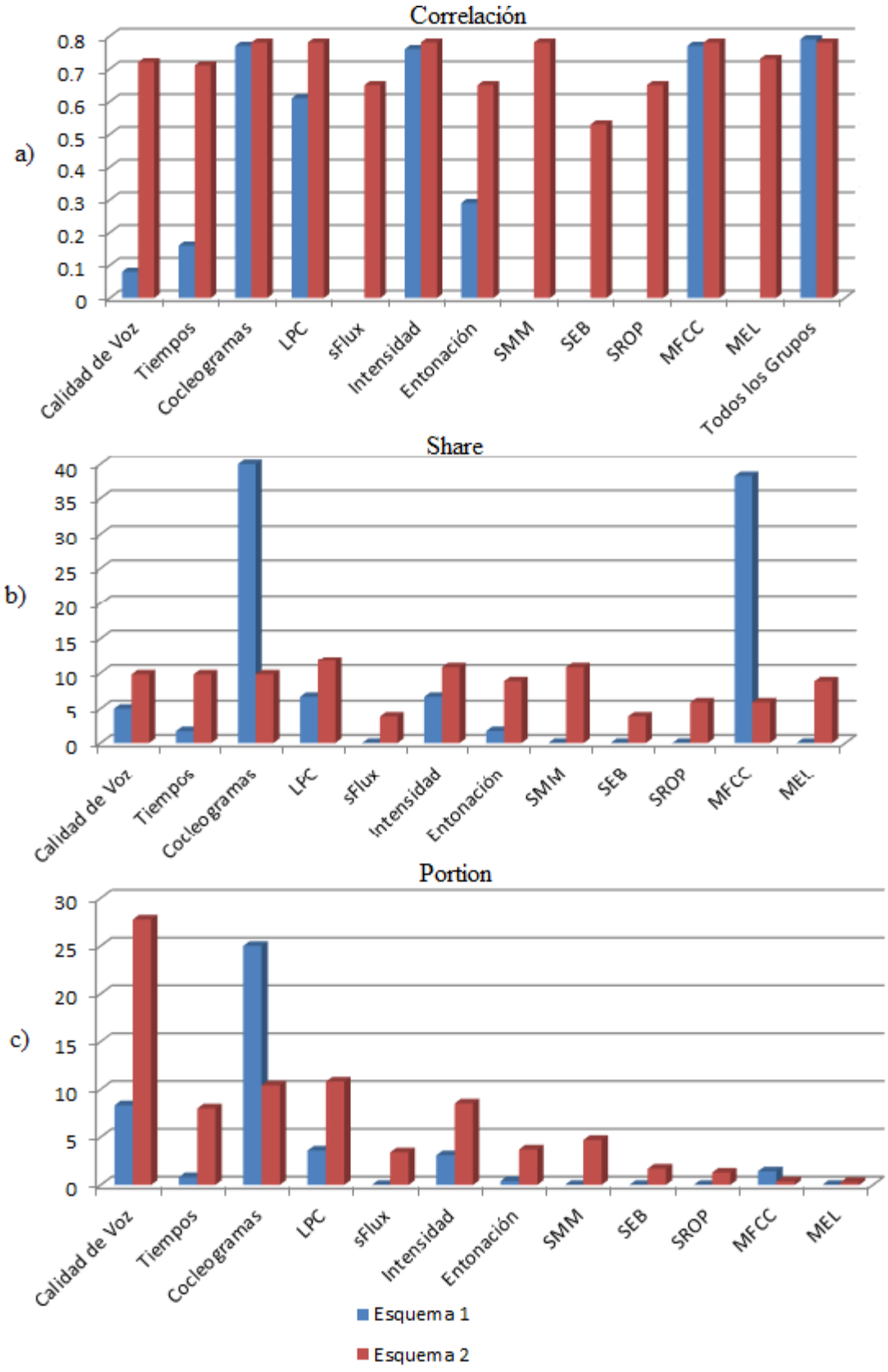


Figura 13 Esquema 1 / Esquema 2 resultados de selección de atributos para Activación

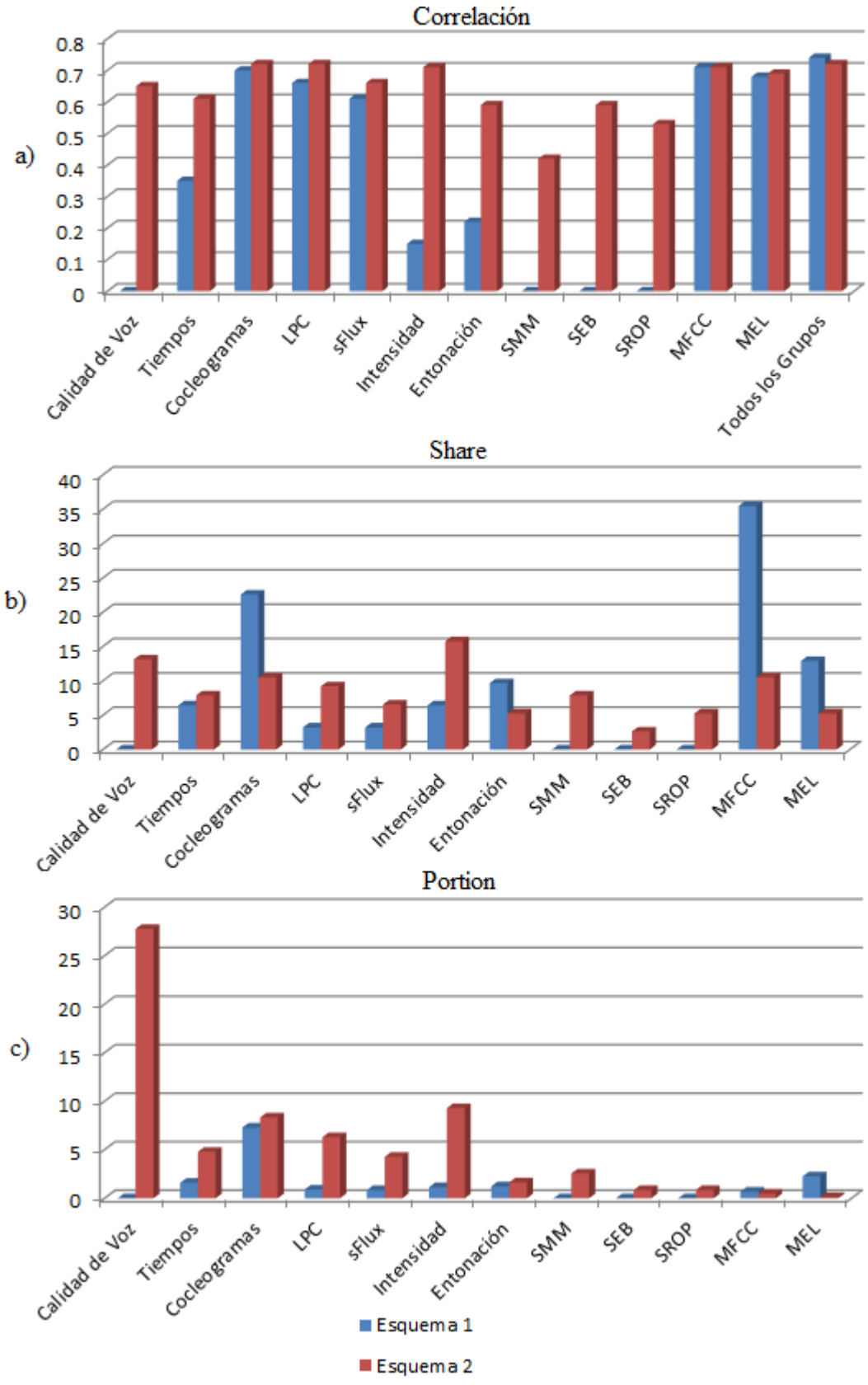


Figura 14 Esquema 1 / Esquema 2 resultados de selección de atributos para Dominación

En los experimentos para Valencia con el Esquema 1 de selección, el grupo con mayor correlación fue MEL (0.5167), contribuyendo al conjunto solución con siete de 62 características, su Portion es muy bajo (0.230), con relación al de otros grupos, ya que hay un total de 3,042 características perteneciendo a este grupo, y su Share podría ser considerado medio (11.29). Estos valores de Share y Portion indican que pocos coeficientes MEL proveen información muy importante. Otro grupo importante fue MFCC con una correlación de 0.4877, indicando que los grupos de información espectral son importantes para estimar Valencia.

En el experimento para Valencia con el Esquema 2 de selección, los mejores grupos fueron LPC (0.5349), Cocleogramas (0.5152) y MEL (0.5129). LPC y Cocleogramas mostraron un Share (11.429 y 14.284) y Portion (7.207 y 10.417) similares, mientras que MEL mostró un Share (7.143) y Portion (0.164) más bajo en comparación con los dos grupos mencionados. Como en el esquema uno, MEL provee pocas características, pero muy importantes. Nos dimos cuenta que para Valencia los grupos de tipo espectral fueron los mejores para los Esquemas 1 y 2. Ningún grupo por sí mismo alcanzó la correlación obtenida por todos los grupos juntos. Podemos ver que los ocho atributos seleccionados en el esquema dos para LPC fueron mucho mejor (0.535) que los 13 atributos seleccionados en el esquema uno para el mismo grupo (0.342). La mejor correlación para Valencia fue obtenida usando el esquema dos alcanzando 0.6232.

Se observó acuerdo en la proporción de características elegidas de cada grupo por cada uno de los dos esquemas de selección. A pesar de que el Esquema 2 asegura incluir características de todos los grupos al final la diferencia entre el número de características seleccionadas por ambos esquemas no fue muy grande, el Esquema 1 conservó 62 y el Esquema 2 conservó 70.

Para Activación el grupo MFCC tiene la mayor correlación en ambos esquemas de selección. MFCC obtuvo 0.7795 en el esquema uno y 0.7897 en el esquema dos. Podemos ver que su Share (38.333 y 5.882) y Portion (1.422 y 0.371) son muy diferentes. Podemos ver además que usando sólo el grupo MFCC se puede obtener un resultado similar al obtenido usando todos los grupos (0.7964 y 0.7870). Estos resultados claramente indican la importancia de este grupo para estimar Activación. Como se esperaba, la Intensidad de la señal fue muy importante para esta primitiva con una correlación de 0.7851 en el esquema dos y 0.7831 en el esquema uno. Se espera que la gente que experimenta un incremento en la actividad o excitación muestre un volumen más alto en su voz.

Otros grupos importantes de características para Activación son Cocleogramas y LPC. Un efecto interesante es que el grupo SpecMaxMin fue importante en el esquema dos con una correlación de 0.7831, con un Share alto, pero en el esquema uno, no fue seleccionada ninguna característica de este grupo.

Se esperaba que el grupo Tiempos fuera importante para Activación ya que, se ha encontrado que entre más rápido se habla se da la impresión de estar más excitado y entre más lento se habla se da la impresión de estar más relajado (Kehrein, 2002). Solamente una característica de Tiempos fue seleccionada en el esquema uno; su correlación fue muy baja (0.167), mientras que en el esquema dos, se seleccionaron diez características obteniendo una correlación relativamente alta (0.7128). En el esquema dos los mejores resultados no fueron obtenidos usando todos los atributos seleccionados (0.787) sino usando solamente los seleccionados mediante el enfoque de fuerza bruta (0.7952). El mejor resultado para Activación fue obtenido usando todos los grupos en el esquema uno (0.7964). Podemos inferir de estos resultados que los grupos más importantes para Activación son MFCC, Cocleogramas e Intensidad.

A diferencia de lo sucedido con Valencia donde el número de características seleccionadas por ambos esquemas es similar, tanto para Activación como para Dominación hubo mayor diferencia. Para Activación el Esquema 1 conservó 62 y el Esquena 2 conservó 102. Para Dominación el Esquema 1 conservó 31 y el Esquena 2 conservó 76. Esto se puede apreciar en las gráficas de Share de estas primitivas donde el Esquema 1 conservó muchas más características de MFCCs y Cocleogramas

En los experimentos realizados con Dominación en el esquema uno los mejores grupos fueron MFCC (0.72) y Cocleogramas (0.702) y en el esquema dos LPC (0.7266) y Cocleogramas (0.7244). En el esquema dos los mejores resultados fueron obtenidos usando sólo las características seleccionadas del grupo LPC (0.7266), mejorando los resultados obtenidos usando todos los grupos (0.7157). En el esquema dos los mejores resultados no fueron obtenidos usando todas las características seleccionadas (0.7157) sino solamente las seleccionadas por el enfoque selectivo (0.726). Los mejores resultados para Dominación fueron obtenidos con el esquema uno usando todos los grupos (0.7437). Podemos inferir que en el caso de Dominación los grupos más importantes son el Espectral, MFCC y Cocleogramas.

En conclusión, y como era de esperarse, el Esquema 1 elige menos características que el Esquema 2 lo que representa una ventaja ya que se ahorra tiempo al extraer características y su subsecuente clasificación. Por otro lado, en cuanto a la calidad de las características seleccionadas no hubo una diferencia clara entre ambos esquemas, ya que la correlación mostrada por ambos conjuntos de características fue similar.

5.4 Comparación de enfoques de extracción de características: selectivo y fuerza bruta

En la literatura se identifican dos enfoques para determinar qué tipo de características extraer de la señal de voz. Los primeros estudios en reconocimiento automático de emociones optaban por un enfoque de extracción de características selectivo. Esto quiere decir que se determinaba que características extraer basándose en el conocimiento de un experto, usualmente con un número pequeño de características que difícilmente rebasaban las cien características. Hoy en día, con el surgimiento de herramientas que permiten extraer un gran número de características y de la disponibilidad de más poder de cómputo es más fácil aplicar un enfoque por fuerza bruta. Uno de los objetivos de esta sección es comparar los resultados obtenidos mediante el enfoque selectivo de extracción de características (Pérez Espinosa & Reyes García, 2009), (Pérez Espinosa, Reyes García, & Villaseñor Pineda, 2010) y el enfoque por fuerza bruta.

La Tabla 13 muestra una comparación entre los enfoques de extracción de características (selectivo contra fuerza bruta) así como la comparación entre los esquemas de selección de características aquí propuestos. En todos los experimentos, ambos enfoques de extracción de características obtuvieron coeficientes de correlación muy similares con excepción del experimento hecho con el esquema uno para Valencia donde los resultados con fuerza bruta (0.5715) fueron mejores que con selectivo (0.4799). Sin embargo, el Share del enfoque selectivo (74.194) fue mucho más alto que el de fuerza bruta (25.806) y la correlación usando ambos grupos fue mayor (0.6189) que la obtenida para cada grupo por separado. Es muy difícil decir qué esquema de extracción de características es mejor ya que ambos esquemas obtuvieron resultados similares. Se comparó clasificando con las mejores características seleccionadas de cada enfoque de extracción por separado, generalmente se obtuvieron mejores resultados de clasificación al unir las características de ambos enfoques en un solo conjunto.

Tabla 13 Extracción de características selectiva vs fuerza bruta – E1 = Esquema 1, E2 = Esquema 2

Enfoque	Total	Seleccionadas E1 | E2		Correlación E1 | E2		Share E1 | E2		Portion E1 | E2	
Valencia									
Selectivo	368	46	27	0.48	0.56	74.19	38.57	12.50	7.34
Fuerza Bruta	6,552	16	43	0.57	0.55	25.81	61.43	0.24	0.68
Ambos	6,920	62	70	0.61	0.62	100	100	0.89	1.04
Activación									
Selectivo	368	56	37	0.79	0.79	93.33	36.28	15.22	10.05
Fuerza Bruta	6,552	4	65	0.76	0.80	6.67	63.72	0.06	1.03
Ambos	6,920	60	102	0.79	0.78	100	100	0.86	1.47
Dominación									
Selectivo	368	13	29	0.72	0.73	41.93	38.16	3.53	7.88
Fuerza Bruta	6,552	18	47	0.73	0.70	58.06	61.84	0.28	0.74
Ambos	6,920	31	76	0.74	0.72	100	100	0.44	1.09

Al combinar las características de ambos enfoques en la estimación de Valencia se consigue aumenta la correlación obtenida cuando se usan por separado. No sucede lo mismo con Activación y Dominación, ya que en esos casos la correlación no aumenta al combinar las características de ambos enfoques.

En conclusión, se obtuvo un desempeño similar para los enfoques de extracción de características por fuerza bruta y selectivo cuando se evaluaron por separado las características pertenecientes a cada enfoque. La utilidad de este hallazgo radica en primer lugar, en que se confirma que las características, elegidas para el enfoque selectivo, descritas en la sección *5.1 Extracción de características acústicas*, ciertamente aportan información importante para la discriminación de emociones. En segundo lugar, nos confirma que se puede aprovechar el poder de cómputo y herramientas de procesamiento de señales actuales para explorar un espacio amplio de propiedades acústicas para encontrar un conjunto confiable de descriptores emocionales. Estos resultados nos sugieren combinar las características de ambos enfoques para estimar Valencia y usar el enfoque selectivo para Activación y Dominación dado que se obtienen los mismos resultados que con el de fuerza bruta pero con un proceso, de extracción y selección de características menos costoso.

5.5 Análisis multilingüe de características

En esta sección estudiamos la importancia de las características dividiéndolas en grupos y trabajando con tres bases de datos una en inglés, otra en alemán y otra en español construida por nosotros mismos para analizar la importancia multilingüe de características, así como saber si estas características tienen diferente importancia para cada idioma.

Como se mencionó en la introducción de este trabajo, algunos autores sugieren que las emociones son independientes de la cultura. Sin embargo, hay un fuerte debate en este punto entre psicólogos quienes dicen que las emociones son universales y quienes dicen que las emociones son dependientes de la cultura (Elfenbein, 2002). Ambos grupos de científicos han aportado evidencia de diferencias y similitudes entre la manera que diferentes culturas expresen las emociones. De hecho, algunos autores han definido expresiones emocionales faciales universales como el psicólogo Paul Ekman. Izard (Elfenbein, 2002) aporta evidencia de que para ciertas culturas reconocer emociones en expresiones faciales de gente de otras culturas es más difícil que hacerlo para gente de su propia cultura. Picard (Picard, 2000) estableció que los patrones expresivos dependen del género, contexto social, expectativas culturales y sociales.

Dado que una emoción en particular es sentida, una variedad de factores influencian la manera en como esa emoción es desplegada. Varios autores han trabajado en el análisis de las características acústicas más importantes desde el punto de vista de la categorización discreta trabajando de manera monolingüe (Xie, Chen, Chen, & Chen, 2005), (Lugger & Yang, 2007), (Batliner A. , et al., 2011) y multilingüe (Polzehl T. a., 2010). Sin embargo, no han sido estudiados con el mismo nivel de profundidad los atributos acústicos desde el punto de vista continuo.

En esta sección estamos interesados en analizar si existen características acústicas que nos permitan estimar el estado emocional de la voz de una persona sin importar el idioma que hable. También discutimos la importancia de estas características, la cantidad de información que proveen y cuáles son más importantes para cada idioma. Para llevarlo a cabo, trabajamos con tres bases de datos de habla emocional, una en inglés, otra en alemán y otra en español.

Extrajimos una variedad de características acústicas y aplicamos técnicas de selección de atributos para encontrar las mejores características de manera monolingüe y multilingüe. Finalmente discutimos por separado cada grupo de características, usando métricas que nos dan una idea de la importancia de cada grupo.

Para comparar el desempeño multilingüe de características acústicas desde el punto de vista de los modelos emocionales continuos usamos tres bases de datos en diferentes idiomas etiquetados con las mismas primitivas emocionales. Las bases de datos que usamos fueron IEMOCAP (Inglés), VAM (Alemán) y EmoWisconsin (Español). En el caso del corpus VAM los valores anotados fueron normalizados a valores continuos entre uno y cinco. Originalmente las primitivas estaban en el rango de -1 a 1 mientras que en IEMOCAP el rango era de 1 a 5.

Todos los resultados de los experimentos de aprendizaje fueron obtenidos usando máquinas de vectores de soporte para regresión y validados mediante validación cruzada de diez pliegues. Las métricas usadas para medir la importancia de cada grupo de características son el coeficiente de correlación de Pearson, Share y Portion.

Los resultados en la Figura 15, Figura 16, Figura 17 están representados en el formato: Resultados para Inglés / Resultados para Alemán / Resultados para Español. Los resultados monolingües para inglés, alemán y español fueron obtenidos entrenando SMOreg con las características seleccionadas para cada idioma y cada primitiva por separado. La evaluación fue hecha mediante validación cruzada de diez pliegues. Los resultados multilingües mostrados en la Figura 18, Figura 19, y Figura 20, fueron obtenidos construyendo clasificadores para cada primitiva. Se muestran los resultados obtenidos usando las muestras de dos idiomas, inglés y alemán, y de los tres idiomas, inglés, alemán y español y el conjunto de características obtenido de la selección de características multilingüe.

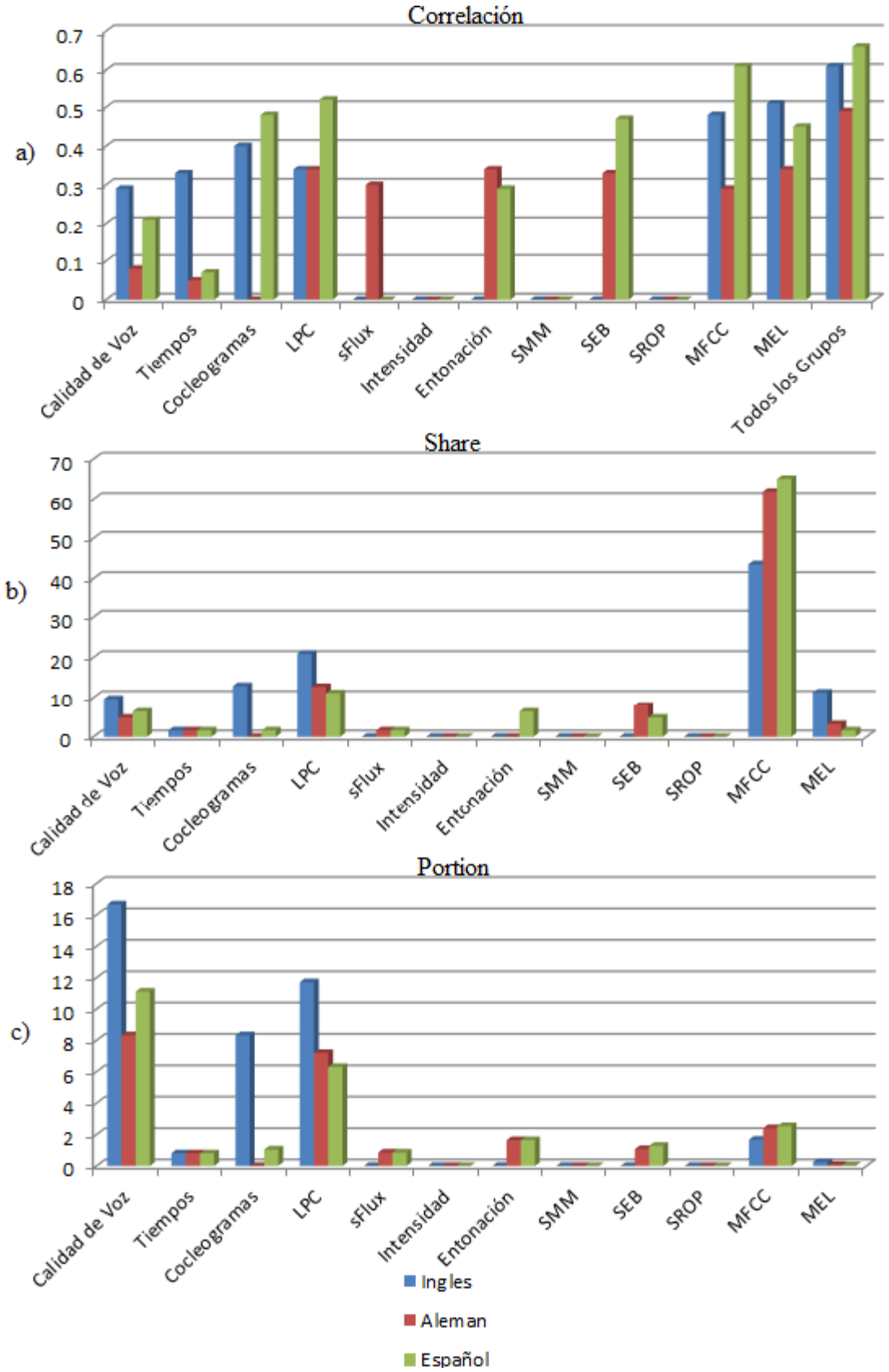


Figura 15 Inglés / Alemán / Español - Valencia

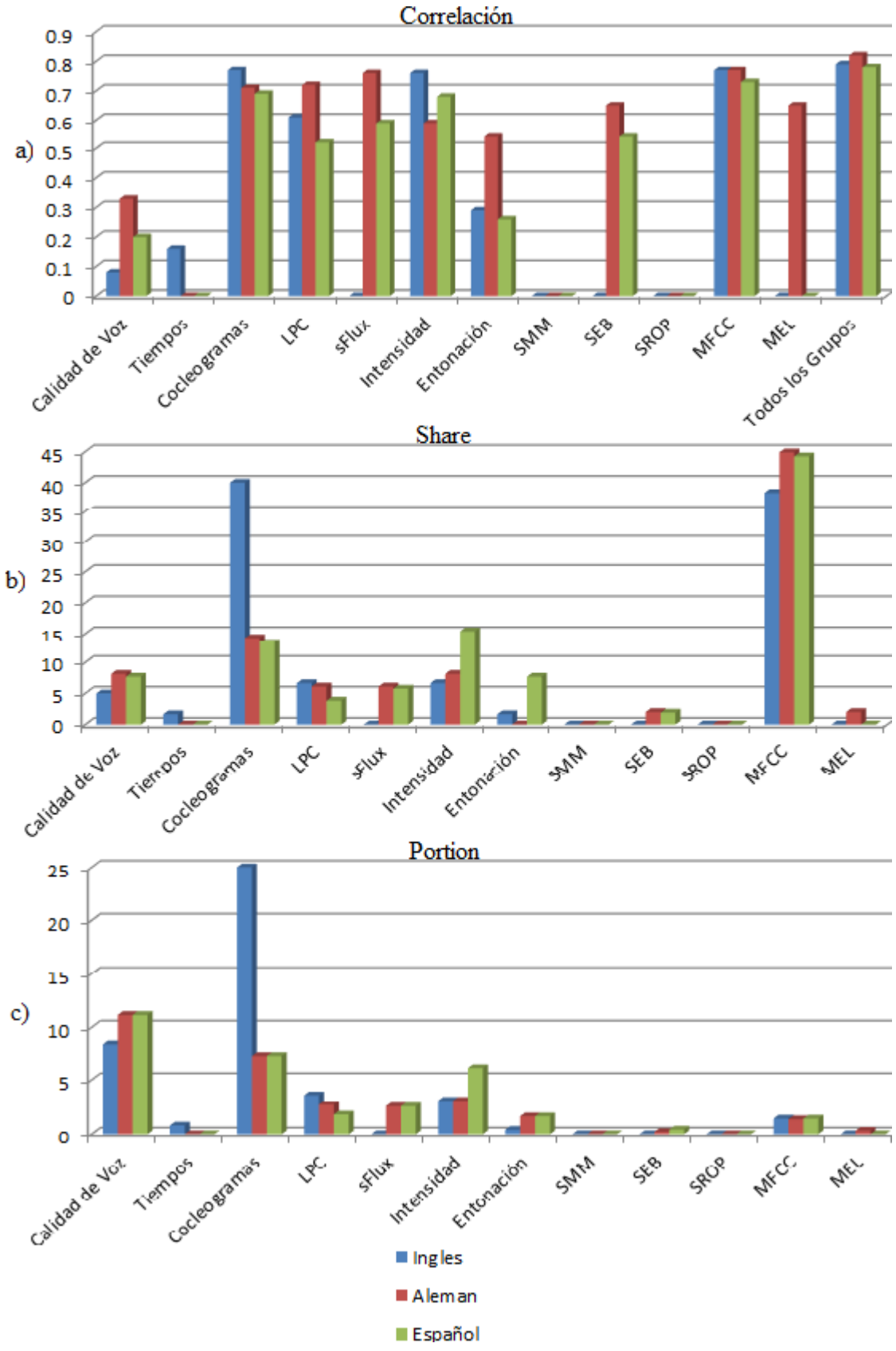


Figura 16 Inglés / Alemán / Español - Activación

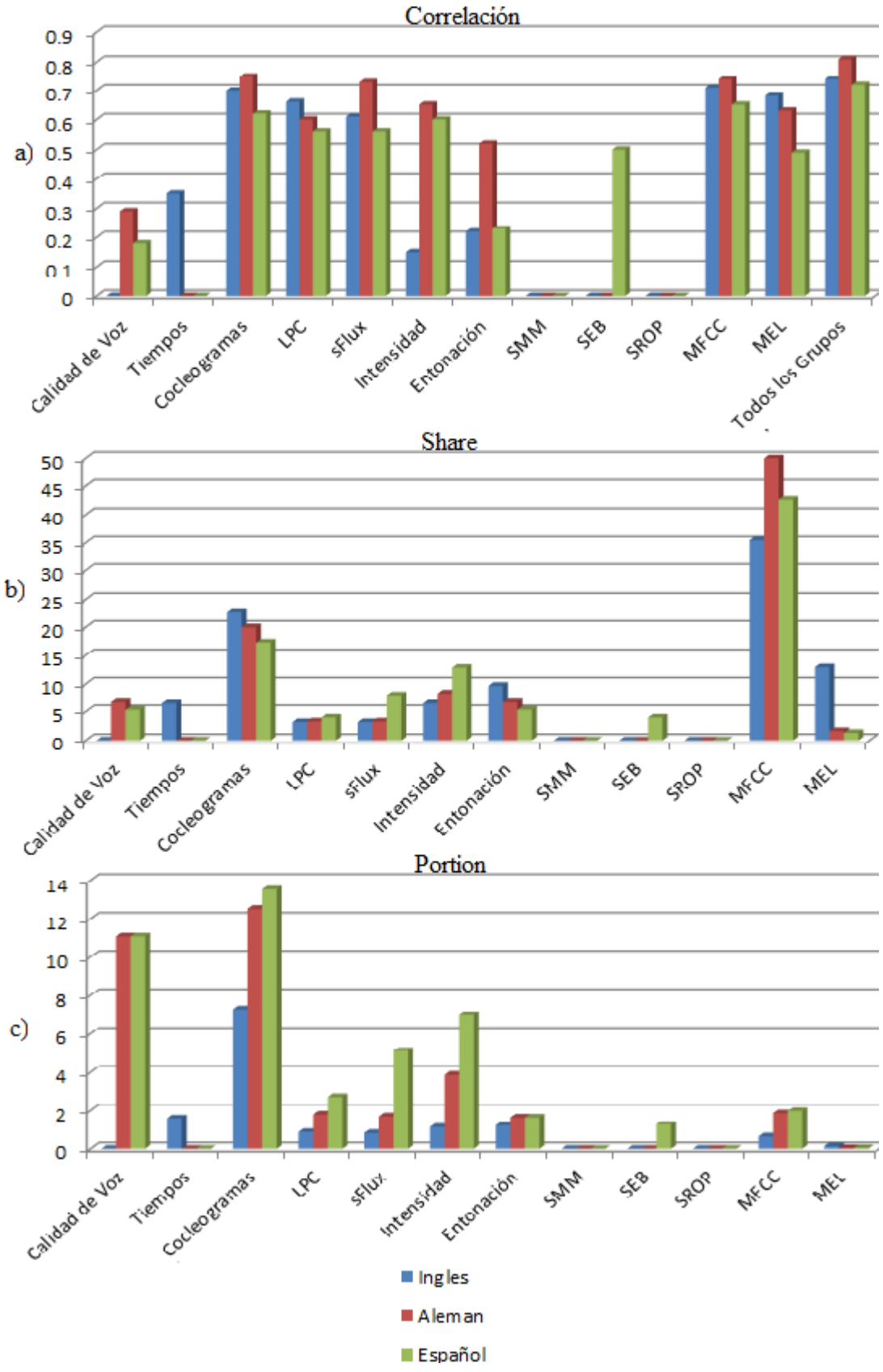


Figura 17 Inglés / Alemán / Español - Dominación

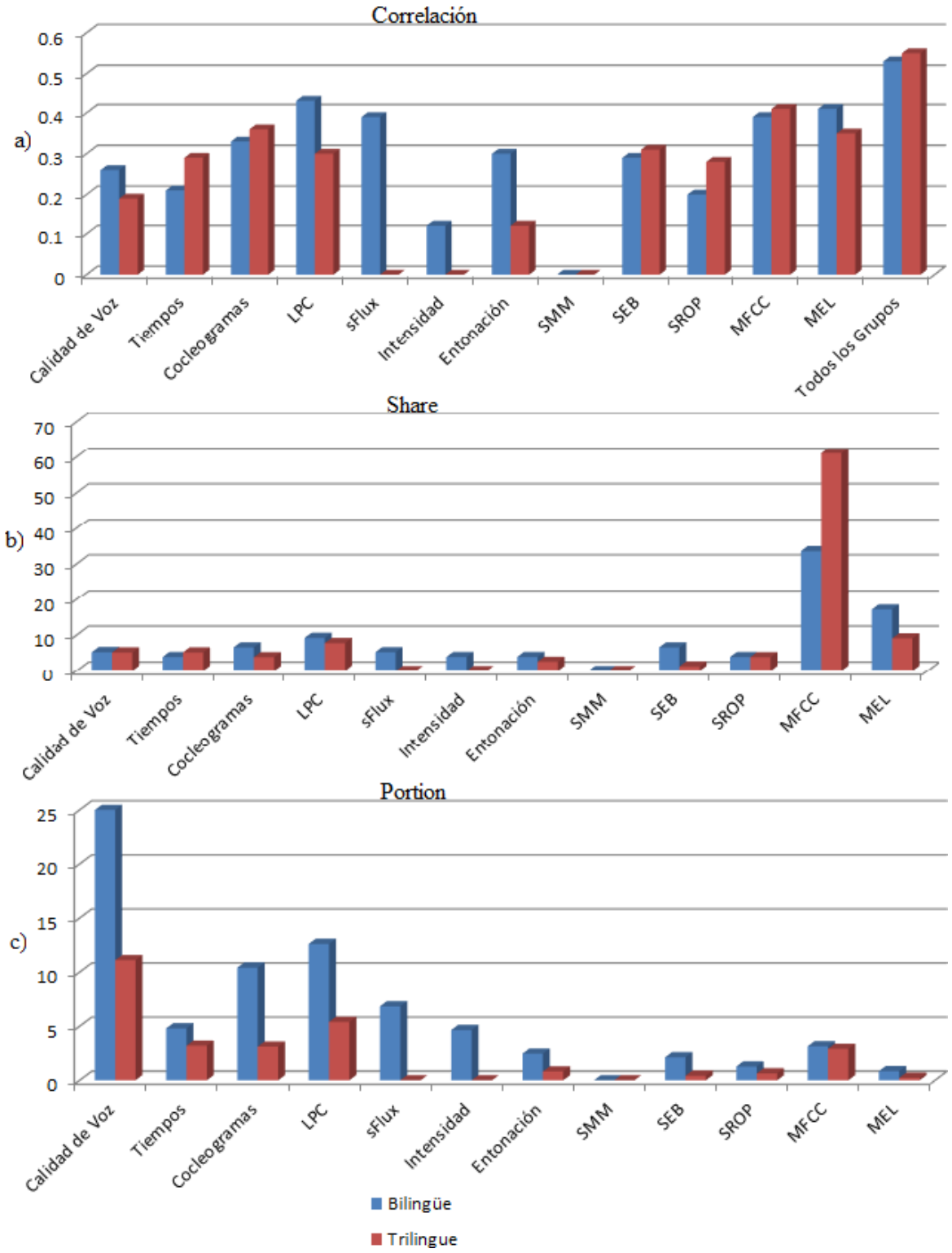


Figura 18 Bilingüe / Multilingüe - Valencia

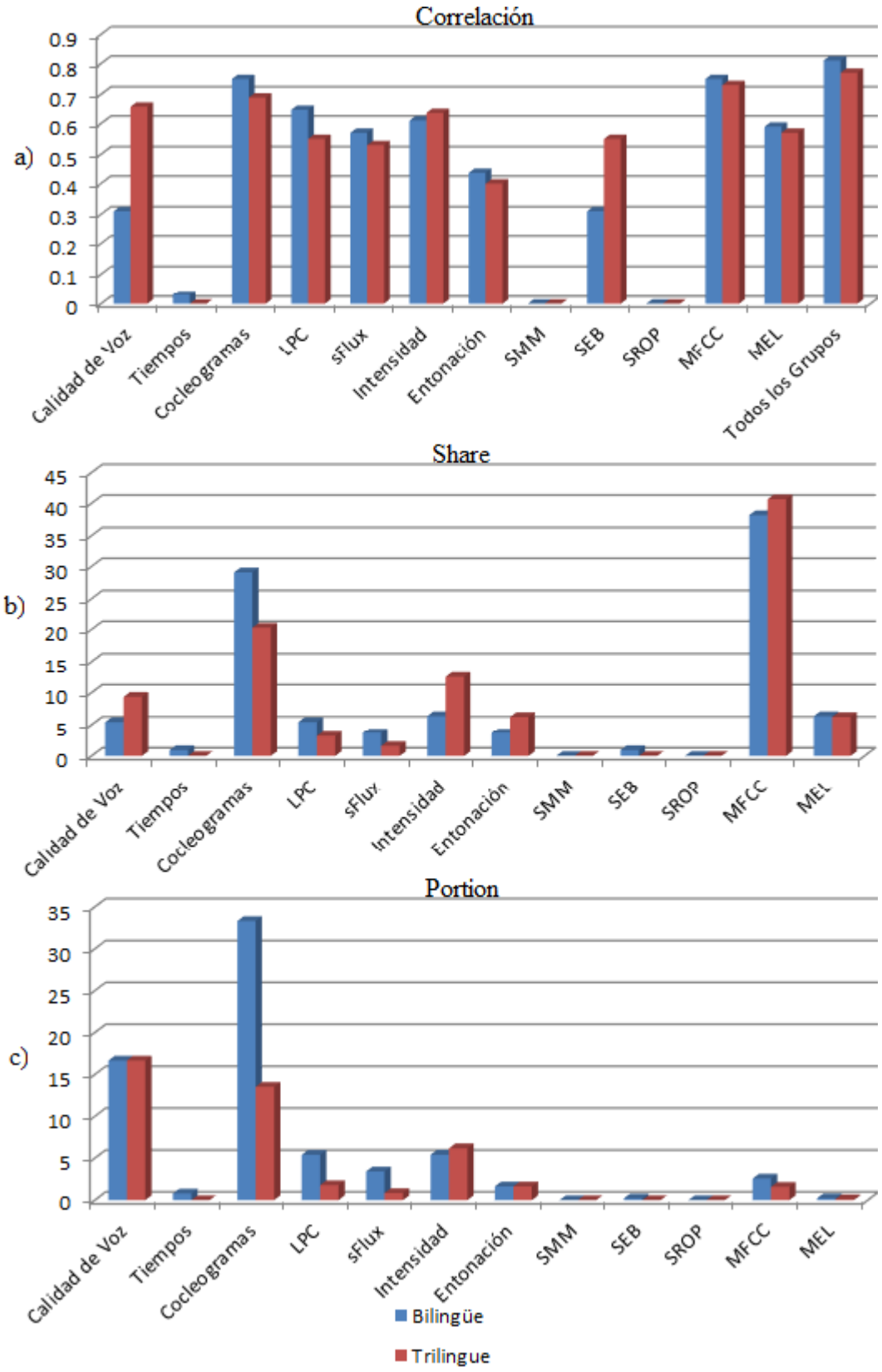


Figura 19 Bilingüe / Multilingüe - Activación

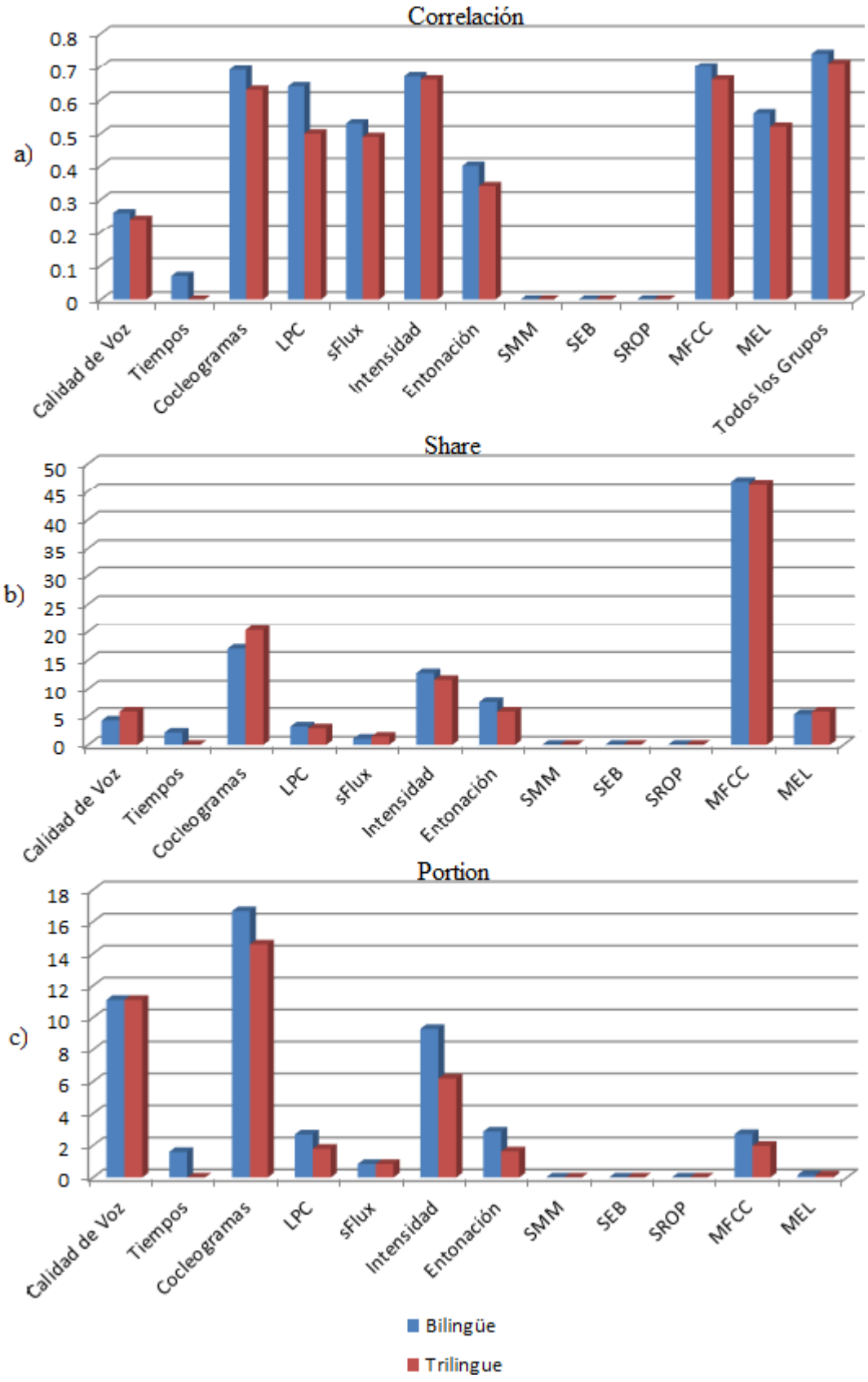


Figura 20 Bilingüe / Multilingüe - Dominación

5.5.1 Resultados de selección de características monolingüe

Los resultados en los tres idiomas coincidieron al indicar el nivel de dificultad que presenta la estimación de cada primitiva. La primitiva con mejores resultados fue Activación con una correlación de 0.80 promediando los tres idiomas, Dominación presentó una correlación promedio de 0.76 y Valencia 0.59. El promedio de la correlación de las tres primitivas es muy similar en los tres idiomas inglés 0.71, alemán 0.71 y español 0.72, por lo que podemos decir que los tres idiomas muestran el mismo nivel de dificultad. En cuanto a las características seleccionadas para cada idioma, de entre los doce grupos analizados, algunos grupos fueron importantes para todos los idiomas y todas las primitivas, como MFCC y LPC.

Para Valencia, obtuvimos mejores resultados en español, alcanzando una correlación de 0.66. En inglés, MEL tuvo el mejor desempeño (0.51). En español MFCC (0.61) y en alemán tres grupos alcanzaron una correlación de 0.34, MEL, Entonación, y LPC. Podemos ver que mientras MEL, LPC y MFCC son importantes para los tres idiomas, Entonación y SEB fueron importantes sólo para alemán y español. Por otro lado sFlux sólo fue importante para alemán. Un aspecto llamativo es que el grupo MEL alcanzó una alta correlación para los tres idiomas, a pesar de tener un Portion muy bajo (0.24/0.06/0.03), es decir con muy pocas características seleccionadas (7/2/1) con respecto al total de características en ese grupo (3,042). En conclusión, podemos decir que en los tres idiomas las propiedades espectrales y en particular los MFCCs son importantes para determinar si una emoción es positiva o negativa. Dados que los MFCCs están ligados estrechamente con la caracterización fonética del habla, podríamos suponer que en los tres idiomas la dicción podría ser alterada al experimentar emociones negativas o positivas.

En el caso de Valencia, es difícil inferir intuitivamente que características prosódicas fueron relacionadas a las emociones negativas y positivas. Por ejemplo, podríamos pensar de emociones negativas como valores opuestos de intensidad, como enojado, que tiene una intensidad alta y triste que tiene una intensidad baja. En emociones positivas con valores opuestos de tiempos, como excitado, en la cual se tiende a hablar más rápidamente y relajado, en la cual se tiende a hablar más lentamente.

Por otro lado, para Activación es razonable pensar que entre más rápido y fuerte hablamos más activos parecemos estar y que entre más lento y bajo hablemos más pasivos parecemos estar (Kehrein, 2002). Por lo tanto, intuitivamente, grupos de características que modelen aspectos prosódicos tales como Intensidad y Tiempos deberían ser más importantes para estimar Activación. En los experimentos podemos confirmar que efectivamente la Intensidad fue importante en los tres idiomas (0.76/0.59/0.68). Sin embargo, Tiempos no dio información valiosa para estimar esta primitiva (0.16/0/0). La Figura 16 a) muestra claramente esta diferencia entre la relevancia de Intensidad y Tiempos y también en relación con los otros grupos de características. Esto nos hace dudar de que las características Tiempos usadas aquí están reflejando adecuadamente los fenómenos relacionados a la velocidad para los idiomas con los que estamos trabajando. Para Activación el mejor grupo fue MFCC (0.77/0.77/0.73), con un porcentaje alto de las características totales en el conjunto final (38.33/44.89/44.23). Podemos ver que este grupo por sí mismo obtiene un desempeño similar al desempeño del conjunto solución (0.79/0.82/0.78). Otros grupos importantes para Activación son Cocleogramas que obtuvieron buenos resultados para los tres idiomas (0.77/0.71/0.69), LPC (0.61/0.72/0.69). Como se esperaba, Intensidad también mostró relevancia (0.76/0.59/0.68). Esta primitiva no mostró diferencias significativas cuando se estimó en diferentes idiomas (0.79/0.82/0.78).

Al igual que para Activación, para Dominación los mejores grupos fueron MFCC y Cocleogramas. Esta primitiva muestra muchas similitudes entre los tres idiomas. Casi todos los grupos en los tres idiomas coincidieron en el grado de importancia. El grupo Intensidad que intuitivamente podría ser bueno para indicar que la persona trata de controlar la situación mostró ser no importante para inglés (0.15), pero bueno para alemán y español (0.65/0.60). Otros grupos importantes para esta primitiva son LPC (0.66/0.90/0.56), sFlux (0.61/0.73/0.56) y MEL (0.68/0.630.49). La diferencia en el número total de características seleccionadas fue considerable, en ingles se seleccionaron 31, en alemán 60 y en español 75. Se mantuvo la proporción de características seleccionadas de cada grupo reflejándose en el Share y Portion.

En conclusión no se observó una tendencia a obtener mejores resultados con alguno de los tres idiomas. Con español se obtuvo la mejor correlación en la estimación de Valencia, mientras que la peor fue obtenida con alemán. Esto puede deberse al ambiente en el que fueron obtenidos los datos. En la generación de los datos en español se indujeron emociones positivas y negativas, por otro lado, está documentado que la base de datos en alemán tiene mayor representación de emociones negativas. En contraparte, en el caso de

la estimación de Activación y Dominación se obtuvieron los mejores resultados con los datos en alemán y lo peores con la base de datos en español. Esto puede deberse al tipo de emociones presentes en cada base de datos, mientras que en alemán el enojo expresado es muy efusivo, en los datos en español las emociones son más mesuradas.

5.5.2 Resultados de selección multilingüe de características

En general el desempeño de los modelos de regresión disminuye cuando los datos incluyen datos en diferentes idiomas. Esto nos puede indicar que la expresión emocional tiene ciertas particularidades en cada idioma lo cual dificulta el descubrimiento de patrones acústicos que se ajusten de manera multilingüe. Sin embargo, hubo una excepción a este comportamiento, ya que el modelo creado con datos multilingües para la estimación de Valencia que obtuvo mejores resultados (0.55) que el modelo monolingüe en alemán (0.49). Los modelos más afectados son el modelo de estimación de Valencia para español que disminuye de 0.66 de manera monolingüe a 0.55 de manera multilingüe y el modelo de estimación de Dominación en alemán que disminuye de 0.81 a 0.71.

En el caso de Valencia, la mejor correlación fue obtenida con MFCC y Cocleagramas. Podemos ver que cuando estimamos Valencia de manera multilingüe, la correlación (0.55) fue menor a la correlación en inglés (0.61) y español (0.66) y mayor a la correlación en alemán (0.49). La selección multilingüe añadió grupos que no habían sido considerados de manera monolingüe tales como Intensidad y SROP.

Para el caso de Activación, la mejor correlación para un grupo fue obtenida por MFCC (0.73), Cocleogramas (0.69), Calidad de Voz (0.66) e Intensidad (0.64), similarmente a lo que había sucedido para inglés podemos ver que la correlación cuando estimamos Activación en el conjunto multilingüe fue buena (0.77) comparado con el resultado tenido estimando Activación en los tres idiomas por separado (0.79/0.82/0.78). Para Dominación el mejor el grupo fue MFCC (0.66), Cocleogramas (0.63) e Intensidad (0.66). Estos resultados coinciden con la selección monolingüe. La correlación cuando se estima Dominación multilingüe (0.71) fue menor que la estimación en inglés (0.74), alemán (0.81) y español 0.72.

En la Figura 18, Figura 19, y Figura 20 se hace una comparación del desempeño de modelos entrenados con un conjunto bilingüe de datos (alemán e inglés) contra modelos entrenados con un conjunto trilingüe de datos (alemán, inglés y español). En el caso de

estimación de Valencia se obtuvieron mejores resultados usando los datos trilingües, en el caso de Activación y Dominación se obtuvieron mejores resultados con los datos bilingües.

En conclusión podemos decir que es mejor entrenar modelos exclusivos para cada idioma. Sin embargo, en aplicaciones que así se lo requieran, es posible entrenar modelos de manera multilingüe que tengan un desempeño aceptable para reconocer emociones en los idiomas en que fueron entrenados. Los resultados mostrados en la estimación de Valencia nos llevan a la conclusión un elemento clave es la diversidad emocional en los datos de entrenamiento. La correlación en la estimación de Valencia en alemán es 0.49 al mezclar alemán e inglés pasa a 0.53 y al mezclarla también con español 0.55. De acuerdo a un trabajo (Wöllmer M. , Eyben, Schuller, Douglas-Cowie, & Cowie, 2009) realizado con el corpus en Aleman (VAM corpus) los valores de Valencia son bajos, es decir, la mayor parte de las muestras en dicho corpus son de emociones negativas, mientras que Activación y Dominación tienen una mejor distribución en el espacio. Al mezclar las muestras del corpus en alemán con emociones positivas como las contenidas en los corpus de inglés y español parece mejorar la precisión en la estimación de Valencia. Por otro lado, la estimación de Activación y Dominación parece dificultarse con la combinación de idiomas y es en los datos en alemán donde se tiene mejor desempeño para la estimación de estas primitivas.

5.5.3 Desempeño interlingüe

Experimento 1

En este experimento realizamos una selección de características de manera interlingüe. Esto es, usamos las características encontradas en un idioma para estimar las primitivas emocionales en los otros dos idiomas. Para evaluar este experimento realizamos validación cruzada de diez pliegues usando sólo datos de un idioma a la vez. La idea de este escenario es analizar si las características que fueron buenas para un idioma en particular también lo son para otro. Por ejemplo, en la gráfica de barras de la Figura 15 a) vemos que la correlación para Valencia en alemán usando todas las características encontradas con datos en alemán es 0.49. Por otro lado, en la Tabla 14 vemos que cuando usamos las características encontradas con datos en inglés la correlación disminuye grandemente hasta 0.33. Para Activación y Dominación, la correlación en esa misma combinación disminuye de 0.82 (ver Figura 16 a) y 0.81 (ver Figura 17 a) hasta 0.78 y 0.77 respectivamente, ver Tabla 14, que es una disminución menor. En todos los casos, la

correlación en la estimación de primitivas disminuye aun cuando, en algunos casos es mucho más notoria la diferencia.

Este experimento deja en evidencia que es más difícil estimar Valencia con los datos en alemán y menos difícil hacerlo con los datos en español. Los resultados obtenidos al entrenar y probar Valencia en español haciendo la selección en inglés (0.63) y alemán (0.61) es mejor que haciendo la prueba, entrenamiento y selección en alemán (0.49). Estos experimentos también evidencian la dificultad de estimar en general Valencia a partir de la voz en relación con Activación y Dominación.

En conclusión podemos decir que las características acústicas más importantes para estimar primitivas emocionales en un idioma también aportan información valiosa para estimar primitivas emocionales en otro.

Tabla 14 Índice de correlación obtenido en estimación monolingüe de primitivas para selección interlingue de características

Selección	Entrenamiento	Prueba	Valencia	Activación	Dominación
Inglés	**Alemán**	**Alemán**	0.3372	0.7867	0.7779
Inglés	**Español**	**Español**	**0.6389**	0.7456	0.6724
Alemán	**Inglés**	**Inglés**	0.5163	0.7678	0.7080
Alemán	**Español**	**Español**	0.6109	0.7709	0.7030
Español	**Inglés**	**Inglés**	0.5382	0.7567	0.7008
Español	**Alemán**	**Alemán**	0.3991	**0.8178**	**0.7997**

Experimento 2

En este experimento se hizo una selección de características de manera multilingüe, es decir, se aplicó el proceso de selección de atributos explicado en la sección *5.2.1 Selección no agrupada de características* al conjunto de muestras formada por la unión de los datos en los tres idiomas. Después, se extrajeron las características obtenidas de manera multilingüe y se entrenaron modelos con las muestras de un idioma y evaluando con las muestras de otro. La idea en este escenario es examinar si los patrones aprendidos para estimar las primitivas emocionales en un idioma pueden ser usados para estimar las primitivas emocionales en otro idioma. En este caso podemos ver en la Tabla 15 que disminuye mucho la exactitud de las estimaciones y que esta tarea es difícil, especialmente cuando usamos los patrones aprendidos en inglés para estimar primitivas emocionales en alemán y español, en sentido inverso, de alemán a inglés el deterioro no es tan evidente.

Tabla 15 Índice de correlación obtenido en estimación interlingue de primitivas para selección multilingüe de características

Selección	Entrenamiento	Prueba	Valencia	Activación	Dominación
Multilingüe	**Inglés**	**Alemán**	0.1292	0.5459	**0.7867**
Multilingüe	**Inglés**	**Español**	-0.0421	0.4126	0.2689
Multilingüe	**Alemán**	**Inglés**	**0.3938**	0.7105	0.6949
Multilingüe	**Alemán**	**Español**	-0.3483	0.4902	0.5200
Multilingüe	**Español**	**Inglés**	-0.2159	0.7554	0.6959
Multilingüe	**Español**	**Alemán**	-0.1933	**0.7884**	0.6838

Tomando en cuenta los resultados del experimento anterior, que nos permite asumir que las características acústicas importantes en un idioma también son importantes en los otros, y tomando en cuenta también los resultados de este experimento, podemos concluir que las primitivas emocionales en diferentes idiomas pueden ser estimadas usando las mismas características acústicas sin embargo, los patrones formados por estas características sí muestran diferencias importantes en cada idioma. Por lo tanto, en la implementación de sistemas de clasificación emocional multilingüe se recomienda usar un mismo módulo de extracción de características pero modelos entrenados específicamente para cada idioma.

Experimento 3

Es este experimento realizamos una selección de características de manera multilingüe. Después, se realizó una validación cruzada de diez pliegues entrenando modelos y evaluándolos con datos de sólo un idioma a la vez. La idea aquí es analizar si las características encontradas en la selección multilingüe son complementarias de alguna manera cuando se estiman muestras de un idioma a la vez. Podemos ver que esto no es completamente cierto ya que aparentemente, hay características que nos ayudan a estimar primitivas emocionales en un idioma, pero de alguna manera afectan la estimación de dichas primitivas en otro idioma dado que disminuye la correlación entre los valores de primitivas estimados y etiquetados. Por ejemplo, Activación, de acuerdo a la Tabla 16, en este escenario es 0.75 en inglés, 0.78 en español y 0.81 en alemán, mientras que haciéndolo monolingüe, de acuerdo a la Figura 16, obtenemos 0.79 en ingles 0.78 en español y 0.82 en alemán. En este ejemplo en español y alemán se obtuvieron resultados muy parecidos seleccionando características de manera monolingüe y multilingüe pero, en inglés disminuyo la correlación.

Tabla 16 Índice de correlación obtenido en estimación mono-lingue de primitivas para selección multilingüe de características

Selección	Entrenamiento	Prueba	Valencia	Activación	Dominación
Multilingüe	**Inglés**	**Inglés**	0.5364	0.7547	0.7129
Multilingüe	**Español**	**Español**	**0.6958**	0.7873	0.7141
Multilingüe	**Alemán**	**Alemán**	0.0253	**0.8193**	**0.8032**

5.6 Selección de muestras

Haciendo una inspección de las muestras de la base de datos IEMOCAP y EMOWisconsin, nos dimos cuenta que se presentan incongruencias en el etiquetado emocional de las muestras, relacionadas con la tarea altamente subjetiva del etiquetado manual de emociones. Dichas incongruencias se presentan cuando varias muestras son etiquetadas con una emoción siguiendo el enfoque discreto pero, son ubicadas en lugares distantes en el espacio tridimensional continuo según su anotación continua. De esta manera podemos identificar muestras congruentes, las cuales muestran acuerdo entre su etiquetado discreto y continuo con respecto a otras muestras y por otro lado, las incongruentes que no muestran acuerdo.

Nuestra hipótesis es que nuestro algoritmo de aprendizaje automático tendría un mejor desempeño si seleccionábamos las muestras más congruentes y apropiadas para representar las propiedades de los enfoques discreto y continuo. Se diseñó un método inspirado en el método de auto entrenamiento explicado en la sección *2.3 Selección de muestras mediante auto-entrenamiento* en el que se parte del conjunto de muestras más confiables, es decir las que mostraron congruencia, con el que se crean modelos de regresión; se van añadiendo al conjunto de entrenamiento muestras identificadas como ambiguas pero que mejoran los modelos de regresión. Este proceso termina cuando ya no hay muestras que mejoren los modelos. Al final se obtienen modelos más representativos y más precisos que los modelos iniciales. Este proceso se representa en la Figura 21.

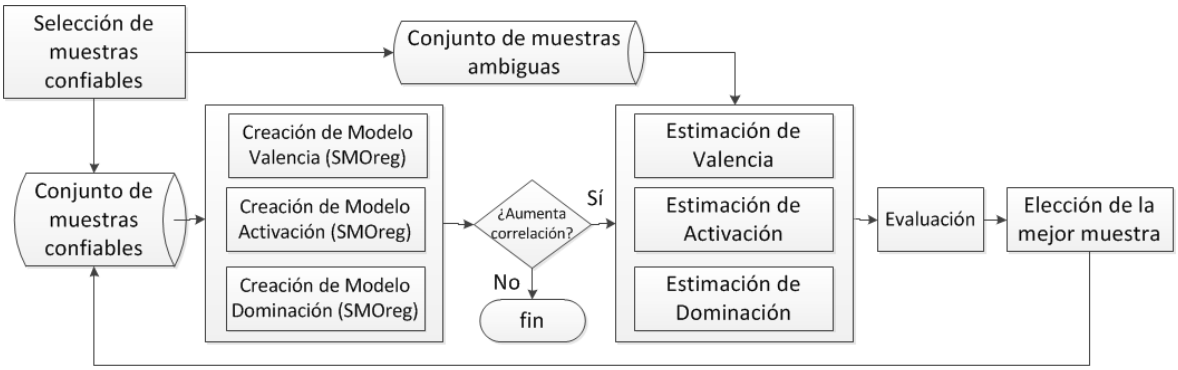


Figura 21 Método de selección de muestras para bases de datos con criterios subjetivos de etiquetado

Para realizar los experimentos de esta sección trabajamos con las bases de datos IEMOCAP Y EMOWisconsin ya que sólo estas dos bases de datos nos permiten trabajar con primitivas emocionales teniendo como referencia emociones discretas. Estas dos bases de datos fueron segmentadas manualmente a nivel de turnos de dialogo (turno de hablante), definidos como segmentos de voz continuos en los que el hablante está hablando activamente. A partir de esta segmentación se obtuvieron 1,819 muestras de la base de datos IEMOCAP repartidas en once clases. De la base de datos EMOWisconsin se obtuvieron 2,039 muestras repartidas en siete clases. Para realizar estos experimentos no usamos todas las clases. Trabajamos sólo con las clases con mayor representación. Para IEMOCAP trabajamos con Enojo (229), Felicidad (135), Tristeza (194) y Neutro (384). Para EMOWisconsin trabajamos con Confianza (911), Dubitativo (514), Nervioso (267), Motivado (72). Elegimos las muestras de estas clases para permitir la comparación de nuestro trabajo con el realizado por (Metallinou, Lee, & Narayanan, 2010) donde se usaron estas clases y para simplificar la demostración de nuestro método. Después de hacer esta selección nos quedamos con 942 muestras para IEMOCAP y 1,764 para EMOWisconsin.

Se aplicó un filtro adicional que consiste en eliminar todas las muestras que tienen la misma anotación para cada una de las tres primitivas, pero diferente anotación para la categoría emocional, ya que las consideramos muestras contradictorias que añaden ruido a nuestro proceso de aprendizaje. Por ejemplo, si las anotaciones para una muestra son (Valencia = 2, Activación = 4, Dominación = 3, Emoción = Enojo) y las anotaciones para otra muestras son (Valencia = 2, Activación = 4, Dominación = 3, Emoción = Felicidad) todas las muestras anotadas con Valencia = 2, Activación = 4 y Dominación = 3 son eliminadas del conjunto de datos. Después de esta selección nos quedamos con 467 muestras 163 para Enojo, 84 para Felicidad, 126 para Neutral y 94 para Tristeza. Las clases Felicidad, Tristeza y Neutro para IEMOCAP y Motivación, y Nerviosismo para

EMOWisconsin fueron sobre-muestreadas mediante SMOTE (Witten & Frank, 2005) para balancear el número de muestras por clase. Después de este sobre-muestreo nuestro conjunto de muestras creció a 645 muestras para IEMOCAP.

Algoritmo 1 Selección de muestras

Entrada: conjunto de N muestras confiables para cada primitiva

Salida: nuevo conjunto de muestras N + n muestras para cada primitiva

Variables locales: *datos_V, datos_A, datos_D* son los conjuntos de muestras confiables para cada primitiva. datos_eliminados son el conjunto de datos que están disponibles pero que no son parte de los datos confiables.

Requiere: $N \neq \emptyset$

```
datos_confiables ← [datos_V, datos_A, datos_D];
[mod_V, mod_A, mod_D] ← genera_modelos_SVM(datos_confiables);
modelos_confiables ← [mod_V, mod_A, mod_D];
while correlacion_promedio > correlacion_anterior do
    for j ← 1 to numero_muestras(datos_eliminados) do
        for i ← 1 to 3 do
            valores_primitivas[i] ← clasifica(modelos[i], datos_eliminados[j]);
        error[j] ← calcula_error(valores_primitivas, valores_reales(datos_eliminados[j]));
    indice_mejor_muestra ← min(error);
    anadir(datos_confiables, datos_eliminados(indice_mejor_muestra));
    eliminar(datos_eliminados(indice_mejor_muestra));
    [mod_V, mod_A, mod_D] ← genera_modelos_SVM(datos_confiables);
    correlacion_anterior ← correlacion_promedio;
    correlacion_promedio ← validacion_cruzada_SVM(mod_V, mod_A, mod_D);
return datos_confiables;
```

Después de este proceso implementamos el mecanismo descrito en el Algoritmo 1, para incorporar las mejores muestras de las que habían sido eliminadas previamente. Se generan tres modelos de predicción entrenados con SVM a partir del conjunto de muestras confiables. Uno para cada primitiva. Los valores de primitivas son estimados con estos modelos para cada muestra del conjunto de muestras eliminadas. Se estima el error entre los valores anotados y estimados por los modelos para cada muestra. Las muestras son evaluadas calculando el promedio de los errores de las tres primitivas. La muestra con el

error promedio mínimo es añadida al conjunto solución, preservando sus anotaciones originales y se elimina del conjunto de muestras eliminadas. Se generan nuevos modelos de estimación con el nuevo conjunto de muestras. El nuevo conjunto de muestras es evaluado mediante validación cruzada de 10 pliegues. Si el promedio de los coeficientes de correlación de Pearson de este modelo es mayor que el previo, el flujo regresa al punto 3, de lo contrario el proceso termina. En el Algoritmo 1 se presenta el pseudocódigo de dicho método.

Este proceso de selección supervisada aprovecha la anotación de emociones discretas y primitivas emocionales para lidiar con la subjetividad de la anotación de emociones. En la Tabla 17 se ilustra la mejoría en los resultados obtenidos sobre las dos bases de datos aplicando las ideas tomadas como base para el diseño del método descrito en esta sección. En primer lugar, no se trabajó con todas las clases, sino solo con las clases que tenían mayor representación en las bases de datos con lo que se logró mejoría en los resultados, es decir aumentó la correlación entre los valores esperados y estimados por el modelo. También es muy notoria la diferencia en la correlación obtenida con los datos confiables y datos eliminados, lo cual confirma que nuestro criterio para elegir muestras confiables es adecuado. Por último se muestra el resultado final del algoritmo de selección de muestras.

Tabla 17 Proceso de selección de muestras para ambas bases de datos

IEMOCAP						
	Neutro	Enojo	Tristeza	Felicidad	Correlación	Total
11 Clases	384	229	194	135	0.5808	1819
4 Clases	384	229	194	135	0.6614	942
Preservados	94	163	126	84	0.7223	467
Eliminados	258	66	100	51	0.3673	475
Balanceo SMOTE	126	163	188	168	0.7246	645
Selección Supervisada	218	167	270	193	0.7366	848
EMOWisconsin						
	Confianza	Inseguridad	Nerviosismo	Motivación	Correlación	Total
7 Clases	911	514	267	72	0.7220	2,039
4 Clases	911	514	267	72	0.7179	1,764
Preservados	349	136	91	41	0.7416	617
Eliminados	562	378	176	31	0.6721	1,147
Balanceo SMOTE	349	136	182	164	0.7789	831
Selección Supervisada	668	363	277	177	0.8088	1,485

5.7 Síntesis y conclusiones de selección de características

En este capítulo llevamos a cabo un estudio acerca de la importancia de diferentes tipos de características acústicas desde el punto de vista de un modelo tridimensional continuo. Analizamos cada primitiva emocional por separado. A través de la identificación de las mejores características para la estimación automática de primitivas emocionales fue posible mejorar la exactitud de la clasificación de emociones. Hasta donde sabemos la importancia de características acústicas no ha sido estudiada antes con este enfoque.

Hemos tomado algunas ideas usadas en el estudio del impacto de características en la clasificación de emociones categóricas cómo Share y Portion, y hemos aplicado estas métricas para el enfoque continuo. Dividimos nuestras 6,920 características acústicas en 12 subgrupos de acuerdo a sus propiedades acústicas. Calculamos algunas métricas para cada grupo para estimar su desempeño en la estimación automática de primitivas emocionales (coeficiente de correlación) en su contribución al conjunto final de características (Share y Portion).

Trabajamos con una base de datos de emociones actuadas, etiquetadas con los dos esquemas de anotación más importantes, continuo y discreto. A pesar de ser actuada esta base de datos fue diseñada tratando de hacerla lo menos artificial posible implementado una interacción por medio de improvisación de diálogos observamos que ambos esquemas de selección de características obtuvieron resultados muy similares y generalmente coinciden en la importancia de los grupos de características.

La principal contribución de este capítulo es el análisis de características acústicas. Este análisis fue basado principalmente en la medida del desempeño mediante la correlación obtenida por los modelos generados de un proceso de aprendizaje automático (SVM). El principal objetivo del análisis es determinar las propiedades acústicas que están más correlacionadas con la presencia de emociones en la voz. Para lograr este objetivo, se agruparon las características de acuerdo al aspecto de la voz que modelan. Por ejemplo, aspectos prosódicos, de calidad de voz espectrales, etc. Para realizar los experimentos de selección se plantearon dos esquemas, uno seleccionando características en cada grupo por separado y otro en el que se seleccionaron teniendo todas las características combinadas.

A través de los experimentos con estos dos esquemas pudimos observar que los subconjuntos de características seleccionadas con un esquema y otro lograban un desempeño similar al entrenar modelos con dichos subconjuntos para predecir primitivas emocionales. Además, se midió el desempeño descomponiendo los subconjuntos de características por grupos. Esto nos permitió descubrir los grupos de características mayor correlacionados con le estimación de primitivas emocionales.

Nos dimos cuenta que los grupos de características espectrales son muy importantes para las tres primitivas. De acuerdo a los resultados las características más importantes para cada primitiva emocional son:

Valencia: MEL - MFCC - Cocleogramas - LPC
Activación: MFCC - Cocleagramas - Intensidad - LPC
Dominación: MFCC – Cocleagramas – LPC – MEL

Claramente MFCC, LPC y Cocleogramas son muy importantes para estimar las primitivas emocionales, ya que aparecen entre las más importantes para las tres primitivas. Estos tres grupos pertenecen a la categoría de información espectral, podemos concluir de este hecho que el análisis espectral es más importante que el análisis prosódico y de calidad de voz para la estimación de primitivas, a excepción del análisis de Activación donde la Intensidad también es muy importante. El grupo de Cocleogramas es un hallazgo interesante; hasta donde sabemos, este grupo no había sido usado antes para reconocimiento de emociones.

Observamos que hay una correlación similar en los resultados para los enfoques de extracción de características por fuerza bruta y selectivo, donde las características seleccionadas pertenecientes a estos conjuntos fueron probadas por separado. Comparando el Portion de los conjuntos selectivo y fuerza bruta se puede decir que el conjunto selectivo tiene menos características pero más importantes. El Share fue más balanceado para ambos enfoques tendiendo a ser un poco más alto para el de fuerza bruta.

Capítulo 6: Estimación Multinivel de Emociones Basada en Interpretación de Primitivas Emocionales

En este capítulo se describe el método propuesto para reconocimiento automático de emociones en voz. La principal característica del modelo que proponemos es que se basa en un modelo emocional continuo. El interés en los enfoques computacionales de modelado emocional continuo ha crecido muy recientemente como lo demuestra el auge de llamados a publicaciones en revistas, conferencias y talleres especializados en este tópico en específico[3]. Nuestro enfoque propone la representación de emociones basada en agrupamiento difuso, con el objetivo de extraer información más completa y descriptiva que otras propuestas basándonos en un modelo tridimensional continuo. La representación es hecha en tres niveles de abstracción.

El primer nivel clasifica emociones en categorías discretas, el segundo trata de estimar la intensidad y mezcla de emociones, y el tercero agrupa emociones en grupos emocionales más amplios. Estos niveles de abstracción nos permiten visualizar desde diferentes puntos de vista el contenido emocional. Cada nivel representa una manera diferente de determinar el estado emocional de acuerdo a los requerimientos del contexto de aplicación. La representación propuesta está basada en las primitivas emocionales Valencia, Activación y Dominación obtenidas por modelos entrenados a partir de características acústicas. Con el objetivo de justificar y dimensionar la contribución de la interpretación multinivel propuesta es necesario considerar escenarios de aplicación del mundo real donde estas alternativas pueden ser usadas de manera práctica.

---

[3]
Call for Papers 1st International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Space, 2010
Call for Papers Elsevier Special Issue Image and Vision Computing Journal (Elsevier) on Affect Analysis in Continuous Input 2011
Call for Papers AVEC 2011: Audio/Visual Emotion Challenge and Workshop - Bridging between modalities, 2011
Call for Papers International Journal of Synthetic Emotions Special Issue on Benefits and Limitations of Continuous Representations of Emotions in Affective Computing, 2011
Call for Papers AVEC 2012: 2nd Audio/Visual Emotion Challenge and Workshop - Facing continuous emotion representation, 2012
Call for Papers 2nd International Workshop on Emotion Representation, Analysis and Synthesis in Continuous Space, 2013
Call for Papers AVEC 2013: 3rd Audio/Visual Emotion Challenge and Workshop – Depression and continuous emotion, 2013

Para ilustrar la necesidad de representación emocional en los tres niveles de abstracción hemos seleccionado algunos trabajos enfocados a solucionar problemas del mundo real usando información emocional en cada uno de los niveles de abstracción que proponemos. El caso de emociones discretas es el más directo de ilustrar. Aquí la idea es identificar emociones discretas muy bien diferenciadas en nuestro espacio emocional. Por ejemplo un sistema de IHC que detecta miedo y enojo. En la Tabla 18 se muestran algunos ejemplos de este tipo de aplicaciones.

Tabla 18 Aplicaciones en el nivel de emociones discretas

Aplicación	Emociones de Interés	Utilidad
Interacción emocional con Robots (Kim, Kwak, Hyun, Kim, & Kwak, 2009)	Enojo, Felicidad, Neutral y Tristeza	Síntesis de emociones en respuesta al estado emocional percibido en el usuario
Mejora de aprovechamiento en educación a distancia (Li, Zhang, & Fu, 2007)	Enojo, Felicidad, Tristeza y Alegría	Mejorar el entorno emocional de aprendizaje el reconocimiento y retroalimentación emocional
Seguridad en el manejo de vehículos (Boril, Sadjadi, Kleinschmidt, & Hansen, 2010)	Neutral, Confianza, Felicidad, Indiferencia, Duda, Confusión, Frustración, Enojo	Medir la capacidad de un conductor para controlar un auto

En el caso de grupo de emociones, como se muestra en l

Tabla 19, podemos pensar en el estudio de reacciones de usuarios en un video juego multi-jugador (Truong, Neerincx, & van Leeuwen, 2008) o en la detección de emociones no deseables en *call centers* (Vidrascu & Devillers, 2005) donde es importante identificar emociones con Activación alta.

En este tipo de aplicaciones no son de interés emociones discretas específicas, sino todo un amplio espectro de emociones o, en otras palabras, una familia de emociones dentro de la misma región en el espacio definido por las primitivas Valencia – Activación – Dominación.

Tabla 19 Aplicaciones en el nivel de grupos de emociones

Aplicación	Regiones de Interés	Utilidad
Análisis de reacción de usuarios en videojuegos (Truong, Neerincx, & van Leeuwen, 2008)	Activación alta	Conocer que el grado de interés y/o emoción que un video juego o algún otro producto despierta en el consumidor.
Monitorización de llamadas en *call centers* (Vidrascu & Devillers, 2005)	Activación alta, Valencia baja	Detectar llamadas problemáticas, en las que muy probablemente un cliente esté exaltado y/o disgustado.
Asistencia a pacientes y Apoyo a médicos (Arnrich, Setz, La Marca, Gerhard, & Ehlert, 2010)	Activación muy alta o Activación muy baja, Valencia baja	Diagnosis y tratamiento de trastorno afectivo bipolar

En el caso de los *call centers*, se podría requerir una región delimitada por valores altos de Activación y bajos de Valencia en el cual se encontrarían emociones como enojo, frustración, molestia o miedo. Es importante notar que en el mismo grupo se consideran emociones con Dominación opuesta como enojo, que tiene Dominación alta y miedo, que tiene Dominación baja.

Para el estudio de reacciones en un videojuego o cualquier otro producto, sería valioso conocer si la reacción emocional es generada en la región de Activación alta, lo cual significaría que el estímulo excita el estado emocional del jugador, en contraste a un estímulo que es inadvertido. Esta región emocional podría incluir emociones como felicidad, excitación, interés, expectación, miedo, enojo, frustración. Cómo podemos ver en el mismo grupo emocional hay estados en extremos opuestos de Valencia como felicidad y enojo, pero que comparten un nivel elevado de Activación. En estos casos, el nivel grupal de representación es muy adecuado.

En la Tabla 20 se muestran algunos ejemplos de aplicación en los que podría ser útil la estimación de mezcla e intensidad emocional sin embargo, el avance en el área de reconocimiento automático de emociones no ha llegado a un nivel de exactitud suficiente para desarrollar este tipo de aplicaciones.

Tabla 20 Posibles aplicaciones en el nivel de mezcla e intensidad

Aplicación	Mezclas de Interés	Intensidades de Interés	Utilidad
Apoyo a psicoterapeutas (Plutchik, 2000)	Felicidad + Confianza	Nivel de confianza Nivel de miedo Nivel de tristeza	Detección de nivel de focalización o vaguedad de emociones
Apoyo a neurólogos (Vera-Muñoz, Pastor-Sanz, Fico, & Arredondo, 2008)	-	Nivel de miedo Nivel de aversión	Diagnosis y seguimiento de enfermedades de Parkinson y Huntington
Apoyo en detección de Neurosis (Sobol-Shikler, 2009)	Alegría + Tensión	-	Detección de mezcla de emociones ambivalentes

Algunos autores (Cacioppo & Berntson, 1994) (Larsen & McGraw, 2011) (Larsen, To, & Fireman, 2007), han demostrado que en ciertas situaciones las personas pueden experimentar más de una emoción a la vez, es decir, una mezcla de emociones. El estudio de este tipo de fenómenos como mezcla e intensidad de emociones podría tener aplicaciones médicas tales como estudios psicológicos de personalidad y psicoterapia, donde se sugiere que rasgos de personalidad pueden ser estudiados en términos de mezclas de emociones (Plutchik, 2000). Plutchik propone que la psicoterapia debe recurrir a un fuerte estímulo emocional para ser eficaz. Argumenta que la mayoría de los síntomas psicológicos se relacionan con emociones que se han tergiversado o vuelto disfuncionales, algunas emociones son demasiado fuertes o persistentes (como el pánico o la depresión), mientras que otras son muy débiles o sutiles (como la confianza y el placer) (Plutchik, 2000). En este caso, la estimación de mezcla y nivel de expresividad emocional puede ser muy útil Sin embargo queda fuera del alcance de este trabajo el comprobar o demostrar la pertinencia de la determinación de mezcla e intensidad de emociones las aplicaciones sugeridas.

Los diferentes niveles de representación emocional propuestos en este capítulo pueden proveer información adicional entre sí. Es muy importante señalar que uno de los aspectos valiosos del método propuesto es que no se necesitan muestras de cada una de las emociones mencionadas en los ejemplos. Lo que se necesita son muestras de diferentes niveles de Valencia, Activación y Dominación, que podrían provenir de otro contexto de aplicación, haciendo que los modelos entrenados sean independientes de los emociones de interés para cada aplicación.

6.1 Creación de modelos de regresión y clasificación

Como muestra el proceso ilustrado en la Figura 22, después de caracterizar la señal de voz se crean modelos de regresión y clasificación emocional. Construimos modelos de predicción usando máquinas *de vectores de soporte (SVM)* a partir de las características acústicas extraídas de la señal de voz. La correlación de la validación cruzada de diez pliegues para estos modelos de predicción es 0.66 para Valencia, 0.82 para Activación, y 0.73 para Dominación en el caso de IEMOCAP y 0.78 para Valencia, 0.85 para Activación y 0.80 para Dominación en el caso de EMOWisconsin. Es importante hacer notar que esta estimación automática no es perfecta; sin embargo, es suficientemente adecuada para nuestros propósitos. Modelamos la relación entre características acústicas y primitivas emocionales usando un enfoque estadístico como SVM, y la relación entre emociones y primitivas emocionales usando lógica difusa.

La elección de ambas técnicas está relacionada con el hecho de que las características acústicas son mediciones objetivas que son bien modeladas con SVM, mientras que las primitivas emocionales son mediciones más subjetivas en las cuales las personas describen lingüísticamente el habla emocional, siendo esta una descripción vaga e imprecisa que puede ser mejor modelada con lógica difusa. Algunos autores han explorado el uso de modelos difusos para reconocimiento y síntesis de emociones. A través de la teoría difusa dichos autores han estudiado la relación entre características acústicas y emociones discretas (Esau, Kleinjohann, & Kleinjohann, 2005), (Giripunje & Bawane, 2007), (Panat & Ingole, 2007) así como la relación entre características acústicas y primitivas emocionales (Grimm M. , Kroschel, Mower, & Narayanan, 2007), (Huang & Akagi, 2005). A diferencia de esos trabajos, nosotros usamos un enfoque difuso para estudiar la interpretación de primitivas emocionales y su relación con emociones, como se explica en la sección 4.4.

En la Figura 22 se muestra el proceso de creación de modelos en el cual se discretizan los datos con el objetivo de usar estos modelos de clasificación junto con los modelos de regresión en la comparación con otros trabajos. La discretización consiste en ordenar las muestras de acuerdo a su valor de primitiva emocional, en rangos que pueden o no incluir un número similar de muestras de acuerdo a la distribución de los datos en la escala continua en que están definidas las primitivas. En el caso de los datos con los que se

trabajó en esta tesis el rango completo en el que están definidas las primitivas emocionales es de uno a diez.

Se recomienda usar rangos con un número de muestras balanceado. Los rangos para discretizar los datos pueden ser, por ejemplo, bajo, medio y alto. Alternativamente, si sólo se está interesado en los extremos, se pueden eliminar las muestras de los rangos intermedios y se crea un modelo de clasificación con las clases bajo y alto de esta manera, al existir diferencias más marcadas entre las muestras de cada rango, se obtienen mejores modelos de clasificación.

En la Tabla 21 se muestra el grado de exactitud de algunos de estos modelos entrenados con diferentes esquemas de selección de características, entrenamiento y prueba probados en la sección*5.5* Análisis multilingüe de características.

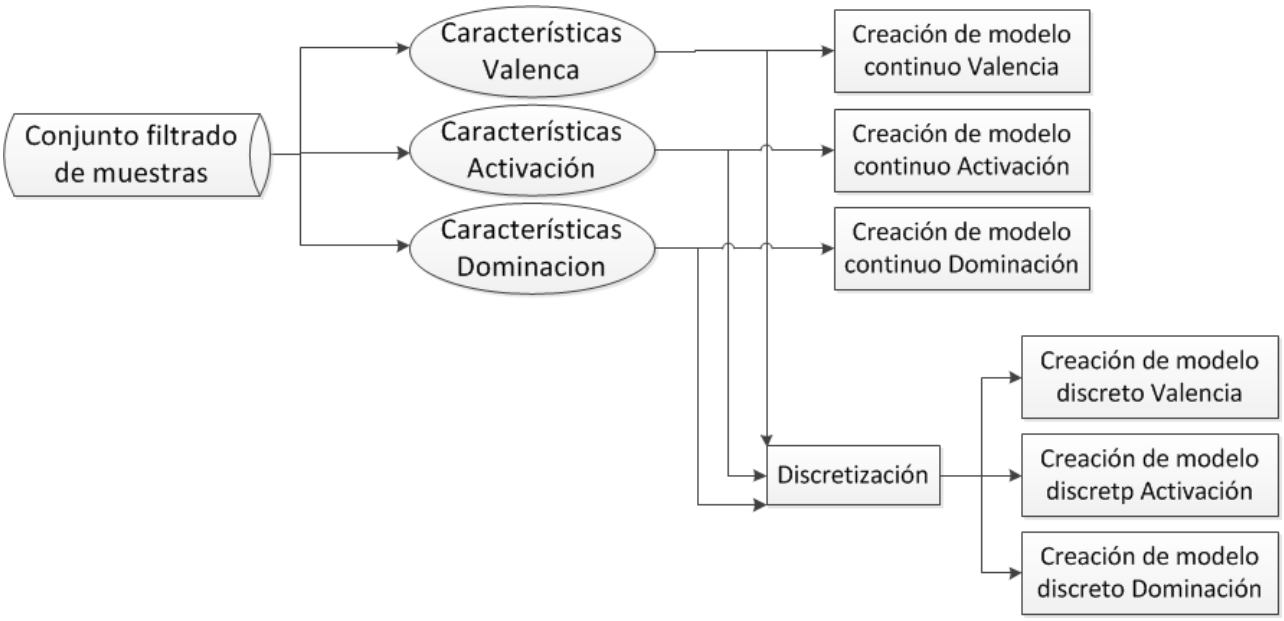


Figura 22 Creación de modelos

Tabla 21 Correlación y precisión para Valencia, Activación, Dominacion

	Valencia		Activación		Dominación	
Selección/Entrenamiento/Prueba	**Correl**	**Precisión**	**Correl**	**Precisión**	**Correl**	**Precisión**
Monolingüe/Español/Español	0.7756	87.93	0.8497	92.05	0.8011	88.67
Multilingüe/Español/Español	0.6958	82.32	0.7873	87.95	0.7141	84.14
Multilingüe/Multilingüe/Multilingüe	0.5539	71.85	0.7722	87.69	0.7100	82.93

6.2 Interpretación multinivel de primitivas

En esta sección se describen los diferentes niveles de interpretación de primitivas emocionales propuestos. La intención de proponer estos niveles de interpretación es aplicarlos para resolver diferentes tipos de problemas del mundo real en los que es de relevancia el reconocimiento emocional. En Tabla 18,

Tabla **19** y Tabla 20 se muestran ejemplos de aplicaciones donde estos niveles de representación pueden ser útiles. La implementación de los algoritmos descritos en este capítulo está basada en agrupamiento difuso el cual que se explica en la siguiente sección *6.2.1 Agrupamiento difuso.*

Para realizar nuestros experimentos iniciales utilizamos dos conjuntos de datos de la base de datos IEMOCAP, ambos conjuntos tienen 848 muestras. El conjunto *A* contiene las muestras anotadas por anotadores humanos con Primitivas Emocionales. El conjunto *E* contiene las mismas muestras pero con Primitivas Emocionales estimadas automáticamente por los modelos construidos con *SVM* y descritos en la sección 5.2. La idea de usar estos dos conjuntos de datos es similar a tener un conjunto de entrenamiento y uno de prueba. Calculamos los centros de los *clusters* usando las muestras en el conjunto *A*, subsecuentemente evaluamos mediante la agrupación de muestras del conjunto *E* con respecto a los centros mencionados.

6.2.1 Agrupamiento difuso

En este trabajo usamos agrupamiento difuso para estimar el contenido emocional en el habla a partir de las estimaciones de primitivas emocionales. El agrupamiento divide las muestras en clases homogéneas o grupos tales que las muestras en la misma clase son tan similares como sea posible. En el agrupamiento difuso, las muestras podrían pertenecer a más de un grupo. Cada muestra tiene grados de membresía indicando el grado al cual pertenece a diferentes grupos. En nuestros experimentos, usamos el *algoritmo Fuzzy c-Means (FCM)* que divide un conjunto de muestras en un conjunto de *C* grupos difusos iniciando con un centro elegido aleatoriamente que es movido en cada iteración. Esta afinación es hecha basada en la minimización de una función objetivo, dicha función se

basa en las medidas de distancia de las muestras hacia el centro del grupo, pesadas por el grado de membresía para ese grupo. Dado un conjunto de *N* muestras FCM genera una lista de *C* centros de grupos y una matriz μ_{cN} especificando el grado de membresía de cada muestra a cada grupo.

Para nuestro modelo hemos usado una versión modificada del algoritmo FCM. Dado que en las bases de datos IEMOCAP y EMOWisconsin las muestras están anotadas con primitivas emocionales y emociones discretas, la modificación consiste en aprovechar este conocimiento para guiar el agrupamiento de acuerdo a ciertas emociones de interés. Esta adaptación es flexible y permite incluir en este proceso el conocimiento de las clases anotadas, además se puede ajustar el peso de estas etiquetas en el proceso de entrenamiento.

Las etiquetas de clase proveen una guía útil durante los procesos de entrenamiento para mejorar el desempeño de *FCM*. Esta idea nos llevó a probar además del algoritmo tradicional *FCM* una versión *supervisada* haciendo algunas modificaciones al algoritmo *SFCM* propuesto en (Kalyani S., 2010). Originalmente *SFCM* intenta desarrollar clasificadores que utilicen muestras etiquetadas y no etiquetadas. En dicho método de clasificación, un conjunto fijo de categorías y muestras etiquetadas con esas categorías son usados para inducir una función de clasificación. El agrupamiento supervisado agrupa datos usando categorías en los datos etiquetados inicialmente así como en extiende y modifica el conjunto de categorías existentes para reflejar irregularidades en el conjunto de datos.

En nuestro caso retomamos la misma idea para trabajar con bases de datos que estén etiquetadas con primitivas emocionales y se puedan utilizar también etiquetas de emociones discretas si es que se cuenta con ellas. En el método de representación multinivel propuesto a continuación se tomó como base el algoritmo FCM , y se incluyeron algunos pasos intermedios, los cuales pueden tomar en cuenta o no las etiquetas de clases si es que se cuentan con ellas, como en el caso del agrupamiento difuso supervisado.

6.2.2 Nivel de representación de emociones discretas

El propósito de este nivel es representar emociones en la forma más convencional. Se clasifican emociones discretas a partir de las estimaciones de primitivas emocionales. En la Figura 23 se muestra la ubicación en el espacio tridimensional de las emociones discretas con más muestras en la base de IEMOCAP y con las cuales estuvimos trabajando en esta sección.

El Algoritmo 2 detalla los pasos a seguir. Se forman *C clusters* usando *FCM* a partir del conjunto de datos *A*, obteniendo la matriz de membresías $UA_{C \times N}$ y la lista de centros CL_{cx1} donde *C* es el número de emociones anotadas en el corpus y *N* el número de muestras (línea 1). Cada muestra de *A* es asignada a un *cluster* (líneas 2 y 3) de acuerdo al valor máximo de membresía. Cada uno de los *C clusters* es relacionado con una de las *C* emociones discretas de acuerdo al número máximo de muestras de cada emoción incluidas en el *cluster* (líneas 4-6). La matriz de membresías *UE* es calculada, de acuerdo a la Formula 7, para *E* con respecto a los centros de *clusters CL* (línea 7). Cada muestra de *E* es clasificada como una emoción (líneas 8 y 9) para evaluar el desempeño de clasificación se compara la emoción asignada por el algoritmo contra la emoción originalmente anotada de forma manual.

La primera fila en la Tabla 22 muestra los resultados cuando se clasifican emociones discretas en *A*. Como puede observarse, la clasificación no es perfecta; esto muestra la naturaleza altamente subjetiva de la anotación manual de emociones. Una buena clasificación en *E* es aún más difícil de alcanzar porque acarrea el error producido por la estimación automática de primitivas; sin embargo, tratamos de modelar esta incertidumbre mediante modelado difuso.

En teoría, la clasificación más alta en *E* alcanzaría los resultados de *A*. La medida de desempeño del clasificador usada es la *cobertura* definida como la división de verdaderos positivos entre la suma de verdaderos positivos más falsos negativos. Como podemos ver, Enojo, y Tristeza tienen la mejor y peor cobertura para ambos conjuntos respectivamente. Los resultados de referencia fueron obtenidos por Metallinou (Metallinou, Lee, & Narayanan, 2010) usando los mismos datos que nosotros usamos para este experimento, es decir, el corpus IEMOCAP.

Algoritmo 2 Representación de emociones discretas

Entrada: conjunto **A** de muestras etiquetadas, conjunto **E** de muestras a clasificar, **C** número de emociones de interés

Salida: emocion_E que es la clase calculada para las muestras en E
Variables locales: clusters_muestras es un arreglo unidimensional que almacena el *cluster* al que pertenece cada muestra de **A**, **num_emociones** es un arreglo unidimensional que almacena el número de muestras de cada clase que pertenecen al *cluster*, **numero_muestras** es un entero que almacena el número de muestras en el conjunto **A**, **emocion_cluster** es un arreglo que almacena la emoción relacionada con cada *cluster*, **UA** es la matriz de membresías de los datos etiquetados, **UE** es la matriz de membresías de los datos a clasificar, **emocion_E** es un arreglo con la emoción asignada a cada muestra en **E**

Requiere: A, E ≠ ∅

```
[UA,CL] ← Fuzzy_Cmeans(A,C);
for i ← 1 to numero_muestras do
    clusters_muestras[i] ← max(UA[1:C,i]);
for i ← 1 to C do
        num_emociones ← cuenta_emociones(clusters_muestras,A[: , C], i);
        emocion_cluster[i] ← max_emocion(emociones_cluster);
UE ← calcula_matriz_membresias(E,CL);
for i ← 1 to numero_muestras(E) do
        emocion_E[i] ← max(UE[1:C,i]);
return emocion_E;
```

Tabla 22 Cobertura para clasificación de emociones discretas

	Enojo	Alegría	Neutro	Tristeza
Conjunto de datos A	91.62	89.12	77.52	71.85
Conjunto de datos E	71.26	56.99	68.81	52.22
Metallinou (Metallinou, Lee, & Narayanan, 2010)	69.68	21.01	35.23	76.84

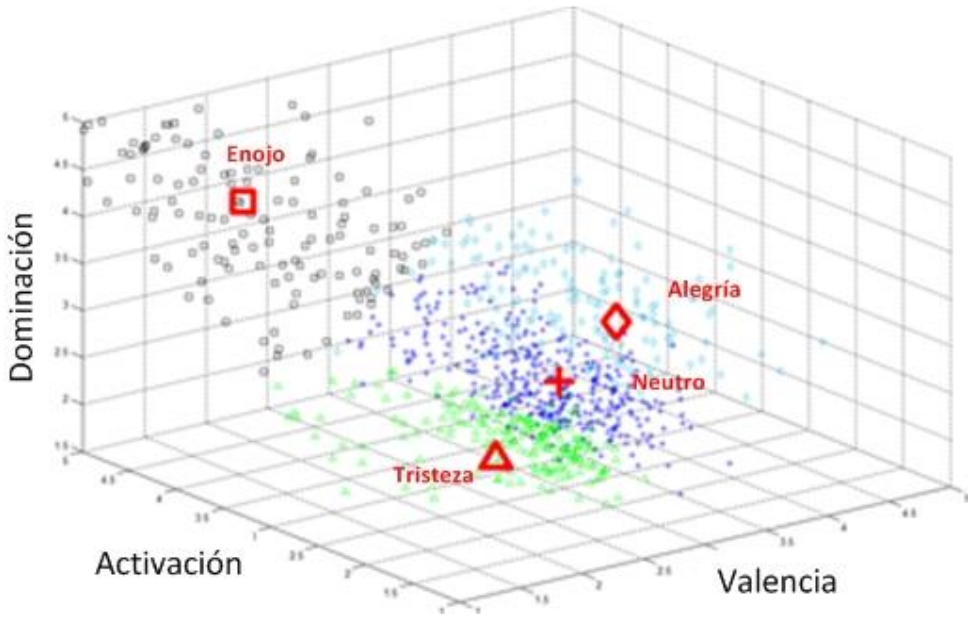


Figura 23 Cuatro emociones discretas ubicadas en el espacio tridimensional: □ Enojo, ◇ Alegría, + Neutro, △ Tristeza

6.2.3 Nivel de representación de expresividad y mezcla

El propósito de este nivel de representación es estimar la composición de un estado emocional en términos de emociones discretas e identificar diferentes niveles de expresividad. De esta manera analizamos emociones presentes en la base de datos usando una granularidad más alta es decir, agrupando en grupos más pequeños, dividiendo emociones discretas en subgrupos de emociones puras, emociones mezcladas y emociones puras con diferentes grados de expresividad. Es importante notar que aun cuando en este trabajo usamos cuatro emociones para ilustrar las propuestas de representación, este enfoque es fácilmente generalizable a un número mayor o menor de emociones discretas o cualquier estado emocional, no necesariamente reconocido como emoción básica; ya que sólo es necesario dar como entrada al algoritmo los centroides de las emociones de interés. Los siguientes pasos funcionan para *C* emociones, donde *C* es mayor que el número de emociones discretas *N*, en este caso particular cuatro. La Figura 24 muestra ejemplos de distribución para siete *clusters* y la Tabla 23 muestra su composición.

Tabla 23 Composición de los *clusters*

Cluster	Composición
□	Representa solamente tristeza
+	Representa enojo poco expresivo
△	Representa solamente alegría
◇	Representa mezcla de tristeza y neutral
○	Representa mezcla de alegría, neutro y tristeza
★	Representa neutral poco expresivo
▽	Representa solamente enojo

El Algoritmo 3 especifica los pasos a seguir. Se forman *D clusters* por *FCM* a partir de *A*, obteniendo la lista de centros CD_{cx1} (línea 2). Se calcula la matriz de membresía *Ucld* entre *CL* y *CD* (línea 4) de acuerdo a Formula 7, es importante notar que esta matriz de membresías no es entre las muestras de entrenamiento y los centros calculados de clase originales, sino entre los centros de clase originales y los centros de grupos nuevos. Se determina la composición de cada *cluster* de acuerdo al método descrito en la sección *6.2.4 Estimación difusa de emociones*(línea 4).

La evaluación se hace calculando *Ue* y mapeando las muestras del conjunto *E* con respecto a *CD*.

Para cada muestra *s* en *E*

La Emoción(*s*) = *c*, donde $UE_{c,m} = \max(UE_{1..C,s})$

Es difícil evaluar objetivamente la mezcla de emociones. Para propósitos de este trabajo evaluaremos la clasificación en grupos duros con las clases generadas en el agrupamiento. Esto no parece la mejor manera de evaluar esta parte ya que estamos evaluando de manera dura una estimación suave. Dejamos para trabajo futuro el encontrar una mejor manera de evaluar esta parte. La cobertura promedio para cinco, seis y siete *clusters* fue 57.20, 58.50, 60.95 respectivamente.

Algoritmo 3 Representación de expresividad y Mezcla

Entrada: conjunto **A** de muestras etiquetadas, conjunto **E** de muestras a clasificar, **C** número de emociones de interés, **D** número de emociones conocidas

Salida: **Ue** lista describiendo la composición de los *clusters* generados (como en la Tabla 23),
Variables locales: **CL** lista de los centros de los *clusters* formados con las emociones de referencia, **CD** lista de los centros de los nuevos *clusters*, **Ucld** matriz de membresías que indica el nivel de pertenencia de nos nuevos *clusters* hacia las emociones de referencia

---

Requiere: A, E ≠ Ø, *D* > *C*

1 [UAc, CL] ← Fuzzy_Cmeans(A,C);

2 [UAd, CD] ← Fuzzy_Cmeans(A,D);

3 Ucld ← calcula_matriz_membresias(CL,CLC);

4 Ue ← determina_composocion(Ucld, A, E, C, D);

5 return Ue;

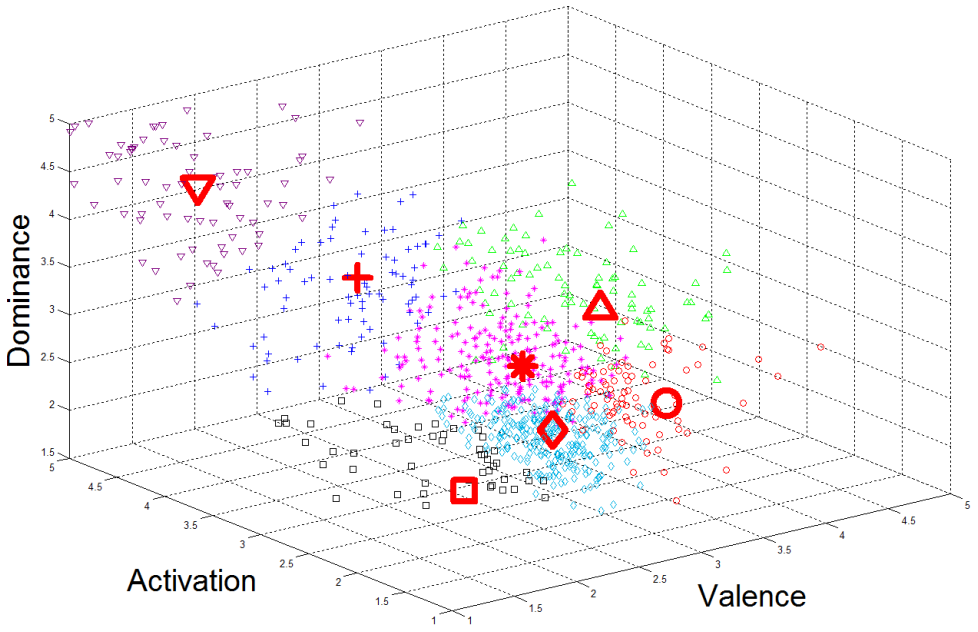


Figura 24 Agrupamiento para siete emociones discretas

6.2.4 Estimación difusa de emociones

Mediante este método se generan umbrales difusos para determinar si una muestra representa un estado emocional dado, si representa un estado emocional de manera intensa, o si es una mezcla de emociones.

Algoritmo 4 Estimación difusa de emociones

Entrada: conjunto **A** de muestras etiquetadas, conjunto **E** de muestras a clasificar, **C** número de emociones de referencia, **N** número de nuevos *clusters*

Salida: composicion_E es una lista que indica de que emociones se compone cada grupo

Variables locales: CLC matriz de centros de nuevos clusters, **CL** matriz de centros de emociones de referencia, **promedios_membresias** matriz que contiene para cada *cluster* el promedio de grado de pertenencia de sus miembros, **emocion_clusters** vector que en cada posición guarda la emoción del *cluster* respectivo, **fis** sistema de inferencia difuso que determina la composición de los *clusters*, **matriz_distancias** matriz bidimensional que especifica a distancia euclidiana entre los centros de las emociones de referencia y los nuevos *clusters*, **U** matriz de membresías de los nuevos *clusters* con relación a las emociones de referencia, **vad_E** matriz que contiene las estimaciones de Valencia, Activación y Dominación para cada muestra del conjunto **E**

Requiere: A, E $\neq \emptyset$

```
[Uclc,CLC] ← Fuzzy_Cmeans(A,N);
[Ucl,CL] ← Fuzzy_Cmeans(A,C);
for i ← 1 to C do
        promedios_membresias[i] ← promedia_membresias(Ucl , A[: , C]);
        emocion_cluster[i] ← max_emocion(promedios_membresias);
[media, varianza] ← calcula_estadisticas(promedios_membresias);
fis ← crea_fis(Ucl, C, media, varianza);
matriz_distancias ← calcula_distacia_euclidiana(CLC, CL);
U ← calcula_matriz_membresias(matriz_distancias);
vad_E ← estimar_vad(E);
composición_E ← evalua_composición(base_reglas, fis, vad_E);
return composicion_E;
```

El método funciona según lo descrito en el Algoritmo 4. Se crea el número de *clusters* nuevos con los datos originales (línea 1) usando FCM. Se crea el número de *clusters* de las clases que sabemos que originalmente están etiquetadas en los datos, a las que llamamos emociones de referencia (línea 2) usando FCM. En el caso de la base de datos IEMOCAP creamos cuatro *clusters*. Se calcula el promedio del grado de membresía de los elementos que pertenecen al *cluster* de acuerdo a las etiquetas originales; de esta manera elegimos que *cluster* representa mejor cada clase y se asignan dichos *clusters* a clases (líneas 3 - 5).

Una vez que ya sabemos a qué emoción representa cada *cluster* creamos automáticamente un sistema de inferencia difuso tipo Sugeno (Sugeno & Kang, 1988) para encontrar automáticamente los umbrales de representación. Se añade una función de membresía de entrada por cada emoción originalmente anotada. Los parámetros para definir las funciones de membresía son la media y la varianza de las membresías de las muestras que pertenecen a la emoción que representa el grupo. La media indica el punto central y la varianza el ancho de la gaussiana (líneas 6 y 7).

Se calcula la matriz de membresía U entre *CL* y *CLC* a partir de las distancias entre ambos conjuntos de centros (líneas 8 y 9) usando la Formula 7, donde *CL* es la lista de centros calculados en la línea 2. Se determina la composición de los *clusters* evaluando el sistema de inferencia difuso (línea 11). Existen 3 posibles salidas: Mezcla de emociones, Emoción con baja intensidad, Emoción pura.

Evaluamos calculando *Ue* con respecto de *CLC*.

Para cada muestra s en *E*

La Emoción(s) = *c*, donde $UE_{c,m} = \max(UE_{1..C,s})$

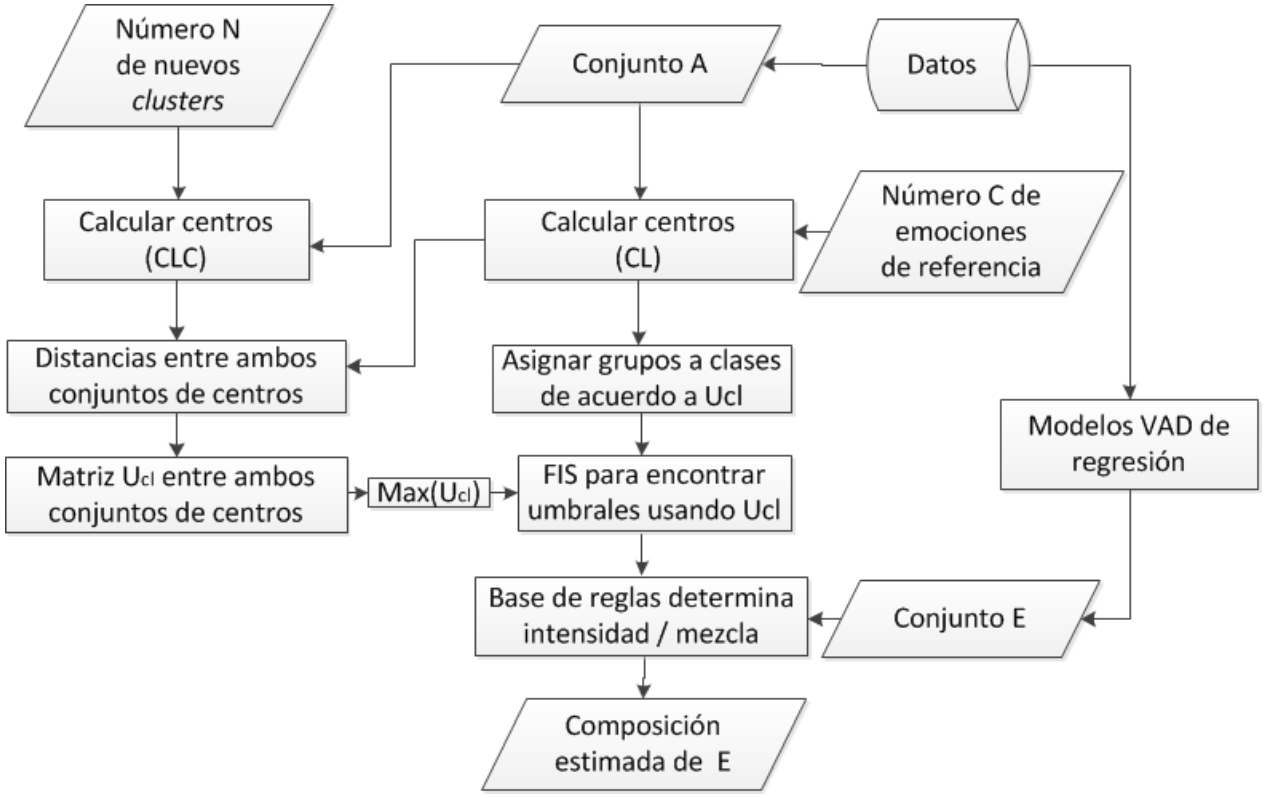


Figura 25 Estimación difusa de emociones

6.2.5 Nivel de representación de grupos

El propósito de esta representación es analizar emociones desde una granularidad más baja agrupando en clases más grandes de acuerdo a los valores de primitivas emocionales. Seguimos los siguientes pasos que funcionan para *C* emociones con *C* menor que el número de Emociones discretas, en este caso particular cuatro. El método funciona según lo descrito en el Algoritmo 5.

Algoritmo 5 Nivel de representación de grupos

Entrada: conjunto **A** de muestras etiquetadas, conjunto **E** de muestras a clasificar, **C** número de emociones de referencia

Salida: emocion_E es una lista que indica de que emociones se compone cada grupo
Variables locales: CLC matriz de centros de nuevos clusters, **clusters_muestras** es un arreglo unidimensional que almacena el *cluster* al que pertenece cada muestra de **A**, **num_emociones** es un arreglo unidimensional que almacena el número de muestras de cada clase que pertenecen al *cluster*, **emocion_cluster** es un arreglo que almacena la emoción relacionada con cada *cluster*, **UE** es la matriz de membresías de los datos a clasificar, **emocion_E** es un arreglo con la emoción asignada a cada muestra en **E**

```
    Requiere: A, E ≠ Ø
1   [UA,CLC] ← Fuzzy_Cmeans(A,C);
2   for i ← 1 to numero_muestras(A) do
3       clusters_muestras[i] ← max(UA[1:C,i]);
4   for i ← 1 to C do
5           num_emociones ← cuenta_emociones(clusters_muestras,A[: , C],i);
6           emocion_cluster[i] ← max_emocion(emociones_cluster);
7   UE ← calcula_matriz_membresias(E,CLC);
8   for i ← 1 to numero_muestras(E) do
9           emocion_E[i] ← max(UE[1:C,i]);
10  return emocion_E;
```

Se forman *C clusters* mediante *FCM* a partir de *A* obteniendo la lista de *clusters CLC* (línea 1). Se determina que emociones fueron incluidas en cada *cluster* contando el número de muestras de cada emoción incluida en cada *cluster* (líneas 2-6). Obtenemos la matriz de membresía *UE* con respecto de *CLC* (línea 7). Cada muestra en *E* es asignada a cada *cluster* (línea 8 y 9).

Para evaluar el desempeño de este agrupamiento de emociones se usan los datos del conjunto E. Se sustituyen la clase de cada muestra de este conjunto por el *cluster* al cual fue asignada dicha clase en las líneas 2 – 6 del algoritmo 5. Por ejemplo, si la emoción enojo fue asignada al Cluster 1, para todas las muestras del conjunto E que pertenezcan a la clase enojo se sustituye el valor de clase “enojo” por “cluster 1”. Posteriormente, se compara estas nuevas etiquetas de clase con las estimadas en las regresadas en la línea 10 del algoritmo 5 para determinar la precisión de la estimación.

Con este algoritmo se generaliza la clasificación de emociones usando datos de entrenamiento con emociones discretas, flexibilizando el uso de datos de entrenamiento en diferentes aplicaciones del mudo real como los presentados en la Tabla 19.

La cobertura para dos *clusters* fue 90.34, el *cluster* uno incluyó Enojo y el dos incluyó Felicidad, Neutro y Tristeza. La cobertura para tres *clusters*, ver Figura 26, fue 75.09, el *cluster* uno incluyó Neutro, Tristeza, el *cluster* dos Felicidad y el *cluster* tres Enojo.

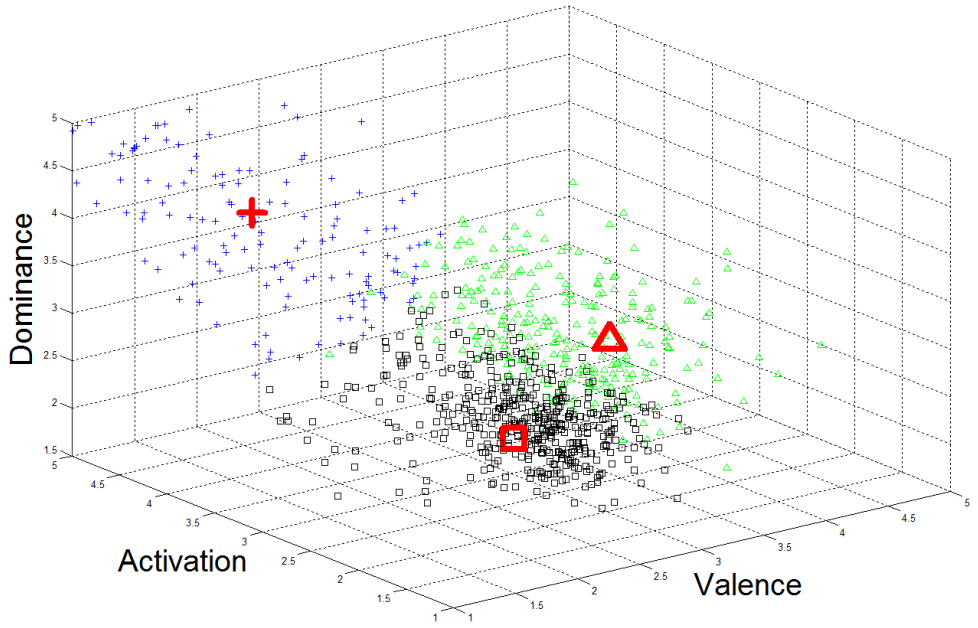


Figura 26 *Clustering* para tres categorías: Enojo, Neutro/Tristeza, Felicidad

6.2.6 Conclusiones de niveles de representación

Como se puede ver en la *Figura 4 Modelo Propuesto* el trabajo realizado en el *Capítulo 5: Caracterización de Voz y Selección de Datos* nos permite generar modelos de clasificación y regresión de primitivas emocionales. En esta parte de la tesis *Capítulo 6: Estimación Multinivel de Emociones Basada en Interpretación de Primitivas Emocionales* se usan las primitivas emocionales para mediante agrupamiento difuso y un modelo tridimensional continuo representar emociones en diferentes niveles de abstracción. Usamos información acústica, con la cual obtenemos una estimación suficientemente buena de Primitivas Emocionales. Información adicional, tal como información visual o biométrica, podría ser usada para mejorar esta estimación. La exactitud de estas

representaciones podría mejorar si la estimación automática de primitivas mejora. La representación propuesta no depende de la información usada para estimar las primitivas sino de las anotaciones emocionales en el corpus. Estas representaciones de emociones pueden ser implementadas para satisfacer las necesidades de interpretar emociones de los usuarios en un amplio espectro de aplicaciones y puede ser adaptado dependiendo del nivel de abstracción que la aplicación requiera.

Capítulo 7: Clasificación Basada en Contexto Emocional

Como se puede apreciar en la Figura 4 en la parte final del método se incluye un módulo que hace una reclasificación de segmentos para corregir posibles errores de clasificación de los modelos de clasificación del módulo de clasificación/regresión. También se puede aplicar este método a la salida del módulo de Nivel de representación de emociones discretas. El cambio en la clasificación de una muestra se hace basándose en el contexto emocional, es decir en la clasificación de segmentos en el vecindario temporal. La información contextual es una manera de modelar el fenómeno en un nivel que se extiende más allá del segmento actual de habla analizado. La mayoría de los trabajos en el reconocimiento de emociones en voz usan características acústicas y algunas características lingüísticas como base para la clasificación de emociones. Una alternativa interesante para mejorar el desempeño de los clasificadores es complementar la información acústica y/o lingüística con características diseñadas para tomar ventaja de la naturaleza del entorno de la conversación.

Los segmentos de habla, usualmente turnos, están incrustados en una estructura más grande, como un dialogo y por lo tanto parece razonable usar evidencia pasada de la actividad del usuario para mejorar la estimación del estado emocional actual. Además, este dialogo también está influenciado por el entorno y el medio en el que se da. De esta manera en el proceso de aprendizaje automático se pueden incluir características del dialogo como ubicación del turno dentro de la conversación y actos del dialogo del turno actual (repetir, reparar, ninguno).

En (Liscombe, Riccardi, & Hakkani-Tür, 2005) se encontró que la exactitud de clasificación muestra mejoría al incluir este tipo de información. Se puede incluso incluir información del usuario como edad, genero, lugar de origen y otros datos si se tiene acceso a ellos dependiendo del contexto de aplicación. Claramente hay algunas características contextuales que son completamente dependientes de la aplicación. Nosotros creemos que la información contextual más importante e independiente de la aplicación es el modelado de la dependencia entre muestras de habla sucesivas.

Dado que tenemos información de la naturaleza de las conversaciones contenidas en el corpus IEMOCAP, es decir sabemos el orden cronológico de las muestras de audio podemos usar esto para modelar el cambio y la influencia de muestras anteriores sobre la muestra actual. En este capítulo proponemos un método basado en campos aleatorios de Markov que utiliza información del estado emocional de muestras anteriores para hacer una corrección sobre la clasificación automática de emociones.

A diferencia de otras propuestas donde se ha aplicado campos aleatorios de Markov en el método de clasificación nosotros lo proponemos como un método de refinamiento de clasificación hecha por el módulo de clasificación/regresión como se muestra en la Figura 4 *Modelo Propuesto*.

La idea de usar CAM en nuestro trabajo es aprovechar la probabilidad de presencia de cierta emoción en muestras sucesivas de habla, donde dicha probabilidad está definida con base en un sistema de vecindad. Se supone que las propiedades físicas dentro del sistema de vecindad no cambian dramáticamente y que tienen coherencia a través del tiempo. Se supone que las propiedades acústicas, y por consiguiente emocionales, dentro de un periodo de tiempo no cambian dramáticamente y muestran coherencia a través del tiempo. Por ejemplo enojado puede cambiar gradualmente a feliz, pero no abruptamente. O en escala continua una emoción positiva se convierte gradualmente en negativa.

Con el objetivo de corroborar esta suposición calculamos la probabilidad de transición entre diferentes emociones. Para calcular estás probabilidades usamos la herramienta *CMU Statistical Language Modeling Toolkit* (Clarkson & R., 1997). El cálculo de estas probabilidades se hizo de manera análoga a como se calculan al construir un modelo de lenguaje para un reconocedor de habla, es decir, contando todas las transiciones entre muestras consecutivas de la base de datos.

Se calculó la probabilidad de cambio entre primitivas emocionales, ver Tabla 24 y Tabla 25, en las cuales el vocabulario de entrada fue el nivel de primitiva emocional discretizado en tres rangos “bajo”, “medio” y “alto”. También se calculó la probabilidad de cambio entre emociones discretas, ver Tabla 26, donde el lenguaje de entrada fueron las emociones del corpus.

Se hizo el cálculo para las bases de datos IEMOCAP y EMOWisconsin. Cómo puede observarse en la Tabla 24 y en la Tabla 25 es más probable que el valor de primitiva emocional de una muestra a otra se mantenga con el mismo valor, o pase de un valor extremo a uno intermedio en lugar de pasar de un valor de un extremo al otro extremo. Por ejemplo, en el corpus IEMOCAP es más probable pasar de Valencia baja a media (0.299) que de Valencia baja a alta (0.021).

Tabla 24 Probabilidades IEMOCAP

	Valencia			Activación			Dominación		
	Bajo	**Medio**	**Alto**	**Bajo**	**Medio**	**Alto**	**Bajo**	**Medio**	**Alto**
Bajo	0.678	0.299	0.021	0.345	0.615	0.036	0.450	0.529	0.017
Medio	0.201	0.682	0.115	0.091	0.779	0.128	0.083	0.755	0.161
Alto	0.027	0.296	0.674	0.039	0.533	0.425	0.012	0.410	0.576

Tabla 25 Probabilidades EMOWisconsin

	Valencia			Activación			Dominación		
	Bajo	**Medio**	**Alto**	**Bajo**	**Medio**	**Alto**	**Bajo**	**Medio**	**Alto**
Bajo	0.391	0.561	0.039	0.355	0.606	0.030	0.316	0.642	0.031
Medio	0.140	0.757	0.100	0.168	0.735	0.094	0.138	0.778	0.081
Alto	0.057	0.552	0.378	0.045	0.555	0.386	0.031	0.574	0.380

En la Tabla 26 también se puede observar que es más probable que dos segmentos contiguos mantengan el mismo estado emocional a que cambien a otro muy distinto. Por ejemplo, es más probable pasar de enojado a frustrado (0.217) que de enojado a feliz (0.005).

Tabla 26 Probabilidades entre emociones EMOWisconsin (E1 – Inseguro, E2 – Motivado, E3 – Nervioso, E4 – Seguro) IEMOCAP (E1 = Enojo, E2 – Excitación, E3 – Frustración, E4, Felicidad, E5 – Neutro, E6 - Tristeza)

	EMOWisconsin				IEMOCAP					
	E1	**E2**	**E3**	**E4**	**E1**	**E2**	**E3**	**E4**	**E5**	**E6**
E1	0.388	0.034	0.123	0.444	0.667	0.003	0.217	0.005	0.071	0.029
E2	0.249	0.105	0.131	0.460	0.001	0.770	0.030	0.093	0.087	0.009
E3	0.276	0.036	0.295	0.376	0.147	0.016	0.650	0.004	0.127	0.050
E4	0.238	0.039	0.123	0.594	0.001	0.016	0.021	0.635	0.111	0.058
E5					0.039	0.049	0.161	0.032	0.680	0.033
E6					0.017	0.016	0.075	0.041	0.059	0.784

La Tabla 27 y la Tabla 28 muestran las probabilidades de transición entre grupos de emociones calculadas usando la herramienta *Statistical Language Modeling Toolkit* (Clarkson & R., 1997). Como podemos observar También resulta más probable que dos muestras consecutivas se mantengan en el mismo grupo emocional.

Tabla 27 Probabilidad de transición entre grupos emocionales - IEMOCAP

	C1	C2	C3	C1	C2	C3	C4
C1	0.65	0.26	0.09	0.71	0.10	0.03	0.16
C2	0.20	0.65	0.14	0.07	0.49	0.21	0.23
C3	0.09	0.20	0.71	0.06	0.27	0.59	0.08
C4				0.13	0.27	0.05	0.54

Tabla 28 Probabilidad de transición entre grupos emocionales - EMOWisconsin

	C1	C2	C3	C1	C2	C3	C4
C1	0.50	0.09	0.40	0.36	0.04	0.38	0.21
C2	0.08	0.46	0.45	0.04	0.46	0.19	0.30
C3	0.21	0.25	0.54	0.22	0.09	0.38	0.31
C4				0.12	0.16	0.32	0.40

7.1 Método para refinamiento de clasificación emocional

A diferencia de otras propuestas donde se ha aplicado *CAM* en el método de clasificación nosotros lo proponemos como un método de refinamiento de clasificación hecha por nuestro método de clasificación. El método aquí propuesto, basado en el método propuesto en (Dutta, 2009) para análisis de imágenes, se puede definir de la siguiente manera:

$$U(\mu|d) = (1-\lambda)\sum_{i=1}^{N}\sum_{j=1}^{C}\left(\mu_{ij}{}^{m}\right)\sqrt{\sum_{l=1}^{L}\left(d_{il}-c_{jl}\right)^{2}} + \lambda\sum_{i=1}^{N}\left(1-\sum_{j=1}^{C}\sum_{k\in N_i}\beta \cdot \sqrt{\mu_{ij}\mu_{kj}}\right)$$

Formula 9 Formula para inclusión de información contextual

Donde:

$U(\mu|d)$ =Energía posterior de valores de membresía μ, de la conversación. La primera parte de la Formula es la energía de observación y la segunda la energía contextual.

λ = Peso para información contextual e información de observación.

μ_{ij} = Valor de membresía de la muestra *i* a la clase *j*. El objetivo del método de CAM-FCM es obtener μ_{ij} que minimice la energía $U(\mu|d)$.

d_{il} = El vector que especifica los valores de Valencia, Activación y Dominación estimados para la muestra *i*.

c_{jl} = Valor medio de las clases, centroide calculado por FCM.

m = Coeficiente de difusión.

β = Peso de los vecinos.

C = Número de Clases.

N = Número total de muestras en la conversación.

L = Número de primitivas emocionales, tres en nuestro caso

N_i = Los índices de los vecinos anteriores a la muestra que serán tomados en cuenta

La energía de observación consiste en calcular la distancia Euclidiana entre el centroide de cada clase emocional y cada punto emocional definido por la Valencia, Activación y Dominación de cada muestra de voz de la conversación. El resultado es una matriz de *C* (número de clases emocionales) por *N* (número de muestras de la conversación). Esta matriz es multiplicada por el valor de membresía (modificado el grado de difusión *m*) de la muestra de voz *i* en el *cluster j*. Este cálculo se hace para todas las muestras de la conversación y todas las clases emocionales y se suman todos los resultados.

La energía contextual toma en cuenta la raíz cuadrada del valor de membresía de la muestra de voz *i* en el *cluster j* multiplicado por el valor de membresía de elocuciones anteriores en el *cluster j*, esto ponderado por la cercanía de la muestra de voz actual hacía las muestras anteriores dentro de un vecindario definido. Este cálculo se hace para las todas las clases emocionales, por cada muestra de voz del vecindario y se suman todos los resultados. El complemento de esta sumatoria es calculado para cada muestra de la conversación. Este cálculo se hace para todas las muestras de la conversación y se suman todos los resultados.

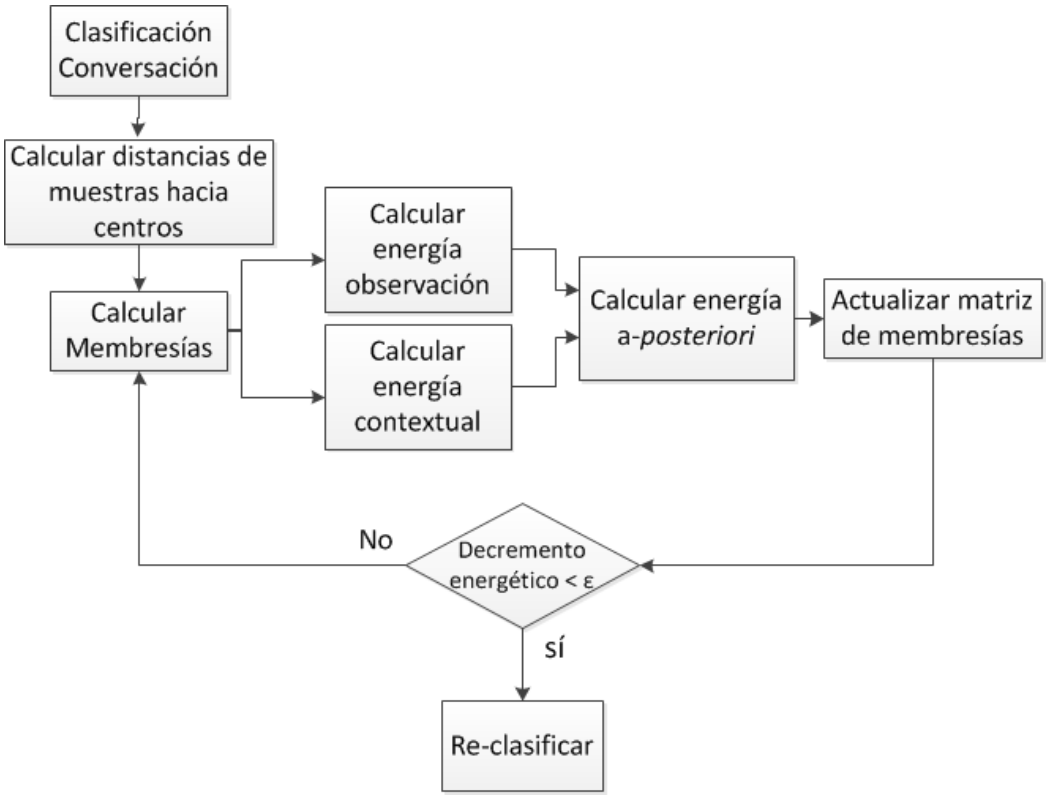


Figura 27 Proceso de reclasificación de segmentos de acuerdo a su contexto en la conversación

La energía contextual mide el grado de variación entre la muestra actual con las muestras anteriores. Si el estado emocional varía mucho de una muestra a otra, la energía contextual es menor que cuando las muestras conservan la misma emoción durante la conversación.

7.2 Algoritmo de optimización

Se usa el esquema de recocido simulado optimizado mediante un algoritmo Metrópolis propuesto en (Dutta, 2009) para alcanzar la convergencia del proceso de reclasificación de muestras mediante la minimización de la función de energía posterior. Este método genera aleatoriamente nuevas configuraciones de membresía a cada clase de cada muestra de voz en la conversación y calcula la función de energía asociada con la nueva configuración. Si la nueva configuración tiene menor energía que la previa la nueva es aceptada. Si la nueva configuración tiene mayor energía que la previa la nueva es aceptada con una probabilidad que decrece con el incremento en la diferencia de la energía de la configuración nueva y previa. El algoritmo inicia con una temperatura alta y decrece de acuerdo a una función de enfriamiento. Cuando el sistema no tiene un decremento significativo en la temperatura el algoritmo termina.

En el Algoritmo 6 se muestra a detalle los pasos a seguir. Se calcula la matriz de membresía de cada muestra clasificada de acuerdo al número de clases. En cada iteración, para cada muestra de voz en la conversación se modifica aleatoriamente su configuración de grado de membresía a cada clase y se calcula la diferencia de energías entre la configuración de membresías previa y la configuración de membresías nueva. Si la diferencia es positiva se reemplaza la configuración de membresías previa por la nueva. En caso que no sea positiva la nueva configuración puede ser aceptada con una cierta probabilidad definida por una constante. Se disminuye la temperatura del sistema hasta alcanzar la convergencia.

Algoritmo 6 Algoritmo de optimización

Entrada: **emociones_conversacion** es un vector bidimensional que contiene los valores de Valencia, Activación y Dominación y la clasificación en orden cronológico de las muestras de una conversación, **C** número de clases, **primitivas_emocionales** valores de Valencia, Activación y Dominación de cada muestra

Salida: **nuevas_emociones_conversacion** es un vector unidimensional que contiene la clasificación emocional nueva

Variables locales: **U** matriz de membresía de las muestras, **U_nueva** matriz de membresías modificada aleatoriamente, **energia_U** la energía posterior calculada sobre la matriz de membresía antes de la modificación aleatoria, **energia_U_nueva** la energía posterior calculada sobre la matriz de membresía después de la modificación aleatoria, **T** temperatura del sistema, **p** probabilidad con la cual puede ser aceptado o no la modificación al grado de membresía

Constantes: **constante_de_disminucion** es la proporción en la cual va disminuyendo la temperatura en cada iteración

Requiere: emociones_conversacion $\neq \emptyset$

```
1   U ← calcula_matriz_membresias(emociones_conversacion, C);
2   for iter ← 1 to max_iteraciones do
3       for n ← 1 to numero_muestras(emociones_conversacion) do
4           U_nueva ← rand(U,n);
5           energía_U ← energía_posterior(U);
6           energía_U_nueva ← energía_posterior (U_nueva);
7           diferencia ← energía_U_nueva - energía_U;
8           p ← min(1 , exp^(−energia_U_nueva) / exp^(−energia_U));
9           if diferencia > 0 then
10              U ← U_nueva;
11          else if rand(1) > p
12              U ← U_nueva;
13      T ← T * constante_de_disminucion;
14  for i ← 1 to numero_muestras(U) do
15      nuevas_emociones_conversaion [i] ← max(U[1:C,i]);
16  return nuevas_emociones_conversaion;
```

7.3 Resultados de inclusión de información contextual

Para probar la eficacia del método para refinamiento de clasificación emocional utilizamos dos enfoques. En el primero trabajamos con emociones discretas seleccionadas de la base de datos IEMOCAP. En este enfoque se probó seleccionando diferente número de emociones. En el caso de cuatro clases las emociones fueron, Alegría, Enojo, Tristeza y Neutro, para 5 clases se añadió Frustración y para seis clases se añadió Excitación. Se eligieron esas emociones por ser las clases con mayor representación en el corpus IEMOCAP. En el caso de las pruebas con primitivas emocionales se usaron todas las muestras disponibles.

La aplicación del método para refinamiento de clasificación emocional se hizo conversación por conversación. En la Tabla 29 se muestran los valores usados en cada variable para estos experimentos.

Tabla 29 Variables usadas en los experimentos con CAM

Variable	Valor
$\wedge$	0.5
B	[0.5, 0.3, 0.1]
M	2

Los resultados mostrados en la Tabla 30 y Tabla 31son el promedio de los resultados de todas las conversaciones contenidas en el corpus.

Como podemos ver en la Tabla 30, para el caso de emociones discretas la clasificación mejora al aplicar el método, reflejándose en un aumento del F-measure calculado después de aplicar el método con relación al F-measure calculado antes de aplicarlo. Sin embargo, en el caso de primitivas emocionales, Tabla 31 , el promedio total del F-Measure no mejoró. Aunque visualmente hay una mejoría en la clasificación emocional, como en el ejemplo que se muestra en la Figura 30, la evaluación objetiva usando las categorías del etiquetado manual no mejora.

Tabla 30 Resultado de incorporación de información contextual usando emociones discretas

Promedio	4 Clases	5 Clases	6 Clases
Intervenciones promedio	46.33	56.62	58.5
F-Measure Original	49.25	42.61	40.01
F-Measure Final	57.31	49.47	44.38

Figura 28 Reclasificación de segmentos, el eje horizontal es el tiempo, la unidad de medición es el número de muestras, el eje vertical es la emoción.

Tabla 31 Resultado de incorporación de información contextual usando primitivas emocionales en rangos de bajo, medio y alto.

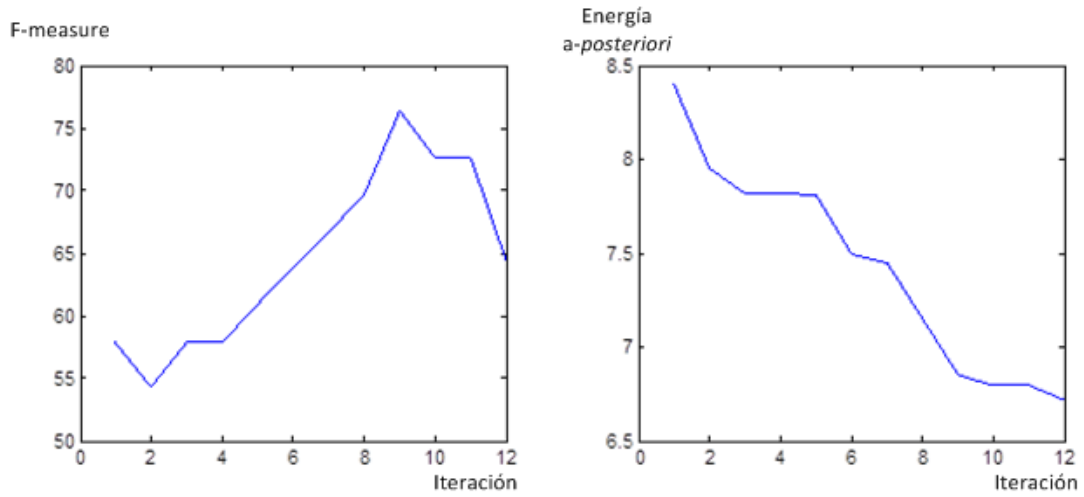
Promedio	Valencia	Activación	Dominación
Intervenciones promedio	32.62	32.62	32.62
F-Measure promedio Original	42.29/61.80	75.19/48.53	61.50/51.38
F-Measure Final	42.25/61.45	73.24/45.57	59.48/48.31

La Figura 28 muestra una conversación de más de 200 muestras en la cual se muestra en la parte superior la clasificación original y en la parte inferior la reclasificación hecha por el método. En la Figura 29 y Figura 30 se muestra el comportamiento del método para refinamiento de clasificación emocional. La base de datos IEMOCAP se compone de muchas conversaciones entre dos personas, cada conversación está segmentada en muestras que son muestras de voz de una sola persona.

En los resultados mostrados en esta sección aplicamos el método para refinamiento de clasificación emocional a cada conversación de la base de datos IEMOCAP por separado. Es decir, el contexto tomado en cuenta para modificar la clasificación de una muestra son

muestras previas y contiguas dentro de la misma conversación. Se usaron 148 conversaciones para realizar estos experimentos.

Figura 29 F-Measure (izquierda) y Evaluación basada en energía (derecha) en cada iteración



En la Figura 29 se muestra una conversación de 31 muestras que convergió en 12 iteraciones. En la gráfica de la izquierda se muestra el F-Measure medido en cada iteración. La gráfica de la derecha muestra la evaluación en energía que esa nueva clasificación produce. Se puede observar que existe correlación entre ambas gráficas, ya que mientras la energía disminuye la clasificación mejora. Sin embargo, en la segunda iteración el F-Measure disminuye y en la novena iteración el F-Measure alcanza su máximo mientras que la energía aún sigue disminuyendo lo suficiente para no alcanzar el criterio de paro. Este comportamiento de desmejora en las primeras iteraciones y el no detenerse en la mejor clasificación, se observó en varias de las conversaciones. En las gráficas mostradas como ejemplo el F-Measure original es 57.24, el máximo 76.41 y el final 64.36.

En la Figura 30 se ejemplifica visualmente la mejoría en la clasificación que produce este método sobre la clasificación inicial. Lo que se aprecia es un “aplanamiento” en algunas regiones de la conversación, por ejemplo en las muestras 1 a 5 en la clasificación automática se alternan las clases bajo y medio; mientras que en la clasificación refinada todas estas muestras se clasifican como medio, lo cual coincide mejor con la clasificación manual. Aun cuando este método siempre tiende a suavizar visualmente estas gráficas, este efecto no siempre resulta en una mejora en la clasificación. Esto se ve reflejado en los resultados promedio mostrados en laTabla 31. La razón podría provenir de la alta

subjetividad a la hora de etiquetar emocionalmente las bases de datos emocionales o podría ser también un efecto del proceso de optimización mediante recocido simulado.

Queda como trabajo futuro de este método tres puntos importantes, buscar una manera de optimizar la parametrización del método ya que resulta complicado encontrar una configuración que tenga buenos resultados para todas las conversaciones, investigar más a fondo el por qué no siempre mejora la clasificación, y buscar la manera de empatar el criterio de paro basado en la disminución de energía con el máximo en la mejora de la clasificación.

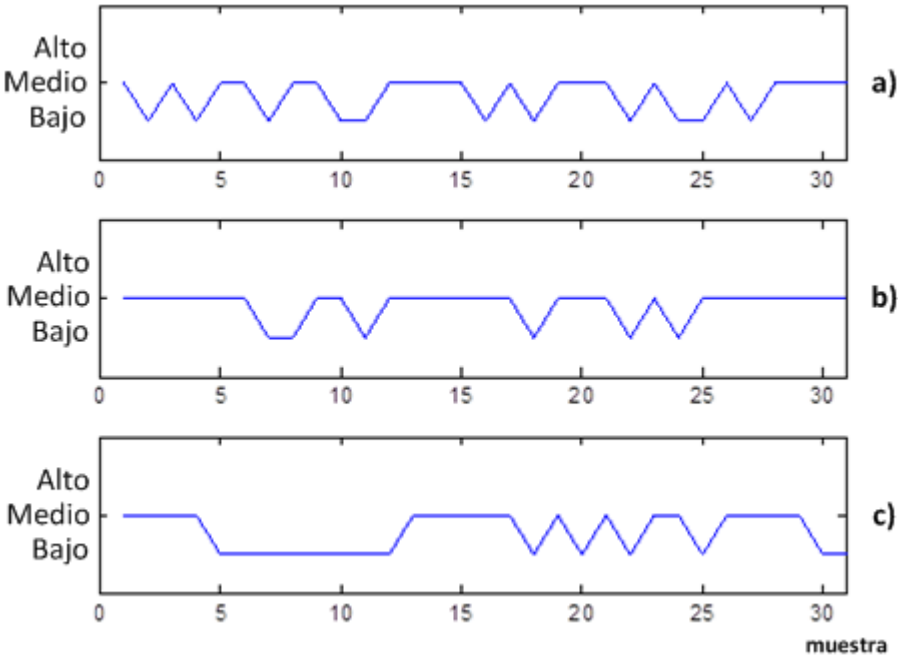


Figura 30 Clasificación en las categorías, alto medio, y bajo para Valencia. a) clasificación automática, b) clasificación corregida mediante contexto, c) clasificación manual

Capítulo 8: Evaluación General

A lo largo de este documento se han presentado diferentes evaluaciones de nuestro trabajo comparándonos con resultados de referencia propios y de otros autores, mostrando mejoría en los resultados y el aporte de nuestra investigación. Sin embargo, esas evaluaciones son parciales ya que no se evalúa el método completo. En esta sección se hace una evaluación del método completo tomando como referencia el trabajo hecho en colaboración por el *Laboratorio de Análisis e Interpretación de Señales*, de la Universidad del Sur de California y el *Instituto para la Comunicación Humano – Computadora* de la Universidad Técnica de Múnich. Cabe señalar que la investigación realizada por estos laboratorios tiene un alto impacto en el área de reconocimiento automático de emociones. Los resultados de dicha colaboración están reportados en (Metallinou, et al., 2012). Se eligió tomar como referencia dicho trabajo debido a que:

1. Usaron la base de datos IEMOCAP la cual se usó ampliamente en los experimentos presentados en esta tesis. Estos datos están públicamente disponibles.

2. Es un trabajo muy reciente (junio de 2012),

3. Sus autores son muy reconocidos en el área,

4. El enfoque propuesto es comparable en varios aspectos con el nuestro como se explica más adelante.

8.1 Antecedentes

El método propuesto en (Metallinou, et al., 2012) está inspirado en dos aspectos emocionales que nosotros también hemos estudiado en nuestro trabajo. El primero es que la expresión emocional humana tiende a evolucionar de forma estructurada dando lugar a ciertos patrones en la evolución emocional. El segundo es que la percepción de una expresión emocional puede ser afectada por manifestaciones emocionales recientes. Por lo tanto, el contenido emocional de las observaciones del pasado y el futuro podrían ofrecer contexto temporal relevante para clasificar el contenido emocional de una observación. En

dicho trabajo, los autores usan información audio-visual para reconocer el contenido emocional en la base de datos IEMOCAP. Asimismo examinan algunos métodos sensibles al contexto que consideran la evolución emocional dentro de una muestra y entre muestras en el transcurso de un diálogo.

8.2 Datos

Para esta evaluación se usaron 4,977 muestras de la base de datos IEMOCAP (Busso, et al., 2008), descrita en la sección 2.3 de este documento. Esta base de datos tiene alrededor de 10,000 muestras. Durante la grabación de las interacciones diádicas, ambos participantes portaban micrófonos, pero sólo uno de ellos portaba marcadores de movimiento facial. Dado que en (Metallinou, et al., 2012) se usa información audiovisual, para realizar sus experimentos los autores seleccionaron sólo las muestras de los participantes que portaban los marcadores de movimiento facial. Es importante señalar que para esta evaluación se usaron datos de las cinco sesiones de grabación de las que está compuesta IEMOCAP. Cuatro de estas cinco sesiones fueron liberadas recientemente, en junio de 2012. En los experimentos previos presentados en esta tesis con IEMOCAP, sólo se contaba con datos de la primera sesión de grabación que fueron liberados en abril de 2010.

8.3 Experimentos

A diferencia de nuestros experimentos anteriores, en los cuales generamos modelos de regresión para estimar los valores continuos de Valencia, Activación y Dominación a partir de características acústicas, para esta evaluación no se estimaron valores continuos, sino que se discretizaron en tres clases cada primitiva. Para Valencia y Activación la clase "bajo" contiene valores en el rango [1, 2], la clase "medio" en el rango (2, 4), y la clase "alto" en el rango [4, 5]. En (Metallinou, et al., 2012) sólo se experimentó con Valencia y Activación y propusieron dichos rangos procurando que hubiera suficientes muestras en cada rango para el entrenamiento de los modelos de clasificación. Nosotros utilizamos los mismos rangos para permitir la comparación de nuestros resultados con los del citado trabajo. En nuestro trabajo también evaluamos Dominación usando los rangos [1, 2.5] para "bajo", (2.5, 4) para "medio" y [4, 5] para "alto". Elegimos dichos rangos siguiendo la misma idea de tener un número balanceado de muestras para las tres clases. En la Tabla 32 se muestra la distribución de muestras para cada clase y primitiva. Como se puede observar

hay un fuerte desbalanceo en las clases ya que la mayor parte de las muestras de voz se localizan en el rango medio de cada primitiva.

Tabla 32 Muestras por clase

	Valencia	Activación	Dominación
Bajo	1,793	557	1,351
Medio	2,201	3,448	2,282
Alto	983	972	1,344

En la Figura 31 se muestra la manera en que están distribuidos los valores de primitivas emocionales entre cada una de las once emociones discretas contenidas en la base de datos IEMOCAP. Por ejemplo, para la categoría Enojo (Ang) la mayoría de segmentos tienen valor bajo en Valencia, Medio y Alto en Activación.

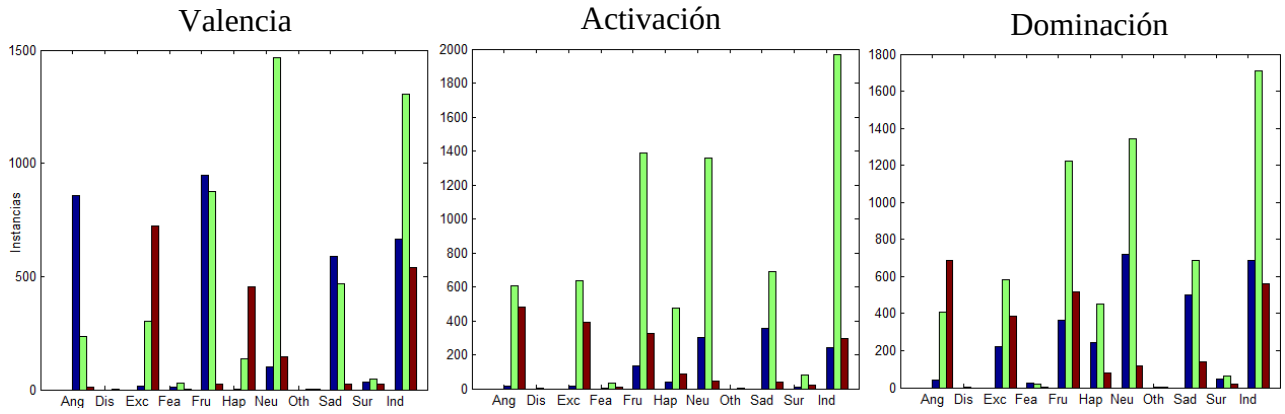


Figura 31 Distribución de rangos de primitivas por categoría emocional. Azul – Bajo, Verde – Medio, Rojo – Alto

En el enfoque propuesto en (Metallinou, et al., 2012) también se experimenta con una técnica de agrupamiento de emociones discretas de acuerdo a sus valores de primitivas emocionales, algo similar a lo que proponemos en la sección *6.2.2 Nivel de representación*

de emociones discretas de este trabajo. En la Figura 32 izquierda se muestra la manera en que se agrupan las muestras en cuatro grupos emocionales de acuerdo a su ubicación en el espacio tridimensional formado por las primitivas, Valencia, Activación, Dominación.

En la Figura 32 derecha se muestra en qué grupos se incluyeron las muestras de cada categoría emocional. Por ejemplo, en el grupo amarillo quedaron la mayoría de las muestras de enojo (Ang) y muchas de frustración. En el grupo azul quedaron casi todas las muestras de Excitación (Exc). El enfoque de agrupamiento propuesto por (Metallinou, et al., 2012) agrupa las muestras en grupos emocionales de acuerdo a su etiquetado manual.

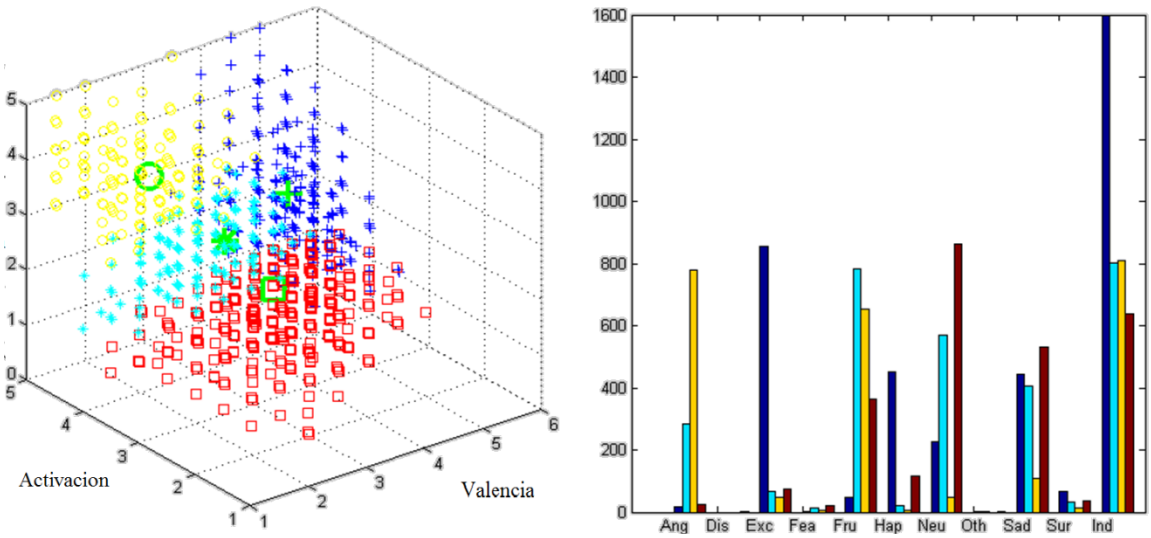


Figura 32 Agrupamiento de muestras en grupos emocionales.

El enfoque propuesto en (Metallinou, et al., 2012) también incluye el modelado de transiciones emocionales, suponiendo que ciertas transiciones emocionales son más probables que otras. Como se abordó en el *Capítulo* 7: Clasificación Basada en Contexto Emocional, nosotros hemos usado CAM para aprovechar esa suposición en el mejoramiento de la clasificación y regresión basadas en información acústica.

La selección de características acústicas y de muestras se hizo aplicando los procedimientos descritos en las secciones *5.2.1 Selección no agrupada de características* y *5.6 Selección de muestras*, respectivamente. En la Tabla 33 se muestran las características usadas por los modelos de clasificación para este experimento.

Tabla 33 Características acústicas usadas para la generación de modelos de clasificación en bajo, medio y alto

Grupo de Características	Total	Valencia	Activación	Dominación
Calidad de Voz	36	4	4	3
Tiempos	125	3	8	3
Cocleogramas	96	4	3	4
LPC	111	2	7	7
sFlux	117	4	2	1
Intensidad	129	7	4	13
Entonación	243	10	9	4
SpecMaxMin	234	1	3	0
SEB	234	4	14	9
SROP	468	8	10	7
MFCC	1,617	72	102	106
MEL	3,042	54	61	75
Total	6,920	173	227	235

8.4 Resultados

A diferencia de los resultados de inclusión de contexto mostrados en la sección *7.3 Resultados de inclusión de información contextual* donde se muestra el promedio de la evaluación de la reclasificación de todas las conversaciones, la evaluación hecha en esta sección fue hecha sobre la totalidad de los segmentos usados para estos experimentos. En la Tabla 34 y Tabla 35 se muestra la comparación de los resultados obtenidos con nuestros métodos de *5.2.1 Selección no agrupada de características, 5.6 Selección de muestras* y *Capítulo* 7: Clasificación Basada en Contexto Emocional contra los resultados obtenidos por Metallinou en (Metallinou, et al., 2012). Los resultados del citado trabajo son

comparables con los nuestros dado que nos aseguramos de usar las mismas muestras de la base de datos IEMOCAP y los mismos rangos de discretización para las primitivas emocionales.

En las dos primeras filas de la Tabla 34 se muestran los mejores resultados obtenidos por Metallinou usando solamente la información de voz y usando la información de voz, rostro y contexto respectivamente. Las siguientes filas de la Tabla 34 muestran los resultados obtenidos progresivamente con nuestros métodos. Metallinou sólo trabajo con Valencia y Activación, nosotros reportamos en la misma tabla los resultados obtenidos para Dominación.

Tabla 34 Resultado de incorporación de información contextual

	Valencia		Activación		Dominación	
	F-Measure	Cobertura	F-Measure	Cobertura	F-Measure	Cobertura
Método propuesto por Metallinou						
Solo voz	49.85	49.99	57.54	61.92	-	-
Voz + Rostro + Contexto	65.12	64.67	54.90	52.28	-	-
Método propuesto en esta tesis						
Selección de Características	57.90	54.26	73.80	54.26	56.10	53.67
Selección de Muestras + Selección de Características	58.30	55.23	75.67	56.22	60.19	57.54
Selección de Muestras + Selección de Características + Contexto	60.16	56.61	73.55	54.49	61.27	54.98

En el caso de Valencia logramos mejores resultados que Metallinou usando únicamente información de voz y mejoramos nuestros propios resultados al incluir información de contexto. Sin embargo estos resultados no fueron tan buenos como los de Metallinou al incluir, además de voz y contexto, información de rostro. Esto confirma que la Valencia es la primitiva emocional más difícil de estimar a partir de voz únicamente y que incluir información adicional, como visual en este caso, puede ayudar mucho.

En el caso de Activación, obtuvimos un F-Measure mucho mayor al de Metallinou pero una cobertura inferior al usar solamente voz. Similarmente, como se aprecia en la última fila de la columna de Activación en la Tabla 34, las medidas de evaluación en ambos trabajos disminuyeron al incluir información contextual pero en nuestro caso no fue tan grande la diferencia en comparación con Metallinou. Esto puede deberse a que

Activación es más difícil de estimar con rostro que con voz y a que al fusionar ambas fuentes de información se afecta el desempeño de usar únicamente voz.

En el caso de Dominación, también obtuvimos mejoría en el F-measure al incluir información contextual como se puede apreciar en la última fila de la columna de Dominación en la Tabla 34.

Tabla 35 Resultado de incorporación de información contextual

	3 clusters		*4 clusters*	
	F-Measure	Cobertura	F-Measure	Cobertura
Metallinou Voz + Rostro (Valencia, Activación)	67.33	66.18	56.54	56.64
Voz (Valencia, Activación)	63.90	63.50	53.90	52.80
Voz (Valencia, Activación, Dominación)	63.90	63.27	55.40	55.35

En la Tabla 35 se hace una comparación en el desempeño del agrupamiento de emociones. Metallinou no hizo una evaluación usando sólo información de voz y como se había mencionado antes, sólo usó Valencia y Activación. En la columna de *4 clusters* se puede observar que añadir Dominación permite diferenciar grupos de emociones que se confunden en el espacio de Valencia y Activación.

Lo que podemos concluir de estos experimentos en primer lugar, es que los resultados del trabajo desarrollado en esta tesis son comparables con los resultados de los mejores resultados del estado del arte. Por otro lado, también podemos concluir que el trabajo hecho en la caracterización y selección de características es muy bueno ya que usando únicamente información acústica se obtuvieron resultados muy cercanos a los obtenidos usando información de voz y rostro en el trabajo tomado como referencia para la comparación de nuestros resultados.

Capítulo 9: Resumen, Conclusiones y Trabajo Futuro

9.1 Resumen del trabajo realizado

En esta sección resumimos el trabajo realizado en esta tesis, los resultados obtenidos se discuten más adelante. El reconocimiento automático de emociones en voz es un área que ha sido estudiada desde hace más de una década. A pesar de que se han obtenido buenos resultados modelando emociones prototípicas, ha sido difícil usar estos modelos en aplicaciones del mundo real, ya que cotidianamente las emociones no son tan marcadamente distinguibles como en las colecciones de datos usadas para crear dichos modelos. Debido a esto, recientemente, ha surgido una tendencia en el estudio de modelos emocionales continuos, con el objetivo de hacer más robusto y flexible el reconocimiento de emociones. Además se ha puesto énfasis en generar colecciones de datos con un contenido emocional más genuino y espontáneo.

En este trabajo proponemos métodos de reconocimiento de emociones en voz que intentan sacar provecho del modelado emocional continuo con el objetivo de facilitar la portabilidad de los modelos de reconocimiento a diferentes aplicaciones del mundo real, independientemente del conjunto de emociones que se desee reconocer. Nuestra hipótesis es que mediante una estimación adecuada de las primitivas emocionales Valencia, Activación y Dominación se puede obtener información suficiente para determinar una gran cantidad de emociones, incluyendo emociones prototípicas, mezcla de emociones, diferentes intensidades o matices de emociones y grupos de emociones semejantes.

De acuerdo a lo anterior, en esta investigación estudiamos ampliamente métodos para selección de atributos con el objetivo de encontrar las características acústicas más relevantes para determinar los niveles de Valencia, Activación y Dominación en la voz. Las propiedades acústicas de la voz analizadas en este trabajo se pueden catalogar como Prosódicas, Espectrales y de Calidad de Voz.

Exploramos dos enfoques de extracción de características. Primero, el enfoque selectivo, mediante el cual diseñamos un conjunto de características de acuerdo a una

revisión del estado del arte y propusimos algunas características que consideramos podrían aportar información paralingüística importante, es decir, información del cómo se dicen las cosas. En el segundo, el enfoque por fuerza bruta, extrajimos una gran cantidad de características relacionadas con procesamiento de voz con la idea de descubrir a los mejores.

Un problema encontrado durante el desarrollo de este trabajo fue la disponibilidad de datos. Existen varias bases de datos anotadas con emociones discretas, pero pocas con primitivas emocionales y de esas ninguna en español. Además, algunas de las bases de datos más usadas para la experimentación en reconocimiento de emociones en voz no contienen emociones genuinas y espontáneas. Por lo tanto, nos dimos a la tarea de generar una base de datos propia, en español, con emociones inducidas, anotada con categorías y primitivas emocionales. Esta base de datos nos permitió hacer un estudio de características acústicas más completo, probando con tres idiomas en tres contextos de obtención de datos diferentes.

Exploramos el uso de información complementaria a la información acústica. Probamos la inclusión de información contextual que se obtiene tomando en cuenta el estado emocional de muestras anteriores durante la misma conversación. Propusimos métodos basados en agrupamiento difuso para representar emociones basándose en la estimación e interpretación de primitivas emocionales.

En general los métodos propuestos obtienen resultados superiores a los reportados en el estado del arte, además de que nuestros métodos ofrecen otras ventajas adicionales como flexibilidad y robustez. Por lo tanto, la hipótesis planteada fue corroborada y los objetivos propuestos fueron alcanzados.

El resto de este capítulo resume el trabajo realizado, y describe las contribuciones de nuestra investigación, las conclusiones derivadas de ella y tópicos de investigación que pueden ser explorados en trabajo futuro. Con el fin de facilitar la lectura de este documento, hemos dividido los aspectos anteriores en secciones independientes.

El trabajo y aportes realizados en esta tesis se pueden agrupar principalmente en dos grandes etapas. La primera es la selección de características acústicas y el segundo el modelado emocional multinivel.

9.1.1 Selección de características

El principal problema para encontrar las mejores características acústicas fue que estuvimos trabajando con un gran número de ellas, casi 7,000, y no con tantas muestras, alrededor de entre 900 y 2,500. Este factor dificultó el buen desempeño de varios métodos de sección de características que probamos. Se propusieron dos maneras de evitar hacer la búsqueda sobre todas las muestras al mismo tiempo y se usó el método *Linear Forward Selection*, que mostró ser adecuado para este tipo de problemas.

Se analizó, mediante el cálculo de tres métricas, el aporte de los grupos de características extraídas de la señal. El análisis fue hecho desde el punto de vista del modelado de emociones continuas, a diferencia de trabajos previos en selección de características para reconocimiento de emociones, los cuales se habían hecho desde el punto de vista del modelado discreto. Este análisis se hizo sobre tres bases de datos, una en inglés, otra en español y otra en alemán, lo cual permitió validar con mayor certeza la relevancia de diferentes grupos de características en el reconocimiento de emociones en la voz humana.

9.1.2 Modelado emocional multinivel

Para el modelado emocional multinivel propusimos un método de reconocimiento de emociones el cual representa fenómenos emocionales con un enfoque más apegado a como ocurren estos fenómenos en la vida cotidiana, dónde las emociones frecuentemente se presentan como una mezcla de emociones y con diferentes intensidades. Para hacer este modelado se usó principalmente agrupamiento difuso mediante *FCM*. Dicho modelado está pensado en diferentes potenciales aplicaciones del reconocimiento automático de emociones.

También se trabajó en una etapa de post-procesamiento de estimación emocional, en la cual se usa la técnica conocida como *Campos aleatorios de Markov* con el objetivo de incluir información del contexto emocional en conversaciones bajo la suposición de que las emociones expresados por las personas no cambian bruscamente de un instante al siguiente.

9.2 Contribuciones

Las principales contribuciones de este trabajo son las siguientes:

- Se contribuyó en el estudio y comprensión de la aplicación de modelos emocionales continuos en el reconocimiento automático de emociones mediante el desarrollo de un método de reconocimiento de emociones basado en el uso de las primitivas emocionales Valencia, Activación y Dominación que representa emociones en tres niveles de abstracción

- Se contribuyó en estudio de la aplicación de técnicas de pre-procesamiento de voz en presencia de ruido y amplitud variante mediante el desarrollo de un método de pre-procesamiento de voz basado en compresión/expansión de audio para mejorar el audio en grabaciones de conversaciones diádicas.

- Se contribuyó el estudio del desempeño de diferentes tipos de características acústicas en el reconocimiento de emociones mediante el diseño de un conjunto de características mediante el enfoque selectivo.

- Se contribuyó en el estudio de técnicas de descubrimiento de mejores características cuando se cuenta con pocas muestras mediante el desarrollo de dos maneras de selección de características combinadas con el método.

- Se contribuyó al área de reconocimiento de emociones en voz con un conjunto de grupos de características emocionales para estimar cada una de las primitivas emociones basándose en un estudio de características. Fuimos los primeros en hacer un análisis de este tipo.

- Se contribuyó en la generación de recursos para el estudio de reconocimiento de emociones en voz con la primera base de datos emocional en español con emociones naturales etiquetadas con primitivas emocionales.

- Se contribuyó con un método heurístico de selección de muestras que es útil cuando se cuenta con etiquetado continuo y discreto y podría ser usado en otras bases de datos que involucren cierto grado de subjetividad en el etiquetado.

- Se contribuyó en la comprensión de las propiedades universales de las emociones mediante el diseñó de un conjunto de grupos de características para encarar el reconocimiento multilingüe de emociones basado en modelos continuos.

- Se contribuyó en la explotación de los modelos emocionales continuos mediante el desarrollo un método de representación emocional multinivel de emociones.

- Se contribuyó en el estudio y explotación de la evolución emocional en conversaciones mediante el desarrollo de un método de incorporación de información contextual

9.3 Conclusiones

Las conclusiones más importantes a las que llegamos a lo largo de esta investigación son:

- Encontramos que estimar Valencia es muy difícil usando únicamente información de voz. Para mejorar dicha estimación es necesario incluir información complementaria, como lingüística, contextual o incluso de otra fuente como puede ser visual.

- Encontramos a través del estudio de características acústicas que Activación y Dominación comparten varias propiedades acústicas ya que a través de los experimentos de selección de características y de regresión realizados mostraron comportamientos similares. Esto nos lleva a la conclusión que no siempre son ejes independientes y existe cierta correlación entre ellos.

- La creación de bases de datos de emociones espontaneas y su respectivo etiquetado emocional es una tarea compleja ya que existe una gran subjetividad lo que dificulta lograr un acuerdo alto entre evaluadores.

- A través de una serie de experimentos con tres bases de datos en inglés, alemán y español pudimos comprobar que el reconocimiento de emociones multilingüe es posible aunque menos preciso que el monolingüe.

- Encontramos que las mismas características que sirven para discriminar emociones en la voz que aportan información para un idioma también lo hacen para los otros.

- Encontramos que los patrones de características acústicas encontrados por los algoritmos de clasificación que muestran buen desempeño en la clasificación emocional en un idioma disminuyen considerablemente su precisión en otro idioma.

- Comprobamos que es valioso considerar información de otras fuentes para complementar la información paralingüística ya que existen fenómenos emocionales no tan distinguibles a través de la voz.

9.4 Trabajo futuro

Como trabajo futuro quedan las siguientes tareas:

1. Implementar un método de optimización de parámetros para el módulo de reclasificación de segmento ya que son alrededor de una decena y es complicado encontrar los valores adecuados para cada uno de ellos.

2. Incorporar otras fuentes de información para mejorar la precisión de los modelos. Dado que los modelos propuestos están basados en el cálculo de primitivas emocionales, es fácil incorporar otras fuentes de información como las incluidas en la base de datos IEMOCAP, información de movimientos faciales, corporales, gestos.

3. Probar las características y modelos desarrollados para detectar y clasificar fenómenos paralingüísticos relacionados, como personalidad, postura, estados de ánimo.

4. Probar con otros idiomas no relacionados con los idiomas probados en esta tesis; tentativamente un idioma no perteneciente a la familia Indo-Europea, como Mandarín, Árabe o Japonés.

5. Estudiar qué otros aspectos, además del idioma, pueden modificar las propiedades acústicas en la expresión emocional como edad, sexo y otros tipos de información contextual.

6. Probar nuestros métodos en datos relacionados con asistencia a pacientes y apoyo a psicólogos y neurólogos, dado que la estimación de mezcla e intensidad emocional tiene aplicación principalmente en esas áreas.

7. Profundizar en caracterización de aspectos prosódicos del habla ya que intuitivamente deberían proveer mayor información de la que mostraron en nuestros experimentos.

Bibliografía

Abrilian, S.; Devillers, L.; Buisine, S.; Martin. (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. *11th international conference human-computer interaction.*

Arnrich, B., Setz, C., La Marca, R., Gerhard, T., & Ehlert, U. (2010). Self Organizing Maps for Affective State Detection. *Machine Learning for Assistive Technologies.*

Averill, J. (1990). Inner feelings, works of the flesh, the beast within, diseases of the mind, driving force, and putting on a show: Six metaphors of emotion and their theoretical extensions. In L. D. E. (Ed.), *Metaphors in the history of psychology* (pp. 104-132). New York: Cambridge University Press.

Batliner, A., Fischer, K., Humber, K., Spliker, J., & Nöth, E. (2003). How to find trouble in communication. *Speech Commun, 40, 1-2*, 117-143.

Batliner, A., Steidl, S., & Noeth, E. (2008). Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus. *Satellite Workshop of LREC 2008 on Corpora for Research on Emotion and Affect.*

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., . . . Amir, N. (2011). Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech. *Comput. Speech Lang., 25*(1), 4-28.

Beale, R., & Peter, C. (2008). The role of affect and emotion in hci. In R. Beale, & C. Peter (Eds.), *Affect and Emotion in Human-Computer Interaction: From Theory to Applications.: LNCS* (1 ed., pp. 1-11). Heidelberg: Springer-Verlag.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International 5:9/10*, (pp. 341-345).

Boril, H., Sadjadi, O., Kleinschmidt, T., & Hansen, J. H. (2010). Analysis and Detection of Cognitive Load and Frustration in Drivers' Speech. *Interspeech'10*, (pp. 502-505). Makuhari, Chiba, Japan.

Bozkurt, E., Erzin, E., Erdem, C. E., & Erdem, T. (2009). Improving Automatic Emotion Recognition from Speech Signals. *Interspeech 2009.* Brighton.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A Database of German Emotional Speech. *Interspeech.* Lissabon.

Busso, C., Bulut, M., Lee, C., Kazemzadeh, A., Mower, E., Kim, S., . . . Narayanan, S. S. (2008, December). IEMOCAP: Interactive emotional dyadic motion capture database. *Journal of Language Resources and Evaluation, 42*(4), 335-359.

Byrne, W. a. (1989). The auditory processing and recognition of speech. *Proceedings of the workshop on Speech and Natural Language* (pp. 325--331). Cape Cod, Massachusetts: Association for Computational Linguistics.

Cacioppo, J. T., & Berntson, G. G. (1994). Relationship between attitudes and evaluative space: A critical review, with emphasis on the separability of positive and negative substrates. *Psychological Bulletin, 115*, 401-423.

Chapelle, O., Scholkopf, B., & Zien, A. (2006). *Semi-Supervised Learning.*

Chávez Garcia, R. O. (2010). *Ordenamiento de imágenes recuperadas utilizando un enfoque de fusión de información multimodal.* Tonantzintla.

Chellappa, R., & Jain, A. K. (1993). *Markov random fields: theory and application.* Academic Press.

Clarkson, P., & R., R. (1997). Statistical Language Modeling Using the CMU-Cambridge Toolkit. *ESCA Eurospeech.* Rhodos, Greece.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, E., & Schrder, M. (2000). An instrument for recording perceived emotion in real time. *ISCA Workshop on Speech and Emotion*, (pp. 19-24).

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., & Schröder, M. (2000). FEELTRACE: An instrument for recording perceived emotion in real time. *ISCA Tutorial and Research Workshop on Speech and Emotion*, (pp. 19-24). Newcastle, Northern Ireland.

Cronbach, L. (1951). Coe cient alpha and the internal structure of tests. *Psychometrika*, 297-334.

Darwin, C. (1872). *The Expression of the Emotions in Man and Animals* (3 ed.). (J. Murray, & P. Ekman, Eds.) London: Oxford University Press.

Devillers, L., & Martin, J.-C. (2008). Emotional Events in Audiovisual Corpora. *LREC'08*, (pp. 1259-1265).

Devillers, L., & Vidrascu, L. (2006). Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs. *Interspeech.* Pittsburgh.

Douglas-Cowie, E., Cowie, R., Sneddon, C. C., Lowry, M. R., Martin, J.-C., Devillers, L., & Batliner, A. (2007). The HUMAINE Database: addressing the needs of the affective computing community. In A. P. Picard (Ed.), *2nd International Conference on Affective Computing and Intelligent Interaction. 4738*, pp. 488-500. Lisbon, Portugal: Springer, LNCS.

Drioli, C., Tisato, G., Cosi, P., & Tesser, F. (2003). Emotions and Voice Quality: Experiments with Sinusoidal Modeling. *VOQUAL'03*, (pp. 127-132).

Dubuisson, T., Dutoit, T., Gosselin, B., & Remacle, M. (2009). On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination. *EURASIP Journal on Advances in Signal Processing, Analysis and Signal Processing of Oesophageal and Pathological Voices, 10.1155/2009/173967*.

Dumouchel, P., Dehak, N., Attabi, Y., Dehak, R., & Boufaden, N. (2009). Cepstral and Long-Term Features for Emotion Recognition. *Interspeech 2009.* Brighton, UK.

Dutta, A. (2009). *Fuzzy c-Means Classification of Multiespectral data incorporating Spatial Contextual information by using Markov Random Field.* Enschede, The Netherlands: International Institute for Geo-Information science and earth observation .

Ekman, P. (1972). *Universals and cultural differences in facial expressions of emotion.* (J. Cole, Ed.) Lincoln: University of Nebraska Press.

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion, 6(3/4)*.

Elfenbein, H. A. (2002). Cross-Cultural Patterns in Emotion Recognition: Highlighting Design and Analytical Techniques. *Emotion, 2*(1), 75 - 84.

Esau, N., Kleinjohann, L., & Kleinjohann, B. (2005). An Adaptable Fuzzy Emotion Model for Emotion Recognition. *EUSFLAT- LFA.*

Eyben, F., Wöllmer, M., & Schuller, B. (2009). openEAR - Introducing the Munich Open-Source Emotion and Affect Recognition Toolkit. *Proc. 4th International HUMAINE Association Conference on Affective Computing and Intelligent Interaction 2009.* Amsterdam.

Eyben, F., Wöllmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., & Cowie, R. (2010, march). On-line emotion recognition in a 3-D activation-valence-time continuum

using acoustic and linguistic cues. *Journal on Multimodal User Interfaces, 3*(1-2), 7-19.

Fell, H. J., & MacAuslan, J. (2003). Automatic Detection of Stress in Speech. *MAVEBA.* Florence, Italy.

Forbes-Riley, K., & Litman, D. J. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. *Proceedings of the Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL 2004)*, (pp. 201-208).

Gabrielli, S., Mayora, O., Bardram, J., & Marcu, G. (2010). Co-Designing Personal HealthCare Solutions for the Treatment of Bipolar Disorder. *NordiCHI 2010 Workshop on Therapeutic Strategies - a Challenge for User Involvement in Design.*

Giripunje, S., & Bawane, N. (2007). ANFIS Based Emotions Recognision in Speech. In N. Bawane (Ed.), *Knowledge-Based Intelligent Information and Engineering Systems* (Vol. 4692, pp. 77-84). Springer Berlin / Heidelberg.

Gómez, R. (1971). *República, Introducción, traducción y notas.* UNAM, Bibliotheca Scriptorum Graecorum et Romanorum Mexicana.

González, G. (1999). *Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in chicanos/latinos.* University of Michigan.

Grant, D., & Berg, E. (1948). A behavioral analysis of degree of reinforcement and ease of shifting to new responses in a weigl-type card-sorting problem. 404-411.

Gravier, G., Sigelle, M., & Cholle, G. (1998). Toward Markov random field modeling of speech. *ICSLP.*

Grimm, M., & Kroschel, K. (2005). Evaluation of natural emotions using self assessment manikins. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, (pp. 381-385). San Juan, Puerto Rico.

Grimm, M., Kroschel, K., & Narayanan, S. (2007). Support vector regression for automatic recognition of spontaneous emotions in speech. *Acoustics, Speech and Signal Processing. ICASSP 2007. IEEE International Conference on*, (pp. IV-1085 - IV-1088).

Grimm, M., Kroschel, K., Mower, E., & Narayanan, S. (2007). Primitives-based evaluation and estimation of emotions in speech. *Speech Communication, 49(10-11)*, 787-800 .

Gunes, H., Schuller, B., Pantic, M., & Cowie, R. (2011). Emotion Representation, Analysis and Synthesis in Continuous Space: A Survey. *Proceedings of IEEE International Conference on Automatic Face and Gesture Recognition (FG'11), EmoSPACE 2011 - 1st International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous spacE.* Santa Barbara, CA, USA.

Gutlein, M., Frank, E., Hall, M., & Karwath, A. (2009). Large-scale attribute selection using wrappers. *Proc IEEE Symposium on Computational Intelligence and Data Mining* (pp. 332-339). IEEE.

Hall, M. A. (1998). *Correlation-based Feature Subset Selection for Machine Learning.* Hamilton, New Zealand.

Herm, O., Schmitt, A., & Liscombe, J. (2008). When calls go wrong: how to detect problematic calls based on log-files and emotions? *INTERSPEECH-2008*, (pp. 463-466).

Hernández, Y., Sucar, L. E., & Conati, C. (2008). An Affective Behaivor Model for Intelligent Tutors. *Intelligent Tutoring Systems (ITS) LNCS, 5091*, 819-821.

Hillsdale, N., & Erlbaum. (1998). Appendix F. Labels describing affective states in five major languages. (K. R. Scherer, Ed.) *Facets of emotion: Recent research*, 241-243.

Huang, C. F., & Akagi, M. (2005). A multi-layer fuzzy logical model for emotional speech. *Eurospeech.*

Iriondo, I. (2008). Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva. *Tesis Doctoral*.

Ishi, C. T., Ishiguro, H., & Hagita, N. (2005). Proposal of acoustic measures for automatic detection of vocal fry. *Interspeech*, (pp. 481-484). Lisbon, Portugal.

James, W. (1884). What is an emotion? *Mind, 19*, 188-205.

James, W. (1890). chapter on Emotion. In *Writings 1879-1899* (pp. 350-365). The library of America.

Ju, E., & Lee, J. (2008). Expressive Facial Gestures From Motion Capture Data. (G. D. Scopigno, Ed.) *EUROGRAPHICS 2008, 27*(2).

Kalyani S., S. K. (2010). Supervised fuzzy C-means clustering technique for security assessment and classification in power systems. *International Journal of Engineering, Science and Technology, 2*(3).

Kandali, A., Routray, A., & Basu, T. (2009). Vocal emotion recognition in five native languages of Assam using new wavelet features. *International Journal of Speech Technology*.

Kehrein, R. (2002). The prosody of authentic emotions. *Speech Prosody Conference*, (pp. 423-426).

Kim, E. H., Kwak, S. S., Hyun, K. H., Kim, S.-H., & Kwak, Y. K. (2009). Design and Development of an Emotional Interaction Robot, Mung. *Advanced Robotics, 23*(6), 767-784.

Kira, K., & Rendell, L. (1992). A Practical Approach to Feature Selection. *Ninth International Workshop on Machine Learning*, (pp. 249-256).

Kockman, M., Burget, L., & Cernocký, J. (2009). Brno Unicersity og Technology System fir Interspeech 2009 Emotion Challenge. *Interspeech.* Brighton, U.K.

Kostoulas, T., Ganchev, T., & Fakotakis, N. (2008). Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data. In *Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction: COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007.* (pp. 235-242). Berlin, Heidelberg: Springer-Verlag.

Lang, P. J. (1980). Behavioral treatment and bio-behavioral assessment: Computer applications. (J. B. Sidowski, Ed.) *Technology in Mental Health Care Delivery Systems*.

Larsen, J. T., & McGraw, A. P. (2011, Jun). Further evidence for mixed emotions. *Journal of Personality and Social Psychology, 100*(6), 1095-1110.

Larsen, J. T., To, Y. M., & Fireman, G. (2007). Children's understanding and experience of mixed emotions. *Psychological Science, 18*, 186-191.

Lee, C. M., & Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. *ICSLP*. Denver, CO, USA.

Lee, C.-C., Mower, E., Busso, C., Lee, S., & Narayanan, S. (2009). Emotion Recognition Using a Hierarchical Binary Decision Tree Approach. *Interspeech.* Brighton, U.K.

Li, W., Zhang, Y., & Fu, Y. (2007). Speech Emotion Recognition in E-learning System Based on Affective Computing. 809-813.

Li, Z. S. (1994). Markov random field models in computer vision. *Lecture Notes in Computer Science, 801*, 361-370.

Lichtenstein, A., Oehme, A., Kupschick, S., & Jürgensohn, T. (2008). Comparing Two Emotion Models for Deriving Affective States from Physiological Data. (C. Peter, & R. Beale, Eds.) *Affect and Emotion in Human-Computer Interaction: From Theory to Applications.: LNCS*, 35-50.

Liebowitz, M. (1983). *The chemistry of love.* Boston: Little, Brown.

Liscombe, J., Riccardi, G., & Hakkani-Tür, D. (2005). Using context to improve emotion detection in spoken dialog systems. *Eurospeech.* Lisboa.

Luengo, I., Navas, E., & Hernáez, I. (2009). Combining spectral and prosodic information for emotion recognition. *Interspeech.* Brighton, U.K.

Luengo, I., Navas, E., & Hernandez, I. (2005, sept.). Reconocimiento automático de emociones utilizando parámetros prosódicos. *Procesamiento del lenguaje natural, 35*, 13-20 1135-5948.

Lugger, M., & Yang, B. (2006). Classification of Different Speaking Groups by Means of Voice Quality Parameters. *Vorträge der ITG-Fachtagung.* Kiel, Germany.

Lugger, M., & Yang, B. (2007). An incremental analysis of different feature groups in speaker independent emotion recognition. *Proceedings of the International Conference on Phonetic Sciences*, (pp. 2149-2152).

Lugger, M., & Yang, B. (2008). Cascaded Emotion Classification via Psychological Emotion Dimensions Using Large Set of Voice Quality Parameters. *IEEE ICASSP*, (pp. 4945-4948). Las Vegas, USA.

Lutz, C., & Miles White, G. (2001). *The Anthropology of Emotions.*

Metallinou, A., Lee, S., & Narayanan, S. S. (2010). Decision level combination of multiple modalities for recognition and analysis of emotional expression. (pp. 2462-2465). Dallas, TX.: IEEE.

Metallinou, A., Wollmer, M., Katsamanis, A., Eyben, F., Schuller, B., & Narayanan, S. (2012, April-June). Context-Sensitive Learning for Enhanced Audiovisual Emotion Classification. *Affective Computing, IEEE Transactions on, 3*(2), 184 - 198.

Montero, J. (2003). *Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano.* Madrid.

Morales, E., & González, J. (02 de 06 de 2009). *Aprendizaje 2*. Obtenido de http://ccc.inaoep.mx/~emorales/Cursos/Aprendizaje2/principal.html

Narayanan, S., Grimm, M., & Kroschel, K. (2008). The vera am mittag german audio-visual emotional speech database. *ICASSP*. Las Vegas, Nevada, U.S.A.

Nasoz, F., Alvarez, K., & Lisetti, L. (2004). Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cogn. Technol. Work, 6*(1), 4-14.

Núñes B., F., Corte S., P., Suarez N., C., Señaris G., B., & Sequeiros, G. (2004). Evaluación perceptual de la disfonía: correlación con los parámetros acústicos y fiabilidad. *Acta otorrinolaringológica española: Organo oficial de la Sociedad española de otorrinolaringología y patología cérvico-facial, 55*(6), 282-287.

Nyhus, E., & Barcelo, E. (2009). The wisconsin card sorting test and the cognitive assessment of prefrontal executive functions: a critical update. *Brain Cogn, 71*(3), 437-451.

Ortony, A., Clore, G. L., & Collins, A. (1988). *The Cognitive Structure of Emotions.* Cambridge: Cambridge University Press.

Osgood, C., May, W., & Miron, M. (1975). *Cross-cultural Universals of Affective Meaning.* University of Illinois Press.

Panat, A. R., & Ingole, V. T. (2007). Affective State Analysis of Speech for Speaker Verification: Experimental Study, Design and Development., *1*, pp. 255-261.

Pérez Espinosa, H., & Reyes García, C. A. (2009). Detection of Negative Emotional State in Speech with Anfis and Genetic Algorithms. *6th International Workshop on Models and Analysis of Vocal Emissions for Biomedical Applications (MAVEBA2009).* Firenze, Italy.

Pérez Espinosa, H., & Reyes García, C. A. (2010). *Reconocimiento de Emociones a Partir de Voz Basado en un Modelo Emocional Continuo.* Tonantzintla, Puebla.: INAOE.

Pérez Espinosa, H., Reyes García, C. A., & Villaseñor Pineda, C. A. (2009). Selección de Atributos para la Estimación de Primitivas en Habla Emocional. *X Encuentro de Investigación* (pp. 207-210). Tonantzintla, Puebla: INAOE.

Pérez Espinosa, H., Reyes García, C. A., & Villaseñor Pineda, L. (2010). Features Selection for Primitives Estimation on Emotional Speech. *IEEE International Conference on Acoustics, Speech and Signal Processing*, (pp. 5138-5141). Dallas, Texas.

Pérez Espinosa, H., Reyes García, C. A., & Villaseñor Pineda, L. (2010). Selección de Atributos Acústicos para Estimación de Primitivas Emocionales en Voz. In M. González Mendoza, & O. Herrera Alcantara (Eds.), *Avances en sistemas inteligentes en México. Sociedad Mexicana de Inteligencia Artificial* (pp. 151-160). Tlaxcala: Sociedad Mexicana de Inteligencia Artificial.

Pérez Espinosa, H., Reyes García, C. A., & Villaseñor Pineda, L. (2011). Multi-Lingual Acoustic Feature Selection for Emotion Estimation using a 3D Continuous Model. *EmoSPACE 2011. 1st International Workshop on Emotion Synthesis, rePresentation, and Analysis in Continuous space.* IEEE.

Pérez Espinosa, H., Reyes García, C. A., & Villaseñor Pineda, L. (2012, January). Acoustic feature selection and classication of emotions in speech using a 3D continuous emotion model. *Biomedical Signal Processing and Control*, *7*(Human Voice and Sounds: From Newborn to Elder), 79–87.

Picard, R. W. (2000). *Affective Computing* (1, July 31 ed.). The MIT Press.

Pitterman, J., & Schmitt, A. (2008). Integrating Linguistic Cues Into Speech-Based Emotion Recognition. *4th IET International Conference on Intelligent Environments.* Seattle, USA.

Pittermann, A., & Pittermann, J. (2006). Getting Bored with HTK? Using HMMs for Emotion Recognition. *8th International Conference on Signal Processing (ICSP).* Guilin, China.

Planet, S., Iriondo, I., Martinez, E., & Montero, J. (2008). True: an online testing platform for multimedia evaluation. *Second International Workshop on EMOTION: Corpora for Research on Emotion and A ect at the 6th Conference on Language Resources & Evaluation (LREC 2008).* Marrakech, Marocco.

Planet, S., Socoró, J., Monzo, C., & Adell, J. (2009). GTM-URL Contribution to the Interspeech 2009 Emotion Challenge. *Interspeech.* Brighton, U.K.

Plutchik, R. (2000). *Emotions in the practice of psychotherapy: clinical implications of affect theories.* American Psychological Association Press.

Polzehl, T. a. (2010). Approaching Multilingual Emotion Recognition from Speech - On Language Dependency of Acoustic/Prosodic Features for Anger Detection. *Proc. of*

the Fifth International Conference on Speech Prosody, 2010. Speech Prosody 2010. Chicago, U.S.A.

Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., & Metze, F. (2009). Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features. *Interspeech 2009.* Brighton, U.K.

Reyes, A. L. (2007). *Un Método para la Identificación del Lenguaje Hablado utilizando Información Suprasegmental.*

Russel, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology, 39*, 1161--1178.

Santiago, K., Reyes G., C. A., & Gomez G., M. (2009). *Conjuntos Difusos tipo 2 aplicados a la Comparación Difusa de Patrones para Clasificación de llanto de infantes con riesgo neurológico.* Tonantzintla, Puebla, México.

Sato, N., & Obuchi, Y. (2007). Emotion Recognition using Mel-Frequency Cepstral Coefficients. *Information and Media Technologies, 2*(3), 835-848.

Scherer, K. R. (2000). Psychological models of emotion. (J. C. Borod, Ed.) *The neuropsychology of emotion*, 137-166.

Scherer, K. R., & Ceschi, G. (1997). Lost luggage: A field study of emotion-antecedent appraisal. *Motivation and Emotion, 21*, 211-235.

Scherer, K., Wallbott, H., & Summerfield, A. (1986). *Experiencing emotion: A cross-cultural study.* Cambridge: University Pres.

Schuller, B., Lang, M., & Rigoll, G. (2005). Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers. *Deutsche Jahrestagung für Akustik, DEGA, Invited Session "Automatische Spracherkennung in gestörter Umgebung"*, (pp. 329-330). Müchen, Germany.

Schuller, B., Steidl, S., & Batliner, A. (2009). The INTERSPEECH 2009 Emotion Challenge. *INTERSPEECH 2009.* Brighton, U.K.

Selting, M. (1994, October). Emphatic speech style mdash; with special focus on the prosodic signalling of heightened emotive involvement in conversation. *Journal of Pragmatics, 22*(3-4), 375-408.

Seol, Y.-S., Kim, D.-J., & Kim, H.-W. (2008). Emotion Recognition from Text Using Knowledge-based ANN. *The 23rd International Technical Conference on Circuits/Systems, Computers and Communications*, (pp. 1569 – 1572). Shimonoseki, Japan.

Sobol-Shikler, T. (2008). *Analysis of affective expression in speech.* University of Cambridge, Computer Laboratory.

Sobol-Shikler, T. (2009). *Analysis of affective expressions in speech.* Tech report, University of Cambridge.

Steidl, S. (2009). *Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech,*. Berlin: Logos Verlag.

Steunebrink, B., Dastani, M., & Meyer, J.-J. (2009). The OCC Model Revisited. In D. Reichardt (Ed.), *Proceedings of the 4th Workshop on Emotion and Computing - Current Research and Future Impact.* Paderborn, Germany.

Sugeno, M., & Kang, G. (1988, oct). Structure identification of fuzzy model. *Fuzzy Sets Systems*, 15-33.

Tóth, S. L., Sztahó, D., & Vicsi, K. (2007). Speech emotion perception by human and machine. *COST2102 International Conference on Nonverbal Features of Human-Human and Human-Machine Interaction*, (pp. 223-236). Patras, Greece.

Truong, K. P., Neerincx, M. A., & van Leeuwen, D. A. (2008). Assessing agreement of observer- and self-annotations in spontaneous multimodal emotion data. (pp. 318-321). ISCA.

Vera-Muñoz, C., Pastor-Sanz, L., Fico, G. L., & Arredondo, M. T. (2008). A Wearable EMG Monitoring System for emotions Assessment. In *Probing Experience: From assessment of user emotions and behaviour to development of products* (pp. 139-148). Springer.

Vidrascu, L., & Devillers, L. (2005). Real-life emotion representation and detection in call centers data. *LNCS, 3784*(1), 739-746.

Vlasenko, B., Schuller, B., Wendemuth, A., & Rigoll, G. (2007). Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing. *Affective Computing and Intelligent Interaction (ACII2007)*, (pp. 139-147). Lisbon.

Vogt, T., & André, E. (2009). Exploring the benefits of dizcretization of acoustic features for speech emotion recognition. *Interspeech.* Brighton, U.K.

Wallach, H. (2004). *Conditional Random Fields: An Introduction.* University of Pennsylvania - Department of Computer & Information Science.

Warrens, M. (2010). Inequalities between multi-rater kappas. *Advances in Data Analysis and Classification, 4*(4), 271-286.

Witten, H. I., & Frank, E. (2005). *Data mining: Practical Machine learning tools and techniques* (2 ed.). San Francisco.

Wöllmer, M., Eyben, F., Keshet, J., Graves, A., Schuller, B., & Rigoll, G. (2009). Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks. *ICASSP*. Taipei, Taiwan.

Wollmer, M., Eyben, F., Reiter, S., Schuller, B., Cox, C., Douglas-Cowie, E., & Cowie, R. (2008). Abandoning emotion classes - towards continuous emotion recognition with modelling of long-range dependencies. *INTERSPEECH-2008*, (pp. 597-600). Brisbane, Australia.

Wöllmer, M., Eyben, F., Schuller, B., Douglas-Cowie, E., & Cowie, R. (2009). Data-driven Clustering in Emotional Space for Affect Recognition Using Discriminatively Trained LSTM Networks. *Interspeech* (pp. 1595-1598). Brighton, UK: ISCA.

Xie, B., Chen, L., Chen, G.-C., & Chen, C. (2005). Statistical Feature Selection for Mandarin Speech Emotion Recognition. In *Advances in Intelligent Computing, Lecture Notes in Computer Science* (Vol. 3644/2005, pp. 591-600). Springer Berlin / Heidelberg.

Zbynik, T. a. (1999). Speech production based on the mel-frequency cepstral coefficients. In I. S. Association (Ed.), *Eurospeech 1999*, (pp. 2335-2338).

Zhou, Z.-H., & Li, M. (2005). Semi-Supervised Regression with Co-Training. In L. P. Saffiotti (Ed.), *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30-August 5, 2005*, (pp. 908-916).

Zhu, X. (2006). *Semi-Supervised Learning Literature Survey.* TR 1530, University of Wisconsin – Madison, Computer Sciences.