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

Humberto Pérez Espinosa, Carlos Alberto Reyes García

Reporte Técnico No. CCC-10-005
4 de mayo de 2010

Coordinación de Ciencias Computacionales
INAOE

Luis Enrique Erro 1
Sta. Ma. Tonantzintla,
72840, Puebla, México.

Contenido

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

Humberto Pérez Espinosa, Carlos Alberto Reyes García

Coordinación de Ciencias Computacionales,
Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro 1, Sta. Ma. Tonantzintla,
72840, Puebla, México
humbertop@inaoep.mx, kargaxxi@inaoep.mx

Resumen. En esta investigación doctoral se trabajará en el reconocimiento de emociones a partir de la señal de voz enfocándose en bases de datos de emociones espontáneas. El reconocimiento de emociones tiene especial importancia en el área de sistemas de interacción humano - computadora, y en sistemas de interacción humano - humano, ya que permite mejorar la calidad de los servicios prestados por estos sistemas, habilitándolos para tomar decisiones importantes basándose en el estado emocional de los usuarios. Para atacar este problema se pretende explorar la utilización de características acústicas principalmente. Se adoptará el enfoque del modelado continuo de emociones, probando técnicas de computación suave y probabilistas para la clasificación de emociones. Este trabajo aportará en la comprensión de los elementos del habla que ayudan a reconocer las emociones y en la creación de un método de reconocimiento de patrones basado en el modelo emocional continuo apropiado para emociones espontáneas. Hasta el momento se ha experimentado con varios tipos de características, incluyendo nuevas características utilizadas en otros campos y se ha recurrido a técnicas de selección de atributos para encontrar las más importantes para nuestros fines. Así mismo, se han clasificado estados emocionales basándose en 2 modelos psicológicos de las emociones, el modelo continuo y el modelo discreto. Los resultados obtenidos son comparables con los mejores resultados en el estado del arte.

1 Introducción

Las emociones son un elemento inherente a los seres humanos. El afecto y la emoción juegan un papel importante en nuestras vidas y están presentes en mucho de lo que hacemos. 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. En la actualidad, los sistemas de Interacción Humano Computadora (IHC) tienden a incorporar sistemas de habla y visión, ya que estos medios son los canales más naturales en la comunicación humana. Uno de los objetivos que persiguen los sistemas de IHC es que la interacción en estos escenarios 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 pueden adaptarse al estado emocional del usuario, como lo hacemos los humanos al conversar, alcanzando una interacción más natural, eficiente y amigable. 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 sobre qué acciones debe seguir un sistema.

A) Un tutorial interactivo (Hernández, et al., 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.

B) Un sistema telefónico de atención automática a clientes que provee asistencia médica a usuarios que llaman pidiendo ayuda (Vidrascu, et al., 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.

C) 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.

D) Las aplicaciones del reconocimiento automático de carga emocional en la voz no se limitan únicamente a la IHC. En la interacción humano – humano, puede usarse para monitorear conversaciones entre agentes y clientes en *call centers* y detectar estados emocionales no deseados (Devillers, et al., 2006). Por ejemplo, un cliente enojado o frustrado o un agente con actitud altanera. De esta manera un inspector de calidad puede tomar decisiones sobre la administración y mejora del personal y de los servicios.

Como muestran estos ejemplos de aplicaciones, mediante el reconocimiento de emociones se puede incrementar el desempeño, la usabilidad y en general la calidad de sistemas de IHC, 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. Para atacar el problema se ha trabajado en dos aspectos principalmente. El primero es desarrollando técnicas de procesamiento y análisis de la señal de voz. En segundo lugar, se ha trabajado en diferentes técnicas de reconocimiento de patrones que tratan de explotar las propiedades del los datos extraídos en el procesamiento de la señal de voz. En esta tesis se abordarán ambos aspectos. Se trabajará en el análisis de la señal de voz para encontrar las características más relevantes en el reconocimiento de emociones y en el desarrollo de un método de clasificación apropiado para enfrentar la complejidad de las características que describen la señal de voz y en cómo aprovechar la información provista por cada tipo de característica para llegar a una buena estimación de emociones Esta tesis abordará el reconocimiento de emociones bajo el enfoque de los modelos emocionales continuos, que como se verá más adelante pueden representar cualquier estado emocional en un espacio multidimensional. Este tipo de modelos representan más adecuadamente la forma en que suceden las emociones en el mundo real, brindan un mayor nivel de generalización y permiten describir la intensidad de las emociones. Esto es muy benéfico ya que las emociones no siempre se generan de manera prototípica, sino se generan como una mezcla de emociones con mayor o menor intensidad. La intención de este trabajo de investigación es el avance hacia el reconocimiento de emociones producidas de manera espontánea. Las principales aportaciones serán, un análisis de la influencia de diferentes tipos de características de la voz, incluyendo la propuesta de características que brinden información novedosa no explotada anteriormente ayudando al reconocimiento de emociones, así como un método diseñado

especialmente para el reconocimiento de emociones en contextos reales basado en un modelo emocional continuo.

2 Antecedentes

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, et al., 2005) (González, 1999). En esta sección se presenta una descripción de los conceptos básicos más importantes relacionados con la teoría de emociones, desarrollados por filósofos, psicólogos, y antropólogos principalmente, así como una síntesis de los trabajos relacionados más influyentes en nuestra propuesta.

2.1 Marco Teórico

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 los estados emocionales 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 las han tratado de responder filósofos, biólogos, psicólogos, y antropólogos. Por ejemplo, el filósofo Platón formuló la doctrina del alma tripartita la cual sugería que el alma tiene una estructura tripartita compuesta por tres áreas: cognición, emoción y motivación. 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 (Darwin, 1998). De este modo las emociones son más o menos las mismas en todos los humanos y en particular independientes de la cultura. Aún cuando los antropólogos afirman que las emociones son productos socioculturales, varios autores han trabajado en demostrar la hipótesis de Darwin. Se ha establecido que el habla es un evento acústico que contiene información importante sobre el funcionamiento del sistema nervioso central, y por lo tanto acarrea información sobre el estado emocional de un individuo. Se han definido expresiones faciales universales. El psicólogo Paul Ekman (Ekman, 1972) establece seis expresiones faciales relacionadas con emociones básicas conocidas como el *Big Six*. William James, psicólogo y filósofo estadounidense, publicó en 1884 un artículo titulado *"What is an emotion?"* (James, 1884). En este trabajo establece que 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.

La definición del término emoción es la base para cualquier tipo de investigación en esta área. Una definición común permite comparar resultados entre diferentes grupos de investigación y evitar malentendidos. La manera en que las emociones son definidas también determina el tipo de fenómenos estudiados en la investigación sobre emociones. Según Scherer (Scherer, 2000) las emociones son definidas como:

Episodios de cambios coordinados en varios componentes (incluyendo al menos activación neuropsicológica, expresión motriz, y sentimientos subjetivos pero posiblemente también 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 a cambios en la situación actual, o un encuentro con un estímulo nuevo. Los eventos internos son, por ejemplo, pensamientos, recuerdos, y sensaciones. Esta definición menciona diferentes características de las

emociones para las cuales, de acuerdo a Scherer, se ha encontrado un creciente consenso en la literatura. Dichas características son:

- Las emociones son de naturaleza episódica y son distintivas. La suposición es que un cambio notable en el funcionamiento del organismo es causado por algunos disparadores externos. 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 consisten de varios componentes, incluyendo la reacción de una triada de emociones, llamada Excitación Psicológica, Expresión Motriz y Sentimiento Subjetivo. Necesariamente los componentes podrían también ser tendencias de acción y procesos cognitivos envueltos en la evaluación de los eventos activadores y en la regulación de procesos emocionales.

Las emociones se diferencian de otros fenómenos afectivos como humor, posturas interpersonales, actitudes o rasgos de personalidad por la intensidad y duración del estado, el grado de sincronización de diferentes sistemas orgánicos durante el estado, la medida en que el cambio de estado se activa o se centra en un acontecimiento o situación y la influencia en el comportamiento. 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.

2.1.1 Modelos Psicológicos de Emoción

A pesar de los muchos intentos tratando de establecer una correspondencia entre emociones y voz no existe un conjunto definido de emociones universalmente aceptado. Hay varios modelos para representar las emociones los cuales son usados para su categorización y organización. Estas categorías difieren dependiendo de las diferentes tareas y aplicaciones. La categorización es principalmente hecha sobre bases subjetivas porque los investigadores no coinciden en determinar un conjunto de etiquetas emocionales. Estos modelos de clasificación difieren principalmente en que algunos usan valores continuos y otros valores discretos para la abstracción de categorías Ambos modelos están relacionados ya que las categorías emocionales discretas pueden ser mapeadas en modelos continuos.

2.1.2 Modelos Discretos

Estos modelos se basan en el concepto de “emociones básicas”, que son la forma más intensa de las emociones, a partir de las cuales se generan todas las demás mediante variaciones o combinaciones de estas. Suponen la existencia de emociones universales, al menos en esencia, que pueden ser distinguidas claramente una de otra por la mayoría de la gente, y 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. La dominación de esta teoría en el reconocimiento automático de emociones puede explicarse por el hecho de que la diferenciación entre emociones es relativamente clara, aún cuando dentro de estos conjuntos de emociones la necesidad de definiciones más detalladas ha sido abordada, por ejemplo distinguiendo entre ira y cólera. Otra explicación puede ser que las representaciones estereotípicas de estas expresiones son asimiladas más fácilmente con el fin de generarlas y reconocerlas, y por lo tanto son más útiles para una construcción rápida de bases de datos y como un punto de partida para

investigación emergente en este campo de investigación. Algunos científicos definen una lista de emociones básicas desde su punto de vista, ver Tabla 1. Como se puede observar en dicha tabla, no existe un criterio único para definir qué emociones forman este conjunto. Los modelos discretos permiten una representación más particularizada de las emociones en las aplicaciones donde solamente se requiere reconocer un conjunto predefinido de emociones. Este enfoque ignora la mayor parte del espectro de expresiones emocionales humanas. 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 validas tanto para emociones extremas como para emociones más sutiles (Sobol-Shikler, 2008). En la Tabla 1, tomada de (Ortony, et al., 1990), se muestran varios conjuntos de emociones básicas en inglés propuestos por distintos autores. Otro de los problemas de estos modelos es la investigación intercultural de emociones y la traducción correcta de términos emocionales o afectivos usados y muchos de estos términos tienen significados connotativos y denotativos diferentes en diferentes idiomas, no hay una solución satisfactoria a este problema (Hillsdale, et al., 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).

2.1.3 Modelos Continuos

En estos modelos los estados emocionales son representados usando un espacio multidimensional continuo. Las emociones son representadas por regiones en un espacio n-dimensional. Los ejes no están relacionados con estados emocionales, sino con primitivas emocionales que son propiedades subjetivas que presentan todas las emociones. Un ejemplo muy conocido es el modelo Arousal-Valence. Este modelo describe las emociones usando un espacio bidimensional. Las emociones son descritas en términos de Valencia y Activación (Steidl, 2009). La Valencia, también llamada placer describe qué tan negativa o positiva es una emoción específica. Activación, también llamada intensidad, describe la excitación interna de un individuo y va desde estar muy tranquila hasta estar muy activa. Otro modelo similar al previo es el modelo tridimensional. Las dos primeras dimensiones son Valencia y Activación mientras la tercera es la energía o 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. Ver Fig. 1. El modelo tridimensional 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. Los modelos continuos permiten mayor flexibilidad en las aplicaciones ya que no se limitan a un conjunto de emociones sino que pueden representar cualquier estado emocional en el espacio multidimensional y trasladarlo a un conjunto de emociones básicas si así se requiere (Grimm, et al., 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 para asignar estados preseleccionados a patrones psicológicos, mientras el enfoque continuo es más adecuado para evaluar la carga emocional (Beale, et al., 2008).

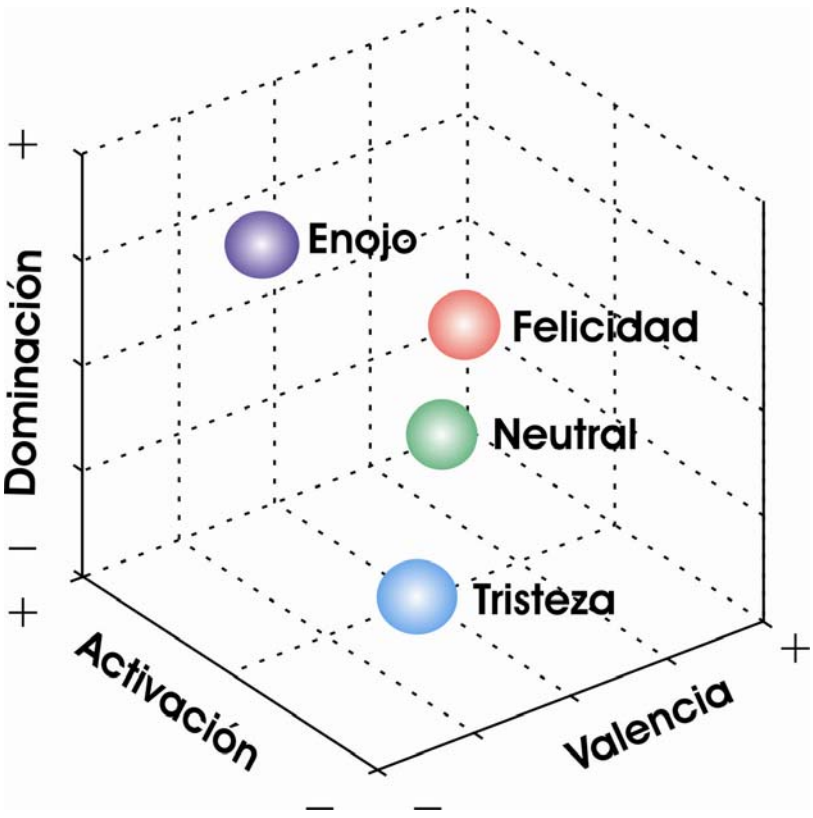

Fig. 1 Modelo Tridimensional Continuo de las Emociones. Valencia – Activación - Dominación

Tabla 1 Conjuntos de emociones básicas propuestos por diferentes autores

Autor	Emociones Básicas	Base de Inclusión
Plutchik	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological process
Ekman, Friesen, Ellsworth	Anger, disgust, fear, joy, sadness, surprise	Universal facial Expressions
Gray	Rage and terror, anxiety, joy	Hardwired
Izard	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James	Fear, grief, love, rage	Bodily involvement
Mowner	Pain, pleasure	Unlearned emotional states
Oatley and Johnson-Laird	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Paksepp	Expectancy, fear, range, panic	Harwired
Tomkins	Anger, interestm contempt, disgust, distress, fear, joy, shame, surprise	Density on neutral firing
Watson	Fear, love, rage	Hardwired
Weiner and Graham	Happiness, sadness	Attribution independent

2.2 Trabajos Relacionados

A partir de la revisión del estado del arte se proponen tres criterios para clasificar los enfoques empleados para resolver el problema de la clasificación automática de emociones a partir de voz. Como se muestra en la Fig. 2, el primer criterio es de acuerdo al tipo de características que se extraen de las bases de datos, el segundo es de acuerdo al tipo de procesamiento de características, el tercero de acuerdo al tipo de modelo emocional que adoptan.

2.2.1 De Acuerdo al Tipo de Información que Utilizan

La información es acústica cuando la extracción se hace únicamente sobre la señal de voz. La mayoría de los trabajos revisados utilizan solamente este tipo de información (Sato, et al., 2007) (Tóth, et al., 2007) (Schuller, et al., 2005) (Luengo, et al., 2005). Las características acústicas suelen agruparse en:

- ***Espectrales*** que describen las propiedades de una señal en el dominio de la frecuencia mediante harmónicos 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

Otro tipo de información es la lingüística, usada en (Lee, et al., 2002) (Devillers, et al., 2006) (Seol, et al., 2008) (Pitterman, et al., 2008), donde la información proviene del texto transcrito a partir de la señal de voz. Es importante hacer notar que en una aplicación real la única forma de obtener información lingüística es mediante un Reconocedor Automático de Habla. Reconocer el habla en vocabularios no restringidos es un problema que no está completamente resuelto. El problema del reconocimiento automático de voz no se abordará en esta tesis;, sin embargo, se estudiará la forma de aprovechar esta información en el caso que se cuente con ella. En la literatura se distinguen principalmente dos enfoques de uso de información lingüística:

- ***Bolsa de Palabras***: En este enfoque (Liscombe, et al., 2005) (Polzehl, et al., 2009) se representa el texto a través de un vector lingüístico. Es una técnica muy usada en categorización automática de texto. Cada palabra agrega una dimensión a un vector lingüístico representando la frecuencia dentro de la elocución. Ya que resultan vectores muy grandes se suele usar alguna técnica de reducción de dimensionalidad.
- ***Palabras Clave***: En este otro enfoque (Lee, et al., 2002) (Wöllmer, et al., 2009), la estrategia es detectar palabras clave para mejorar la clasificación. Para identificar las palabras claves se suele usar el concepto de "palabra relevante". Una palabra relevante con respecto a una categoría es aquella que aparece más frecuentemente en esa categoría que en otras partes del corpus y es considerada como una medida de distancia de las palabras nulas cuya frecuencia relativa en cada clase es la misma.

Las características lingüísticas están muy relacionadas con el contexto en el que se emplean, por lo que el uso de información lingüística para la clasificación de emociones está fuertemente ligado al corpus, siendo difícil usar los mismos diccionarios de palabras emocionales en diferentes aplicaciones, lo que limita su portabilidad entre aplicaciones. La información de contexto (Liscombe, et al., 2005) (Herm, et al., 2008) (Forbes-Riley, et al., 2004) se obtiene principalmente a partir de la voz o de datos de los registros del sistema, como puede ser el tipo de usuario, género, edad, motivo de la interacción, tipo de diálogo que está sosteniendo o algún otro tipo de información sobre la interacción entre el usuario y el sistema que pudiera ayudar a la clasificación

de emociones. Este tipo de información generalmente es sólo un complemento en sistemas que usan información acústica y/o lingüística. Sin embargo, se ha sugerido el uso de información de contexto como única fuente para detectar emociones positivas (Herm, et al., 2008). Las características contextuales en algunos casos se extraen semiautomáticamente y son dependientes de la aplicación (Liscombe, et al., 2005), lo que limita su portabilidad entre aplicaciones.

2.2.2 De Acuerdo al Tipo de Procesamiento de Características

En esta categoría se distinguen dos enfoques el Modelado Dinámico de características y el Modelado Estático.

- **Modelado dinámico:** Se emplean características como tono, energía, MFCCs y sus derivativas etc. con modelos de clasificación dinámicos como Hidden Markov Models (Pittermann, et al., 2006) 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 energía, velocidad (Vlasenko, et al., 2007).

- **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, energía, o coeficientes espectrales, y de la aplicación de funciones estadísticas, como media, desviación estándar quantiles, sobre las características, lo cual resulta en vectores de características del mismo tamaño para todas las elocuciones (Vogt, et al., 2009) (Planet, et al., 2009) (Lee, et al., 2009).

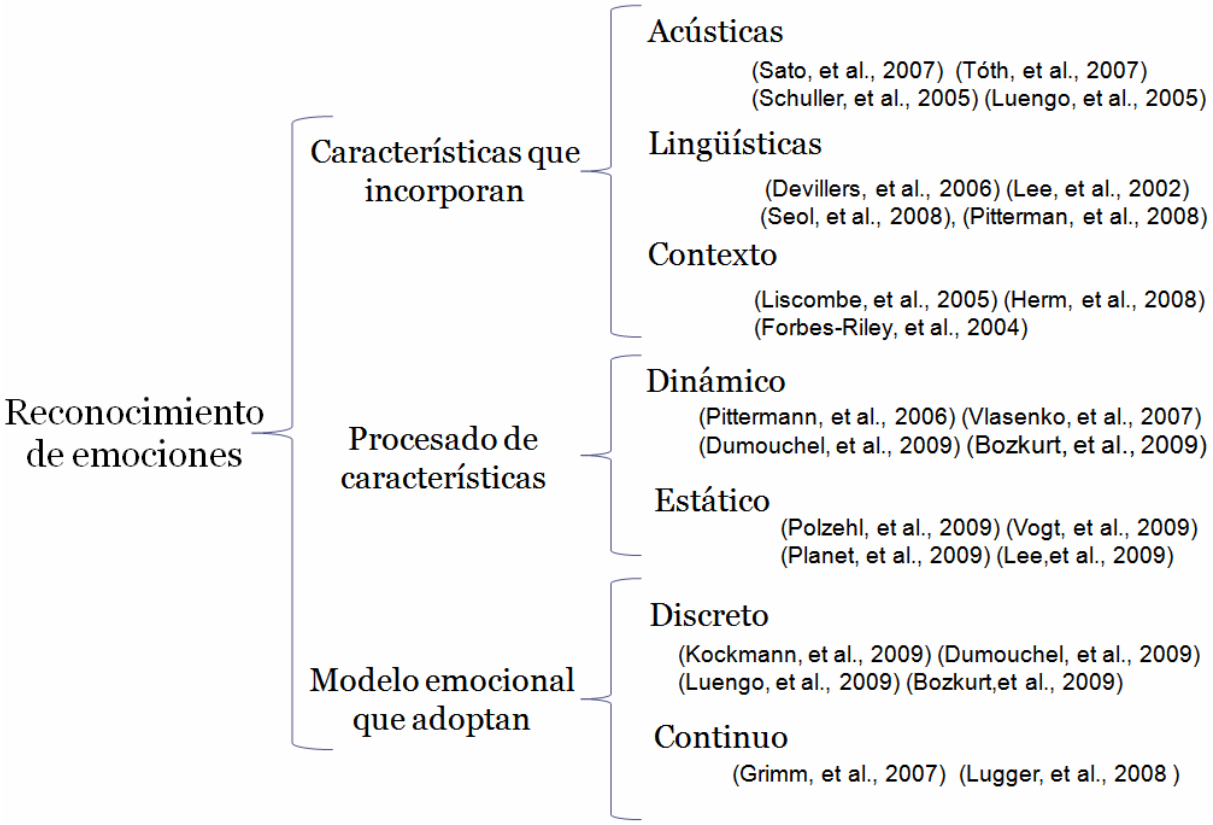

Fig. 2 Agrupamiento de trabajos relacionados por 3 criterios diferentes

En (Vlasenko, et al., 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.
Ambos tipos de procesamiento, dinámico y estático, capturan diferentes propiedades del habla emocional. El procesamiento dinámico captura información de cómo evolucionan las características a través del tiempo, mientras que el procesamiento estático evita un sobre modelado fonético al aplicar funciones estadísticas sobre los LLDs en periodos de tiempo. En el reconocimiento de emociones en voz es más común el procesado estático. Sin embargo, el procesado dinámico ha cobrado fuerza en los últimos años (Dumouchel, et al., 2009) (Bozkurt, et al., 2009) mostrando buenos resultados aplicando técnicas como Gaussian Mixture Models.

2.2.3 De Acuerdo al Modelo Emocional que Adoptan

Como se vio en la sección 2.1.1, los modelos emocionales pueden dividirse en modelos continuos y modelos discretos, los trabajos revisados hasta el momento pueden clasificarse en trabajos que distinguen:

- **Emociones Específicas**: En este tipo de trabajos solamente se desea detectar una emoción específica, como decepción, confianza, frustración, enojo, de acuerdo a un dominio de aplicación. Por ejemplo, se puede desear identificar tensión (Fell, et al., 2003) en un sistema

de cobranza donde suelen surgir conflictos entre el agente que solicita un pago y el cliente o frustración en un sistema automatizado de información, donde los clientes frecuentemente no logran obtener la información que necesitan.

- **Un conjunto de emociones básicas:** Este tipo de trabajos se basan en un conjunto de emociones básicas, por ejemplo el "Big Six" de Ekman (alegría, enojo, tristeza, sorpresa, disgusto, miedo), con una clasificación binaria Positivo – Negativo, o con otro conjunto de emociones. Ver tabla 1. En esta categoría caen la mayoría de los trabajos encontrados en la literatura (Kostoulas, et al., 2008) (Pittermann, et al., 2006) (Kockman, et al., 2009) (Luengo, et al., 2009).

- **Un espacio continuo de emociones:** En esta categoría se predicen primitivas emocionales de acuerdo a un modelo multidimensional como pueden ser Valencia – Dominación - Activación o Valencia - Interacción (Grimm, et al., 2007) (Lugger, et al., 2008). En algunos trabajos se aplican los modelos continuos para clasificar emociones básicas. Este tipo de enfoque es el que creemos puede tener mejores resultados de acuerdo a las propiedades que presentan las emociones reales. Sobre este enfoque se va a desarrollar el trabajo de esta tesis.

Lichtenstein (Lichtenstein, et al., 2008) realizó un experimento para comparar ambos enfoques, discreto y continuo, en cuanto a cuál es más adecuado para estimar estados emocionales 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.

2.2.4 Trabajos Relacionados más Importantes

A continuación se presenta un análisis de los 3 trabajos más influyentes en nuestra propuesta.

Predicción de Primitivas Emocionales (Grimm, et al., 2007)**:** En este trabajo se emplea un modelo emocional multidimensional. Se usa un estimador lógico difuso y una base de reglas derivada a partir de características acústicas de la señal de voz tales como melodía, energía, velocidad y características espectrales. Se prueban una base de datos actuada y también una base de emociones auténticas. Se predicen primitivas emocionales que son mapeadas hacia un conjunto de categorías emocionales usando un clasificador *K Vecinos más cercanos (Knn).* A partir de los archivos de audio se extraen características acústicas. Se etiqueta esta base de datos con primitivas emocionales. Este etiquetado se hace a través de una evaluación por escuchas humanos. El modelo tridimensional usado se compone de las primitivas; Valencia, Activación y Dominación. Se etiquetan usando una técnica denominada *Self Assessment Manikins (SAMs)* (Grimm, et al., 2005)**.** Una vez teniendo las características acústicas y habiendo etiquetado el corpus con primitivas emocionales se calculó la correlación que existe entre ambas, con el fin de establecer reglas que establezcan que combinaciones de valores en las características acústicas corresponden a cierto grado de una primitiva emocional. El proceso de clasificación consiste en extraer las características acústicas de las instancias 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. Se realizan varios experimentos con subconjuntos de instancias de su corpus de entrenamiento. Se alcanza 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 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, et al., 2008).

Clasificación de Estados Emocionales Espontáneos en Niños (Steidl, 2009)**:** En este trabajo se crea el corpus FAU Aibo que está diseñado para realizar investigación orientada a pasar de emociones básicas o prototípicas, a emociones que aparecen en escenarios más realistas donde dichas emociones son principalmente estados emocionales sutiles, y mezclas de diferentes estados. Se trata de un corpus de habla espontánea en alemán, con tintes emocionales de niños en la edad de 10 a 13 años que interactúan con el robot Aibo de Sony. La idea es combinar un corpus de habla infantil con habla emocional *natural* 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. Once estados emocionales se etiquetaron a nivel de palabra y frase. Se llevaron a cabo experimentos de clasificación en tres niveles de segmentación: nivel de palabra, nivel de turno, y nivel de bloque intermedio. Los mejores resultados se obtuvieron en el nivel de bloque intermedio donde se alcanzó una tasa promedio de reconocimiento de casi el 70% para 4 clases, Enojo, Enfático, Neutral y Maternal. Aplicando la medida de entropía propuesta por el autor para la evaluación de descodificadores, el desempeño de la clasificación a nivel de palabra mejoró ligeramente con relación a la media del etiquetado hecho por evaluadores humanos. Se propone un conjunto de características acústicas y lingüísticas. El desempeño de las características lingüísticas es ligeramente peor que el de las características acústicas. Se obtuvo una mejora mediante la combinación de ambas fuentes de información. Las características acústicas se agruparon en prosódicas, espectrales, y de calidad de voz. La energía y la duración de las características basadas en las características prosódicas y espectrales MFCC, son las características acústicas más relevantes en este escenario. Los modelos de unigramas y bolsa de palabras fueron las características lingüísticas más relevantes. Este trabajo aborda la problemática que existe al pasar del estudio de emociones actuadas a emociones espontáneas mediante la creación de un corpus de emociones inducidas. El corpus generado muestra espontaneidad emocional ya que las emociones en los niños suelen ser más fidedignas que las de los adultos. Sin embargo, los experimentos de clasificación hechos en este trabajo se basan únicamente en modelos emocionales discretos dejando abierto la interrogante de cómo podrían mejorar los resultados obtenidos al aplicar el enfoque de los modelos emocionales continuos.

Clasificación de Emociones en Cascada usando Dimensiones Emocionales (Lugger, et al., 2008)**:** En este trabajo se clasifican emociones básicas independientemente del hablante. Se trabaja con la base de datos en alemán llamada *Berlin Emotional* que consta de 6 emociones básicas: tristeza, aburrimiento, neutral, ansiedad, felicidad e ira. Se usan características prosódicas y de calidad de voz. Se hace la observación de que el conjunto de características óptimo depende fuertemente de las emociones a ser clasificadas y a partir de esto se realiza una clasificación en cascada de 3 fases basada en el modelo psicológico emocional continuo. En la primera etapa, 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 clasificaron 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. Este trabajo encara el problema de continuidad de emociones mediante una clasificación en cascada inspirada en un modelo emocional continuo. Sin embargo, se trabajó sobre una base de datos actuada y etiquetada con emociones básicas basándose en una categorización manual de los niveles correspondientes de cada emoción con las 3 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.

2.2.5 Ubicación de la Propuesta

El método propuesto incluirá la estimación automática de primitivas emocionales y el mapeo de emociones básicas hacia un modelo tridimensional continuo. Se trabajará sobre bases de datos de emociones espontáneas etiquetadas con primitivas emocionales. Nuestro trabajo estará basado en un modelo continuo tridimensional cuyas primitivas son Valencia, Activación y Dominación. Será necesario contar con una base de datos etiquetada con emociones básicas para validar el mapeo entre ambos modelos emocionales. Estas bases de datos podrían ser la misma con ambos tipos de etiquetado o bases de datos diferentes. Se explorará el uso tanto de procesamiento estático como dinámico de características fusionando ambos tipos de información. Se emplearán principalmente características acústicas, que incluyen características prosódicas, de calidad de voz, y espectrales. Se explorará el uso de características lingüísticas como complemento o soporte a las características acústicas. La Tabla 2 resume las principales diferencias de nuestra propuesta en comparación con los trabajos relacionados más cercanos.

Tabla 2. 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	Espectral	Texto
Grimm 07	Espontánea	Continuo	Continuo	Estático	√		√	
Steidl 09	Espontánea	Discreto	Discreto	Estático	√	√	√	√
Lugger 08	Actuada	Discreto	Continuo	Estático	√	√	√	
Propuesta	Espontánea	Continuo/ Discreto	Continuo	Estático/ Dinámico	√	√	√	√

3 Planteamiento de la Propuesta

3.1 Problemática

El área de reconocimiento automático de emociones ha sido un área de investigación muy activa en los últimos años, no obstante aun no hay una solución clara para este problema. Diversos inconvenientes 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, sin embargo, no se han tenido buenos resultados al trasladar el conocimiento extraído de estas bases de datos a contextos reales (Steidl, 2009).
Las características de las bases de datos actuadas son:

1. Los "Retratos de emociones" representando expresiones prototípicas e intensas facilitan la búsqueda de correlación acústica y la subsecuente clasificación automática.
2. Las grabaciones en estudio con alta calidad eliminan problemas en el procesamiento de la señal. Por ejemplo ruido o reverberación.
3. Se puede garantizar una cantidad balanceada de ejemplos por clase.

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. En otros casos, existen elocuciones con una carga emocional muy ligera, cercana a un estado emocional neutro. Además, 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 no es posible asegurar una cantidad balanceada de 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. Sin embargo, se ha descubierto que entre más nos acercamos a un escenario realista, menos fiable es la prosodia como un indicador del estado emocional del hablante (Batliner, et al., 2003), por lo tanto es necesario encontrar características que complementen la información que proporciona el aspecto prosódico de la voz.

Finalmente, hasta la fecha la mayoría de los trabajos 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 prácticamente necesario repetir todo el trabajo si se desea agregar una nueva emoción o se desea trabajar con otro corpus. Además los modelos discretos no parecen ser los más adecuados para trabajar con emociones espontáneas ya que no representan apropiadamente el traslape 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 y lingüísticas, además de emplear un modelo continuo

que nos permita apegarnos más a la realidad. Nuestro trabajo estará basado en un modelo continuo tridimensional cuyas primitivas son Valencia, Activación y Dominación. Se emplearán principalmente características acústicas, que incluyen características prosódicas, de calidad de voz, y espectrales. Se explorará el uso de características lingüísticas como complemento o soporte a las características acústicas. Además se explorará el uso tanto de procesamiento estático como dinámico de características.

3.2 Preguntas de Investigación

- ¿Qué características acústicas son útiles para reconocer emociones en el habla independientemente del dominio de aplicación?
- ¿Cuáles de esas características son más útiles para un modelo emocional continuo?
- ¿De qué forma podemos emplear un modelo emocional continuo en él diseño de un método de reconocimiento de emociones aplicable a emociones espontáneas?
- ¿Cómo se pueden aprovechar las ventajas de los modelos emocionales continuos para estimar con una buena precisión la carga emocional en la voz?
- ¿El uso de modelos continuos mejorará el reconocimiento de emociones con respecto al uso de modelos discretos? ¿Cuánto puede mejorar el reconocimiento de emociones en voz usando modelos continuos con respecto al uso de modelos discretos?

3.3 Objetivo General

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

3.4 Objetivos Específicos

1. Identificar diferentes tipos de características relevantes en el reconocimiento de emociones.
2. Determinar el aporte de cada tipo de característica al reconocimiento de primitivas emocionales en bases de datos de emociones espontáneas
3. Diseñar una arquitectura de fusión de los diferentes tipos de características para identificar las primitivas de un modelo emocional continuo
4. Diseñar un esquema de reconocimiento de patrones basado en un modelo emocional continuo
5. Estudiar la relación entre primitivas en nuestro modelo emocional y determinar la manera de interpretarlas para ubicar emociones básicas en aplicaciones especificas.

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

3.5 Contribuciones

1. Identificación de un conjunto de características acústicas para soportar un modelo emocional continuo.

2. Un método de reconocimiento de emociones que tome provecho de las características propuestas en el punto anterior, que esté basado en el modelo emocional continuo brindando información más detallada sobre el estado emocional y que proporcione mayor flexibilidad y facilidad para transferir de una aplicación a otra.

3. Un clasificador emocional adecuado al reconocimiento de emociones espontáneas

4 Metodología

Las tareas listadas abajo no se realizaran en el estricto orden en el que aparecen. Existen tareas que se pueden realizar en paralelo como lo muestra el calendario de actividades mostrado en la sección 8. La metodología se divide en tres componentes. En el Componente 1 se estudian las características que se usarán. En el Componente 2 se desarrolla el método propuesto para el reconocimiento de emociones basado en un modelo emocional continuo. Dicho método se ilustra en las Figuras 3, 4 y 5. En el Componente 3 se evalúan los resultados de los dos componentes previos.

Componente 1

---

1. **Identificar grupos de características acústicas usadas hasta el momento mediante la revisión del estado del arte.**
    a. Hacer una recopilación de características extraídas de la señal de voz que hayan sido propuestas en los trabajos en esta área publicados hasta el momento.
    b. Al mismo tiempo hacer una relación de los métodos de clasificación empleados con cada conjunto de características.
    c. Hacer una lista de bases de datos utilizadas en dichos trabajos poniendo especial atención en bases de datos de emociones espontáneas para tratar de obtenerlas.

2. **Estudiar métricas de calidad de voz y articulación usadas en medicina y comprobar la viabilidad de aplicación:**

    a. Realizar un estudio sobre estándares y metodologías de medición de calidad y otros aspectos en la de voz en áreas médicas.
    b. Estudiar la viabilidad de extraer las medidas subjetivas usadas en estos estándares médicos de manera automática.
    c. Adoptar características acústicas para la clasificación de emociones basadas en este estudio.

3. **Estudiar métodos basados en información lingüística**
    a. Hacer un estudio sobre trabajos enfocados a extraer emociones a partir de texto.

b. Estudiar la viabilidad de aplicar las técnicas y características propuestas en estos trabajos sobre las bases de datos que se tengan disponibles para experimentar.
c. Calcular el aporte de la información lingüística a la clasificación de emociones.
d. Estudiar la manera de fusionar la información lingüística con la acústica para mejorar la clasificación

4. **Proponer grupos de características representativas, analizar sus propiedades y experimentar en diferentes contextos de aplicación:**

a. Agrupar características de acuerdo a alguna taxonomía como puede ser la presentada en la sección 2.2.1 de tal manera que al momento de probar se haga de una manera más ordenada o metodológica y se pueda explicar a qué grupo o clasificación de características pertenecen las características utilizadas en cierto experimento.
b. Hacer un estudio de la importancia de características individuales y de grupos de características basado en la aplicación de técnicas de selección de atributos y/o de reducción de dimensionalidad
c. Crear un sistema basado en scripts que extraiga estas características automáticamente para probarlas con diferentes bases de datos.
d. Conseguir las bases de datos de emociones espontáneas y actuadas para comparar.
e. Experimentar con las características en diferentes bases de datos.
f. Experimentar con diferentes clasificadores de acuerdo a los enfoques de procesamiento de características estático y dinámico. Probar Hidden Markov Models y Gaussian Mixture Models.
g. Probar diferentes formas de fusionar la información obtenida con el procesamiento estático y dinámico.
h. Explorar nuevas características basándose en los hallazgos de pasos previos.

Componente 2

---

5. **Crear un modelo que asocie medidas objetivas extraídas de la señal de audio con primitivas emocionales usadas en evaluaciones perceptivas:**

a. Hacer un estudio del estado del arte para definir qué modelo emocional continuo será el más conveniente para usar en esta tesis. Se definirá la dimensionalidad y las primitivas usadas.
b. Proponer un método para descubrir la relación existente entre las características estudiadas en pasos previos con las primitivas emocionales del modelo continuo propuesto.
c. Experimentar con técnicas difusas y probabilísticas para determinar dicha relación.
d. Diseñar un esquema de mapeo entre las características acústicas y primitivas emocionales en un modelo continuo a partir del estudio de esa relación.
e. Probar otras técnicas de computación suave para mejorar el desempeño de los modelos difusos. Entre otras técnicas podrían emplearse Algoritmos Genéticos.
f. Experimenta usando un clasificador para cada primitiva.
g. Diseñar un método de reconocimiento de patrones adecuado a los hallazgos de pasos anteriores

6. **Generar un mapeo entre coordenadas en un modelo continuo y emociones básicas:**

a. Hacer una revisión del estado del arte para determinar según diferentes autores que niveles de primitivas emocionales se esperan para las emociones básicas siendo analizadas de acuerdo a la base de datos con la que se esté probando.
b. Diseñar un método de transformación de los valores esperados de las primitivas emocionales, según el paso anterior, hacia una categoría emocional específica. Este método no dependerá de un conjunto fijo de emociones sino que aceptará diferentes emociones dependiendo de la aplicación o de la base de datos con la que se esté probando.
c. Probar ANFIS (sistema adaptativo de inferencia neuro difuso) para diseñar un método que genera automáticamente reglas que relacionen primitivas emocionales con emociones básicas a partir de un entrenamiento basado en una base de datos de emociones básicas

7. Crear un esquema que fusione la información obtenida con cada grupo de características

a. Implementar un esquema de filtrado y pesado de características dependiendo de la información disponible y de las características de la base de datos.
b. Incorporar procesamiento dinámico y estático de características, fusionando las predicciones hechas por cada tipo de procesamiento

Componente 3

---

8. Evaluar el sistema de acuerdo a los lineamientos de HUMAINE Association:

a. Hacer una evaluación bajo los estándares de HUMAINE Association (Schuller, et al., 2009).

9. Evaluar el desempeño en otros contextos:

a. Llevar a cabo evaluaciones sobre diferentes bases de datos tanto de emociones reales, como actuadas con el fin de evaluar el alcance del sistema.
b. Hacer una evaluación subjetiva con personas no especializadas o no entrenadas.

5 Avances

En resumen se tienen avances en los siguientes puntos:

- Hemos realizado una serie de experimentos con técnicas de clasificación difusas como *ANFIS* y *FPMT2*, y con técnicas probabilistas como *Bayes y Suppport Vector Machines* con el objetivo de medir su desempeño en la clasificación de emociones en 2 bases de datos de emociones espontáneas y una de emociones actuadas. Creamos un método basado en Algoritmos Genéticos para seleccionar atributos y afinar los parámetros de configuración de ANFIS.
- Hemos extraído y probado más de 350 características acústicas. Se ha identificado un conjunto de características que muestran buen desempeño para predecir primitivas emocionales en un modelo emocional continuo. Se han incorporado características propuestas por nosotros y probado otras propuestas por otros autores. Se han evaluado mediante técnicas de selección de atributos.

- Hemos realizado experimentos usando un reconocedor de voz basado en Hidden Markov Models, probando de esta manera el procesamiento dinámico de características obteniendo buenos resultados.
- Hemos estudiado el uso de información lingüística así como su fusión con información acústica.
- Se ha estudiado la influencia de diferentes grupos de características acústicas del habla en la estimación de emociones básicas y de las primitivas emocionales que definen un modelo emocional tridimensional continuo.
- Con los resultados obtenidos hemos podido ubicar el desempeño tanto de nuestras características acústicas, como de las técnicas de selección y clasificación empleadas en relación con el estado del arte.
- Nos hemos comparado contra otros autores mediante el uso de bases de datos y metodologías de evaluación estándar. Se ha alcanzado un buen desempeño en los experimentos hechos hasta la fecha, siendo estos comparables con los mejores del estado del arte.
- Hemos realizado experimentos para comprobar la viabilidad de nuestro método propuesto obteniendo resultados alentadores.

5.1 Bases de Datos

Hasta el momento hemos experimentado con tres bases de datos dos con emociones espontáneas y una con emociones actuadas. Las tres bases de datos están en alemán.

FAU Aibo: La primera es la base de datos FAU Aibo descrita en (Steidl, 2009). Este es un corpus con grabaciones de niños interactuando con el robot mascota Aibo de Sony. 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.

VAM: La segunda base de datos usada es llamada VAM corpus y está descrita en (Narayanan, et al., 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 elocuciones 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.

Berlin Emotional Speech: Esta base de datos está en alemán. Fue grabada por actores. Las grabaciones tuvieron lugar en una cámara sin eco de la Universidad Técnica de Berlín.Consta de 816 frases, 10 actores, 7 estados emocionales, enojo, aburrimiento, disgusto, miedo, felicidad, tristeza y neutro. 10 frases diferentes.

5.2 Características Propuestas

Proponemos 3 grupos de características: Prosódicas, Espectrales y de Calidad de Voz. Las características prosódicas fueron subdivididas en Tiempos en la Elocución, Contorno Melódico, Energía. Cabe señalar que otros autores también han usado características contextuales (Liscombe, et al., 2005) (Seol, et al., 2008) las cuales no incluimos. Diseñamos un conjunto de características que representa varios aspectos de la voz, incluyendo los tradicionales que tienen que ver con la prosodia como duración, pitch y energía y otros que han mostrado buenos resultados en tareas relacionadas como reconocimiento de habla, reconocimiento de hablante, clasificación de llanto de bebés, creemos que pueden dar buenos resultados como son características espectrales y de calidad de voz. Existen trabajos en los que se proponen muchas características, más de 1,000, o muy pocas, 10 o 15. Creemos que es importante encontrar la cantidad adecuada, representando bien los aspectos relevantes de la voz para esta tarea. Tener demasiadas características puede añadir ruido al clasificar. Pensamos que las características que proponemos resumen bien los aspectos de la voz que son importantes para encontrar emociones; proponemos un conjunto de 369 características. Para calcular las características se ha usando el software PRAAT (Boersma).

Características Prosódicas: La prosodia es una fuente de información muy rica en el procesamiento del habla ya que tiene una función paralingüística muy importante que complementa el mensaje lingüístico con una intención determinada la cual refleja una actitud o estado emocional del hablante. También conlleva una función extralingüística que aporta información sobre las características del locutor, como su edad, su sexo, su estatus socioeconómico, etc. A continuación se definen las propiedades acústicas de los sonidos del habla relacionadas con la expresión de prosodia.

Tiempos en la Elocución: Algunos autores extraen este tipo de características de corpus etiquetados con duraciones en algún nivel de segmentación. Debido a que los corpus con los que estamos trabajando no cuentan con este tipo de etiquetado se proponen dos tipos de medición de tiempos. La primera basada en la duración de las silabas y la segunda basada en la duración de pausas insertadas en los periodos con voz. Para obtener la duración de las silabas se utilizo la técnica descrita en (Wempe, et al., 2007), donde se detectan sílabas automáticamente sin necesidad de una transcripción. Para esto picos de intensidad que son precedidos por "*dips"* en la intensidad son considerados como sílabas potenciales que en un proceso posterior se confirman o se descartan. Después de detectar silabas se calcula el tiempo del tiempo total con presencia de voz en cada grabación. Finalmente, la velocidad de la voz para cada grabación se obtiene dividiendo el total de sílabas detectadas entre el tiempo total de voz. A partir de este procedimiento se calcularon los siguientes parámetros:

- *speech rate:* número de silabas entre tiempo total con voz en la elocución.
- *syllable duration mean:* a duración promedio de las sílabas en la elocución.
- *syllable duration standard deviation:* La desviación estándar de las duraciones de las silabas en la elocución.

El segundo tipo de características de tiempos se extrajo a partir del cálculo de tiempos de periodos de silencio y de voz en la elocución. Las características obtenidas son:

- *pause to speech ratio*: se obtiene dividiendo el número total de pausas entre el número total de segmentos con voz.
- *pause duration mean:* El promedio de duración de las pausas en una elocución.
- *pause duration std*: Es la desviación estándar de la duración de las pausas en la elocución.

- *speech duration mean*: El promedio de duración de los segmentos con voz en una elocución.
- *speech duration std*: La desviación estándar de los segmentos con duración en una elocución.

Contorno Melódico en la Elocución: Se obtuvo el pitch mediante el método de correlación y se calcularon las siguientes medidas estadísticas:

- *Pitch Average*
- *Pitch Standard Deviation*
- *Pitch Range*
- *Mínimum Pitch Point*
- *Máximum Pitch Point*
- *Pitch QuartRange*
- *Pitch 25 % quantile*
- *Pitch 75 % quantile*
- *Pitch Median*

Contorno de Energía: Los parámetros de energía describen características de la amplitud de la señal. Las siguientes medidas estadísticas fueron calculadas sobre el contorno de energía de la elocución.

- *Intensity Average*
- *intensity_std*
- *intensity_quartlow 25 % quantile*
- *intensity_quartup 75 % quantile*
- *intensity_min*
- *intensity_max*
- *intensity_range rango entre minimo y maximo*
- *intensity_range_q rango entre quantiles*

Calidad de Voz: Hemos incluido las siguientes características como descriptores de la calidad de voz, elegimos estas características por haber sido relacionadas con la escala GRBAS en trabajos relacionados (Dubuisson, et al., 2009) (Núñes B., et al., 2004) (Lugger, et al., 2006) (Ishi, et al., 2005). Gran parte de estas características nunca se habían usado en reconocimiento de emociones. Por ejemplo las diferencias de energía entre bandas de frecuencia y el radio entre bandas de frecuencia fueron usadas por (Dubuisson, et al., 2009) para discriminar entre voces patológicas y normales. Los incrementos y decrementos en picos de energía fueron usados por (Ishi, et al., 2005) para la detección automática de "*vocal fry*".

- *Jitter*
- *Shimmer*
- *fraction of locally unvoiced frames*
- *number of voice breaks*
- *degree of voice break*
- *harmony autocorrelation mean*
- *noise-to-harmonics ratio mean*
- *harmonics-to-noise ratio mean*
- *harmonicity mean*
- *power_rising_mean*
- *power_rising_std*
- *power_falling_mean*
- *power_falling_std*
- *harmonicity standard deviation*
- *harmonicity min*
- *harmonicity max*
- *energy difference between frequency bands 60-400 Hz, 400-2000Hz, 2000-5000Hz, 5000-8000H*
- *frequency bands ratio 60-400 Hz, 400-2000Hz, 2000-5000Hz, 5000-8000Hz*

La articulación también es un parámetro importante para medir la calidad de la voz. Como características articulatorias se incluyeron algunas medidas estadísticas de los primeros 4 formantes:

- *Formant1std*
- *Formant1mean*
- *Formant2mean*
- *Formant3std*
- *Formant4std*
- *Formant4mean*

- *Formant2std*
- *Formant3mean*

Características Espectrales: Incluimos 5 tipos de representaciones espectrales. Algunas de estas representaciones nunca se habían usado en reconocimiento de emociones como los Cocleagramas que se han empleado en clasificación de llanto de bebé (Santiago, et al., 2009) y otros se han estudiado muy poco para esta tarea como los Wavelets (Kandali, et al., 2009).

1) Representación de la señal en el dominio de la frecuencia mediante una transformada rápida de Fourier

   - *Skewness*
   - *kurtosis*
   - *std*
   - *centroid*

2) Long-Term Average Spectrum: Representa la densidad de energía espectral como una función de la frecuencia Es expresado en dB/Hz relativo a 2710-5 Pa.
   - *slope*
   - *min*
   - *std*
   - *max*
   - *mean*

3) Wavelets: Wavelets Representación de la señal mediante wavelets. Es una alternativa a la transformada de Fourier. La transformada Wavelet permite una buena resolución en las bajas frecuencias.
   - *variance*
   - *min*
   - *std*
   - *max*
   - *mean*
   - median

4) MFCC: Los Coeficientes Cepstrales de Frecuencia Mel (MFCC, por sus siglas en inglés), son coeficientes que representan el habla basándose en la percepción auditiva del ser humano. Las siguientes funciones estadísticas fueron calculadas para 16 coeficientes:
   - *variance*
   - *min*
   - *std*
   - *max*
   - *mean*
   - median

5) Cocleagrama: Un cocleograma representa la excitación de los filamentos del nervio auditivo de la membrana basilar, la cual se encuentra en la cóclea, en el oído interno. Esta excitación se representa como una función en el tiempo (en segundos) y frecuencia Bark. Las siguientes funciones estadísticas fueron calculadas para 16 coeficientes:

   - *variance*
   - *min*
   - *std*
   - *max*
   - *mean*
   - median

6) LPC: Linear Predictive Coding es utilizado principalmente para la compresión de datos y el análisis de una señal digital de habla. Este método está basado en el hecho de que una señal que transporta mensajes no es completamente aleatoria, es decir, existe una correlación entre muestras sucesivas, debido a que la voz mantiene sus propiedades prácticamente invariantes durante determinados intervalos de tiempo. LPC utiliza esta característica para reducir la cantidad de datos cuando se guarda la información de la señal. Las siguientes funciones estadísticas fueron calculadas para 16 coeficientes:

   - *variance*
   - *min*
   - *std*
   - *max*
   - *mean*
   - median

5.3 Selección de Características

A continuación se describen los métodos de selección de atributos que hemos usado hasta el momento en nuestros experimentos.

Subset Evaluation: Evalúa el valor de un subconjunto de atributos considerando la habilidad predictiva individual de cada característica junto con el grado de redundancia entre ellas. Se prefieren los subconjuntos de características que están altamente correlacionadas con la clase y bajamente correlacionadas entre ellas (Witten, et al., 2005). Usamos una búsqueda genética para encontrar el mejor subconjunto.

Relief Attribute Evaluation: Evalúa el valor de un atributo muestreando repetidamente una instancia y considerando el valor del atributo dado para la instancia más cercana de la misma y de diferente clase (Witten, et al., 2005). Como método de búsqueda usamos un método de ranqueo el cual ordena atributos de acuerdo a sus evaluaciones individuales.

Principal Components Analysis (PCA): Realiza un análisis de componentes principales y una transformación de los datos. Es usando en conjunto con un método de ranqueo. La reducción de dimensionalidad es completada escogiendo suficientes eigenvectores para alcanzar cierto porcentaje en la varianza de los datos (Witten, et al., 2005).

InfoGainAttributeEval: Evalúa los atributos de acuerdo a su ganancia de información. Considera un atributo a la vez.

5.4 Método de Reconocimiento de Emociones Propuesto

El método propuesto tiene dos fases una de entrenamiento y otra de aplicación. La primera fase se compone de dos sub-métodos. El primero lo denominamos Método General de Estimación de Primitivas. *E*ste método no sufrirá ninguna modificación al ser llevado de una aplicación a otra. El segundo sub-método es denominado Método Particular de Clasificación de Emociones Básicas. Este sub-modelo se adapta de una aplicación a otra de acuerdo a las categorías emocionales que se requiera clasificar. La segunda fase, denominada fase de aplicación se compone de dos partes. La primera está formada por los modelos de predicción de primitivas entrenados en el Método General de Estimación de Primitivas. La segunda parte está formada por el Sistema de Inferencia Difuso entrenado en el Método Particular de Clasificación de Emociones Básicas. El método propuesto se ilustra en las Figuras 3, 4 y 5.

Fase de Entrenamiento

A) Método General de Estimación de Primitivas. (Ver Fig. 3)

1. **DB Primitivas Emocionales:** La entrada de este método es un corpus con la característica de que las instancias están etiquetadas con valores en una escala continua para las primitivas emocionales: Valencia, Activación y Dominación. Por el momento el único corpus que tenemos con estas características es el corpus VAM con el cual estaremos trabajando.
2. **Extracción de Características:** Este módulo de nuestro método realiza las siguientes tareas:

- Preprocesa al audio con el fin de eliminar ruido, eliminar silencios largos, eliminar dependencias del hablante, y otros factores que afecten la fidelidad de las características acústicas extraídas.
- Extrae las características descritas en la sección 5.3. Hasta el momento se tiene un conjunto de 369 características. Sin embargo, como se verá más adelante, se ha detectado que estas características no son buenas para estimar la primitiva Valencia por lo que se seguirá trabajando en buscar características que funcionen mejor para estimar dicha primitiva.
- Forma 3 grupos de características seleccionando las mejores para estimar cada primitiva emocional.

3. **Entrenamiento SVM:** Mediante Máquinas de Vectores de Soporte se entrenan 3 modelos para predecir cada una de las primitivas emocionales. Las instancias están conformadas por características acústicas cuyo valor a predecir son cada una de las primitivas.
4. **Modelos D, A y V**: El resultado del módulo 3 son tres modelos, uno para predecir cada una de las primitivas emocionales. La salida del Método General de Estimación de Primitivas es un conjunto de datos que está conformado por las predicciones de Valencia, Activación y Dominación para cada una de las instancias dadas como entrada al método.

B) Método Particular de Clasificación de Emociones Básicas. (Ver Fig. 4)

5. **BD Emociones Básicas**: La entrada de este método es una base de datos que está etiquetada con emociones básicas. Hasta el momento hemos probado dos corpus con esta característica, el corpus FAU Aibo y el corpus Berlin Emotional. Esta base de datos nos servirá para validar la ubicación de emociones básicas en el espacio tridimensional continuo.
6. **Transformación**: Este modulo substituye las etiquetas de emociones básicas por valores que corresponden a etiquetas lingüísticas, por ejemplo, si una instancia está etiquetada como enojo se sustituirá dicha etiqueta por 3 etiquetas:
    a. bajo para la primitiva valencia
    b. alto para la primitiva activación
    c. alto para la primitiva dominación
7. **Extracción de Características**: Extrae las mismas características acústicas que el modulo 2.
8. **Entrenamiento SVM**: Se entrenan 3 modelos para predecir cada una de las primitivas emocionales. Las instancias están formadas por las características acústicas cuyo valor a predecir son las etiquetas generadas en el paso 6.
9. **Modelos D, A y V**: El resultado del módulo 8 son tres modelos, uno para cada una de las primitivas emocionales. Usando estos modelos se genera un conjunto de datos que está conformado por las predicciones de Valencia, Activación y Dominación para cada una de las instancias dadas como entrada al método. Además el conjunto de instancias generado mantiene las etiquetas de emociones básicas con el que contaba originalmente la base de datos.
10. **Entrenamiento ANFIS:** Mediante este módulo se generan reglas que nos sirven para mapear emociones básicas en el espacio tridimensional. La entrada son las primitivas emocionales estimadas por los modelos del módulo 9 junto con la emoción básica que definen. La salida es un conjunto de reglas y funciones de membresía que servirán para clasificar emociones básicas a partir de los valores de las primitivas emocionales.
11. **Clasificación mediante Modelo FIS**: La salida del módulo 10 y del Método Particular de Clasificación de Emociones Básicas es un Sistema de Inferencia Difuso. La entrada de este módulo son los valores de las 3 primitivas emocionales estimadas por el Método General de Estimación de Primitivas y la salida es la emoción básica a la cual se asemeja más.

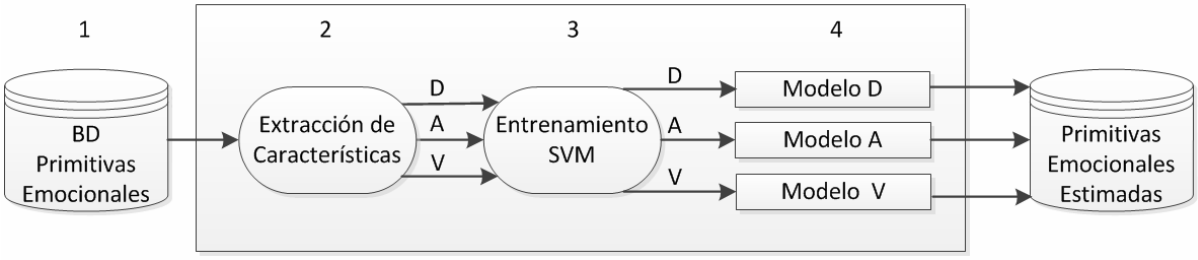

Fig 3. Método General de Estimación de Primitivas

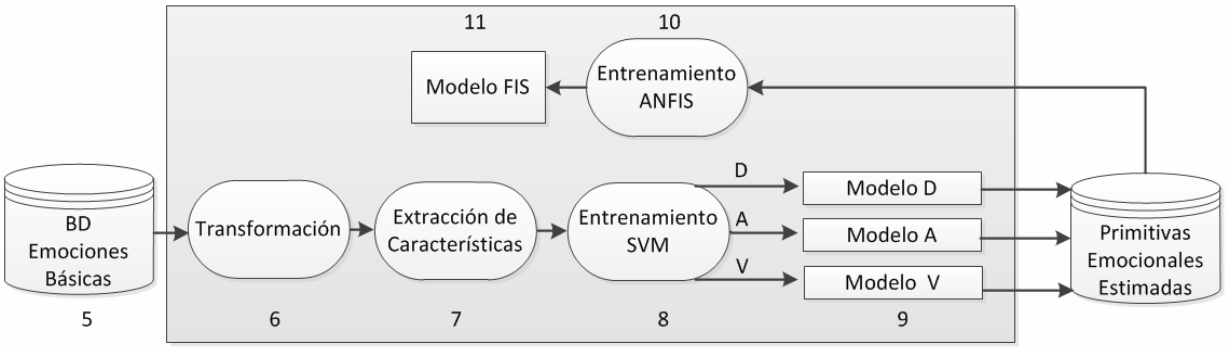

Fig 4. Método Particular de Clasificación de Emociones Básicas

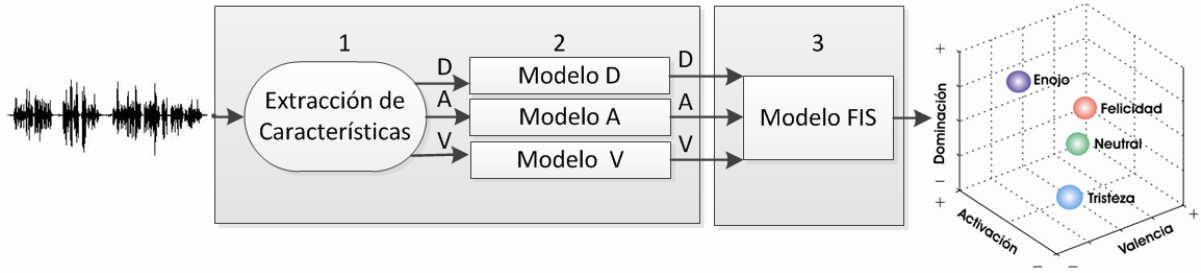

Fig 5. Fase de Aplicación

Fase de Aplicación (Ver Fig. 5)

1. **Extracción de Característica:** La entrada de este módulo es una señal de voz. Extrae las mismas características que el módulo 2 de la fase de entrenamiento.
2. **Modelos D, A y V:** Estiman el valor de Dominación, Activación y Valencia de la señal dada de acuerdo al entrenamiento hecho por el Método General de Estimación de Primitiva.

3. Clasificación de Emociones: Ubica el estado emocional de acuerdo a las emociones básicas que se requieran de acuerdo al entrenamiento hecho por el Método Particular de Clasificación de Emociones Básicas.

6 Experimentos

Experimento 1: Análisis de Características Acústicas Clasificando 2 y 5 Categorías Emocionales

Objetivos:

- Comprobar la viabilidad de los pasos 1 y 2 de la metodología
- Probar el conjunto de características propuesto hasta el momento
- Integrar nuevas características usadas en otros campos
- Aplicar técnicas de selección de atributos para identificar características valiosas
- Compararse con los mejores resultados del *INTERSPEECH 2009 Emotion Challenge* clasificando 2 y 5 clases

Configuración del Experimento:

Se clasificaron 2 y 5 clases emocionales del corpus FAU Aibo, los detalles de esta base de datos se dan en la sección 5.1. Incrementamos el número de características usadas. Pasamos de 21 características acústicas usadas en experimentos anteriores a 369. La mayor parte de estas características son espectrales. Aún cuando los mejores resultados en experimentos anteriores se obtuvieron clasificando con ANFIS y Support Vector Machines, para este experimento usamos solamente Support Vector Machines ya que los recursos que ANFIS requiere crecen de manera exponencial de acuerdo al número de características usadas. No descartamos el uso de ANFIS ya que hay técnicas para optimizar su uso, sin embargo, esa posibilidad la exploraremos más adelante. Se usaron técnicas de selección de atributos. Debido a que las clases de este corpus están desbalanceadas se probó SMOTE como técnica de balaceo de clases, la cual genera muestras sintéticas de la clase minoritaria a partir de un número definido de vecinos más cercanos. En la prueba con 2 clases tenemos las clases No Negativa y Negativa. El número de instancias por clase es No Negativa 3,357 y Negativa 6,596 Para la prueba con 5 clases las clases son: Enojo, Enfático, Neutral, Positivo y Resto. También, usamos SMOTE para balancear las clases. En este caso el número de instancias por clase son: Neutral 5,586, Enfático 2,092, Enojado 881, Resto 720 y Positivo 674. El principal criterio de comparación con otros trabajos es el promedio no ponderado del recuerdo. Este criterio fue sugerido por los organizadores del INTERPEECH 2009 Emotion Challenge y se calcula de la siguiente manera:

$$\frac{(CC\_c1 / TI\_c1) + \ldots + (CC\_cN / TI\_cN)}{N}$$

CC_ci: Clasificados Correctamente de la clase i
TI: Número de Instancias
N: Número de clases

Resultados:

En la Tabla 3 se muestran algunos de los resultados obtenidos en experimentos con 2 clases previos a este, en los cuales se estuvo usando un conjunto de 21 características acústicas. El mejor promedio no ponderado del recuerdo fue 0.6041 el cual hemos superado usando nuestro nuevo conjunto de características. Con el conjunto de características actual, en el caso de la clasificación de Negativo y No Negativo los mejores resultados los obtuvimos usando todos los atributos y balanceando el

corpus con SMOTE. Obtuvimos un recuerdo promedio no ponderado de 0.7566, Ver tabla 4, lo cual se asemeja al mejor resultado del INTERSPEECH 2009 emotion challenge que fue 0.7590. En dicho trabajo (Polzehl, et al., 2009) usaron características lingüísticas, prosódicas, espectrales / cepstrales y de calidad de voz. Usaron Information Gain para seleccionar 320 de entre 1,500 características. Igualaron el número de instancias de ambas clases dividiendo en dos algunas de las instancias etiquetadas como Negativos. Como clasificadores emplearon Multi Layer Perceptron y SVM. Similarmente, en la clasificación de 5 clases obtuvimos el mejor resultado con todos los atributos y con el corpus balanceado con SMOTE. En este caso se llegó a un recuerdo no ponderado de 0.4564, ver Tabla 5, lo cual se acerca al mejor resultado del INTERSPEECH 2009 emotion challenge que fue 0.4827. En dicho trabajo (Lee, et al., 2009) usaron las características propuestas por los organizadores del concurso. Para balancear el corpus se hizo un ajuste de umbral de decisión basado en la distribución de clases. Como métodos de clasificación usaron Bayesian Logistic Regresion y SVM con los cuales crearon un árbol de clasificadores binarios en el que la clase de un clasificador es pasada al siguiente.

Tabla 3. Experimentos Previos clasificando 2 clases con 21 características

Recuerdo	
Promedio no Ponderado	**Promedio Ponderado**
0.6041	0.7231
0.5990	0.7207
0.5962	0.7190

Tabla 4. Resultados Clasificando 2 clases Negativa y No Negativa.

Técnica de Selección de Atributos	**Recuerdo**	
	Promedio no Ponderado	**Promedio Ponderado**
Sin Balanceo		
Ninguna	**0.7186**	0.7769
SubsetEval	0.6918	0.7587
ReliefAttribute	0.6994	0.7653
Balanceado Por SMOTE		
Ninguna	**0.7566**	0.7566
SubsetEval	0.7495	0.7495
ReliefAttribute	0.7526	0.7526
Referencia		
Polzehl, et al., 2009	***0.7590***	***0.7600***

En las Tablas 6 y 7 se muestran las 20 mejores características encontradas por los selectores de atributos. En el experimento con 2 clases se encontraron 49 características diferentes. 9 características aparecieron en dos de los selectores y una, intensity_average, apareció en los tres. En el experimento con 5 clases se encontraron también 49 características diferentes. 9 características aparecieron en dos de los selectores y una, pause_to_speech_ratio, apareció en los tres.

Tabla 5. Resultados Clasificando 5 clases: Enojo, Enfático, Neutral, Positivo y Resto.

Técnica de Selección de Atributos	Recuerdo		Precisión	
	Promedio no Ponderado	Promedio Ponderado	Promedio no Ponderado	Promedio Ponderado
Sin Balanceo				
Ninguna	**0.3596**	0.6354	0.5871	0.6148
SubsetEval	0.3138	0.6211	0.5165	0.5781
ReliefAttribute	0.3371	0.6290	0.5088	0.5805
Balanceado Por SMOTE				
Ninguna	**0.4564**	0.5925	0.6022	0.5939
SubsetEval	0.4133	0.5715	0.6293	0.6018
ReliefAttribute	0.4419	0.5860	0.6244	0.6040
Referencia (LogReg - SVM)	***0.4827***	***0.4882***	***0.3945***	***0.5811***

Tabla 6. Mejores Características con 2 Clases

SubsetEval	Info Gain	ReliefAttribute
coclea_max1	coclea_max11	coclea_var2
coclea_prom15	intensity_average	pause_to_speech_ratio
wl_var	ltas_mean	mean_noise
syllable_duration_std	wl_var	mfcc_std4
syllable_duration_mean	wl_std	coclea_std2
pitch_std	coclea_max10	harmonicity_mean
lpc_var4	coclea_max12	mean_autocorrelation
formant3_stdev	ltas_std	mfcc_min4
wl_max	intensity_max	mfcc_var4
mfcc_std2	wl_max	coclea_max11
mfcc_std13	wl_min	mfcc_min3
energy_diff2	coclea_var13	mfcc_max15
coclea_max4	coclea_var12	formant2_stdev
coclea_std5	coclea_std13	coclea_max10
coclea_max6	coclea_std12	ltas_mean
number_voice_breaks	coclea_max13	intensity_average
mean_harmonics	coclea_max9	coclea_max4
coclea_var15	coclea_std5	coclea_max2
intensity_average	coclea_var15	mfcc_max16
wl_std	coclea_var11	coclea_var4

Tabla 7. Mejores Características 5 Clases

SubsetEval	Info Gain	ReliefAttribute
ltas_max	ltas_max	pause_to_speech_ratio
number_voice_breaks	number_voice_breaks	mfcc_std4
pause_to_speech_ratio	pause_to_speech_ratio	coclea_var2
coclea_max1	coclea_max11	coclea_std2
ltas_mean	intensity_average	mean_noise
pause_duration_mean	wl_var	harmonicity_mean
syllable_duration_std	ltas_mean	mfcc_var4
pitch_average	wl_std	mean_autocorrelation
pitch_quartup	coclea_max10	mfcc_min1
energy_diff1	coclea_max12	mfcc_min4
intensity_average	ltas_std	mfcc_min3
pitch_range	intensity_max	coclea_var4
intensity_std	wl_max	pitch_range
wl_var	wl_min	pitch_max
jitter	coclea_max13	coclea_std4
mfcc_prom9	coclea_std12	coclea_max2
coclea_var5	coclea_std13	pitch_quartup
coclea_var12	coclea_var12	mfcc_std3
coclea_min15	coclea_var13	coclea_prom2
coclea_std15	intensity_std	mfcc_std2

Conclusiones:

- Las características espectrales incorporadas y la inclusión de más características prosódicas y de calidad de voz mejoran mucho la clasificación. Clasificando 2 clases se pasó de un recuerdo no ponderado de 0.6041 a uno de 0.7186 usando SVM, con todos los atributos y sin balancear.
- Los resultados obtenidos son comparables con los resultados de los trabajos ganadores del INTERSPEECH 2009 emotion challenge, lo cual indica que hemos obtenido buenos resultados comparables con los del estado del arte.
- No existió acuerdo entre las características más importantes encontradas por los diferentes selectores de características.
- No se logró mejorar los resultados mediante selección de atributos.

Experimento 2. Incorporación de características lingüísticas

Objetivos:

- Comprobar la viabilidad del paso 3 de la metodología
- Incluir información que complemente la información acústica extraída de la voz.
- Probar si es de utilidad fusionar la información lingüística extraída a partir de las transcripciones con la información lingüística.
- Probar técnicas usadas en categorización de texto.

Configuración del Experimento:

Tradicionalmente el reconocimiento de emociones se ha hecho únicamente sobre información acústica extraída de la señal de voz. Sin embargo, varios autores (Lee, et al., 2002) (Liscombe, et al., 2005) (Polzehl, et al., 2009) han mostrado que usando información lingüística se puede mejorar la clasificación basada en características acústicas. En este experimento exploramos el uso de información lingüística, extraída de las transcripciones del corpus FAU Aibo para clasificar emociones Negativas y No Negativas. Se prueban tres métodos de selección de características y dos métodos de balanceo. Se usa la técnica conocida como *Bolsa de Palabras*, en este enfoque se representa el texto de manera numérica, en nuestro caso, la representación de información se basa en la especificación de presencia o ausencia de palabras en cada clase. Bolsa de Palabras es una técnica muy usada en categorización automática de texto. Cada palabra agrega una dimensión a un vector lingüístico representando la presencia o ausencia dentro de la elocución. Ya que resultan vectores muy grandes se suele usar alguna técnica de reducción de dimensionalidad. Nosotros hemos eliminado las "Stop Words".

Se tienen las siguientes estadísticas de nuestro corpus:

Instancias	9,959	
Palabras totales	16,196	
No Negativas	11,834	
Negativas	4,362	
Palabras distintas	793	
Palabras más repetidas		
Aibo	2,017	(sustantivo)
Nach	1,191	(stopword prep.)
Links	1,091	(adverbio)
Dich	630	(pronombre)
Stopwords buscadas	129	
Stopwords Eliminadas	77	
Total de palabras en el vocabulario Final	716	

Utilizamos dos formas diferentes de balancear las clases ya que el corpus está desbalanceado. Se cuentan con 3,358 ejemplos de frases negativas y 6,601 ejemplos de frases no negativas. Se representó cada instancia con un vector de 716 dimensiones correspondientes al tamaño del vocabulario. Se marcó con un "1" la presencia de la palabra correspondiente en el ejemplo y con un "0" la ausencia de ella.

Para balancear las clases se probaron dos métodos:

- SMOTE: El cual genera muestras sintéticas de la clase minoritaria a partir de un número definido de vecinos más cercanos
- Remuestreo: Se toman de cada clase el número de muestras que tenga la clase minoritaria

Se probaron tres técnicas de selección de atributos, con el objetivo de reducir la dimensionalidad de los vectores de características usando solo las palabras que aportan mayor información para discriminar entre ambas clases emocionales.

Resultados:

Se usó el 80% de las muestras para entrenar y el 20% para probar. En el caso de SMOTE se clasificaron 13,317 muestras 6,601 no negativas y 6,716 negativas. En el caso de remuestreo se clasificaron 6,716 muestras, 3,358 de cada clase. Los mejores resultados se obtuvieron cuando se usó el total de 716 atributos sin usar ninguna técnica de selección de atributos. Ninguna técnica de

selección dio mejores resultados en general. Se obtuvieron mejores resultados usando SMOTE como técnica de balanceo. En la Tabla 8 se muestra una comparación entre el desempeño logrado clasificando sólo con características acústicas, sólo con características lingüísticas y fusionando ambos tipos de características. Los mejores resultados se obtuvieron prediciendo la clase de las instancias con características lingüísticas e incorporando esta predicción como una característica más al conjunto de características acústicas. En esta misma tabla se muestran los resultados de referencia obtenidos por (Polzehl, et al., 2009) donde podemos observar que también obtuvieron una mejora en el desempeño de la clasificación al fusionar información acústica y lingüística.

Tabla 8. Resultados de clasificar emociones fusionando información acústica y lingüística

Características	Recuerdo		Precisión		Referencia (Keyword Spotting –MLP/SVM) Recuerdo	
	Promedio no Ponderado	Promedio Ponderado	Promedio no Ponderado	Promedio Ponderado	Promedio no Ponderado	Promedio Ponderado
Acústicas	0.7284	0.7284	0.7415	0.7415	0.7530	0.7440
Lingüísticas	0.8137	0.8137	0.8145	0.8145	0.7120	0.7030
Combinación	**0.8705**	0.8749	0.8750	0.8750	0.7590	0.7600

Conclusiones:

- La información lingüística extraída de las transcripciones del corpus FAU Aibo es útil para discriminar entre los estados emocionales negativo y no negativo.
- Para esta tarea la selección de atributos no ofrece ningún beneficio por lo que es mejor usar el vocabulario completo
- Algunas de las palabras con mayor poder de discriminación como el sustantivo “Aibo” o los verbos “ir” y ”venir” son exclusivas o muy relacionadas con el contexto

Experimento 3: Selección de Características para la Estimación de Primitivas

Objetivos:

- Comprobar la viabilidad de los pasos 4 y 5 de la metodología
- Probar una versión inicial de nuestro Método General de Estimación de Primitivas
- Probar nuestras características, clasificadores y métodos de selección en una tarea más enfocada a los objetivos de la tesis como es el estimar primitivas emocionales que definen un modelo emocional continuo.
- Comparar nuestros resultados con los resultados de un trabajo relacionado (Grimm, et al., 2007)

Configuración del Experimento:

En este experimento se estudia la influencia de diferentes grupos de características acústicas del habla en la estimación de primitivas emocionales que definen un modelo emocional tridimensional continuo. Se extrae un conjunto de características propuestas a partir de un corpus de emociones espontáneas. Se usó el corpus VAM, el cual está etiquetado con primitivas emocionales. Este corpus está en alemán. Estas características representan distintos aspectos importantes de la voz

permitiendo predecir con un grado de exactitud aceptable las primitivas emocionales Valencia, Activación, y Dominación. Se aplican algunas técnicas de selección de atributos y de reducción de dimensionalidad de las características propuestas para encontrar el mejor subconjunto para estimar cada una de las primitivas y se muestran los resultados obtenidos.

Selección de Características: Se realizó una selección de atributos mediante *Subset Evaluation* y algoritmos genéticos así como con *Relief Attribute Evaluation* con renqueo de atributos (Witten, et al., 2005). También, se aplicó la técnica de reducción de dimensionalidad llamada *PCA*. El desempeño de los subconjuntos de características fue evaluado por el clasificador *SVR*.

Subset Evaluation: Buscamos los mejores subconjuntos de características para cada primitiva de manera separada Se realizó una validación cruzada de 10 pliegues, obteniendo de esta forma 10 subconjuntos diferentes. Los conjuntos finales de características se obtuvieron tomando las características que estuvieron presentes en más del 80% de los subconjuntos.

Relief Attribute Evaluation: Se buscó por separado los mejores subconjuntos para cada primitiva. Realizamos una validación cruzada de 10 pliegues. Probamos con diferentes números de *n* mejores atributos para mejorar el desempeño.

Principal Components Analysis (PCA): Se probó tomando diferente número de *n* mejores eigenvectores y probando su desempeño en la clasificación.

Tabla 9 Resultados de evaluación de subconjunto

Selector de atributos	# Atributos	Valencia	Activación	Dominación
CfsSubsetEval	45	0.3575	**0.8121**	0.7665
CfsSubsetEval	57	0.3711	0.7818	**0.7927**
ReliefAttEval	57	**0.4415**	0.7923	0.7803
PCA	62	0.4013	0.8049	0.7858
Todas las características	252	0.3682	0.7897	0.7673

Tabla 10 Evaluación Correlación/Error

Primitiva Emocional	Bagging SVM	Bagging PaceRegression
Valencia	0.4421/0.1339	**0.4515/0.1319**
Activación	0.8063/0.1596	**0.8149/0.1563**
Dominación	0.7882/0.1450	**0.7990/0.1418**
Promedio	0.6789/0.1462	**0.6885/0.1433**
Referencia	0.6000/0.2400	

Resultados:

Los resultados de la evaluación del desempeño de los subconjuntos de características encontrados se muestran en la Tabla 9. El mejor subconjunto de características para predecir la primitiva Valencia se encontró utilizando *Relief Attribute Evaluation*; el subconjunto consta de 57 características. Para la primitiva Activación el mejor subconjunto se encontró utilizando *Subset Evaluation* y se compone de 45 características.

Tabla 11. Características Seleccionadas para cada primitiva ordenadas por importancia de grupo. Elementos como mfcc_mean 5 significa que la media del 5to y 6to coeficientes se seleccionaron

Valencia	Activación	Dominación
Espectral	**Espectral**	**Espectral**
ltas_max	*fft_centroid*	*fft_centroid*
tas_min	*wavelet_min*	*wavelet_min*
fft_std	*wavelet_std*	*wavelet_std*
fft_centroid	*mfcc_median 2,3,9*	*mfcc_median 1,2,3,9*
wavelet_var	*mfcc_min3*	*mfcc_min 3,7,9,11,12*
wavelet_max	*mfcc_var 3,5,9,12-16*	*mfcc_var 3,5,910,12-16*
wavelet_min	*mfcc_max 4,7,8,11,13*	*mfcc_max4,7,8,10,11,13*
wavelet_std	*mfcc_std 5*	*mfcc_std 5,6*
mfcc_median 1,3,4,9,11,13	*mfcc_mean 6,9,14*	*mfcc_mean 6,9,11,14*
mfcc_mean 1,3-6,9,11,13	*cochlea_mean 1,9,13*	*cochlea_mean 1,3,9,13*
mfcc_var 2,6,10,12-15	*cochlea_med 2,5,7*	*cochlea_med 2,5,7*
mfcc_max 1,3,5,7,10-15	*cochlea_max 10,16*	*cochlea_max 1, 10,16*
mfcc_min 2,3,12	*cochlea_var 15*	*cochlea_var 15*
mfcc_std 2,4,10,12-14	**C. Melódico**	*cochlea_std 2*
	Mean	*cochlea_med 2,5,7*
Calidad de Voz	*25_quantile*	**C. Melódico**
h_t_n_r_mean	*75_quantile*	*Mean*
f1_std	*Median*	*25_quantile*
f1_mean	**Calidad de Voz**	*75_quantile*
f2_50_quantile	*voice_breaks*	*Median*
	h_autocorrelation	**Calidad de Voz**
C. Melódico	*h_t_n_r_mean*	*voice_breaks*
75_quantile	*f2_mean*	*h_autocorrelation*
quantile_range		*h_t_n_r_mean*
Median	**Energía**	*f2_mean*
	Std	**Energía**
Tiempos	*25_quantile*	*std*
voice_duration_mean	*Minimum*	*25_quantile*
	quantile_range	*minimum*
		quantile_range

Por último, para la Dominación se llegó a un subconjunto de 74 características utilizando S*ubset Evaluation*. Para mejorar los resultados de clasificación obtenidos con nuestros subconjuntos de características finales se probó con otros clasificadores como el *Pace Regression* y con ensambles de clasificadores como *Bagging.* Los resultados se muestran en la Tabla 10. Usando un ensamble de Clasificadores *Pace Regression* se llegó a estimar las primitivas emocionales con una correlación promedio de 0.6885 y con un error medio absoluto de 0.1433. Estos resultados mejoran los reportados (Grimm, et al., 2007), donde también se estimaron primitivas emocionales del mismo corpus obteniendo una correlación promedio de 0.6000 y un error medio de 0.2400. En la Tabla 11 se muestran las características seleccionadas para cada una de las primitivas. Para la Valencia fueron seleccionadas características de todos los tipos excepto de energía. Las características que estuvieron mejor clasificadas por el selector Relief Attribute Evaluation fueron las espectrales, siguiendo las de calidad y las prosódicas. En el caso de la activación, sí se incluyeron características

de energía, pero ninguna de tiempos de la elocución. Para la activación, los atributos de contorno melódico estuvieron en promedio mejor clasificados que los de calidad de voz. Esto a diferencia del subconjunto de valencia, donde los atributos de calidad tuvieron más importancia que los de contorno melódico. Finalmente, para el subconjunto de Dominación se eligieron las mismas características para contorno melódico, calidad de voz y Energía que en el caso de Activación. La diferencia entre ambos subconjuntos fueron los atributos espectrales seleccionados. En el caso de la dominación se eligieron más atributos espectrales.

Conclusiones:

- Es viable usar conjuntos de características diferentes para cada primitiva emocional, ya que para cada primitiva hay ciertas características del habla que tienen un mayor peso de discriminación.
- En el caso de la Valencia, los grupos de características más importantes son: Espectrales, Contorno Melódico, y Calidad de la voz; son menos importantes los de tiempos de elocución y las de energía. Esto puede explicarse porque existen emociones que se ubican en extremos opuestos del eje de Valencia y que presentan un nivel energético similar, y/o emociones que están cercanas en el eje de valencia y sin embargo tienen un nivel de energía muy diferente.
- En la caso de la Activación, los grupos de características más importantes son: Espectrales, Energía, Contorno Melódico y Calidad de Voz. La energía toma mayor importancia para esta primitiva ya que el nivel de energía en la voz parece estar más relacionada con que tan relajada o excitada esté una persona.
- La Dominación comparte con la Activación el orden de importancia en los grupos de características. Sin embargó, parece necesitar más información sobre la distribución de energía en los diferentes rangos de frecuencia de la voz. Esto se podría explicar debido a que el hablante modifica su timbre o para establecer cierta intencionalidad en el habla.
- El conjunto de características es confiable ya que se obtuvieron buenos resultados de clasificaciones incluso mejorando los resultados obtenidos en (Grimm, et al., 2007).
- Predecir valencia es más difícil que predecir Activación y Dominación
- Las características más relevantes para predecir primitivas emocionales difieren de las más importantes para clasificar emociones básicas.

Experimento 4: Ubicación de Emociones básicas en Espacio Emocional Continuo Mediante Estimación de Primitivas

Objetivos:

- Comprobar la viabilidad del paso 6 de la metodología
- Mostrar la viabilidad de usar un modelo continuo para mapear emociones básicas en un espacio continuo.
- Comprobar que mediante el uso de primitivas emocionales se puede llegar a emociones básicas.
- Comprobar la exactitud de predicción y extracción de primitivas emocionales de nuestro Método General de Estimación de Primitivas

Configuración del Experimento:

Se entrenó un modelo para predecir primitivas emocionales a partir del corpus VAM. Usando este modelo se predijeron primitivas emocionales para las instancias del corpus Berlin Emotional. Usando estas predicciones como atributos se entrenó un modelo para predecir las clases emocionales con las cuales está etiquetado el corpus Berlin Emotional. A partir de la ubicación en 2 dimensiones de emociones básicas hechas por (Scherer, 2001) Activación / Valencia. (Cowie, et al., 2000) Activación / Valencia y (Bänziger, et al., 2005) Dominación / Valencia construimos la tabla 12 que ubica un conjunto de 7 estados emocionales básicos en un espacio tridimensional. En la Fig. 6 se ilustra la ubicación de dichas emociones. La hipótesis de este experimento es que al predecir las primitivas emocionales las instancias de esta base de datos serán ubicadas de acuerdo a la posición de la clase emocional correspondiente en la Fig 6. Se realizó una clasificación entre todos los pares de emociones con el fin de evaluar su ubicación en el espacio tridimensional dada por las primitivas emocionales predichas.

Tabla 12. Ubicación de emociones en espacio 3D según Scherer - Cowi - Banzinger

Emoción	Valencia	Activación	Dominación
enojo	1.25	8.5	9
felicidad	8.25	6.5	6
aburrimiento	3	1	2.5
miedo	2.75	8	1.5
tristeza	1.5	3.5	2.5
disgusto	1	7.5	7
neutral	5	5	5

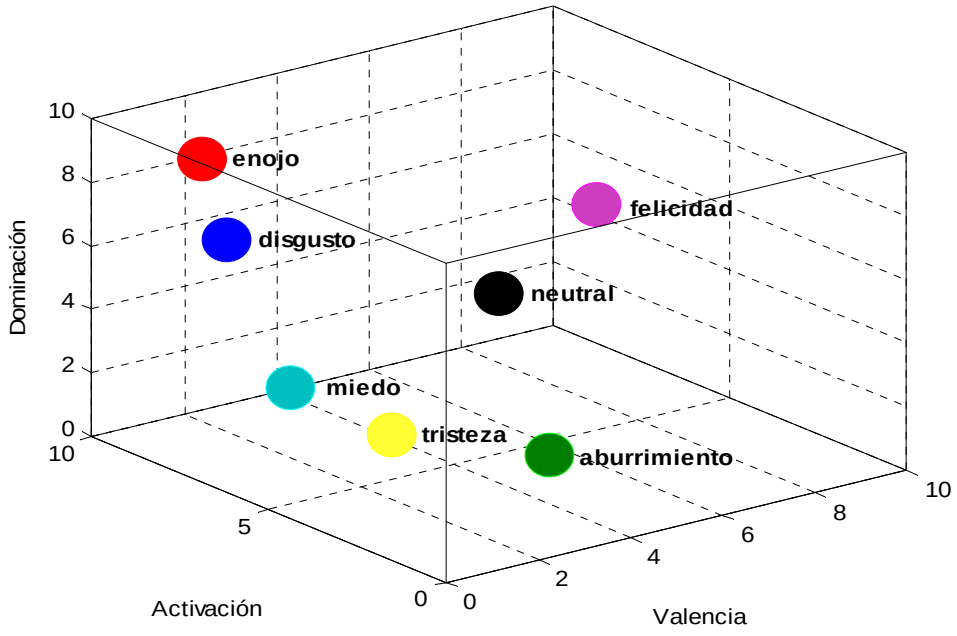

Fig 6. Ubicación de emociones en espacio 3D según Scherer - Cowi - Banzinger

Resultados:

En las primeras 4 filas de la tabla 13 se muestran los pares de emociones que al clasificarse obtuvieron mejor desempeño. En las últimas 4 filas se muestran los pares de emociones que al clasificarse tuvieron peores resultados. Puede observarse que en general, las emociones que tienen mayor distancia entre ellas tienen una mejor exactitud de clasificación, mientras que las emociones que tienen menor distancia tienen mala exactitud de clasificación. En la Tabla 14 se muestran la exactitud de clasificación y la distancia euclidiana entre todos los pares de emociones. Se encuentran resaltadas las combinaciones donde no se cumplió la observación hecha anteriormente. Es decir, para la combinación Felicidad – Enojo donde la distancia fue grande (7.8) la clasificación fue muy mala (57.7). Por otro lado, para combinación Tristeza – Miedo donde la distancia fue menor (7.0) la clasificación fue buena (73.9). En el caso de Felicidad – Enojo, podemos ver en la Tabla 12 que estas dos emociones están separadas principalmente por su valor de Valencia, mientras que Tristeza – Miedo están separadas principalmente por su valor de Activación. Esto nos lleva a pensar que el mal desempeño en la predicción de Valencia, como se mostró en el experimento 3, está afectando la distinción de emociones que difieren principalmente en esta primitiva.

Tabla 13. Distancia Euclidiana entre Emociones y Exactitud de Clasificación

	Emociones	**Distancia**	**Exactitud**
Buen Desempeño de Clasificación	Enojo vs Aburrimiento	10.08	75.3
	Felicidad vs Aburrimiento	8.37	66.9
	Enojo vs Tristeza	8.20	70.9
	Felicidad vs Tristeza	8.17	73.4
Mal Desempeño de Clasificación	Tristeza vs Disgusto	6.04	56.3
	Disgusto vs Miedo	5.79	56.3
	Aburrimiento vs Tristeza	2.92	58.1
	Enojo vs Disgusto	2.25	55.2

Tabla 14. Matriz de Distancia Euclidiana y Exactitud de Clasificación entre pares de Emociones

Exactitud/Distancia	Enojo	Aburrimiento	Disgusto	Miedo	Felicidad	Tristeza
Enojo	100/0	75.3/10.1	55.2/2.2	62.3/7.6	57.7/7.8	70.9/8.2
Tristeza	75.3/10.1	100/0	67.7/8.1	73.9/7.0	66.9/8.3	58.1/2.9
Disgusto	55.2/2.2	67.7/8.1	100/0	56.3/5.7	51.1/7.3	56.3/6.0
Miedo	62.3/7.6	73.9/7.0	56.3/5.7	100/0	52.1/7.2	67.9/4.7
Felicidad	57.7/7.8	66.9/8.3	51.1/7.3	52.1/7.2	100/0	73.4/8.1
Tristeza	70.9/8.2	58.1/2.9	56.3/6.0	67.9/4.7	73.4/8.1	100/0

Conclusiones:

- A pesar de trabajar con 2 corpus diferentes el mapeo de emociones básicas en un espacio emocional continuo es congruente.
- El modelado continuo de emociones parece ser una opción viable para la categorización de estados emocionales.
- El cálculo de la valencia está afectando fuertemente el desempeño de clasificación

Experimento 5: Generación de Reglas para Mapear Emociones Básicas en un Espacio Emocional Tridimensional Continuo

Objetivos:

- Comprobar la viabilidad del paso 6 de la metodología.
- Probar una versión inicial de nuestro Método Particular de Clasificación de Emociones Básicas
- Probar un método para mapear emociones básicas en un espacio tridimensional continuo
- Incorporar información conocida sobre la ubicación de emociones básicas en un espacio tridimensional continuo

Configuración del Experimento:

Se usó el corpus Berlin Emotional, del cual se extrajeron características acústicas. Como se mencionó anteriormente, este corpus está etiquetado con emociones básicas. Se sustituyeron las etiquetas de emociones básicas por etiquetas lingüísticas, como se explico en la sección 5.4 módulo 6 de la fase de entrenamiento del Método Particular de Clasificación de Emociones Básicas. Se entrenaron 3 modelos para predecir cada una de las primitivas emocionales. Las instancias están formadas por las características acústicas cuyo valor a predecir son las etiquetas lingüísticas generadas en el paso anterior. Aplicando dichos modelos se predijeron los valores para las primitivas emocionales para todas las instancias del corpus usando un 10 – fold cross validation. Se creó un Sistema de Inferencia para mapear las primitivas emocionales a emociones de acuerdo a las siguientes reglas basadas en la Tabla 12:

- Si activación es alto y dominación es bajo entonces la emoción es Miedo
- Si activación es alto y dominación es alto y valencia es bajo entonces la emoción es Disgusto/Enojo
- Si activación es alto y dominación es alto y valencia es alto entonces la emoción es Felicidad
- Si activación es bajo y dominación es bajo y valencia es alto entonces la emoción es Aburrimiento/Neutro
- Si activación es bajo y dominación es bajo y enojo es bajo entonces la emoción es tristeza

A diferencia del Método Particular de Clasificación de Emociones Básicas propuesto, en este experimento se implemento una versión inicial del método generando manualmente estas reglas en lugar de generar un sistema de inferencia difuso a través de un entrenamiento con ANFIS.

Resultados:

En la Tabla 15 se muestran los resultados de la clasificación para 5 y 7 emociones hechos por nuestras primitivas emocionales y el sistema de inferencia que construimos manualmente en comparación con la clasificación hecha con características acústicas y Máquinas de Vectores de Soporte. Cabe señalar que las características acústicas usadas en este experimento son las características que nosotros proponemos y que se han explicado en la sección 5.2. Como hemos mencionado anteriormente el principal criterio de comparación que estamos usando es el promedio no ponderado del recuerdo. Como se observa en la Tabla 15, clasificando 5 clases se obtuvieron mejores resultados con nuestra base de reglas. Por el contrario, clasificando 7 clases se obtuvieron mejores resultados usando características acústicas y SVM.

Tabla 15. Comparación entre resultados obtenidos mediante una versión preliminar del Método Particular de Clasificación de Emociones Básicas y Clasificación con Características Acústicas

	Primitivas Emocionales -Base de Conocimiento				Características Acústicas - SVM			
# Clases	Recuerdo		Precisión		Recuerdo		Precisión	
	No Ponderado	Ponderado	No Ponderado	Ponderado	No Ponderado	Ponderado	No Ponderado	Ponderado
5 Clases	**0.8623**	0.8796	0.8709	0.8819	0.8431	0.8704	0.8551	0.8624
7 Clases	0.7760	0.7681	0.7901	0.7940	**0.8413**	0.8590	0.8359	0.8491

Conclusiones:

- Es viable utilizar un conjunto de reglas basadas en conocimiento previo sobre el valor de las primitivas emocionales para mapear emociones básicas en un espacio tridimensional continuo.
- Esta primera versión de nuestro Método Particular de Clasificación de Emociones Básicas obtuvo resultados comparables con los de la clasificación basada en características acústicas.

7 Conclusiones

En este documento se propone un método de clasificación de emociones a partir de la voz basado en un modelo emocional continuo, que incluye el estudio de características apropiadas para esta tarea y un esquema para reconocer emociones automáticamente. Este método intenta eliminar las limitantes de los métodos basados en emociones básicas actuales. Dentro de los avances conseguidos hasta el momento se ha hecho un estudio con más de 350 características incluyendo información acústica y lingüística. Se ha planteado un método de reconocimiento de emociones en dos fases. La primera fase, de entrenamiento, incluye dos sub-métodos, el Método Particular de Clasificación de Emociones Básicas, que sirve para generar un Sistema de Inferencia el cual mapea emociones básicas en un espacio tridimensional continuo, y el Método General de Estimación de Primitivas, que sirve para estimar primitivas emocionales a partir de una señal de voz mediante modelos entrenados con SVM. La segunda fase es la de aplicación y está compuesta por los modelos entrenados por los sub-métodos de la fase de entrenamiento.

Se han realizado experimentos para validar la viabilidad de lo propuesto en esta tesis. Mediante estos experimentos nos hemos comparado contra los resultados del estado del arte clasificando emociones básicas y prediciendo primitivas emocionales obteniendo resultados alentadores. Con base en los avances logrados hasta el momento y en los resultados obtenidos en nuestros experimentos concluimos que se pueden lograr los objetivos planteados según la metodología y el plan de trabajo propuestos.

Referencias

Bänziger T., Tran V. and Scherer K. R. The Geneva Emotion Wheel: A tool for the verbal report of emotional reactions [Conference] // ISRE 2005, Conference of the International Society for Research on Emotions. - Bari, Italy : [s.n.], 2005.

Batliner A. [et al.] How to find trouble in communication [Journal] // Speech Commun. - 2003. - Vols. 40, 1-2. - pp. 117-143.

Beale R. and Peter C. The role of affect and emotion in hci [Book Section] // Affect and Emotion in Human-Computer Interaction: From Theory to Applications.: LNCS. - [s.l.] : Heidelberg: Springer-Verlag, 2008. - 1.

Bhuta T., Patrick L. and Garnett J. Perceptual evaluation of voice quality and its correlation with acoustic measurements [Journal] // Journal of Voice. - 2004. - Vol. 18 (3). - pp. 299-304.

Boersma P. Praat, a system for doing phonetics by computer [Conference] // Glot International 5:9/10. - pp. 341-345.

Bozkurt E. [et al.] Improving Automatic Emotion Recognition from Speech Signals [Conference] // Interspeech 2009. - Brighton : [s.n.], 2009.

Cowie R. [et al.] FEELTRACE: An instrument for recording perceived emotion in real time [Conference] // ISCA Tutorial and Research Workshop on Speech and Emotion. - Newcastle, Northern Ireland : [s.n.], 2000. - pp. 19-24.

Darwin C. The Expression of the Emotions in Man and Animals [Book] / ed. Murray John and Ekman Paul. - London : Oxford University Press, 1998. - 3.

Devillers L. and Vidrascu L. Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs [Conference] // Interspeech. - Pittsburgh : [s.n.], 2006.

Dubuisson T. [et al.] On the Use of the Correlation between Acoustic Descriptors for the Normal/Pathological Voices Discrimination [Journal] // EURASIP Journal on Advances in Signal Processing, Analysis and Signal Processing of Oesophageal and Pathological Voices. - 2009. - Vol. 10.1155/2009/173967.

Dumouchel P. [et al.] Cepstral and Long-Term Features for Emotion Recognition [Conference] // Interspeech 2009. - Brighton, UK. : [s.n.], 2009.

Ekman P. An argument for basic emotions [Journal] Cognition and Emotion. - 1992. - Vol. 6(3/4).

Ekman P. Universals and cultural differences in facial expressions of emotion [Conference] // Nebraska Symposium on Motivation / ed. Cole J.R.. - Lincoln : University of Nebraska Press, 1972.

Fell H. J. and MacAuslan J. Automatic Detection of Stress in Speech [Conference] // MAVEBA. - Florence, Italy : [s.n.], 2003.

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

González G.M. Bilingual computer-assisted psychological assessment: an innovative approach for screening depression in chicanos/latinos [Report]. - [s.l.] : University of Michigan, 1999.

Grimm M. [et al.] Primitives-based evaluation and estimation of emotions in speech [Journal] // Speech Communication. - 2007. - Vols. 49(10-11). - pp. 787-800 .

Grimm M. and Kroschel K. Evaluation of natural emotions using self assessment manikins [Conference] // IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). - San Juan, Puerto Rico : [s.n.], 2005. - pp. 381-385.

Herm O., Schmitt A. and Liscombe J. When calls go wrong: how to detect problematic calls based on log-files and emotions? [Conference] // INTERSPEECH-2008. - 2008. - pp. 463-466.

Hernández Y., Sucar L. E. and Conati C. An Affective Behaivor Model for Intelligent Tutors [Journal] // Intelligent Tutoring Systems (ITS) LNCS. - 2008. - Vol. 5091. - pp. 819-821.

Hillsdale NJ and Erlbaum Appendix F. Labels describing affective states in five major languages [Journal] // Facets of emotion: Recent research / ed. Scherer K. R.. - 1998. - pp. 241-243.

Iriondo I. Producción de un corpus oral y modelado prosódico para la síntesis del habla expresiva [Journal]. - Barcelona : [s.n.], 2008. - Vol. Tesis Doctoral.
Ishi C. T., Ishiguro H. and Hagita N. Proposal of acoustic measures for automatic detection of vocal fry [Conference] // Interspeech. - Lisbon, Portugal : [s.n.], 2005. - pp. 481-484.
James W. What is an emotion? [Journal] // Mind. - 1884. - Vol. 19. - pp. 188-205.
Jang R. ANFIS: Adaptive-Network-Based Fuzzy Inference Systems [Journal] // IEEE Transactions on Systems, Man, and Cybernetics. - May 1993. - 3 : Vol. 23. - pp. 665-685.
Kandali A., Routray A. and Basu T. Vocal emotion recognition in five native languages of Assam using new wavelet features [Journal] // International Journal of Speech Technology. - 2009.
Kockman M., Burget L. and Cernocký J. Brno Unicersity og Technology System fir Interspeech 2009 Emotion Challenge [Conference] // Interspeech. - Brighton, U.K. : [s.n.], 2009.
Kostoulas T., Ganchev T. and Fakotakis N. Study on Speaker-Independent Emotion Recognition from Speech on Real-World Data [Book Section] // Verbal and Nonverbal Features of Human-Human and Human-Machine Interaction: COST Action 2102 International Conference, Patras, Greece, October 29-31, 2007.. - Berlin, Heidelberg : Springer-Verlag, 2008.
Lazarus R. Passion and reason: Making Sense of our emotions [Book]. - New York : Oxford University Press, 1994.
Lee C. M. and Pieraccini R. Combining acoustic and language information for emotion recognition [Conference] // ICSLP. - Denver, CO, USA : [s.n.], 2002.
Lee C-C. [et al.] Emotion Recognition Using a Hierarchical Binary Decision Tree Approach [Conference] // Interspeech. - Brighton, U.K. : [s.n.], 2009.
Lichtenstein A. [et al.] Comparing Two Emotion Models for Deriving Affective States from Physiological Data [Journal] // Affect and Emotion in Human-Computer Interaction: From Theory to Applications.: LNCS / ed. Peter C. and Beale R.. - Heidelberg : Springer-Verlag., 2008. - pp. 35-50.
Liscombe J., Riccardi G. and Hakkani-Tür D. Using context to improve emotion detection in spoken dialog systems [Conference] // Eurospeech. - Lisboa : [s.n.], 2005.
Luengo Gil and Iker Reconocimiento automático de emociones utilizando parámetros prosódicos [Journal] // Procesamiento del lenguaje natural. - sept. 2005. - Vol. 35. - pp. 13-20 1135-5948.
Luengo I., Navas E. and Hernáez I. Combining spectral and prosodic information for emotion recognition [Conference] // Interspeech. - Brighton, U.K. : [s.n.], 2009.
Lugger M. and Yang B. Cascaded Emotion Classification via Psychological Emotion Dimensions Using Large Set of Voice Quality Parameters [Conference] // IEEE ICASSP. - Las Vegas, USA : [s.n.], 2008. - pp. 4945-4948.
Lugger M. and Yang B. Classification of Different Speaking Groups by Means of Voice Quality Parameters [Conference] // Vorträge der ITG-Fachtagung. - Kiel, Germany : [s.n.], 2006.
Narayanan S., Grimm M. and Kroschel K. The vera am mittag german audio-visual emotional speech database [Conference] // ICASSP. - Las Vegas, Nevada, U.S.A. : [s.n.], 2008.
Núñes B. F. [et al.] Evaluación perceptual de la disfonía: correlación con los parámetros acústicos y fiabilidad [Journal] // Acta otorrinolaringológica española: Organo oficial de la Sociedad española de otorrinolaringología y patología cérvico-facial. - 2004. - 6 : Vol. 55. - pp. 282-287.
Ortony A. and Turner T.J. What's basic about basic emotions? [Journal] // Psychological Review. - 1990. - pp. 315-331.
Ortony A., Clore G. L. and Collins A. The Cognitive Structure of Emotions [Book]. - Cambridge : Cambridge University Press, 1988.
Pitterman J. and Schmitt A. Integrating Linguistic Cues Into Speech-Based Emotion Recognition [Conference] // 4th IET International Conference on Intelligent Environments. - Seattle, USA : 2008.
Pittermann A. and Pittermann J. Getting Bored with HTK? Using HMMs for Emotion Recognition [Conference] // 8th International Conference on Signal Processing (ICSP). - Guilin, China : [s.n.], 2006.
Planet S. [et al.] GTM-URL Contribution to the Interspeech 2009 Emotion Challenge [Conference] // Interspeech. - Brighton, U.K. : [s.n.], 2009.
Polzehl T. [et al.] Emotion Classification in Children's Speech Using Fusion of Acoustic and Linguistic Features [Conference] // Interspeech 2009. - Brighton, U.K. : [s.n.], 2009.
Santiago K., Reyes G. C. A: and Gomez G. M.P. Conjuntos Difusos tipo 2 aplicados a la Comparación Difusa de Patrones para Clasificación de llanto de infantes con riesgo neurológico [Book]. - Tonantzintla, Puebla, México : [s.n.], 2009.

Sato Nobuo and Obuchi Yasunari Emotion Recognition using Mel-Frequency Cepstral Coefficients [Journal] // Information and Media Technologies. - 2007. - 3 : Vol. 2. - pp. 835-848.
Scherer K. R. Emotion [Book Section] // Introduction to Social Psychology: A European Perspective / book auth. Hewstone M. and Stroebe W.. - Blackwell, Oxford : [s.n.], 2001.
Scherer K. R. Psychological models of emotion [Journal] // The neuropsychology of emotion / ed. Borod J. C.. - Oxford, New York : Oxford University Press, 2000. - pp. 137-166.
Schuller B., Lang M. and Rigoll G. Robust Acoustic Speech Emotion Recognition by Ensembles of Classifiers [Conference] // Deutsche Jahrestagung für Akustik, DEGA, Invited Session "Automatische Spracherkennung in gestörter Umgebung". - Müchen, Germany : [s.n.], 2005. - pp. 329-330.
Schuller B., Steidl S. and Batliner A. The INTERSPEECH 2009 Emotion Challenge [Conference] // INTERSPEECH 2009. - Brighton, U.K. : [s.n.], 2009.
Seol Yong-Soo, Kim Dong-Joo and Kim Han-Woo Emotion Recognition from Text Using Knowledge-based ANN [Conference] // The 23rd International Technical Conference on Circuits/Systems, Computers and Communications. - Shimonoseki, Japan : [s.n.], 2008. - pp. 1569 – 1572.
Sobol-Shikler T. Analysis of affective expression in speech [Report] / Computer Laboratory ; University of Cambridge. - 2008.
Steidl S. Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech, [Book]. - Berlin : Logos Verlag, 2009.
Tóth Sz. L, Sztahó D. and Vicsi K. Emotion perception by human and machine [Conference] // COST2102 International Conference on Nonverbal Features of Human-Human and Human-Machine Interaction. - Patras, Greece : [s.n.], 2007. - pp. 223-236.
Vidrascu L. and Devillers L. Real-life emotion representation and detection in call centers data [Journal] // LNCS. - 2005. - Vol. 3784. - pp. 739-746.
Vlasenko B. [et al.] Frame vs. Turn-Level: Emotion Recognition from Speech Considering Static and Dynamic Processing [Conference] // Affective Computing and Intelligent Interaction (ACII2007). - Lisbon : [s.n.], 2007. - pp. 139-147.
Vogt T. and André E. Exploring the benefits of dizcretization of acoustic features for speech emotion recognition [Conference] // Interspeech. - Brighton, U.K : [s.n.], 2009.
Wempe T. and Jong N. Automatic measurement of speech rate in spoken dutch [Journal] // ACLC Working papers. - 2007. - Vol. 2.
Witten H. I. and Frank E. Data mining: Practical Machine learning tools and techniques [Book]. - San Francisco : [s.n.], 2005. - 2.
Wöllmer M. [et al.] Robust discriminative keyword spotting for emotionally colored spontaneous speech using bidirectional LSTM networks [Conference] // ICASSP. - Taipei, Taiwan : [s.n.], 2009.