Procesamiento del Lenguaje Natural, núm. 35 (2005), pp. 13-20 recibido 21-04-2005; aceptado 01-06-2005

Reconocimiento automático de emociones utilizando parámetros prosódicos

Iker Luengo
UPV/EHU
Alda. Urquijo s/n
ikerl@bips.bi.ehu.es

Eva Navas
UPV/EHU
Alda. Urquijo s/n
eva@bips.bi.ehu.es

Inmaculada Hernáez
UPV/EHU
Alda. Urquijo s/n
inma@bips.bi.ehu.es

Jon Sánchez
UPV/EHU
Alda. Urquijo s/n
ion@bips.bi.ehu.es

Resumen: Este artículo presenta los experimentos realizados para identificar automáticamente la emoción en una base de datos de habla emocional en euskara. Se han construido tres clasificadores diferentes: uno utilizando características espectrales y GMM, otro con parámetros prosódicos y SVM y el último con características prosódicas y SVM. Se extrajeron 86 características prosódicas y posteriormente se aplicó un algoritmo para seleccionar los parámetros más relevantes. El mejor resultado se obtuvo con el primero de los clasificadores, que alcanzó una precisión del 98.4% cuando se utilizan 512 componentes gaussianas. El clasificador construido con los 6 parámetros prosódicos más relevantes alcanza una precisión del 92.3% a pesar de su simplicidad, demostrando que la información prosódica es de gran importancia para identificar emociones.
Palabras clave: Habla emocional, reconocimiento de emociones

Abstract: This paper presents the experiments made to automatically identify emotion in an emotional speech database for Basque. Three different classifiers have been built: one using spectral features and GMM, other with prosodic features and SVM and the last one with prosodic features and GMM. 86 prosodic features were calculated and then an algorithm to select the most relevant ones was applied. The first classifier gives the best result with a 98.4% accuracy when using 512 mixtures, but the classifier built with the best 6 prosodic features achieves an accuracy of 92.3% in spite of its simplicity, showing that prosodic information is very useful to identify emotions.
Keywords: Emotional speech, emotion recognition

1 Introducción

Con el progreso de las nuevas tecnologías y la introducción de sistemas interactivos, se ha incrementado enormemente la demanda de interfaces amigables para comunicarse con las máquinas. Dado que la voz es el medio de comunicación más natural para los humanos, es necesario proporcionar a estas interfaces la capacidad de generar y reconocer el habla. Ya se han desarrollado diferentes sistemas de este tipo, desde avatares y modernos juguetes interactivos a sistemas automáticos de servicio al cliente, y existe una gran labor de investigación vinculada a este campo (Hozjan y Kačič, 2003)(Petrushin, 2000)(Seppänen, Väyrynen y Toivanen, 2003).

Actualmente, uno de los mayores objetivos en este tipo de interfaces es la naturalidad en la comunicación hombre máquina. Para ello se necesitan sistemas de síntesis de habla de gran calidad y gramáticas de reconocimiento más flexibles. Pero, dado que los humanos tendemos a expresar nuestro estado emocional a través de la voz, este tipo de interfaces naturales necesita también tener la capacidad de generar habla emocional y de reconocer el estado emocional del hablante.

El objetivo de este trabajo es realizar una primera aproximación hacia un sistema automático de reconocimiento de emociones en euskara. Más concretamente, el objetivo es

ISSN: 1135-5948

comparar el comportamiento de los reconocedores utilizando distintos conjuntos de parámetros (características espectrales clásicas y parámetros prosódicos) y diferentes sistemas clasificadores como los modelos de componentes gaussianas (*Gaussian Mixture Models*, GMM) o las máquinas de vectores soporte (*Support Vector Machines*, SVM).

El artículo está organizado de la siguiente manera: en la próxima sección se describe la base de datos que se ha utilizando en estos experimentos. Seguidamente se resume el proceso de extracción de las características prosódicas. Posteriormente, se describen los diferentes experimentos de reconocimiento que se han llevado a cabo, junto con los resultados obtenidos. Finalmente se presentan las conclusiones de este trabajo.

2 Descripción de la base de datos

En estos experimentos se ha utilizado una base de datos de habla emocional en euskara grabada por la Universidad del País Vasco (Navas et al., 2004). Esta base de datos incluye las seis emociones que han sido consideradas como básicas (Cowie y Cornelius, 2003)(Scherrer, 2003): alegría, asco, ira, miedo, sorpresa y tristeza y que han sido utilizadas tanto en trabajos de reconocimiento de emociones (Lay Nwe, Wei Foo y De Silva, 2003) como de generación de las mismas (Boula de Mareüil, Célérier y Toen, 2002).

La grabación la realizó una actriz profesional de doblaje a lo largo de dos días. Las grabaciones se llevaron a cabo en un estudio profesional, con una frecuencia de muestreo de 32 kHz y 16 bits por muestra. En las grabaciones se utilizó un laringógrafo, para obtener la señal de pulso glotal sincronizada con la señal de voz y poder calcular la curva de entonación con mucha precisión.

El corpus textual que se grabó en la base de datos está dividido en dos partes diferentes. En la primera de ellas los textos son los mismos para todas las emociones. Esta parte se equilibró fonéticamente y se grabó también en estilo neutro para poder utilizarlo como referencia. La segunda parte contiene textos relacionados semánticamente con la emoción. Estos textos son por lo tanto diferentes para cada una de las emociones consideradas. En esta parte no se incluyó el estilo neutro.

El hecho de tener un texto común a todas las emociones facilita la comparación de las características acústicas de cada una de ellas, ya que el contenido fonético es el mismo en todos los casos. Sin embargo, puede ser más difícil para la actriz expresar las emociones con suficiente naturalidad. Los textos con contenido relacionado con la emoción facilitan esta tarea, pero dificultan la comparación posterior de los rasgos que caracterizan cada emoción.

Para verificar si las emociones fueron expresadas correctamente, se sometió la base de datos a un proceso de evaluación subjetiva, en el que se evaluaron separadamente las dos partes del corpus: la de textos neutros o parte común y la de textos relacionados semánticamente con la emoción o parte específica. En este proceso participaron 15 personas, todas ellas nativos o con nivel fluido de euskara. El grado medio de correlación entre las respuestas de los evaluadores fue de 0.67. Los resultados de esta evaluación, que se recogen en la Figura 1, mostraron que todas las emociones son correctamente identificadas al menos en el 70% de los casos, excepto el asco, que ha sido una emoción difícil de identificar también en otras lenguas (Iida et al., 2003)(Burkhardt y Sendlmeier, 2000).

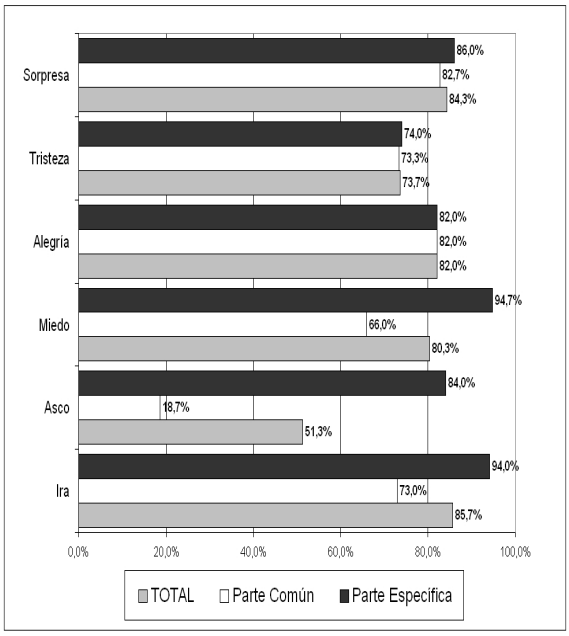

Figura 1: Resultado de la evaluación subjetiva de la base de datos

Uno de los objetivos de estos experimentos es comprobar si las características prosódicas pueden ser utilizadas para diferenciar el estilo neutro del habla emocionada y más aún para identificar la emoción. Por ello en este trabajo

se ha utilizado únicamente la primera parte de la base de datos que tiene textos comunes, ya que es la que contiene estilo neutro.

El corpus correspondiente a esta parte de la base de datos contiene números, palabras aisladas y frases de distinta duración. La Tabla 1 muestra un resumen de este contenido. En total se dispone de 97 grabaciones para cada emoción con una duración global de unos siete minutos por emoción.

Tipo de elemento	Cantidad
Números aislados	21
Palabras aisladas	21
Frases enunciativas cortas	10
Frases interrogativas cortas	5
Frases enunciativas medias	22
Frases interrogativas medias	8
Frases largas	10
Total de elementos por emoción	97

Tabla 1: Cantidad y tipo de los elementos de la base de datos por emoción

3 Extracción de los parámetros prosódicos

Es conocido que existe una relación entre la información prosódica y la expresión de emociones en el habla y que rasgos como la intensidad, la curva de frecuencia fundamental, la velocidad de locución son características importantes en la discriminación de emociones en la voz (Iriondo et al., 2000)(Lay Nwe, Wei Foo y De Silva, 2003) (Montero et al., 1999)(Mozziconacci, 2000).

Para este primer experimento sólo se han utilizado características relacionadas con el pitch y la energía. En la Figura 2 se muestra el proceso de extracción de estas características. Las curvas de entonación y energía se extraen de las grabaciones, tanto en escala lineal como logarítmica. Se calculan las curvas de primera y segunda derivada, ya que la velocidad de cambio del pitch y la energía puede proporcionar nueva información que sea útil para el reconocimiento. Finalmente se calculan distintos parámetros estadísticos a partir de todas estas curvas, utilizando la información de actividad vocal (VAD) y la de sonoridad (U/UV), ambas obtenidas asimismo a partir de las grabaciones.

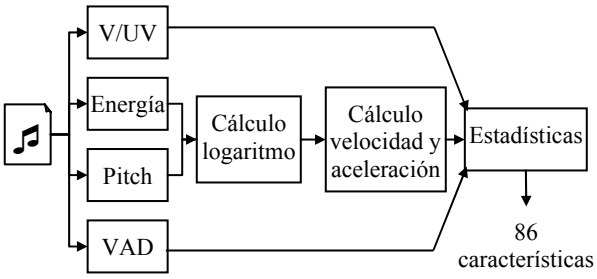

Figura 2: Diagrama de extracción de las características prosódicas

3.1 Obtención de la curva de entonación

Utilizando la señal del laringógrafo, se obtuvo una curva de entonación con una gran precisión para cada una de las frases. Los valores de pitch se calcularon como la inversa del tiempo que transcurre entre dos instantes consecutivos de cierre de la glotis. Las curvas de entonación se estimaron con una frecuencia de muestreo de 1 kHz, es decir, se obtuvo una muestra de pitch cada milisegundo.

La información de sonoridad (V/UV) se extrajo detectando los segmentos en los que existía información de cierre de la glotis y los que carecían de esta información.

3.2 Obtención de la curva de energía

Las grabaciones se analizaron con ventanas de 25 ms, cada 10 ms y se calculó el valor medio de la potencia de cada trama enventanada. Este proceso proporciona una muestra de potencia cada 10 ms. Todas las curvas se normalizaron al valor medio del estilo neutro.

3.3 Estimación de la actividad vocal

Para poder rechazar las tramas en las cuales no hay información vocal, es necesario realizar un detector de la actividad vocal (*Vocal Activity Detector*, VAD). De este modo, el nivel de ruido presente en las tramas de silencio no corromperá las características calculadas. Se implementó un VAD basado en el cálculo de la desviación espectral a largo plazo (*Long Term Spectral Deviation*, LTSD) entre las tramas vocales y las de ruido. El sistema implementado está basado en el presentado por Ramírez et al. (2004), en el que se utiliza un umbral de decisión adaptativo para conseguir la mayor eficiencia para cada nivel de ruido.

3.4 Estimación del *Jitter* y el *Shimmer*

El *Jitter* y el *Shimmer* están relacionados con las microvariaciones de pitch y de las curvas de potencia respectivamente. Por lo tanto pueden ser estimadas a partir de la velocidad de cambio de la pendiente de estas curvas. En este trabajo, el *Jitter* y el *Shimmer* se han estimado como el número de cruces por cero de las curvas derivadas. El resultado se normalizó al número de tramas utilizadas para el cálculo (tramas sonoras para el *Jitter* y tramas con actividad vocal para el *Shimmer*), para tener en cuenta la duración de la frase.

3.5 Cálculo de las características prosódicas

Como no existe un conocimiento *a priori* de las características que van a proporcionar un mejor resultado en el reconocimiento de emociones, se consideró más adecuado calcular un gran número de parámetros diferentes y descartar posteriormente aquéllos que resultaran redundantes.

Una vez se han obtenido las curvas de pitch y energía, así como sus primeras y segundas derivadas, se calcularon distintos parámetros estadísticos para cada una de ellas:

- Valor medio
- Varianza
- Valor máximo
- Valor mínimo
- Rango
- Sesgo
- Kurtosis

El cálculo de las características relacionadas con el pitch se realizó teniendo en cuenta únicamente las tramas en las que existía valor de pitch y descartando las tramas sordas. De manera similar, las características relacionadas con la energía se calcularon utilizando sólo las tramas para las cuales el VAD detectó actividad vocal.

Para cada frase se obtuvieron 12 curvas (pitch, energía, sus versiones logarítmicas y la primera y segunda derivada de cada una de ellas) y para cada curva se calcularon los siete parámetros estadísticos mencionados. De este modo se obtuvieron 84 características para cada frase. Añadiendo el *Jitter* y el *Shimmer* se obtienen los 86 parámetros prosódicos por frase finalmente utilizados en los experimentos.

4 Experimentos y resultados

Se han realizado tres experimentos de clasificación diferentes: el primero de ellos, con un sistema GMM tradicional y características espectrales; el segundo, con SVM y características prosódicas y el último con un sistema GMM y características prosódicas.

La precisión de cada sistema se calculó con una prueba *Jack-knife*. Primeramente, las frases de cada emoción se aleatorizaron y después se dividieron en cinco grupos. Esta aleatorización asegura que los bloques estén equilibrados, ya que la base de datos contiene diferentes tipos de elementos (desde palabras aisladas a frases de larga duración). Se entrenaron cinco sistemas diferentes y se realizaron cinco tests siguiendo un método de validación cruzada (*cross-validation*). Cuando se completaron los cinco tests, se calculó la matriz general de confusión y la tasa de precisión global. Esta precisión se estimó como la razón del número de frases correctamente clasificadas al número total de frases presentes en los tests.

4.1 GMM con características espectrales

Las grabaciones de la base de datos se convirtieron en coeficientes Mel-Cepstrum (MFCC), con primeras y segundas derivadas, utilizando tramas de 25 ms cada 10 ms, ventana Hamming y un factor de preénfasis de 0.97.

Para entrenar y probar los GMM se utilizó el software HTKv3 (Young et al., 2000). El entrenamiento de los modelos de una componente se realizó con tres ciclos de entrenamiento de Baum-Welch. Los modelos de varias componentes se entrenaron a partir del conjunto de dos componentes gaussianas, en un proceso iterativo en el cual el número de componentes gaussianas se incrementa en dos y se aplican dos iteraciones de reestimación de Baum-Welch hasta que se alcanza el número deseado de componentes gaussianas.

La Figura 3 muestra la precisión obtenida en este experimento para cada número de componentes gaussianas. Como era de esperar, la precisión del sistema se incrementa al aumentar el número de componentes gaussianas, hasta que se alcanza el nivel de saturación. La Tabla 2 muestra la matriz de confusión para el caso de 512 componentes gaussianas.

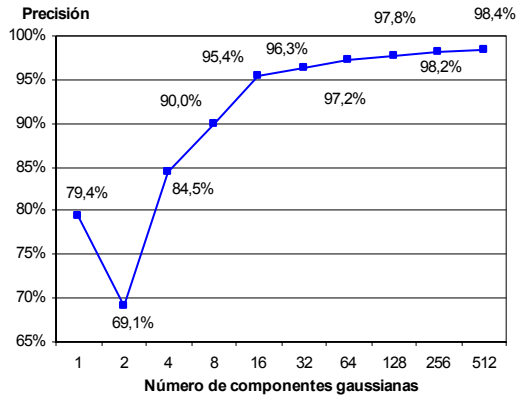

Figura 3: Precisión en el reconocimiento con parámetros MFCC y GMM, para un número diferente de componentes gaussianas

		ENTRADA						
		Ira	Mied	Sor	Asc	Aleg	Tris	Neu
SALIDA	Ira	97	-	-	-	-	-	1
	Mied	-	96	-	-	-	-	-
	Sor	-	1	97	-	-	-	-
	Asc	-	-	-	93	-	-	-
	Aleg	-	-	-	-	93	-	-
	Tris	-	-	-	1	-	97	1
	Neu	-	-	-	3	4	-	95
	Prec(%)	100	99.0	100	95.9	95.9	100	97.9

Tabla 2: Matriz de confusión con MFCC y GMM de 512 componentes gaussianas

4.2 SVM con características prosódicas

Para el entrenamiento y las pruebas de los SVM se utilizó la librería de funciones LibSVM (Chang y Lin, 2005). Para la clasificación multiclase se utilizó un kernel gaussiano y una aproximación uno contra uno.

En un primer experimento, se utilizaron las 86 características prosódicas calculadas y se alcanzó una precisión total del 93.50%. De todos modos es de esperar que muchas de estas características prosódicas sean redundantes, sobre todo si se considera que en su cálculo se han utilizado dos versiones diferentes de las mismas curvas, una en escala lineal y otra en escala logarítmica. Por ello se implementó un sistema de selección de características.

Para la selección de las características se utilizó un método *wrapper* adelante 3-atrás 1. Durante este proceso, en cada paso se selecciona el parámetro que maximiza la precisión del sistema. Esta precisión se obtiene entrenando un clasificador completamente nuevo con un *Jack-knife* test. Una vez se han realizado tres selecciones de parámetros consecutivas, se elimina el parámetro menos útil, es decir, aquél que tras ser eliminado reduce en menor medida la precisión del clasificador.

Los resultados de este experimento se muestran en la Figura 4. Aunque los resultados que se obtienen utilizando un menor número de parámetros son ligeramente peores que cuando se utilizan los 86, el coste computacional necesario para la extracción de los parámetros y el entrenamiento de un sistema de tal complejidad puede no merecer la pena. Utilizando únicamente 6 parámetros, se obtiene una precisión del 92.32%, sólo un 1.18% menos que usando los 86 parámetros.

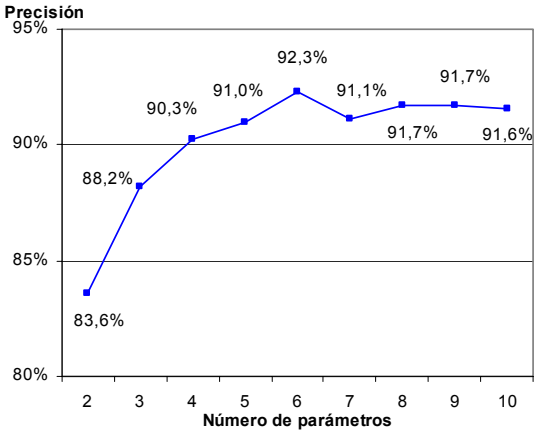

Figura 4: Precisión del reconocimiento con parámetros prosódicos y SVM, para distinto número de parámetros

La Tabla 3 presenta la matriz de confusión para el caso en el que se utilizan únicamente los seis mejores parámetros, que son:

- Valor medio del pitch en escala lineal
- Valor medio de la energía en escala lineal
- Varianza del pitch en escala lineal
- Sesgo del pitch en escala logarítmica
- Rango del pitch en escala logarítmica
- Rango de la energía en escala logarítmica

		ENTRADA						
		Ira	Mied	Sor	Asc	Aleg	Tris	Neu
SALIDA	Ira	92	-	-	1	2	-	-
	Mied	-	94	9	-	-	-	-
	Sor	-	3	88	-	-	-	-
	Asc	-	-	-	80	-	4	3
	Aleg	2	-	-	-	88	-	1
	Tris	2	-	-	10	-	93	1
	Neu	1	-	-	6	7	-	92
	Prec(%)	94.9	96.9	90.7	82.5	90.7	95.9	94.9

Tabla 3: Matriz de confusión con SVM y los 6 mejores parámetros prosódicos

		ENTRADA						
		Ira	Mied	Sor	Asc	Aleg	Tris	Neu
SALIDA	Ira	89	2	4	-	4	-	-
	Mied	-	90	8	-	-	-	-
	Sor	1	5	83	-	-	-	-
	Asc	2	-	-	73		14	1
	Aleg	4	-	-	-	82	-	8
	Tris	-	-	-	14	-	83	1
	Neu	1	-	-	10	11	-	87
	Prec(%)	91.8	92.8	87.4	75.3	84.5	85.6	89.7

Tabla 5: Matriz de confusión con GMM y los 6 mejores parámetros prosódicos

4.3 GMM con características prosódicas

En este experimento se utilizó también el software HTKv3 para el entrenamiento y test de los GMM. Debido al uso de un test *Jack-knife* y a que se obtuvo un único vector de características para cada frase, sólo se disponía de unos 80 vectores para el entrenamiento del modelo. Por ello, dada la falta de material de entrenamiento, se utilizaron modelos de una sola componente gaussiana. Estos modelos fueron creados con tres iteraciones de entrenamiento Baum-Welch.

Se realizaron dos experimentos, uno con los 86 parámetros calculados y otro sólo con los 6 parámetros que dieron mejor resultado en el clasificador SVM. Mientras el sistema con 86 parámetros obtuvo una precisión del 84.79%, el de 6 características mejoraba este resultado con una precisión del 86.71%. Este aumento de la precisión cuando se utilizan menos parámetros puede ser debido a que los modelos con muchos parámetros están demasiado adaptados a los datos de entrenamiento y no son capaces de generalizar correctamente. La Tabla 4 y la Tabla 5 muestran las matrices de confusión en estos dos casos.

		ENTRADA						
		Ira	Mied	Sor	Asc	Aleg	Tris	Neu
SALIDA	Ira	88	3	4	1	4	-	-
	Mied	1	89	13	-	-	1	-
	Sor	1	5	78	-	-	-	-
	Asc	3	-	-	76	1	7	2
	Aleg	4	-	-	3	68	-	8
	Tris	-	-	-	7	-	89	1
	Neu	-	-	-	10	24	-	86
	Prec(%)	90.7	91.8	82.1	78.4	70.1	91.8	88.7

Tabla 4: Matriz de confusión con GMM y los 86 parámetros prosódicos

5 Conclusiones

Los resultados de los trabajos de reconocimiento de emociones son difíciles de comparar, porque se utilizan bases de datos muy distintas. Algunos trabajos utilizan voz actuada, mientras otros recogen emociones verdaderas; unos utilizan bases de datos multilocutor y otros no; el conjunto de emociones básicas consideradas no es el mismo en todos los casos…

Posiblemente el trabajo de Hozjan y Kačič, (2003) es el más próximo al presentado en este artículo, ya que se utilizó voz actuada y un único locutor, con el mismo conjunto de emociones básicas. En su artículo Hozjan y Kačič describen el uso de la base de datos multilingüe INTERFACE (Tchong et al., 2000) para entrenar y probar un clasificador de emociones dependiente del locutor basado en 144 parámetros estadísticos calculados a partir de diferentes características prosódicas. Obtenían una precisión de entre el 60% y el 90% dependiendo del idioma considerado, utilizando redes neuronales. Se han realizado otros trabajos de reconocimiento de emociones entre los cuales cabe destacar los siguientes. Petrushin (2000) alcanza una precisión del 63.5% en entornos multilocutor considerando cuatro emociones y estilo neutro en una base de datos de habla actuada. Utiliza estadísticos de parámetros relacionados con la entonación, la energía y los formantes de los sonidos de cada frase. Seppänen, Väyrynen y Toivanen (2003) obtienen una precisión del 80.7% con sólo tres emociones y estilo neutro, calculando estadísticos de entonación y energía y utilizando la técnica kNN (*k Nearest Neighbor*). Finalmente, Nogueiras et al. (2001) alcanzan un resultado del 82.5% de precisión en un entorno multilocutor, considerando las mismas emociones que se han tenido en cuenta en

nuestro trabajo. Para ello utilizaron HMM y la evolución temporal de parámetros prosódicos.

En cuanto a la capacidad de los parámetros prosódicos de diferenciar las emociones, a partir de los resultados obtenidos en este trabajo es claro que los clasificadores tradicionales basados en GMM y características espectrales alcanzan mejores resultados que los construidos con parámetros prosódicos. De todos modos, es de remarcar que un clasificador sencillo basado en SVM con tan solo 6 parámetros prosódicos alcanza una precisión del 92.32%, únicamente un 6% menos que un sistema GMM-MFCC de 512 componentes gaussianas. Este incremento del error puede ser compensado con la reducción del tiempo de entrenamiento y prueba del sistema. Además, considerando que ambos sistemas utilizan parámetros de distinto tipo, es de esperar que la fusión de ambos proporcione mejores resultados, especialmente, si se trabaja en entornos más complejos, tipo multilocutor, multisesión u otros.

Otro aspecto destacable es la incapacidad de los evaluadores para identificar el asco, frente a los resultados obtenidos por los sistemas automáticos. El asco es una emoción poco frecuente, para la que los humanos estamos poco entrenados en comparación con otras emociones más habituales. En otras palabras, para realizar la prueba de evaluación los evaluadores no fueron entrenados previamente, como lo ha sido el sistema automático.

Dado que en este trabajo se ha utilizado voz actuada, las emociones pueden estar sobreactuadas, y los resultados de clasificación en condiciones reales pueden ser peores. Aun así, es un buen punto de partida para analizar cómo la información prosódica puede colaborar en las tareas de reconocimiento de emociones.

Estos resultados son muy esperanzadores. Como los parámetros prosódicos a largo plazo parecen estar muy poco correlados con las características espectrales a corto plazo (Campbell, Reynolds y Dunn, 2003), es de esperar que el uso combinado de los dos tipos de parámetros reduzca todavía más el error de clasificación.

Los trabajos futuros incluyen considerar otros parámetros, como la duración de los sonidos para estimar la velocidad de locución. También se realizará fusión de expertos, para que la información espectral que cambia con el tiempo y los parámetros prosódicos a largo plazo trabajen juntos para reducir el error del sistema.

6 Agradecimientos

Este trabajo ha sido parcialmente financiado por el Ministerio de Ciencia y Tecnología (TIC2003-08382-C05-03) y la Universidad del País Vasco (UPV-0147.345-E-14895/2002).

Bibliografía

Boula de Mareüil, P., Célérier, P. y J., Toen. 2002. Generation of Emotions by a Morphing Technique in English, French and Spanish. En *Proceedings of Speech Prosody*. páginas 187-190.

Burkhardt, F. y W. F. Sendlmeier. 2000. Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis. En *Proceedings of ISCA Workshop on Speech and Emotion,* páginas 151-156.

Campbell, J., Reynolds, D. y R. Dunn. 2003. Fusing High and Low Level Features for Speaker Recognition, *Proceedings of Eurospeech'03*, páginas 2665-2668.

Cowie, R. y R. Cornelius. 2003. Describing the Emotional States that are Expressed in Speech. *Speech Communication* 40(1,2): 2-32.

Chang, Ch. y Ch. Lin. 2005. *LIBSVM: a Library for Support Vector Machines*, software disponible en http://www.csie.ntu.edu.tw/~cjlin/libsvm

Hozjan, V. y Z. Kačič. 2003. Improved Emotion Recognition with Large Set of Statistical Features. En *Proceedings of Eurospeech'03*, páginas 133-136, Ginebra.

Iida, A., Campbell, N., Higuchi, F. y M. Yasumura. 2003. *A Corpus-based Speech Synthesis System with Emotion*. Speech Communication 40(1,2): 161-187.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J., Bernardos, D., Oliver, J., Tena, D. y L. Longhi. 2000. Validation of an Acoustical Modelling of Emotional Expression in Spanish using Speech Synthesis Techniques. En *Proceedings of ISCA Workshop on Speech and Emotion*, páginas 161-166.

Lay Nwe, T., Wei Foo, S. y L. De Silva. 2003. Speech Emotion Recognition Using Hidden Markov Models. *Speech Communication* 41(4): 603-623.

Montero, J.M., Gutierrez-Arriola, J., Palazuelos, S., Enríquez, E., Aguilera, S. y J.M. Pardo. 1999. Emocional Speech Síntesis: from Speech Database to TTS. En *Proceedings of ICPhS'99*, páginas 957-960.

Mozziconacci, S. 2000. The Expression of emotion Considered in the Framework of an Intonation Model. En *Proceedings of ISCA Workshop on Speech and Emotion*, páginas 45-52.

Navas, E., Hernáez, I., Castelruiz, A. y I. Luengo. 2000. Obtaining and Evaluating an Emotional Database for Prosody Modelling in Standard Basque. *Lecture Notes on Artificial Intelligence* 3206: 393-400.

Nogueiras, N., Moreno, A., Bonafonte, A. y J. Mariño. 2001. Speech Emotion Recognition Using Hidden Markov Models, En *Proceedings of Eurospeech'01*, páginas 2679-2682.

Petrushin, V.A. 2000. Emotion Recognition in Speech Signal: Experimental Study, Development and Application. En *Proceedings of ICSLP'00*, páginas 222-225, Denver.

Ramirez, J., Segura, J., Benitez, C., de la Torre, A. y A. Rubio. 2004. Efficient Voice Activity Detection Algorithms Using Long Term Speech Information. *Speech Communication* 42: 271-287.

Seppänen, T., Väyrynen, E y J. Toivanen. 2003. Prosody Based Classification of Emotions in Spoken Finnish. En *Proceedings of Eurospeech'03*, páginas 717-720, Ginebra.

Scherrer, K.R.. 2003. Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication* 40: 227-256.

Tchong, C., Toen, J., Kacic, Z., Moreno, A. y A. Nogueiras. 2000. Emotional speech synthesis database recordings. Informe Técnico. IST–1999–No 10036–D2, INTERFACE Project.

Young, S., Odell, J., Ollason, D., Valchev, V. y P. Woodlans. 2000. *The HTK book*, Cambridge University, Cambridge.