Mejora del Funcionamiento de Sistemas de Diálogo Hablado Mediante Reconocimiento del Estado Emocional de Usuarios

Enhancement of Spoken Dialogue Systems by Means of User Emotion Recognition

Ramón López-Cózar
Dpto. de LSI, CITIC-UGR
Universidad de Granada
rlopezc@ugr.es

Jan Silovsky
Institute of Information Technology,
Technical University of Liberec,
Czech Republic
jan.silovsky@tul.cz

David Griol
Dpto. de Informática Universidad
Carlos III de Madrid
dgriol@inf.uc3m.es

Resumen: Este artículo propone una nueva técnica para mejorar el funcionamiento de sistemas de diálogo hablado mediante el reconocimiento del estado emocional de los usuarios. La técnica se basa en el uso de dos módulos de *fusión* para combinar predicciones emocionales. El primer módulo emplea varios métodos de fusión para combinar predicciones generadas por clasificadores que procesan distintos tipos de información relacionada con cada frase pronunciada por el usuario. Estas predicciones constituyen la entrada del segundo módulo de fusión, el cual emplea un determinado método de fusión para combinar las predicciones generadas por el primer módulo, y obtener así la predicción de mayor probabilidad. Esta predicción representa la decisión final de nuestra técnica acerca del estado emocional del usuario. Hemos realizado experimentos considerando dos categorías emocionales ('No-Negativo' y 'Negativo') y clasificadores que procesan información prosódica, acústica, léxica y relacionada con actos del diálogo. Los resultados obtenidos usando un corpus emocional creado en nuestra Universidad muestran que el primer módulo de fusión mejora notablemente las tasas de reconocimiento de los clasificadores, así como el funcionamiento de un sistema de reconocimiento de referencia. El segundo módulo de fusión, que representa la novedad de nuestro trabajo, permite incrementar las tasas de reconocimiento del primer módulo en un porcentaje del 2,25% absoluto.
Palabras clave: Sistemas de diálogo hablado, reconocimiento de emociones, computación afectiva.

Abstract: In this paper we propose a new technique to enhance the performance of spoken dialogue systems by means of recognising users' emotional states. The technique employs two fusion modules that combine emotional predictions. The former employs a number of fusion methods to combine predictions made by classifiers that deal with different types of information regarding each sentence uttered by the user. These predictions are the input to the second fusion modules, which employs a fusion method to combine the predictions and obtain the most likely emotional category. This category represents the final decision of our technique regarding the emotional state of the user. We have carried out experiments considering two emotional categories ('Non-negative' and 'Negative') and classifiers to deal with information regarding prosody, acoustics, lexical items and dialogue acts. The results obtained employing an emotional corpus collected in our University show that the first fusion module clearly outperforms the classifiers, and so it does regarding a baseline system. The second fusion module, which represents the novelty of our study, enables enhancing the accuracy of the former fusion method by 2.25% absolutely.
Keywords: Spoken dialogue systems, emotion recognition, affective computing.

This research has been funded by the project HADA TIN2007-64718 of the Ministry of Education and Science of Spain, and the grant no. OC09066 (project COST 2102) of the Ministry of Education, Youth and Sports of the Czech Republic.

1 Introducción

La Computación Afectiva es un nuevo paradigma de la interacción persona-ordenador que tiene un gran interés en la actualidad (Piccard, 1997). Su objetivo es lograr que la comunicación con el ordenador sea lo más natural y amigable posible para el usuario. Para ello, incorpora mecanismos cuya finalidad es reconocer y/o generar emociones en el usuario.

El reconocimiento de emociones del usuario es una tarea que entraña una gran dificultad, debido a varias razones. Por una parte, no disponemos de un conocimiento ampliamente aceptado por la comunidad científica acerca de cómo los seres humanos procesamos las emociones. Por otra parte, no existe una diferencia clara entre algunos tipos de emociones, lo que provoca que en muchas ocasiones una misma emoción sea interpretada de forma distinta por personas diferentes.

La Computación Afectiva se ha aplicado a diversas tareas para intentar mejorar el servicio ofrecido por el ordenador. Por ejemplo, se ha empleado para diseñar el comportamiento de los agentes animados típicamente empleados en sistemas de diálogo multimodal (López-Cózar y Araki, 2005). También se ha utilizado en sistemas automáticos de tutorización, para lograr que estos sistemas se adapten mejor al estado emocional de los alumnos, y se incremente por tanto su ritmo de aprendizaje (Ai et al. 2006).

Asimismo, este nuevo paradigma de interacción se ha aplicado en sistemas de atención automática de llamadas de emergencia, con objeto de reconocer estados emocionales relacionados con estrés, dolor, temor o pánico (Devillers y Vidrascu, 2006). Otras aplicaciones incluyen interacción con robots (Bänziger y Scherer, 2005) y juegos de ordenador (Klein et al. 2002).

La Computación Afectiva también se ha aplicado en sistemas de diálogo hablado, área en que se centra nuestra investigación (Ang et al. 2002; Lee et al. 2002; Liscombe et al. 2005). El objetivo es detectar estados emocionales negativos de los usuarios para adaptar convenientemente el comportamiento de los sistemas. Cuando estos sistemas se usan en *call-centers* para proporcionar atención telefónica de forma automática, un objetivo es determinar el momento en que se debe transferir la llamada telefónica a un operador humano, en caso de que ésta no pueda ser atendida satisfactoriamente por el sistema. Por ejemplo, podría pensarse que la transferencia se debe realizar si el sistema se ve obligado a solicitar al usuario un mismo dato varias veces de forma consecutiva (p.e. su número de teléfono). Esta situación se suele dar en la práctica cuando el sistema no comprende el dato, o bien, lo reconoce con escaso valor de confianza.

Sin embargo, las expectativas y la paciencia de los usuarios de estos sistemas varían notablemente de unos usuarios a otros. Por ejemplo, hay usuarios que admiten proporcionar un mismo dato al sistema varias veces. En cambio, para otros usuarios, tener que repetir un mismo dato más de tres veces es algo totalmente inaceptable. Ello muestra que es necesario tener en cuenta otros factores que permitan decidir con mayor grado de acierto el momento exacto en que se debe realizar la transferencia de la llamada telefónica.

Consideramos que un factor determinante en el funcionamiento correcto de los sistemas es el cambio en el estado emocional del usuario, y en consecuencia, en este artículo proponemos una técnica para optimizar el reconocimiento de estados emocionales. Nuestro estudio se centra en el reconocimiento de dos categorías emocionales: 'No-Negativo' y 'Negativo'. Otros trabajos existentes en la literatura proponen distinguir entre un mayor número de categorías emocionales. Por ejemplo, Morrison et al. (2007) consideran seis categorías: ira, disgusto, temor, felicidad, tristeza y sorpresa.

No obstante, para la tarea que nos atañe (detección del momento más adecuado en que un sistema de diálogo debe transferir una llamada telefónica a un operador humano) hemos agrupado las diferentes emociones en las dos categorías indicadas. La categoría 'No-negativo' representa el estado emocional en que asumimos se encuentra el usuario cuando no experimenta problemas en la interacción con el sistema de diálogo. La categoría 'Negativo' representa el estado emocional en que se encuentra el usuario cuando comienza a experimentar problemas de comunicación con el sistema.

Consideramos que la detección del cambio entre ambos estados emocionales permite implementar sistemas de diálogo hablado que se adapten en mayor medida a los usuarios, optimizando en consecuencia el servicio ofrecido por los mismos.

El resto del artículo está organizado de la siguiente forma. En la sección 2 se presenta la técnica que proponemos para mejorar el reconocimiento del estado emocional de los usuarios de sistemas de diálogo hablado, basada en el uso de dos módulos de *fusión* para realizar la combinación de información. La sección 3 describe los clasificadores que hemos utilizado en los experimentos para procesar información prosódica, acústica, léxica y relacionada con actos del diálogo. La sección 4 presenta los experimentos llevados a cabo. En primer lugar discute los métodos de fusión de información empleados, seguidamente describe el corpus de frases usado, y en tercer lugar analiza los resultados obtenidos. Finalmente, la sección 5 presenta las conclusiones y comenta algunas líneas de trabajo futuro.

2 Técnica propuesta para mejorar el reconocimiento de emociones

La técnica que proponemos para mejorar el funcionamiento de los métodos actuales de reconocimiento de emociones de usuarios de sistemas de diálogo hablado se basa en el uso de un conjunto de clasificadores $\Omega = \{C_1, C_2, \ldots, C_m\}$, que reciben como entrada vectores de características obtenidos de cada frase pronunciada por el usuario.

Los clasificadores generan *predicciones* acerca del estado emocional del usuario, que tienen el formato siguiente: (h_i, p_i), $i = 1...S$, donde h_i es una categoría emocional (p.e., 'No-Negativo' y 'Negativo'), p_i es la probabilidad de que la frase haya sido pronunciada por un usuario que se encuentra en el estado emocional h_i, y S es el número total de categorías emocionales consideradas.

Las predicciones obtenidas constituyen la entrada de un módulo llamado Fusión-0, el cual emplea n métodos de fusión, F_{0i}, $i = 1\ldots n$, para generar otras predicciones. Estas nuevas predicciones son vectores que tiene el formato siguiente: $(h_{0j,k}, p_{0j,k})$, donde $j = 1\ldots n$, $k = 1\ldots S$, $h_{0j,k}$ es una categoría emocional y $p_{0j,k}$ es la probabilidad de que la frase haya sido pronunciada por un usuario que se encuentra en dicha categoría emocional.

El módulo Fusión-1 recibe las predicciones generadas por Fusión-0, y usando un determinado método de fusión (F_F) genera el vector (h_F, p_F), donde h_F es la categoría emocional que tiene mayor probabilidad (p_F). Esta categoría emocional representa el estado emocional del usuario hallado por la técnica propuesta. La mejor combinación de métodos de fusión a usar en Fusión-0 ($F_{01}, F_{02},\ldots,F_{0j}$, $1 \leq j \leq n$), así como el mejor método de fusión a usar en Fusión-1 (F_F) deben ser determinados experimentalmente.

3 Clasificadores

Describimos a continuación los cuatro clasificadores que hemos usado en los experimentos para procesar información prosódica, acústica, léxica y relacionada con actos del diálogo. Como se ha discutido en la sección anterior, la salida de cada clasificador constituye la entrada del módulo Fusión-0.

3.1 Clasificador prosódico

El clasificador prosódico analiza estadísticas globales de pitch y energía, así como características obtenidas de los segmentos de voz y no voz de las frases. Como resultado, crea un vector n-dimensional para cada frase de entrada, que constituye una entrada del módulo Fusión-0.

Tras realizar diversos experimentos para determinar el conjunto óptimo de características, hemos decidido usar un total de 11 características: media, mínimo y máximo de *pitch*; media de las derivadas del *pitch*; media y varianza de los valores absolutos del *pitch*; máximo de la energía; media de los valores absolutos de las derivadas de la energía; correlación entre las derivadas del *pitch* y la energía; longitud media de los segmentos de voz; y longitud del mayor segmento monótono.

El clasificador usa modelos mixtos Gausianos (Guassian Mixture Models, GMMs) dependientes del sexo de los usuarios para representar las categorías emocionales (Neiberg et al. 2006). Para calcular la probabilidad del vector de entrada n-dimensional (x) que representa a la frase de entrada, dada una categoría emocional λ, usamos la siguiente expresión:

$$P(x|\lambda) = \sum_{l=1}^{Q} w_l P_l(x) \qquad (1)$$

es decir, empleamos una combinación lineal con pesos w_l de Q densidades Gausianas unimodales $P_l(x)$. Definimos la función de densidad $P_l(x)$ de la siguiente forma:

$$P_l(x) = \frac{1}{\sqrt{(2\pi)^n \det \Sigma_l}} \exp\left(-\frac{1}{2}(x-\mu_l)' \Sigma_l^{-1}(x-\mu_l)\right) \quad (2)$$

donde los μ_l's son vectores de valores medios y los Σ_l's son matrices de covarianzas. Para determinar la categoría emocional que el clasificador proporciona como respuesta (h), usamos la siguiente expresión:

$$h = \arg\max_S P\left(x \middle| \lambda^S\right) \quad (3)$$

donde λ^S representa los modelos para las dos categorías emocionales consideradas ('No-negativo' y 'Negativo'), y la función *max* se calcula empleando el algoritmo EM (Expectation-Mazimization). Para calcular el valor de las probabilidades p_i de las predicciones usamos la siguiente expresión:

$$p_i = \beta_i / \sum_{k=1}^{2} \beta_k \quad (4)$$

donde β_i es la probabilidad de h_i, y los β_k son las probabilidades de las dos categorías emocionales consideradas.

3.2 Clasificador acústico

Para implementar el clasificador acústico hemos modelado los patrones de las frases de entrada mediante GMMs dependientes del sexo de los locutores, al igual que en el caso del clasificador prosódico. No obstante, la entrada para el clasificador acústico es una secuencia de vectores de características $x = \{x_1,\ldots,x_T\}$, en lugar de un vector n-dimensional. Asumiendo independencia estadísticas de los vectores en x, podemos calcular la probabilidad de una categoría emocional λ de la siguiente forma:

$$P(x|\lambda) = \sum_{t=1}^{T} \log P(x_t|\lambda) \quad (5)$$

Para determinar la categoría emocional que proporciona el clasificador como salida, empleamos Eq. (4), y para calcular la probabilidad de dicha categoría usamos Eq. (5).

3.3 Clasificador léxico

En experimentos previos con el sistema de diálogo (Autorreferencia) hemos observado que palabras similares desde un punto de vista acústico (p.e. *sesenta* y *setenta*), así como palabras que se ven claramente afectadas por el acento andaluz de los usuarios (p.e. *seis*), tienen mayor probabilidad de ser reconocidas incorrectamente, lo cual suele provocar el enfado de los usuarios del sistema.

Por consiguiente, el objetivo del clasificador léxico es detectar de forma automática este tipo de palabras potencialmente *problemáticas*, mediante un análisis de las salidas del reconocedor de habla del sistema. Para ello, usamos el enfoque propuesto por Lee y Narayanan (2005), que se basa en el concepto de *relevancia emocional*. La relevancia emocional de una palabra, dada una categoría emocional, se define como la información mutua entre la palabra y la categoría.

Si asumimos que W es una frase reconocida por el sistema de diálogo, compuesta por n palabras: $W = w_1\ w_2\ \ldots w_n$, y E es un conjunto de categorías emocionales: $E = \{e_1,e_2,\ldots,e_S\}$, entonces la relevancia emocional de una palabra w_i respecto a una categoría emocional e_j se puede definir como sigue:

$$salience(w_i, e_j) = P(e_j \mid w_i) \times mutual\_Information(w_i, e_j) \quad (6)$$

Para reconocer el estado emocional de la frase de entrada, consideramos que cada palabra de la misma es independiente de las demás, y calculamos un valor de activación a_k como sigue:

$$a_k = \sum_{m=1}^{n} I_m w_{mk} + w_k \quad (7)$$

donde $K = 1\ldots S$ y n es el número de palabras en W. I_m es un indicador que tiene el valor 1 si w_k es una palabra relevante para la categoría emocional (es decir, $salience(w_i, e_j) \neq 0$) y tiene el valor 0 en otro caso; w_{mk} es un peso de conexión entre la palabra y la categoría emocional; y w_k representa un factor de influencia (*bias*). Calculamos el peso de la conexión (w_{mk}) de la siguiente forma:

$$w_{mk} = mutual\_Information(w_m, e_k)$$

y el valor del factor de influencia (w_k) como sigue:

$$w_k = \log P(e_k)$$

Finalmente, la categoría emocional que proporciona el clasificador como salida (h) es aquélla que tiene un mayor valor de activación a_k:

$$h = \arg\max_k(a_k) \quad (8)$$

Para calcular el valor de las probabilidades p_i de cada predicción emocional del clasificador usamos la siguiente expresión:

$$p_i = a_i / \sum_{j=1}^{2} a_j \quad (9)$$

donde a_i representa el valor de activación de h_i, y las a_j's son los valores de activación para las dos categorías emocionales consideradas ('No-Negativo' y 'Negativo').

3.4 Clasificador de actos del diálogo

En varias teorías acerca del análisis de diálogos hablados persona-ordenador se considera que cada tipo de *prompt* generado por un sistema de diálogo puede asociarse a un tipo de acto del diálogo (*dialogue act*), por ejemplo: *saludar*, *despedirse*, *proporcionar información* y *confirmar*. Asimismo, es bien conocido que los usuarios de sistemas de diálogo suelen sentirse incómodos cuando los sistemas les solicitan una vez tras otra un mismo dado, por ejemplo, su número de teléfono. Por consiguiente, la misión del clasificador de actos del diálogo es reconocer el estado emocional 'Negativo' mediante la detección de repeticiones consecutivas de *prompts* del sistema, es decir, repeticiones de actos del diálogo. La categoría emocional que proporciona el clasificador como salida se calcula de la siguiente forma:

$$E_n = \arg\max_k P(E_k \mid AD_{n-(L*2-1)}, ..., AD_{n-1}) \quad (10)$$

De acuerdo con esta expresión, la categoría emocional correspondiente a un turno n del usuario en el diálogo (E_n) es aquélla que maximiza la probabilidad a posteriori dada una secuencia de actos del diálogo AD_i's (esto es, una secuencia de *prompts* del sistema de diálogo). En esta expresión, L representa la longitud de la secuencia. Obsérvese que si $L = 1$ la decisión acerca de E_n depende tan sólo del anterior *prompt* del sistema en el diálogo.

4 Experimentos

En los experimentos hemos evaluado el comportamiento de la técnica propuesta considerando:

i) Dos categorías emocionales ('No-Negativo' y 'Negativo');
ii) Cuatro clasificadores: C_1 = clasificador prosódico, C_2 = clasificador acústico, C_3 = clasificador léxico, y C_4 = clasificador de actos del diálogo;
iii) Tres métodos de fusión de información: media de probabilidades (MP), producto de probabilidades (PP) y votación (V) (Morrison et al. 2007).

Cuando los métodos de fusión se usan en el módulo Fusión-0 reciben como entradas las predicciones de los clasificadores. En cambio, cuando se usan en Fusión-1, reciben como entradas las predicciones de Fusión-0 usando varias combinaciones de dichos métodos. Los métodos MP y PP combinan predicciones calculando la media y el producto, respectivamente, de las probabilidades existentes en las predicciones que reciben como entradas. En cambio, el método de votación (V) combina las predicciones contando el número de clasificadores o métodos de fusión (según se use en Fusión-0 o Fusión-1) que consideran una determinada categoría emocional h_i como la más probable. Usando este método, la probabilidad p_i para dicha categoría emocional se calcula usando la siguiente expresión:

$$P(h_i \mid X, Y) = Vh_i / \sum_{j=1}^{2} Vh_j \quad (11)$$

donde X e Y representan las dos categorías emocionales consideradas, Vh_i es el número total de votos para h_i, y los Vh_j's representan el número de votos para cada categoría emocional.

4.1 Corpus de frases

El corpus de frases usado en los experimentos ha sido construido a partir de un corpus de 440 diálogos telefónicos entre alumnos de nuestra Universidad y el sistema de diálogo (Autorreferencia). Cada diálogo de este corpus fue almacenado en un fichero de traza que contiene cada *prompt* del sistema (p.e. *¿Te gustaría comer algo?*), el tipo del *prompt* (p.e. *Preguntar*C*omerAlgo*), el nombre del fichero

de muestras de voz que almacena la respuesta del usuario para dicho *prompt*, y la salida del reconocedor de habla tras analizar esta respuesta. El corpus tiene un total de 7.923 frases (ficheros de voz), de las cuales un 50,3% han sido grabadas por locutores masculinos y el resto por locutores femeninos.

Para entrenar y evaluar el funcionamiento de los clasificadores hemos dividido el corpus en dos particiones disjuntas: una para entrenamiento (5.983 frases correspondientes al 75% de los diálogos) y otra para test (1.985 frases correspondientes al 25% restante de diálogos). La división se ha realizado de tal forma que ambas particiones tengan frases representativas de los 18 tipos de frases existentes en el corpus: pedidos de productos de comida rápida, números de teléfono, códigos postales, direcciones de usuarios en nuestra ciudad, etc.

Las frases han sido etiquetadas por cuatro anotadores (2 hombres y 2 mujeres), quienes seleccionaban de forma aleatoria las frases para evitar ser influenciados por el contexto del diálogo. Los anotadores han asignado una etiqueta a cada frase ('Neutro', 'Cansado' o 'Enfadado') de acuerdo con el estado emocional del usuario percibido por el anotador.

Para realizar los experimentos presentados en este artículo, hemos asignado a las frases las etiquetas 'No-Negativo' o 'Negativo' de acuerdo con la opinión mayoritaria de los anotadores. Si ésta era 'Neutro' la etiqueta asignada ha sido 'No-Negativo', en otro caso la etiqueta ha sido 'Negativo'. Como resultado, observamos que en el corpus existe un 81% de frases anotadas con la etiqueta 'No-Negativo' y un 19% de frases anotadas con la etiqueta 'Negativo'. Ello muestra que el corpus está claramente desequilibrado en cuanto a porcentajes de categorías emocionales.

4.2 Entrenamiento de los clasificadores

Para realizar el entrenamiento de los clasificadores hemos analizado cada diálogo del corpus de test para localizar en él, desde el principio hasta el final, la siguiente información:

i) Cada tipo de *prompt* del sistema de diálogo;
ii) El fichero de muestras de voz que contiene la respuesta del usuario para dicho *prompt*;
iii) La salida del reconocedor de habla del sistema (frase en modo texto) tras analizar dicho fichero de muestras de voz.

Los tipos de *prompts* (p.e. *PreguntarComerAlgo*) han sido usados para crear secuencias de actos del diálogo de longitud L, $1 \leq L \leq 10$, con las cuales se ha probado el funcionamiento del clasificador de actos del diálogo a fin de determinar el valor óptimo de este parámetro. Los experimentos realizados muestran que el funcionamiento óptimo de este clasificador se obtiene con el valor $L = 4$.

Los ficheros de muestras de voz se han utilizado para realizar el entrenamiento de los clasificadores prosódico y acústico, mientras que las salidas del reconocedor de habla se han empleado para el entrenamiento del clasificador léxico.

4.3 Resultados experimentales

A efectos de poder comparar el funcionamiento de nuestra técnica con un sistema de referencia, hemos considerado un *baseline* que asigna a cada frase de entrada la etiqueta mayoritaria ('Neutro'). Dado que el 81% de las frases de nuestro corpus han sido anotadas con dicha etiqueta, la tasa de acierto de este sistema de referencia es 81%.

Para evaluar el comportamiento de nuestra técnica hemos usado el corpus de diálogos de test, proporcionando a los clasificadores la información existente en cada diálogo (tipos de *prompts*, ficheros de muestras de voz y salidas del reconocedor de habla) de forma análoga a como se ha descrito en la sección 4.2 para el entrenamiento. El funcionamiento de cada clasificador y de cada módulo de fusión se ha almacenado en un fichero de traza para poder analizar posteriormente su comportamiento.

La Tabla 1 muestra los resultados medios de funcionamiento de los clasificadores. Se puede observar que el funcionamiento del clasificador prosódico y el del acústico es un ligeramente inferior al del *baseline* (81%).

Clasificador	**%**
Prosódico	80,27
Acústico	79
Léxico	85,8
Actos del diálogo	91,81

Tabla 1. Resultados de los clasificadores.

En cambio, el clasificador léxico y el de actos de diálogo presentan un funcionamiento mejor. En base a estos resultados, podría pensarse que los dos primeros clasificadores no deberían ser utilizados, dado que funcionan peor que el *baseline*. No obstante, el clasificador prosódico reconoce correctamente el 86,95% de las frases anotadas con la etiqueta 'Negativo', y el clasificador acústico reconoce el 82,96% de dichas frases, mientras que el *baseline* reconoce el 0% de las mismas. Ello indica que los resultados mostrados en la tabla están claramente afectados por el desequilibrio del corpus de fases comentado en la sección 4.1.

La Tabla 2 muestra los resultados medios obtenidos por el módulo Fusión-0 al combinar las salidas de diversas combinaciones de clasificadores[2]. La combinación se ha realizado mediante los tres métodos de fusión considerados: media de probabilidades (MP), producto de probabilidades (PP) y votación (V).

Método fusión	**Clasificadores**	**%**
MP	Acu+Pro	84,15
	Lex+Pro	85,04
	AD+Pro	90,49
	Acu+Lex+Pro	89,20
	Acu+AD+Pro	90,24
	AD+Lex+Pro	90,02
	Acu+AD+Lex+Pro	**90,49**
	Media	88,66
PP	Acu+Pro	84,15
	Lex+Pro	85,16
	AD+Pro	91,49
	Acu+Lex+Pro	89,17
	Acu+AD+Pro	91,33
	AD+Lex+Pro	90,06
	Acu+AD+Lex+Pro	**92,23**
	Media	89,08
V	Acu+Pro	88,64
	Lex+Pro	86,40
	AD+Pro	88,20
	Acu+Lex+Pro	88,76
	Acu+AD+Pro	88,91
	AD+Lex+Pro	88,47
	Acu+AD+Lex+Pro	**89,04**
	Media	88,35

Tabla 2. Resultados usando Fusión-0.

[2] En la tabla, Pro, Acu, Lex y AD hacen referencia a los cuatro clasificadores utilizados: Prosódico, Acústico, Léxico y de Actos del Diálogo, respectivamente.

Como se puede observar, PP es el método de fusión que mejores resultados ha proporcionado, con un porcentaje medio de reconocimiento de 89,08%. La mejor tasa de reconocimiento (92,23%) se ha obtenido cuando Fusión-0 utiliza las predicciones de los cuatro clasificadores. Por consiguiente, usando dicha configuración, el módulo ha mejorado el funcionamiento del *baseline* en 11,23% absoluto (de 81% a 92,23%).

Analizando los ficheros de traza de este módulo observamos que el método MP permite reconocer el 95,75% de las frases anotadas con la etiqueta 'No-Negativo' y el 85,37% de las frases anotadas con la etiqueta 'Negativo'. En cambio, el método PP permite reconocer el 95,93% de las frases anotadas con la etiqueta 'No-Negativo' y el 88,91% de las frases anotadas con la etiqueta 'Negativo'.

La Tabla 3 muestra los resultados obtenidos al usar el módulo Fusión-1 para combinar las predicciones generadas por el módulo Fusión-0. Hemos usado en Fusión-1 los tres métodos de fusión empleados en Fusión-0 (MP, PP y V). Las entradas al módulo Fusión-1 han sido las predicciones generadas por varias combinaciones de estos métodos de fusión, generadas mediante Fusión-0: MP+PP, MP+V, PP+V y MP+PP+V. En todos los casos, el módulo Fusión-0 ha recibido las predicciones generadas por los cuatro clasificadores, dado que esta es la configuración que proporciona los mejores resultados de acuerdo con la Tabla 2.

Como se puede observar en la Tabla 3, el funcionamiento óptimo del módulo Fusión-1 se ha obtenido al usar PP para combinar las predicciones generadas por el módulo Fusión-0 usando MP y PP (94,48%).

Métodos de fusión en Fusión-0	**Métodos de fusión en Fusión-1**		
	MP	PP	V
MP+PP	93,68	**94,48**	93,53
MP+V	93,20	93,23	93,20
PP+V	93,44	94,38	93,20
MP+PP+V	93,23	94,l36	93,17
Media	93,40	94,11	93,28

Tabla 3. Resultados usando Fusión-1.

Esta mejora respecto al mejor comportamiento de Fusión-0 (92,23%) se obtiene porque Fusión-1 combina predicciones emocionales obtenidas empleando diversos métodos de

fusión (MP y PP), y por consiguiente, se beneficia de las respectivas ventajas de estos métodos. MP permite obtener clasificaciones menos sensibles a errores, y es especialmente apropiado cuando se dispone de espacios de características correlacionados, en los cuales los clasificadores cometen errores de forma independiente. En nuestro caso, dichos espacios se corresponden con información prosódica y acústica.

El método PP permite obtener buenos resultados cuando se usan espacios de características independientes, y es particularmente útil cuando los clasificadores cometen errores pequeños. En nuestro estudio, los espacios de características se corresponden con información prosódica, acústica, léxica y de actos del diálogo.

5 Conclusiones y trabajo futuro

Comparando los resultados mostrados en las Tablas 1 y 2 se observa que Fusión-0 permite mejorar el comportamiento de los clasificadores funcionando asiladamente. Fusión-0 también mejora claramente el funcionamiento del *baseline*, cuya tasa de reconocimiento es 81%.

Comparando los resultados mostrados en las Tablas 2 y 3 para el método PP, se observa que el uso de Fusión-1 permite incrementar las tasas de reconocimiento de Fusión-0 en 2,25% absoluto (desde 92,23% a 94,48%). El módulo Fusión-1 representa la novedad de nuestro trabajo, en comparación con otros trabajos existentes en la literatura.

En trabajos futuros tenemos previsto evaluar el comportamiento de nuestra técnica empleando fuentes de información no consideradas actualmente, por ejemplo, estilo de habla, tipo de tarea realizada por el sistema de diálogo e información extra-lingüística, por ejemplo, emisión de sonidos de disgusto al empezar a hablar.

También tenemos previsto incorporar pesos en el proceso de fusión de información. En los experimentos realizados hemos asumido que todos los clasificadores y métodos de fusión tienen la misma relevancia cuando se realiza la combinación de las predicciones emocionales. No obstante, puede ser interesante tener en cuenta el acierto de cada clasificador y método de fusión para cada estado del diálogo, y pesar en función de dicho acierto su contribución a la decisión final acerca del estado emocional del usuario.

Bibliografía

Ai, H., Litman, D. J., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A. 2006. Using system and user performance features to improve emotion detection in spoken tutoring systems. Actas de Interspeech, pp. 797-800.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., Stolcke, A. 2002. Prosody-based automatic detection of annoyance and frustration in human-computer dialog. Actas de ICSLP, pp. 2037-2039.

Bänziger, T., Scherer, K. R. 2005. The role of intonation in emotional expressions. Speech Communication, 46, pp. 252-267.

Devillers, L., Vidrascu, L. 2006. Real-life emotions detection with lexical and paralinguistic cues on human-human call center dialogs. Actas de Interspeech, pp. 801-804.

Klein, J., Moon, Y., Picard, R.W. 2002. This computer responds to user frustration: theory, design and results. Interacting with Computers, 14(2), pp. 119-140.

Lee, C. M., Narayanan, S. S., Pieraccini, R. 2002. Combining acoustic and language information for emotion recognition. Actas de ICSLP, pp. 873-876.

Lee, C. M., Narayanan, S. S. 2005. Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, vol. 13(2), pp. 293-303.

Liscombe, J., Riccardi, G., Hakkani-Tür, D. 2005. Using context to improve emotion detection in spoken dialogue systems. Actas de Interspeech, pp. 1845-1848.

López-Cózar, R., Araki, M. 2005. Spoken, Multilingual and Multimodal Dialogue Systems. Development and Assessment. John Wiley & Sons Publishers.

López-Cózar, R., Callejas, Z. 2005. Combining Language Models in the Input Interface of a Spoken Dialogue System. Computer Speech and Language, 20, pp. 420-440.

Morrison, D., Wang, R., De Silva, L. C. 2007. Ensemble methods for spoken emotion recognition in call-centres. Speech Communication, vol. 49(2) pp. 98-112.

Neiberg, D., Elenius, K., Laskowski, K. 2006. Emotion recognition in spontaneous speech using GMMs. Actas de Interspeech, pp. 809-812.

Piccard. R. 1997. Affective Computing. MIT Press.

Tax, D., Van Breukelen, M., Duin, R., Kittler, J. 2000. Combining multiple classifiers by averaging or multiplying. Pattern Recognition, vol. 33, pp. 1475-1485.