F^2 – New Technique for Recognition of User Emotional States in Spoken Dialogue Systems

Ramón López-Cózar
Dept. of Languages and Computer Systems, CTIC-UGR, University of Granada, Spain
rlopezc@ugr.es

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

David Griol
Dept. of Computer Science Carlos III University of Madrid, Spain
dgriol@inf.uc3m.es

Abstract

In this paper we propose a new technique to enhance emotion recognition by combining in different ways what we call *emotion predictions*. The technique is called F^2 as the combination is based on a double fusion process. The input to the first fusion phase is the output of a number of classifiers which deal with different types of information regarding each sentence uttered by the user. The output of this process is the input to the second fusion stage, which provides as output the most likely emotional category. Experiments have been carried out using a previously-developed spoken dialogue system designed for the fast food domain. Results obtained considering three and two emotional categories show that our technique outperforms the standard single fusion technique by 2.25% and 3.35% absolute, respectively.

1 Introduction

Automatic recognition of user emotional states is a very challenging task that has attracted the attention of the research community for several decades. The goal is to design methods to make computers interact more naturally with human beings. This is a very complex task due to a variety of reasons. One is the absence of a generally agreed definition of emotion and of qualitatively different types of emotion. Another is that we still have an incomplete understanding of how humans process emotions, as even people have difficulty in distinguishing between them. Thus, in many cases a given emotion is perceived differently by different people.

Studies in emotion recognition made by the research community have been applied to enhance the quality or efficiency of several services provided by computers. For example, these have been applied to spoken dialogue systems (SDSs) used in automated call-centres, where the goal is to detect problems in the interaction and, if appropriate, transfer the call automatically to a human operator.

The remainder of the paper is organised as follows. Section 2 addresses related work on the application of emotion recognition to SDSs. Section 3 focuses on the proposed technique, describing the classifiers and fusion methods employed in the current implementation. Section 4 discusses our speech database and its emotional annotation. Section 5 presents the experiments, comparing results obtained using the standard single fusion technique with the proposed double fusion. Finally, Section 6 presents the conclusions and outlines possibilities for future work.

2 Related work

Many studies can be found in the literature addressing potential improvements to SDSs by recognising user emotional states. A diversity of speech databases, features used for training and recognition, number of emotional categories, and recognition methods have been proposed. For example, Batliner et al. (2003) employed three different databases to detect troubles in communication. One was collected from a single experienced actor who was told to express anger because of system malfunctions. Other was collected from *naive* speakers who were asked to read neutral and emotional sentences. The third database was collected using a WOZ scenario designed to deliberately provoke user reactions to system malfunctions. The study focused on detecting two emotion categories: emotional (e.g. anger) and neutral, employing classifiers that dealt with prosodic, linguistic, and discourse information.

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 281–288,
The University of Tokyo, September 24-25, 2010.

Liscombe et al. (2005) made experiments with a corpus of 5,690 dialogues collected with the "How May I Help You" system, and considered seven emotional categories: positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative, and very other negative. They employed standard lexical, prosodic and contextual features.

Devillers and Vidrascu (2006) employed human-to-human dialogues on a financial task, and considered four emotional categories: anger, fear, relief and sadness. Emotion classification was carried out considering linguistic information and paralinguistic cues.

Ai et al. (2006) used a database collected from 100 dialogues between 20 students and a spoken dialogue tutor, and for classification employed lexical items, prosody, user gender, beginning and ending time of turns, user turns in the dialogue, and system/user performance features. Four emotional categories were considered: uncertain, certain, mixed and neutral.

Morrison et al. (2007) compared two emotional speech data sources The former was collected from a call-centre in which customers talked directly to a customer service representative. The second database was collected from 12 non-professional actors and actresses who simulated six emotional categories: anger, disgust, fear, happiness, sadness and surprise.

3 The proposed technique

The technique that we propose in this paper to enhance emotion recognition in SDSs considers that a set of classifiers $\Omega = \{C_1, C_2, \ldots, C_m\}$ receive as input feature vectors f related to each sentence uttered by the user. As a result, each classifier generates one emotion prediction, which is a vector of pairs (h_i, p_i), $i = 1 \ldots S$, where h_i is an emotional category (e.g. Angry), p_i is the probability of the utterance belonging to h_i in accordance with the classifier, and S is the number of emotional categories considered, which forms the set $E = \{e_1, e_2, \ldots, e_S\}$.

The emotion predictions generated by the classifiers make up the input to the first fusion stage, which we call Fusion-0. This stage employs n fusion methods called F_{0i}, $i = 1 \ldots n$, to generate other predictions: vectors of pairs $(h_{0j,k}, p_{0j,k})$, $j = 1 \ldots n$, $k = 1 \ldots S$, where $h_{0j,k}$ is an emotional category, and $p_{0j,k}$ is the probability of the utterance belonging to $h_{0j,k}$ in accordance with the fusion method F_{0j}.

The second fusion stage, called Fusion-1, receives the predictions provided by Fusion-0 and generates the pair $(h_{11,1}, p_{11,1})$, where $h_{11,1}$ is the emotional category with highest probability, $p_{11,1}$. This emotional category is determined employing a fusion method called F_{11}, and represents the user's emotional state deduced by the technique. The best combination of fusion methods to be used in Fusion-0 (F_{01}, $F_{02}, \ldots, F_{0j}$, $1 \leq j \leq n$) and the best fusion method to be used in Fusion-1 (F_{11}) must be experimentally determined.

3.1 Classifiers

In the current implementation our technique employs four classifiers, which deal with prosody, acoustics, lexical items and dialogue acts regarding each utterance.

3.1.1 Prosodic classifier

The input to our prosodic classifier is an n-dimensional feature vector obtained from global statistics of pitch and energy, and features derived from the duration of voiced/unvoiced segments in each utterance. After carrying out experiments to find the appropriate feature set for the classifier, we decided to use the following 11 features: pitch mean, minimum and maximum, pitch derivatives mean, mean and variance of absolute values of pitch derivatives, energy maximum, mean of absolute value of energy derivatives, correlation of pitch and energy derivatives, average length of voiced segments, and duration of longest monotonous segment.

The classifier employs gender-dependent Gaussian Mixture Models (GMMs) to represent emotional categories. The likelihood for the n-dimensional feature vector (x), given an emotional category λ, is defined as:

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

i.e., a weighted linear combination of Q unimodal Gaussian densities $P_l(x)$. The density function $P_l(x)$ is defined as:

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

where the μ_l's are mean vectors and the Σ_l's covariance matrices. The emotional category

deduced by the classifier, h, is decided according to the following expression:

$$h = \arg\max_{S} P\left(x\middle|\lambda^{S}\right) \quad (1)$$

where λ^S represents the models for the emotional categories considered, and the *max* function is computed employing the EM (Expectation-Maximization) algorithm. To compute the probabilities p_i for the emotion prediction of the classifier we use the following expression:

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

where β_i is the log-likelihood of h_i, S is the number of emotional categories considered, and the β_k's are the log-likelihoods of these emotional categories.

3.1.2 Acoustic classifier

Prosodic features are nowadays among the most popular features for emotion recognition (Dellaert et al. 1996; Luengo et al. 2005). However, several authors have evaluated other features. For example, Nwe et al. (2003) employed several short-term spectral features and observed that Logarithmic Frequency Power Coefficients (LFPCs) provide better performance than Mel-Frequency Cepstral Coefficient (MFCCs) or Linear Prediction Cepstral Coefficients (LPCCs). Experiments carried out with our speech database (which will be discussed in Section 4) have confirmed this observation. However, we have also noted that when we used the first and second derivatives, the best results were obtained for MFCCs. Hence, we decided to use 39-feature MFCCs (13 MFCCs, delta and delta-delta) for classification.

The emotion patterns of the input utterances are modelled by gender-dependent GMMs, as with the prosodic classifier, but each input utterance is represented employing a sequence of feature vectors $x = \{x_1,\ldots,x_T\}$ instead of one n-dimensional vector. We assume mutual independence of the feature vectors in x, and compute the log-likelihood for an emotional category λ as follows:

$$P\left(x\middle|\lambda\right) = \sum_{t=1}^{T} \log P\left(x_t\middle|\lambda\right)$$

The emotional category deduced by the classifier, h, is decided employing Eq. (1), whereas Eq. (2) is used to compute the probabilities for the prediction, i.e. for the vector of pairs (h_i, p_i).

3.1.3 Lexical classifier

A number of previous studies on emotion recognition take into account information about the kinds of word uttered by the users, assuming that there is a relationship between words and emotion categories. For example, swear words and insults can be considered as conveying a negative emotion (Lee and Narayanan, 2005). Analysis of our dialogue corpus (which will be discussed in Section 4) has shown that users did not utter swear words or insults during the interaction with the Saplen system. Nevertheless, there were particular moments in the interaction at which their emotional state changed from Neutral to Tired or Angry. These moments correspond to dialogue states where the system had problems in recognising the sentences uttered by the users.

The reasons for these problems are basically two. On the one hand, most users spoke with strong southern Spanish accents, characterised by the deletion of the final *s* of plural words, and an exchange of the phonemes *s* and *c* in many words. On the other hand, there are words in the system's vocabulary that are very similar acoustically.

Hence, our goal has been to automatically find these words by means of a study of the speech recognition results, and deduce the emotional category for each input utterance from the emotional information associated with the words in the recognition result. To do this we have followed the study of Lee and Narayanan (2005), which employs the information-theoretic concept of *emotional salience*. The emotional salience of a word for a given emotional category can be defined as the mutual information between the word and the emotional category. Let W be a sentence (speech recognition result) comprised of a sequence of n words: $W = w_1\ w_2\ \ldots w_n$, and E a set of emotional categories, $E = \{e_1, e_2, \ldots, e_S\}$. The mutual information between the word w_i and an emotional category e_j is defined as follows:

$$mutual\_Information(w_i, e_j) = \log \frac{P(e_j \mid w_i)}{P(e_j)}$$

where $P(e_j \mid w_i)$ is the posterior probability that a sentence containing the word w_i implies the emotional category e_j, and $P(e_j)$ represents the prior probability of the emotional category.

Taking into account the previous definitions, we have defined the emotional salience of the word w_i for an emotional category e_j as follows:

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

After the salient words for each emotional category have been identified employing a training corpus, we can carry out emotion recognition at the sentence level, considering that each word in the sentence is independent of the rest. The goal is to map the sentence W to any of the emotional categories in E. To do this, we compute an activation value a_k for each emotional category as follows:

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

where $k = 1 \ldots S$, n is the number of words in W, I_m represents an indicator that has the value 1 if w_k is a salient word for the emotional category (i.e. $salience(w_i,e_j) \neq 0$) and the value 0 otherwise; w_{mk} is the connection weight between the word and the emotional category, and w_k represents bias. We define the connection weight w_{mk} as:

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

whereas the bias is computed as: $w_k = \log P(e_k)$. Finally, the emotional category deduced by the classifier, h, is the one with highest activation value a_k:

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

To compute the probabilities p_i's for the emotion prediction, we use the following expression:

$$p_i = a_i / \sum_{j=1}^{S} a_j$$

where a_i represents the activation value of h_i, and the a_j's are the activation values of the S emotional categories considered.

3.1.4 Dialogue acts classifier

A dialogue act can be defined as the function performed by an utterance within the context of a dialogue, for example, greeting, closing, suggestion, rejection, repeat, rephrase, confirmation, specification, disambiguation, or help (Batliner et al. 2003; Lee and Narayanan, 2005; Liscombe et al. 2005).

Our dialogue acts classifier is inspired by the study of Liscombe et al. (2005), where the sequential structure of each dialogue is modelled by a sequence of dialogue acts. A difference is that they assigned one or more labels related to dialogue acts to each user utterance, and did not assign labels to system prompts, whereas we assign just one label to each system prompt and none to user utterances. This decision is made from the examination of our dialogue corpus. We have observed that users got tired or angry if the system generated the same prompt repeatedly (i.e. repeated the same dialogue act) to try to get a particular data item. For example, if it had difficulty in obtaining a telephone number then it employed several dialogue turns to obtain the number and confirm it, which annoyed the users, especially if they had employed other turns previously to correct misunderstandings. Hence, our dialogue act classifier aims to predict these negative emotional states by detecting successive repetitions of the same system's prompt types (e.g. prompts to get the telephone number).

In accordance with our approach, the emotional category of a user's dialogue turn, E_n, is that which maximises the posterior probability given a sequence of the most recent system prompts:

$$E_n = \arg\max_k P(E_k \mid DA_{n-(L*2-1)}, \ldots, DA_{n-3}, DA_{n-1})$$

where the prompt sequence is represented by a sequence of dialogue acts (DA_i's) and L is the length of the sequence, i.e. the number of system's dialogue turns in the sequence. Note that if $L = 1$ then the decision about E_n depends only on the previous system prompt. In other words, the emotional category obtained is that with the greatest probability given just the previous system turn in the dialogue. The probability of the considered emotional categories

given a sequence of dialogue acts is obtained by employing a training dialogue corpus.

By means of this equation, we decide the most likely emotional category for the input utterance, selecting the category with the highest probability given the sequence of dialogue acts of length L. This probability is used to create the pair (h_i, p_i) to be included in the emotion prediction.

3.2 Fusion methods

In the current implementation our technique employs the three fusion methods discussed in this section. When used in Fusion-0, these methods are employed to combine the predictions provided by the classifiers. When used in Fusion-1, they are used to combine the predictions generated by Fusion-0.

3.2.1 Average of probabilities (AP)

This method combines the predictions by averaging their probabilities. To do this we consider that each input utterance is represented by feature vectors $x^1,\ldots,x^m$ from feature spaces $X^1,\ldots,X^m$, where m is the number of classifiers. We also assume that each input utterance belongs to one of S emotional categories h_i, $i = 1\ldots S$. In each of the m feature spaces a classifier can be created that approximates the posterior probability $P(h_i \mid x^k)$ as follows:

$$f_i^k(x^k) = P(h_i \mid x^k) + \varepsilon_i^k(x^k)$$

where $\varepsilon_i^k(x^k)$ is the error made by classifier k. We estimate $P(h_i \mid x^k)$ by $f_i^k(x^k)$ and assuming a zero-mean error for $\varepsilon_i^k(x^k)$, we average all the $f_i^k(x^k)$'s to obtain a less error-sensitive estimation. In this way we obtain the following mean combination rule to decide the most likely emotional category:

$$P(h_i \mid x^1, \ldots, x^m) = \frac{1}{m}\sum_{k=1}^{m} f_i^k(x^k)$$

3.2.2 Multiplication of probabilities (MP)

Assuming that the feature spaces $X^1,\ldots,X^m$ are different and independent, the probabilities can be written as follows:

$$P(x^1, \ldots, x^m \mid h_i) = P(x^1 \mid h_i) \times P(x^2 \mid h_i) \times \ldots \times P(x^m \mid h_i)$$

Using Bayes rule we can obtain the following equation, which we use to decide the most likely emotional category for each input utterance (represented as feature vectors $x^1,\ldots,x^m$):

$$P(h_i \mid x^1, \ldots, x^m) = \frac{\prod_k P(h_i \mid x^k) / P(h_i)^{m-1}}{\sum_{i'}\left\{\prod_{k'} P(h_{i'} \mid x^{k'}) / P(h_{i'})^{m-1}\right\}}$$

3.2.3 Unweighted vote (UV)

This method combines the emotion predictions by counting the number of classifiers (if used in Fusion-0) or fusion methods (if used in Fusion-1) that consider an emotional category h_i as the most likely for the input utterance. If we consider three emotional categories X, Y and Z, h_i is decided as follows:

$$h_i = \begin{cases} X & \text{if } \sum_{j=1}^{m} X_j \geq \sum_{j=1}^{m} Y_j \text{ and } \sum_{j=1}^{m} X_j \geq \sum_{j=1}^{m} Z_j \\ Y & \text{if } \sum_{j=1}^{m} Y_j \geq \sum_{j=1}^{m} X_j \text{ and } \sum_{j=1}^{m} Y_j \geq \sum_{j=1}^{m} Z_j \\ Z & \text{if } \sum_{j=1}^{m} Z_j \geq \sum_{j=1}^{m} X_i \text{ and } \sum_{j=1}^{m} Z_j \geq \sum_{j=1}^{m} Y_j \end{cases}$$

where m is the number of classifiers or fusion methods employed (e.g., in our experiments, X = Neutral, Y = Tired and Z = Angry). The probability p_i for h_i to be included in the emotion prediction is computed as follows:

$$P(h_i \mid X, Y, Z) = Vh_i / \sum_{j=1}^{3} Vh_j$$

where Vh_i is the number of votes for h_i, and the Vh_j's are the number of votes for the 3 emotional categories. If we consider two emotional categories X and Y, the most likely emotional category h_i and its probability p_i are analogously computed (e.g., in our experiments, X = Non-negative and Y = Negative).

4 Emotional speech database

Our emotional speech database has been constructed from a corpus of 440 telephone-based dialogues between students of the University of Granada and the Saplen system, which was

previously developed in our lab for the fast food domain (López-Cózar et al. 1997; López-Cózar and Callejas, 2006). Each dialogue was stored in a log file in text format that includes each system prompt (e.g. *Would you like to drink anything?*), the type of prompt (e.g. AnyFoodOrDrinkToOrder?), the name of the voice samples file (utterance) that stores the user response to the prompt, and the speech recognition result for the utterance. The dialogue corpus contains 7,923 utterances, 50.3% of which were recorded by male users and the remaining by female users.

The utterances have been annotated by 4 labellers (2 male and 2 female). The order of the utterances has been randomly chosen to avoid influencing the labellers by the situation in the dialogues, thus minimising the effect of discourse context. The labellers have initially assigned one label to each utterance, either <NEUTRAL>, <TIRED> or <ANGRY> according to the perceived emotional state of the user. One of these labels has been finally assigned to each utterance according to the majority opinion of the labellers, so that 81% of the utterances are annotated as 'Neutral', 9.5% as 'Tired' and 9.4% as 'Angry'. This shows that the database is clearly unbalanced in terms of emotional categories.

To measure the amount of agreement between the labellers we employed the Kappa statistic (*K*), which is computed as follows (Cohen, 1960):

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

where *P(A)* is the proportion of times that the labellers agree, and *P(E)* is the proportion of times we would expect the labellers to agree by chance. We obtained that $K = 0.48$ and $K = 0.45$ for male and female labellers, respectively, which according to Landis and Koch (1977) represents *moderate agreement*.

5 Experiments

The main goal of the experiments has been to test the proposed technique using our emotional speech database, and employing:

- Three emotional categories (Neutral, Angry and Tired) on the one hand, and two emotional categories (Non-negative and Negative) on the other. The experiments employing the former category set will be called *3-emotion* experiments, whereas those employing the latter category will be called *2-emotion* experiments.
- The four classifiers described in Section 3.1, and the three fusion methods discussed in Section 3.2.

In the 3-emotion experiments we consider that an input utterance is correctly classified if the emotional category deduced by the technique matches the label assigned to the utterance. In the 2-emotion experiments, the utterance is considered to be correctly classified if either the deduced emotional category is Non-negative and the label is Neutral, or the category is Negative and the label is Tired or Angry.

To carry out training and testing we have used a script that takes as its input a set of labelled dialogues in a corpus, and processes each dialogue by locating within it, from the beginning to the end, each prompt of the Saplen system, the voice samples file that contains the user's response to the prompt, and the result provided by the system's speech recogniser (sentence in text format). The type of each prompt is used to create a sequence of dialogue acts of length *L*, which is the input to the dialogue acts classifier. The voice samples file is the input to the prosodic and acoustic classifiers, and the speech recognition result is the input to the lexical classifier. This procedure is repeated for all the dialogues in the corpus.

Experimental results have been obtained using 5-fold cross-validation, with each partition containing the utterances corresponding to 88 different dialogues in the corpus.

5.1 Performance of Fusion-0

Table 1 sets out the average results obtained for Fusion-0 considering several combinations of the classifiers and employing the three fusion methods. As can be observed, MP is the best fusion method, with average classification rates of 89.08% and 87.43% for the 2 and 3 emotion experiments, respectively. The best classification rates (92.23% and 90.67%) are obtained by employing the four classifiers, which means that the four types of information considered (acoustic, prosodic, lexical and related to dialogue acts) are really useful to enhance classification rates.

Fusion Method	Classifiers	2 emot.	3 emot.
AP	Aco, Pro	84.15	82.46
	Lex, Pro	85.04	82.71
	DA, Pro	90.49	87.48
	Aco, Lex, Pro	89.20	86.17
	Aco, DA, Pro	90.24	88.56
	DA, Lex, Pro	90.02	88.02
	Aco, DA, Lex, Pro	**90.49**	**88.32**
	Average	88.66	86.25
MP	Aco, Pro	84.15	82.86
	Lex, Pro	85.16	83.71
	DA, Pro	91.49	89.78
	Aco, Lex, Pro	89.17	87.91
	Aco, DA, Pro	91.33	89.23
	DA, Lex, Pro	90.06	87.82
	Aco, DA, Lex, Pro	**92.23**	**90.67**
	Average	89.08	87.43
UV	Aco, Pro	88.64	85.19
	Lex, Pro	86.40	83.01
	DA, Pro	88.20	84.92
	Aco, Lex, Pro	88.76	85.54
	Aco, DA, Pro	88.91	85.89
	DA, Lex, Pro	88.47	85.61
	Aco, DA, Lex, Pro	**89.04**	**87.56**
	Average	88.35	85.39

Table 1: Performance of Fusion-0 (results in %).

5.2 Performance of Fusion-1

Table 2 shows the average results obtained when Fusion-1 is used to combine the predictions of Fusion-0. The three fusion methods are tested in Fusion-1, with Fusion-0 employing four combinations of these methods: AP,MP; AP,UV; MP,UV; and AP,MP,UV. In all cases Fusion-0 uses the four classifiers as this is the configuration that provides the highest classification accuracy according to the previous section.

Comparison of both tables shows that Fusion-1 clearly outperforms Fusion-0. The best results are attained for MP, which means that this method is preferable when the data contain small errors (emotion predictions generated by Fusion-0 with accuracy rates around 90%).

To find the reasons for these enhancements we have analysed the confusion matrix of Fusion-1 using MP. The study reveals that for the 2-emotion experiments this fusion stage works very well in predicting the Non-negative category, very slightly enhancing the classification rate of Fusion-0 (96.58% vs. 95.93%), whereas the classification rate of the Negative category is the same as that obtained by Fusion-0 (88.91%). Overall, the best performance of Fusion-1 employing MP (94.48%) outdoes that of Fusion-0 employing AP (90.49%) and MP (92.23%).

Regarding the 3-emotion experiments, our analysis shows that using MP, Fusion-1 slightly lowers the classification rate of the Neutral category obtained by Fusion-0 (97.79% vs. 97.9%), but slightly raises the rate of the Tired category (93.62% vs. 93.26%), and the Angry category (77.49% vs. 76.81%). Overall, the performance of Fusion-1 employing MP (94.02%) outdoes that of Fusion-0 employing AP (88.32%) and MP (90.67%).

Fusion methods used in Fusion-0	Fusion method used in Fusion-1 (2 emotions)			Fusion method used in Fusion-1 (3 emotions)		
	AP	MP	UV	AP	MP	UV
AP,MP	93.68	**94.48**	93.53	91.77	**94.02**	90.96
AP,UV	93.20	93.23	93.20	91.65	93.13	90.10
MP,UV	93.34	94.38	93.20	91.27	93.98	89.48
AP,MP,UV	93.23	94.36	93.17	91.57	93.97	89.06
Average	93.40	94.11	93.28	91.57	93.78	89.90

Table 2: Performance of Fusion-1 (results in %).

6 Conclusions and future work

Our experimental results show that the proposed technique is useful to improve the classification rates of the standard fusion technique, which employs just one fusion stage. Comparing results in **Table 1** and **Table 2** we can observe that for the 2-emotion experiments, Fusion-1 enhances Fusion-0 by 2.25% absolute (from 92.23% to 94.48%), while for the 3-emotion experiments, the improvement is 3.35% absolute (from 90.67% to 94.02%). These improvements are obtained by employing AP and MP in Fusion-0 to combine the emotion predictions of the four classifiers, and using MP in Fusion-1 to combine the outputs of Fusion-0.

The reason for these improvements is that the double fusion process (Fusion-0 and Fusion-1) allows us to benefit from the advan-

tages of using different methods to combine information. According to our results, the best methods are AP and MP. The former allows gaining maximally from the independent data representation available, which are the input to Fusion-0 (in our study, prosody, acoustics, speech recognition errors, and dialogue acts). The latter provides better results when the data contain small errors, which occurs when the predictions provided by Fusion-0 are the input to Fusion-1.

Future work will include testing the technique employing information sources not considered in this study. The sources we have dealt with in the experiments (prosodic, acoustic, lexical, and dialogue acts) are those most commonly employed in previous studies. However, there are also studies that suggest using other information sources, such as speaking style, subject and problem identification, and non-verbal cues.

Another future work is to test the technique employing other methods for classification and information fusion. For example, it is known that people are usually confused when they try to determine the emotional state of a speaker, given that the difference between some emotions is not always clear. Hence, it would be interesting to investigate the performance of the technique employing classification algorithms that deal with this vague boundary, such as fuzzy inference methods, and using boosting methods for improving the accuracy of the classifiers.

Finally, in terms of application of the technique to improve the system-user interaction, we will evaluate different dialogue management strategies to enable the system's adaptation to negative emotional states of users (University students). For example, a dialogue management strategy could be as follows: i) if the emotional state is Tired begin the following prompt apologising, and transfer the call to a human operator if this state is recognised twice consecutively, and ii) if the emotional state is Angry apologise and transfer the call to a human operator immediately.

Acknowledgments

This research has been funded by Spanish project HADA TIN2007-64718, and the Czech Grant Agency project no. 102/08/0707.

References

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. Proc. of Interspeech, pp. 797-800.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E. 2003. How to find trouble in communication. *Speech Communication*, vol. 40, pp. 117-143.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educat. Psychol. Measurement, vol. 20, pp. 37-46.

Dellaert, F., Polzin, T., Waibel, A. 1996. Recognizing emotion in speech. Proc. of ICSLP, pp. 1970-1973.

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

Landis, J. R., Koch, G. G. 1977. The measurement of observer agreement for categorical data. *Biometrics*, vol. 33, pp. 159-174.

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. Proc. of Interspeech, pp. 1845-1848.

López-Cózar, R., García, P., Díaz, J., Rubio, A. J. 1997. A voice activated dialog system for fast-food restaurant applications. Proc. of Eurospeech, pp. 1783-1786.

López-Cózar, R., Callejas, Z. 2006. Combining Language Models in the Input Interface of a Spoken Dialogue System. *Computer Speech and Language*, 20, pp. 420-440.

Luengo, I., Navas, E., Hernáez, I, Sanchez, J. 2005. Automatic emotion recognition using prosodic parameters. Proc. of Interspeech, pp.493-496.

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.

Nwe, T. L., Foo, S. V., De Silva, L. C. 2003. Speech emotion recognition using hidden Markov models. *Speech Communication*, vol. 41(4), pp. 603-623.