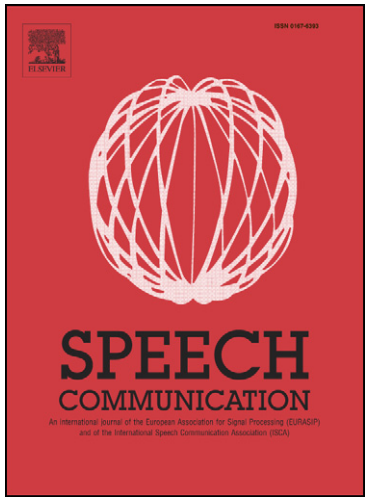
Enhancement of Emotion Detection in Spoken Dialogue Systems by Combining Several Information Sources

Ramón López-Cózar, Jan Silovsky, Martin Kroul



Enhancement of Emotion Detection in Spoken Dialogue Systems by Combining Several Information Sources

Ramón López-Cózar, Jan Silovsky, Martin Kroul

PII: S0167-6393(11)00007-0
DOI: 10.1016/j.specom.2011.01.006
Reference: SPECOM 1964

To appear in: *Speech Communication*

Received Date: 1 April 2010
Revised Date: 7 January 2011
Accepted Date: 11 January 2011

Please cite this article as: López-Cózar, R., Silovsky, J., Kroul, M., Enhancement of Emotion Detection in Spoken Dialogue Systems by Combining Several Information Sources, *Speech Communication* (2011), doi: 10.1016/j.specom.2011.01.006

Enhancement of Emotion Detection in Spoken Dialogue Systems by Combining Several Information Sources

Ramón López-Cózar[1], Jan Silovsky[2], Martin Kroul[2]
[1]Dept. of Languages and Computer Systems, Faculty of Computer Science,
University of Granada, Spain, E-mail: rlopezc@ugr.es
[2]Institute of Information Technology and Electronics, Faculty of Mechatronics, Technical
University of Liberec, Czech Republic,
E-mail: {jan.silovsky, martin.kroul}@tul.cz

Abstract

This paper proposes a technique to enhance emotion detection in spoken dialogue systems by means of two modules that combine different information sources. The first one, called Fusion-0, combines emotion predictions generated by a set of classifiers that deal with different kinds of information about each sentence uttered by the user. To do this, the module employs several methods for information *fusion* that produce other predictions about the emotional state of the user. The predictions are the input to the second information fusion module, called Fusion-1, where they are combined to deduce the emotional state of the user. Fusion-0 represents a method employed in previous studies to enhance classification rates, whereas Fusion-1 represents the novelty of the technique, which is the combination of emotion predictions generated by Fusion-0. One advantage of the technique is that it can be applied as a posterior processing stage to any other methods that combine information from different information sources at the decision level. This is so because the technique works on the predictions (outputs) of the methods, without interfering in the procedure used to obtain these predictions. Another advantage is that the technique can be implemented as a modular architecture, which facilitates the setting up within a spoken dialogue system as well as the deduction of the emotional state of the user in real time. Experiments have been carried out considering classifiers to deal with prosodic, acoustic, lexical and dialogue acts information, and three methods to combine information: multiplication of probabilities, average of probabilities and unweighted vote. The results show that the technique enhances the classification rates of the standard fusion by 2.27%

and 3.38% absolute in experiments carried out considering two and three emotion categories, respectively.

Keywords: Adaptive spoken dialogue systems, combination of classifiers, information fusion, emotion detection, human-computer interaction.

1 Introduction

Research on affective computing is a very challenging field that aims to design methods to make computers interact more naturally with human beings. Given the importance of emotions in human communication, many attempts have been made to include detection of human emotions in computers, which is a very difficult 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 (Ortony et al., 1990).

Automatic detection of user emotions has been applied to a variety of applications to enhance the quality or efficiency of the service provided by the computer. One of these is concerned with decision processes (Petrushin, 2000; Plutchik, 1994), for example, to select the behaviour of the animated agent typically employed by multimodal dialogue systems (López-Cózar and Araki, 2005). A second application is concerned with spoken tutoring systems which adapt to the emotions of students (Ai et al., 2006). The goal in this application is to increase the learning rate as previous studies suggest that there is a relationship between emotions and learning speed. The automatic detection of user emotions has also been applied to medical-emergency applications to detect stress, pain, fear or panic (Devillers and Vidrascu, 2006), to the interaction with robots (Kanda et al., 2004; Lee et al., 2009; Polzehl et al., 2009; Steidl, 2009), and to computer games, where the goal is to detect whether the user is enthusiastic or bored (Klein et al., 2002; Kuncheva et al., 2001; Scheirer et al., 2002).

Another kind of application, which is the focus of this paper, is call centres automated by means of spoken dialogue systems (Möller, 2004). Using these systems, nowadays users can

use spontaneous speech to access a number of telephone-based services. For example, they can make flight or train bookings and get information about bank accounts, traffic conditions or weather forecasts. The goal of emotion detection in this application is to detect problems in the interaction and transfer the call automatically to a human operator (Ang et al., 2002; Lee et al. 2002; Liscombe et al., 2005). For example, a typical communication problem occurs when the system tries to confirm time after time a word uttered by the user (e.g. a city name when making a flight booking) because the confirmation uttered by the user is misunderstood by the system. One key question is to decide at which point of the dialogue (i.e. after how many confirmation attempts) the call should be transferred. This decision can be based on the number of confirmation attempts made by the user. However, users differ from each other, and hence their demands, expectations and patience using a spoken dialogue system are unpredictable. For example, when dealing with a system misunderstanding, one user may be willing to repeat the same data three times, whilst for another user this may be unacceptable. The change in the emotional state of the user can be a better indicator of when the system is not fulfilling his expectations. Thus, dialogue systems that employ emotion detection techniques to decide whether to keep on interacting with the user or to transfer the call to a human operator at a particular moment can provide a better service.

The remainder of the paper is organised as follows. Section 2 presents a review of the literature on emotion detection in the field of spoken dialogue systems. Section 3 describes the proposed technique, discussing differences and similarities with previous studies. Section 4 focuses on the experiments. It first presents a study of the independence of the information sources employed. Then it describes the speech database and the classifiers. Next, it discusses the results obtained considering the standalone performance of the classifiers as well as those obtained employing the two fusion modules. The section ends addressing the statistical significance of our results and comparing them with previous studies available in the literature. Finally, section 5 presents the conclusions and discusses possibilities for future work.

2 Related work

There is a number of studies in the literature on automatic emotion detection that have been conducted to recognise emotions considering various types of features that can be extracted from a speech signal. Many studies have dealt with prosodic features typically derived from parameters like pitch frequency, loudness, energy contours or speaking rate (Dellaert et al., 1996; Lee et al., 2001; Ang et al., 2002; Luengo et al., 2005). Good performance was also demonstrated by systems based on modelling short-term acoustic features, e.g. Mel-Frequency Cepstral Coefficients (MFCC) or Logarithmic Frequency Power Coefficients (LFPC) (Nwe et al., 2003). Many authors have used a variety of methods based on pattern classification, for example, maximum likelihood Bayes classification, kernel regression (Dellaert et al., 1996), linear discriminant classification (Lee et al., 2002), k-nearest neighbourhood (Lee et al., 2001; Morrison et al., 2007), Bayesian networks (Barra-Chicote et al., 2009), neural networks (Huber et al., 2000; Batliner et al., 2003; Xu et al., 2009), support vector machines (Devillers and Vidrascu, 2006) and decision trees (Ai et al., 2006; Lee et al., 2009).

Other information sources have been considered in addition to acoustic/prosodic features. For example, Ang et al. (2002), Lee et al. (2002) and more recently Polzehl et al. (2009) in the INTERSPEECH 2009 Emotion Challenge (Schuller et al., 2009), have employed linguistic information. Other authors have taken into account the contextual nature that the dialogue structure provides, such as dialogue acts (e.g. *repeat*, *repair*, *rejection*, *ask start over*) and discourse markers (e.g. repetition of the same subdialogue). Following this direction, Hastie et al. (2002) employed system's dialogue acts, Batliner et al. (2003) and Lee and Narayanan (2005) used discourse features, Ai et al. (2006) employed sequential features, and Liscombe et al. (2005) considered contextual information.

Some studies make use of hybrid or ensemble methods. For example, Dellaert et al. (1996) employed an ensemble of several classifiers; Petrushin (2000) used ensembles of neural networks; and Nakatsu et al. (1999) proposed an ensemble of eight neural networks, each

trained for a particular kind of emotion. More recently, Morrison et al. (2007) proposed using two existing ensemble classification methods; on the one hand a stacked generalisation, and on the other a variation of majority voting. Other methods have been proposed but these have been less broadly adopted. For example, Nwe et al. (2003) studied emotion detection employing a discrete hidden Markov model as a classifier.

Another issue addressed differently by the research community is which emotions to detect. Given the difficulty of this task, in order to improve annotation and detection accuracy, many studies consider a discrete number of emotion categories. For example, Huber et al. (2000) and Yacoub et al. (2003) considered 'anger' and 'neutral'; Lee et al. (2001), Lee and Narayanan (2003) and Litman and Forbes-Riley (2006) distinguished between 'non-negative' and 'negative'; Ang et al. (2002) took into account 'neutral', 'annoyed' and frustrated'; and Batliner et al. (2003) considered 'emotional' and 'neutral'. Other researchers have addressed a greater number of categories. For example, Ai et al. (2006) distinguished among 'uncertain', 'certain', 'mixed' and 'neutral'; Devillers and Vidrascu (2006) considered 'anger', 'fear', 'relief' and 'sadness', and Liscombe et al. (2005) carried out experiments considering seven emotion categories: 'positive/neutral', 'somewhat frustrated', 'very frustrated', 'somewhat angry', 'very angry', 'somewhat other negative', and 'very other negative'.

Many researchers have employed data collected from real interactions between users and systems. For example, Lee et al. (2001) used a corpus of human-to-machine dialogues and carried out gender-specific experiments, taking into account that pitch-related features are very different between female and male, especially the mean, maximum, and minimum of F0. Lee et al. (2002) employed a speech database collected from real users engaged in a spoken dialogue with a call centre application. Ang et al. (2002) employed a database collected from real users to detect frustration and annoyance in natural human-computer dialogues. Lee and Narayanan (2005) employed a database obtained from real users engaged in a spoken dialogue with a machine agent over the telephone. Liscombe et al. (2005) employed a dialogue corpus collected

with the "How May I Help You" system, and Ai et al. (2006) used a database collected from dialogues between students and a spoken dialogue tutor.

Other authors have used data obtained from Wizard of Oz (WOZ) scenarios in order to simulate human-to-machine communication in which the users become angry and/or frustrated with the system (Huber et al., 2000; Klein et al., 2002). A number of authors have used data collected from actors. For example, Yacoub et al. (2003) used a speech database recorded by 8 actors expressing 15 emotions, whereas Morrison et al. (2007) employed a database that contains 720 utterances collected from 12 non-professional actors and actresses who simulated six emotion categories: 'anger', 'disgust', 'fear', 'happiness', 'sadness', and 'surprise'.

Given the variety of information sources available, as discussed above, another issue addressed differently by the research community is how to combine information to optimise emotion detection. One approach is to make the combination at the feature level, employing vectors comprised of features extracted from the different information sources. For example, this method was followed by Lee et al. (2001), Yacoub et al. (2003), Devillers and Vidrascu (2006), Morrison et al. (2007), Lugger and Yang (2009) and Luengo et al. (2009), the latter in the INTERSPEECH 2009 Emotion Challenge. The drawback of this method is the growth of dimensionality as the number of features increases. Another approach combines the information at the decision level, in which case a classifier is employed to decide the emotion of each source, and a final decision is made by combining the hypotheses made by different classifiers. For example, this approach has been followed by Lee et al. (2002), Batliner et al. (2003), Lee and Narayanan (2005) and Bozkurt et al. (2009), the latter in the INTERSPEECH 2009 Emotion Challenge.

3 The proposed technique

The technique proposed in this paper to enhance emotion detection in spoken dialogue systems is inspired by previous methods that function at the decision level. The technique considers that a set of classifiers $\Omega = \{C_1, C_2, \ldots, C_m\}$ receive as input classifier-specific feature vectors v_k, $k =$

1...*m*, related to each sentence uttered by the user. Employing a classification algorithm, each classifier generates one emotion prediction. This prediction is a vector of pairs (h_i, p_i), $i = 1 \ldots S$, where h_i is an emotion category (e.g. 'Neutral', 'Tired' or 'Angry'), p_i is the probability of the utterance belonging to h_i according to the classification algorithm, and S is the number of emotion categories considered. The technique employs two fusion processes. The first of these takes as its input the predictions made by m classifiers, whereas the second uses the predictions made by n fusion methods.

The motivation for the technique is that we have observed in previous experiments that some fusion methods do not provide the correct emotion category for a given input utterance, while others do. Therefore, we considered that just as the combination of the predictions of classifiers is useful to increase the classification rate, the combination of the predictions of fusion methods could be useful as well. One advantage of the technique is that it can be applied as a posterior processing stage to any other methods that combine information from different information sources (e.g. related to acoustics, lexical items or discourse) at the decision level. This is so because the technique works on the outputs (predictions) of the methods, without interfering in the procedure used to obtain these predictions.

Fig. 1 shows the particular application of the technique that we have used in our experiments. We have considered four classifiers ($m = 4$) that will be described in detail in Section 4.3, and three fusion methods ($n = 3$). As can be observed, we have used a script that takes as its input a set of labelled dialogues[1] in a test corpus, and processes each dialogue by locating within it, from the beginning to the end, each prompt of the Saplen system (López-Cózar and Callejas, 2005), 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 goal of this system, which is running in our lab for experimental purposes, is to answer telephone calls related to the fast food domain.

[1] Fig. 2 shows a sample labelled dialogue.

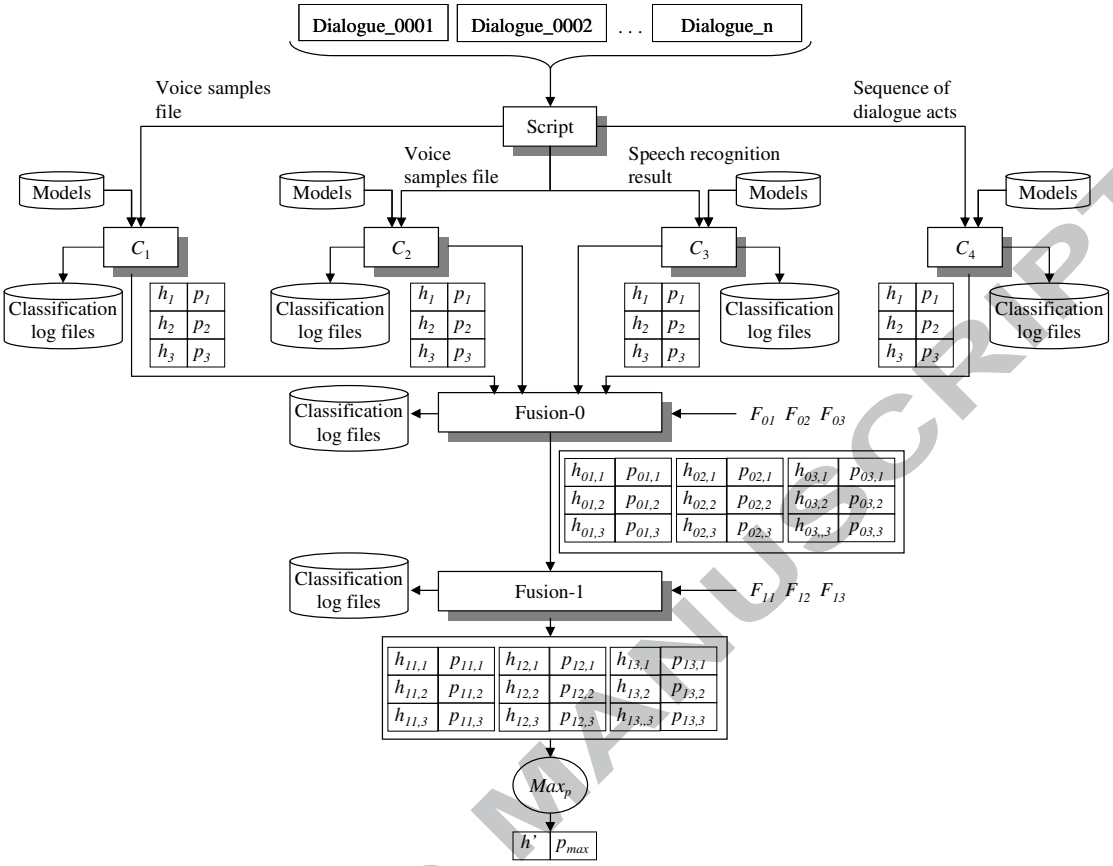


Fig. 1. Usage of the proposed technique (C_1 = Prosodic classifier, C_2 = Acoustic classifier, C_3 = Lexical classifier, C_4 = Dialogue acts classifier).

The voice samples file is the input to the prosodic and acoustic classifiers. In our implementation, both classifiers contain a module that extracts the required features from the samples in order to carry out classification. The speech recognition result is the input to the lexical classifier, and 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. This procedure is repeated for all the dialogues in the test corpus. By using the architecture shown in the figure, the technique is tested just as if it had been used during the system's interaction with the users. Moreover, it makes it possible to deduce the emotional state of the user in real time, given that once the ASR result is available, the overall delay introduced in the response time of the dialogue system is smaller than 1 sec in a PC Core 2 Duo at 2.66 GHz. This delay includes the process for feature extraction needed by the acoustic and prosodic classifier, the analysis of the features made by

these classifiers, the analysis of the recognition result carried out by the lexical classifier, and the analysis of the history of previous system prompts made by the dialogue act classifier.

As can be observed in the figure, each classifier analyses its input and provides one prediction, i.e., a vector that contains three (h_i, p_i) pairs. These pairs indicate how likely it is that the input utterance belongs to the emotion categories 'Neutral', 'Tired' and 'Angry', respectively. The predictions of the classifiers are the input to the Fusion-0 module, which combines them employing three fusion methods *F*: Average of probabilities (AP), Multiplication of probabilities (MP) and Unweighted Vote (UV) (Kittler et al., 1998; Roli et al., 2004; Le et al., 2005; Morrison et al., 2007). The output of Fusion-0 is three predictions, one per fusion method. Each prediction contains three pairs ($h_{0j,k}$, $p_{0j,k}$) which indicate how likely it is that according to each fusion method, the input utterance belongs to the emotion categories 'Neutral', 'Tired' and 'Angry'. These predictions are the input to the Fusion-1 module, where they are combined again using the same fusion methods *F*. The output of this module is three predictions, each containing three pairs ($h_{1j,k}$, $p_{1j,k}$). Finally, the deduced emotion category, *h'*, is determined by inspecting these predictions and selecting the most likely emotion category.

3.1 Comparison between the proposed technique and previous studies on emotion detection for spoken dialogue systems

3.1.1 Comparison considering the integration within the architecture of spoken dialogue systems

A number of previous studies on emotion detection for spoken dialogue systems focus on techniques for enhancing detection rates but do not address explicitly the integration of the techniques into the architecture of these systems. Moreover, they do not address how the deduced emotion can be employed to affect the system's interaction with the user (Yacoub et al., 2003; Morrison et al., 2007). Other studies discuss modules to integrate the proposed techniques into spoken dialogue systems. For example, Batliner et al. (2003) proposed a module called MOUSE (Monitoring Of User State Emotion) which is placed between the language

understanding and the dialogue management modules of a spoken dialogue system. The input it receives is the speech signal together with the semantic representation obtained by the understanding component. If MOUSE recognises an utterance as 'normal' then it signals 'no trouble' and further normal dialogue processing is initiated. However, if it recognises the utterance as 'indicating trouble' a specific action is initiated, for example, hand over to a human operator or make the system apologise.

Similarly as Batliner et al. (2003), the technique that we propose was conceived to be implemented as an additional module of a spoken dialogue system. The input to this module can be extracted in real time from the user interaction and the dialogue history, whereas its output (the deduced emotional state of the user) can be used to influence the system's performance. For example, a dialogue management strategy using the deduced emotion category could be as follows: i) if the deduced emotional category is 'Neutral' do not initiate any particular action to address the state; ii) if the category is 'Tired' begin the following prompt apologising, and transfer the call to a human operator if this state is detected twice consecutively; iii) if the category is 'Angry' apologise and transfer the call to the human operator immediately.

3.1.2 Comparison considering the fusion of information sources

As discussed in Section 2, many previous studies carried out emotion detection combining information sources at the feature level. Thus, they employ vectors comprised of features extracted from a number of information sources, which are the input to classification algorithms. For example, Liscombe et al. (2005) employed vectors of 80 features related to lexical, prosodic, dialogue acts, and contextual information; Huber et al. (2000) used word-related vectors comprised of 121 features, and sentence-level vectors comprised of 27 features; and Batliner et al. (2003) employed vectors comprised of 91 prosodic features and 30 parts-of-speech (POS) features.

Other studies propose classification techniques that make the fusion at the decision level. For example, Lee et al. (2002) combined acoustic and language information employing an

"OR" logical combiner according to which, the final decision was 'Emotional' if either the acoustic or the language input was considered as 'Emotional' ('Neutral' otherwise). Also working at the decision level, Lee and Narayanan (2005) combined acoustic, language and discourse information, averaging the confidence scores provided by classifiers, and considered as the final arbiter the emotion category with the highest average score.

The technique that we propose works at decision level since, as discussed above, the decision for each input utterance is made considering the separate predictions, i.e. decisions, of a set of classifiers.

3.1.3 Comparison considering the number of fusion stages

Most recent studies addressing the combination of emotion information provided by a set of classifiers obtain the deduced emotion category in just one fusion stage. This is the case of the great majority of studies that work at the decision level. Only a few studies consider two fusion stages. Among these, Morrison et al. (2007) used stacking generalisation to combine predictions from multiple classifiers. This method takes the predicted target categories of several 'base' or 'level-0' classifiers, and uses these to train a meta-learner or level-1 classifier. The meta-learner uses the level-0 predictions and the target categories to determine which classifiers are correct or incorrect, and generates a higher level prediction based on this. The technique we propose is inspired by this study. The similarity between the two is that the input to the first fusion stage is the results of a set of classifiers. The difference lies in the second fusion stage: while Morrison et al. (2007) used this fusion stage to decide which classifiers are correct or incorrect, our technique uses this fusion stage to combine the hypotheses generated by the first fusion stage.

4 Experiments

The goal of the experiments is to test the proposed technique employing:

i) Three emotion categories ('Neutral', 'Angry', 'Tired') on the one hand, and two emotion categories ('Non-negative', 'Negative') on the other. The experiments

employing the former category set will be called '3-emotion' experiments throughout the paper, whereas those employing the latter category set will be called '2-emotion' experiments.

ii) The four classifiers shown in Fig. 1, which will be described in Section 4.3: C_1 = Prosodic classifier, C_2 = Acoustic classifier, C_3 = Lexical classifier, and C_4 = Dialogue acts classifier.

iii) The three fusion methods mentioned in Section 3: Average of probabilities (AP), Multiplication of probabilities (MP) and Unweighted Vote (UV) (Kittler et al., 1998; Roli et al., 2004; Le et al., 2005; Morrison et al., 2007).

In the 3-emotion experiments we have considered that an input utterance is correctly classified if the deduced emotion category matches the label assigned to the utterance in a previous annotation procedure, which will be described in Section 4.2. In the 2-emotion experiments, the utterance is considered to be correctly classified if either the deduced emotion category is 'Non-negative' and the label is 'Neutral', or the category is 'Negative' and the label is 'Tired' or 'Angry'.

4.1 Study of the independence between information sources

Studying to what extent the four information sources considered in the experiments (prosodic, acoustic, lexical and related to dialogue acts) are independent from each other is important, since if two information sources are highly correlated, the corresponding classifiers are likely to give a similar decision on a given input data. It is known that the use of classifiers using information sources dependent on each other cannot improve the performance, it can even worsen performance. To make such a study we have calculated the *Q-statistic* measure on the test corpus (1,981 utterances). This measure computes the similarity between two classifiers C_i and C_j as follows (Kuncheva et al., 2001):

$$Q_{ij} = \frac{N_{11}N_{00} - N_{01}N_{10}}{N_{11}N_{00} + N_{01}N_{10}} \quad (1)$$

where N_{11} is the number of both classifiers making the correct decision; N_{10} is the number of C_i being correct and C_j being incorrect; N_{01} is the number of C_i being incorrect and C_j being correct; and N_{00} is the number of both classifiers making an incorrect decision. $N = N_{11} + N_{10} + N_{01} + N_{00}$ is the total number of data. Q_{ij} is in the range [-1,1]. For statistically independent classifiers $Q_{ij} = 0$, and the higher the absolute value of Q_{ij}, the more dependent the classifiers are. The results obtained from the study are shown in Table 1.

Table 1. Q-statistic for pairwise classifiers.

Classifiers	2 emotion categories	3 emotion categories
Pro+Aco	0.4789	0.4474
Pro+Lex	0.4322	0.2260
Pro+DA	0.3327	0.1699
Aco+Lex	0.5017	0.2589
Aco+DA	0.2350	0.0344
Lex+DA	0.6842	0.6647

As expected, for each classifier pair, the Q-statistic is higher for the 2-emotion than for the 3-emotion experiments, given that in the former case there are more coincidences in the predictions of the classifiers because there is no distinction between 'Tired' and 'Angry'.

The lowest Q-statistic values (0.2350 and 0.0344) are attained for the acoustic and dialogue acts classifiers, which shows that, in general, the users of the Saplen system did not acoustically emphasise their emotions; they were not actors simulating emotions (see Section 2), and were asked to interact with the system as naturally as possible. Hence, in many cases the dialogue act classifier predicts one emotion and the acoustic classifier predicts a different one.

Also as expected, the results show that the classifiers are not completely independent from each other given that sometimes there is coincidence in their deduced emotion category. In particular, the highest Q-statistic values (0.6842 and 0.6647) are attained for the lexical and the dialogue acts classifiers. This shows that there is a high coincidence in the emotion predicted by these two classifiers. The reason is that when the system has problems in understanding the user, it prompts reiteratively for the same data, and thus the dialogue acts classifier predicts the

emotion category 'Negative' quite accurately. In these problematic situations the users typically correct system misunderstandings uttering words such as 'No', 'Incorrect' and 'Said', which enables the lexical classifier in correctly predicting the 'Negative' emotion category as well. As will be discussed in Section 4.4.5.1, the experimental results prove that the level of dependence is low enough as to suggest us employ the four classifiers. However, it must be noted that level of dependence between information sources might be domain-dependent, given that the range and strength of emotional expressions may be different across domains. For example, we have observed that in our experiments in the fast food domain, the negative emotional states are a consequence of failures of the dialogue system. However, this might not be necessarily the case in other domains, e.g. tutoring systems, where emotional reactions can be more complex.

4.2 Description and annotation of the speech database

The speech database used in the experiments was constructed from a corpus of 440 telephone-based dialogues between users (students of our University) and the Saplen system. 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 orthographic transcriptions of user responses were included manually in the log files after the collection of the corpus.

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 assigned one label to each utterance, either '<NEUTRAL>', '<TIRED>' or '<ANGRY>' according to the perceived emotional state of the user.

The classification of emotions by a set of labellers is problematic because a given emotion can be perceived differently by different labellers. This is why studies on emotion

detection usually employ statistical measures to provide information about the inter-labeller agreement, such as the Cohen's Kappa statistic (Cohen, 1960). To measure the amount of agreement between the labellers in our experiments we have employed the Kappa statistic (K), which is computed as follows:

$$K = \frac{P(A) - P(E)}{1 - P(E)} \quad (2)$$

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 (Carletta, 1996; Cohen, 1960). We have obtained that the average K = 0.45, which according to Landis and Koch (1977) represents 'moderate agreement'. This low value for K shows the difficulty in labelling naturally occurring dialogues.

To carry out the 3-emotion experiments we assigned one label to each utterance according to the majority opinion of the labellers. If there was a tie in the labelling of an utterance (e.g. Neutral-Neutral-Tired-Tired), the utterance did go through a second stage of labelling in which the most experienced labeller in the group decided the label to be assigned to the utterance. We found that 81% of the utterances were annotated with the label 'Neutral', 9.5% with the label 'Tired' and 9.4% with the label 'Angry'. This shows that the experimental database is clearly unbalanced in terms of emotion categories.

Fig. 2 shows an excerpt of a labelled dialogue (translated from Spanish into English), where the number of dialogue turn is within brackets and the emotion category is within angle brackets.

(1) Saplen: Good evening, welcome to our restaurant … Would you like to have anything?
PromptType: AnyFoodOrDrinkToOrder?

(2) User: Yes, uhm, I would like to have one ham sandwich, one green salad and one beer, please <NEUTRAL>

(3) Saplen: What size would you like for the beer? Normal or large.
PromptType: GetDrinkSize

(4) User: Large <NEUTRAL>

(5) Saplen: OK, one ham sandwich, one green salad, and one large beer. Please say your telephone number.
PromptType: GetTelephone#

(6) User: 1 2 3 4 5 6 7 8 9 <NEUTRAL>

(7) Saplen: I guess you said 1 2 3 4 5 3 7 8 9. Is this correct?
PromptType: ConfirmTelephone#

(8) User: No, I said 1 2 3 4 5 6 7 8 9 <NEUTRAL>

(9) Saplen: OK, I have to confirm your telephone number again. Did you say 1 2 3 4 5 6 7 8 9? Please say yes or no.
PromptType: ConfirmTelephone#

(10) User: Yes <NEUTRAL>

(11) Saplen: OK, please say your post code.
PromptType: GetPostCode

(12) User: 1 8 0 3 4 <NEUTRAL>

(13) Saplen: I guess you said 1 8 0 6 4. Is this correct?
PromptType: ConfirmPostalCode

(14) User: No, I said 1 8 0 3 4 <TIRED>

(15) Saplen: I'm sorry, I have to confirm your post code again. Did you say 1 8 0 6 4? Please say yes or no.
PromptType: ConfirmPostalCode

(16) User: No, 1 8 0 3 4 <ANGRY>
…

Fig. 2. Excerpt of a labelled dialogue.

It can be observed that according to the labellers' annotation, in most dialogue turns the user is in the 'Neutral' emotional state. In turn (8) a system misunderstanding is corrected, but according to the annotation, the labellers have not perceived any change in the user's emotional state because of the system malfunction. However, in turns (14) and (16) they have perceived changes to 'Tired' and 'Angry' due to the system's difficulty in understanding the post code.

To train and test the classifiers we have divided the dialogue corpus into two disjoint corpora: one for training (5,942 utterances corresponding to 75% of the dialogues) and the other for testing (1,981 utterances corresponding to the 25% remaining dialogues). The division is made in such a way that both sets contain utterances representative of the 18 different utterance types in the dialogue corpus in accordance with the user dialogue act: product orders, telephone numbers, post codes, addresses, queries, confirmations, amounts, food names, ingredient names, drink names, sizes, tastes, temperatures, street names, building numbers, building floors,

apartment letters, and error indications. Table 2 sets out a distribution of the database in terms of utterances associated with each emotion category.

Table 2. Distribution of utterances in the speech database.

2 emotion categories				3 emotion categories			
	Whole corpus	Training partition	Test partition		Whole corpus	Training partition	Test partition
Non-negative	6,427	4,820	1,607	Neutral	6,427	4,820	1,607
Negative	1,496	1,122	374	Tired	752	564	188
				Angry	744	558	186

We think that the number of utterances from the same speaker in both sets is very small, given that no dialogues were split between training and test sets, and most speakers who participated in the collection of the database made just one call to our call centre.

4.3 Classifiers

Given that a large number of features indicating emotional states are present in the speech signal, one problem of emotion detection is to differentiate those that can be attributed to emotional behaviour from others that may be simply characteristics of spontaneous conversational speech. Several types of features are described in the literature, such as acoustic (e.g. Mel-frequency cepstral coefficients), prosodic, lexical (e.g. swear words and insults), related to the discourse (e.g. repetitions and rectifications), and non-verbal (e.g. laughter and mouth noise). In these experiments we have used four classifiers to employ some of these information sources.

4.3.1 Prosodic classifier

Features derived from pitch and energy, usually called 'prosodic features', have been successfully applied to emotion detection in many studies, e.g. Dellaert et al. (1996). In the experiments presented in this paper we have used global statistics of pitch and energy and features derived from the duration of voiced/unvoiced segments, to create one feature vector per utterance that forms the input to the prosodic classifier. To extract the pitch contours we have used the Praat software[2]. After carrying out additional experiments to find the appropriate

[2] http://www.praat.org

feature set for the classifier, we decided to use the following 11 features: pitch mean, minimum and maximum, mean of pitch derivatives, 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 emotion categories (Neiberg et al., 2006). The emotion category deduced by the classifier, h', is decided according to the maximum likelihood criterion. To compute the scores p_i for the emotion prediction of the classifier (see Fig. 1) we have used the following expression:

$$p_i = \beta_i / \sum_{k=1}^{S} \beta_k \tag{3}$$

where β_i is the log-likelihood of category h_i and S is the number of emotion categories considered.

4.3.2 Acoustic classifier

Like prosodic features, acoustic features are well established and widely used in the task of emotion detection. 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 Coefficients (MFCCs) or Linear Prediction Cepstral Coefficients (LPCCs). Additional experiments carried out with our speech database have confirmed this observation. However, we also observed 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 the evaluation of the classifier.

The emotion patterns of the input utterances are modelled by gender-dependent GMMs, as with the prosodic classifier, but each input utterance is represented by a sequence of feature vectors instead of one vector. The emotion category deduced by the classifier, h', is again decided by employing the maximum likelihood criterion, whereas Eq. (3) is used to compute the scores for the prediction, i.e. for the vector of pairs (h_i, p_i).

4.3.3 Lexical classifier

A number of previous studies on emotion detection take into account information about the kinds of words uttered by the users, assuming that there is a relationship between words and emotions. For example, swear words and insults can be considered as conveying a negative emotion (Lee et al., 2002; Lee and Narayanan, 2005). Analysis of our dialogue corpus has shown that the 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 understanding the sentences uttered by the users.

The reasons for the understanding problems are basically two-fold. 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. For example, the word 'sesenta_y_cinco' (sixty-five) sounds very similarly as the word 'setenta_y_cinco' (seventy-five), which sometimes caused the confusion of the speech recogniser. Because of these problems, there were sentences uttered by the users that could not be understood by the system. Hence, the users had to initiate error-recovery subdialogues, generally starting their next dialogue turns with expressions such as 'No, I said …' (see sample dialogue in Fig. 2). As the error-correction requires extra dialogue turns, the emotional state of the users changed from 'Neutral' to 'Angry' or 'Tired', especially when they had to correct consecutive system misunderstandings, or they had corrected other system errors in previous dialogue turns.

Words that are acoustically similar to others or that are considerably affected by the users' accents are more likely to be misrecognised, which usually causes negative emotional states of the users. Hence, our goal has been to automatically find these words by means of a study of the speech recognition results, and deduce the emotion category for each input utterance from the emotional information associated with the words in the utterance. 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 emotion category can be defined as the mutual information between the word and the emotion 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 H a set of emotion categories, $H = \{h_1, h_2, \ldots, h_S\}$. The mutual information between the word w_j and an emotion category h_i, $mutual\_Information(w_j, h_i)$, is defined as follows (Cover and Thomas, 1991):

$$mutual\_Information(w_j, h_i) = \log \frac{P(h_i \mid w_j)}{P(h_i)} \tag{4}$$

where $P(h_i \mid w_j)$ is the posterior probability that a sentence containing the word w_j implies the emotion category h_i, and $P(h_i)$ represents the prior probability of the emotion category. Note that the mutual information is positive only if $P(h_i \mid w_j) > P(h_i)$. Taking into account the previous definitions, the emotional salience of the word w_j for an emotion category h_i has been defined as:

$$salience(w_j, h_i) = P(h_i \mid w_j) \times mutual\_Information(w_j, h_i) \tag{5}$$

When the salient words for each emotion category have been identified employing the training corpus, we have carried out the emotion detection at the sentence level, considering that each word in a sentence is independent of the rest. The goal has been to map the sentence W to any of the emotion categories in H. To do this, we compute an activation value a_i for each emotion category as follows:

$$a_i = \sum_{j=1}^{n} I_{ji} c_{ji} + b_i \tag{6}$$

where $i = 1 \ldots S$, S is the number of emotion categories, n is the number of words in W, I_{ji} represents an indicator that has the value 1 if w_j is a salient word for the emotion category (i.e. $salience(w_j, h_i) \neq 0$) and the value 0 otherwise; c_{ji} is the connection weight between the word and the emotion category, and b_i represents bias specific for each emotion category. The connection

weight is defined as: $c_{ji} = mutual\_Information(w_j, h_i)$, whereas the bias is computed as: $b_i = \log P(h_i)$. Finally, the deduced emotion category, *h'*, is the one with highest activation value a_i:

$$h' = \arg\max_i (a_i) \tag{7}$$

To compute the scores p_i for the emotion prediction, we use the following expression:

$$p_i = a_i / \sum_{k=1}^{S} a_k \tag{8}$$

where a_i represents the activation value of h_i.

4.3.4 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 assigned 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 reiteratively (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 acts 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 emotion category of a user's dialogue turn n, $h'(n)$, is that which maximises the posterior probability given a sequence of the most recent system prompts:

$$h'(n) = \arg\max_i P(h_i(n) \mid DA_{n-(L*2-1)}, ..., DA_{n-7}, DA_{n-5}, DA_{n-3}, DA_{n-1}) \quad (9)$$

where the prompt sequence is represented by a sequence of dialogue acts (DA_j's) and L is the length of the history being considered, i.e. the number of recent system's dialogue turns in the evaluated sequence. In essence, the method uses dialogue-level n-grams of speech acts, also used by Taylor et al. (1998) and Stolcke et al. (2000). Note that if $L = 1$ then the decision about $h'(n)$ depends only on the previous system prompt. In other words, the emotion category obtained is that with the greatest probability given just the previous system turn in the dialogue. The probability of the considered emotion categories given a sequence of dialogue acts is obtained employing a training dialogue corpus.

In accordance with Eq. (9), the decision on the emotion category $h'(n)$ for the first four user turns in a dialogue, considering $L = 3$, is computed as follows:

$$h'(2) = \arg\max_i P(h_i(2) \mid DA_1)$$

$$h'(4) = \arg\max_i P(h_i(4) \mid DA_1, DA_3)$$

$$h'(6) = \arg\max_i P(h_i(6) \mid DA_1, DA_3, DA_5)$$

$$h'(8) = \arg\max_i P(h_i(8) \mid DA_3, DA_5, DA_7)$$

Employing this approach, we decide the most likely emotion category for the input utterance by selecting the one with the highest probability given the sequence of dialogue acts of length L. This probability is used to create the (h_i, p_i) pair to be included in the emotion prediction (see Fig. 1).

4.4 Performance of the classifiers and the fusion modules

As discussed in Section 4.2, 81% of the utterances in the speech database have been annotated as 'Neutral'. Hence, in the experiments we have considered a baseline that classifies each input utterance as 'Neutral' in the 3-emotion experiments, and as 'Non-negative' in the 2-emotion experiments. Obviously, the classification rate[3] of this classifier is 81% (note that the rate is 100% for the 'Neutral' utterances, 0% for the 'Tired' utterances and 0% for the 'Angry' utterances).

4.4.1 Performance of the prosodic classifier

Firstly, we have employed the 5,942 utterances from 330 training dialogues to create the gender-dependent Gaussian Mixture Models (GMMs) representative of the three emotion categories: 'Neutral', 'Tired' and 'Angry'. Secondly, we have used the 1,981 utterances from 110 dialogues in the test corpus to evaluate the performance of the classifier. The classifier uses an automatic procedure to decide the gender of the speaker based on an analysis of the voice signal in the input utterance. This gender detector is also based on GMMs and uses the same short-term acoustic features as the acoustic classifier, achieving an accuracy of around 92%.

According to the gender decision, the classifier used the appropriate gender-dependent GMMs. Table 3 sets out the results obtained in the testing, where the labels in the rows correspond to the correct emotion categories and the labels in the columns represent the emotion categories deduced by the proposed technique. As can be observed, the best performance for the 2-emotion experiments is attained for the utterances labelled as 'Negative' (86.89%), whereas for the 3-emotion experiments it is attained for the utterances labelled as 'Angry' (82.25%). The classification rates are 72.69% and 71.47% for the 2 and 3-emotion experiments, respectively, which means that the classifier performs worse than the baseline. These results show that in general it is very difficult for the classifier to distinguish between the three emotion categories

[3] The classification rate has been computed by dividing the total number of correctly classified utterances by the total number of analysed utterances.

considered. This happens because the users of the Saplen system uttered sentences naturally, without emphasising any kind of emotion, which makes it very difficult for the classifier to correctly deduce the emotion.

Table 3. Confusion matrices of the prosodic classifier (results in %).

2 emotion categories		
	Non-negative	Negative
Non-negative	**69.38**	30.61
Negative	13.10	**86.89**
Correctly classified: 72.69		

3 emotion categories			
	Neutral	Tired	Angry
Neutral	**69.38**	14.62	15.99
Tired	15.42	**78.72**	5.85
Angry	10.21	7.52	**82.25**
Correctly classified: 71.47			

However, the classifier enables the classification of 'Tired' and 'Angry' utterances with rates around 79% and 82%, respectively, whereas the baseline cannot classify these utterances at all. The reason why the rates of the classifier are lower is that it incorrectly classifies some 'Neutral' utterances as 'Tired' or 'Angry', and the proportion of 'Neutral' utterances is much larger in the speech database (as discussed in Section 4.2), which causes a big impact on the classification rate.

4.4.2 Performance of the acoustic classifier

Employing the models representative of the three emotion categories created for this classifier, as well as the automatic procedure for gender detection, the acoustic classifier analysed the test utterances and obtained the results set out in Table 4. It can be observed that the best performance for the 2-emotion experiments is achieved for the utterances labelled as 'Negative' (83.15%), whereas for the 3-emotion experiments it is attained for the utterances labelled as 'Tired' (78.19%).

Table 4. Confusion matrices of the acoustic classifier (results in %).

2 emotion categories		
	Non-negative	Negative
Non-negative	**72.68**	27.31
Negative	16.84	**83.15**
Correctly classified: 74.65		

3 emotion categories			
	Neutral	Tired	Angry
Neutral	**72.68**	8.46	18.85
Tired	14.36	**78.19**	7.44
Angry	19.89	7.52	**72.58**
Correctly classified: 73.19			

The classification rates are 74.65% and 73.19% for the 2 and 3-emotion experiments respectively, which means that the classifier works worse than the baseline. Similarly as what happens with the prosodic classifier, these results show the difficulty of the classifier in distinguishing the three emotion categories, given that the users of the Saplen system uttered sentences naturally, without artificially emphasising any emotion. However, the classification rates for the 'Tired' and 'Angry' utterances are around 78% and 72%, respectively, whereas the baseline always fails in classifying these sentences.

4.4.3 Performance of the lexical classifier

Following the approach discussed in Section 4.3.3, we have computed the emotional salience for the words in the results of the Saplen system's speech recogniser corresponding to the 5,942 training sentences. Table 5 set out a partial listing of the words (translated from Spanish into English) with the largest salience values for each emotion category. It can be observed that the words with largest salience value for the 'Neutral' category are concerned with product orders (e.g. 'Two', 'Salads', 'Order', 'Want', 'Curry' or 'Cheese'). This happens because the Saplen system did not have much trouble in understanding product orders, and thus the users were mostly in the 'Neutral' emotional state when they uttered product orders, which is typically at the beginning of the dialogue.

Table 5. Most salient words for each emotion category.

Neutral	Salience	Tired	Salience	Angry	Salience
Beer	0.1037	Sixty-eight	0.9346	Sixty-seven	1.0167
Two	0.0969	Sixty-seven	0.9346	Eighty-three	1.0165
Salads	0.0866	Sixty-three	0.9346	Twenty-one	0.4421
Cake	0.0845	Six	0.4990	Seventy-four	0.4419

One	0.0747	Fifty-eight	0.3038	Ninety-five	0.3884
Order	0.0745	Seventy-eight	0.2971	Sixty-four	0.3578
Sandwich	0.0695	Eleven	0.2817	Forty-five	0.3200
Four	0.0656	Twenty-six	0.2627	Fifty-eight	0.2630
Curry	0.0606	Forty-five	0.2588	Sixty-five	0.2328
Cheese	0.0549	Ninety-nine	0.2398	Sixty-eight	0.2168
Orange	0.0522	Seven	0.1746	Ninety-nine	0.2080
Fanta	0.0519	No	0.1620	Eleven	0.1992
Sandwiches	0.0496	Incorrect	0.1581	No	0.1941
Ham	0.0467	Said	0.1454	Incorrect	0.1911
Loin	0.0436	Sixty-five	0.1370	Said	0.1329

On the contrary, the greatest salience values for the 'Tired' and 'Angry' categories are related to digit pairs (e.g. 'Twenty-six'), which are employed by the users to provide their telephone numbers[4]. This happens because the system had problems in understanding some telephone numbers, motivated by the strong southern Spanish accents of most users. Hence, the users had to repeat the telephone number several times, which provoked a change in their emotional state from 'Neutral' to 'Tired' or 'Angry'. It can also be observed that the words 'No', 'Incorrect' and 'Said', which are typically employed by the users to correct system errors, appear in the negative emotion categories (either 'Tired' or 'Angry'). This occurs because users are mostly in a negative emotional state when they utter sentences of the form 'No, I said …' to correct system misunderstandings.

Using the information about emotional salience learnt from the training, the lexical classifier analysed the utterances in the test corpus, obtaining the results set out in Table 6.

Table 6. Confusion matrices of the lexical classifier (results in %).

2 emotion categories		
	Non-negative	Negative
Non-negative	**94.15**	5.84
Negative	36.36	**63.63**
Correctly classified: 88.38		

3 emotion categories			
	Neutral	Tired	Angry
Neutral	**94.15**	3.98	1.86
Tired	19.68	**76.59**	3.72
Angry	55.37	40.32	**4.30**
Correctly classified: 84.04			

[4] In Spain people typically use digits and combinations of digits instead of isolated digits to utter telephone numbers (e.g. "nine eight seven sixty-five forty-three twenty-one").

The classification rates are 88.38% and 84.04% for the 2-emotion and 3-emotion experiments, respectively. From these results it follows that the classifier outperforms the baseline by approximately 7% absolute for the 2-emotion experiments (88.38% vs. 81%) and by 3% absolute for the 3-emotion experiments (84.04% vs. 81%). This shows that it is relatively easy for the classifier to distinguish between 'Non-negative' and 'Negative' emotions, but it is more difficult to differentiate between 'Tired' and 'Angry'. This difficulty arises because there is no correlation between speech recognition errors and a specific negative emotion category, as given a misrecognised sentence, the emotional state of the users changes from 'Neutral' to any of the two negative categories.

4.4.4 Performance of the dialogue acts classifier

Following the method described in Section 4.3.4, we have employed the 330 training dialogues to compute the posterior probabilities of the three emotion categories in our speech database ('Neutral', 'Tired' and 'Angry') given sequences of system's dialogue acts of length L, $1 \leq L \leq 10$. Table 7 shows a partial list of the most likely sequences of system's dialogue acts associated with a negative emotion category ('Tired' or 'Angry') considering $L = \{1, 2, 3, 4\}$.

Table 7. Partial list of the most likely sequences of system's dialogue acts associated with a negative emotion category.

***L* = 1**	
AnythingElse?	0.87
ConfirmBuilding#	0.74
ConfirmTelephone#	0.58
GetTelephone#	0.44

GetBuilding#	0.12
GetPostCode	0.09
***L* = 2**	
AnythingElse?, AnythingElse?	0.92
ConfirmBuilding#, ConfirmBuilding#	0.87
ConfirmCancelProduct, AnythingElse?	0.78
ConfirmTelephone#, GetTelephone#	0.73
AnyThingToDrink?, AnythingElse?	0.71
ConfirmDrinkOrder, AnythingElse?	0.69
***L* = 3**	
AnythingElse?, AnythingElse?, AnythingElse?	0.95
ConfirmBuilding#, ConfirmBuilding#, ConfirmBuilding#	0.93
ConfirmCancelProduct, AnythingElse?, AnythingElse?	0.90
GetBuilding#, ConfirmBuilding#, ConfirmBuilding#	0.88
AnyFoodOrDrinkToOrder?, AnythingElse?, AnythingElse?	0.87
ConfirmBuilding#, GetBuilding#, ConfirmBuilding#	0.84
***L* = 4**	
AnyFoodOrDrinkToOrder?, AnythingElse?, AnythingElse?, AnythingElse?	0.97
AnythingElse?, AnythingElse?, AnythingElse?, AnythingElse?	0.95
ConfirmTelephone#, ConfirmTelephone#, ConfirmTelephone#, ConfirmTelephone#	0.93
AnythingElse?, AnythingElse?, AnyFoodOrDrinkToOrder?, AnythingElse?	0.93
ConfirmBuilding#, ConfirmBuilding#, GetBuilding#, ConfirmBuilding#	0.92
ConfirmBuilding#, GetBuilding#, ConfirmBuilding#, ConfirmBuilding#	0.90

It can be observed that the system's dialogue acts that are most likely to induce a negative emotional state of the user are related, on the one hand, to confirmation prompt types (e.g. 'ConfirmBuilding#', 'ConfirmTelephone#', 'ConfirmCancelProduct', 'ConfirmDrinkOrder'), whilst on the other, to prompts to obtain from the user data comprised of digits (e.g. 'GetBuilding#', 'GetTelephone#', 'GetPostCode'). This happens because of the difficulty of the dialogue system in correctly understanding confirmations, telephone numbers and post codes, especially when the sentences were uttered by users with strong southern Spanish accents. Because of this difficulty, sometimes the affirmative confirmation 'sí' (yes) was recognised as 'seis' (six), and the negative confirmation 'no' was recognised as 'dos' (two). These confirmation errors made the system's misunderstanding of the confirmation. Hence, the system had to employ additional dialogue turns to get the confirmation, which was irritating for the users. Regardless of the users' accent, as discussed in Section 4.3.3, there are words in the system's vocabulary that are very similar acoustically, which caused the confusion of the speech recogniser. This problem forced some users to employ additional dialogue turns to make the system understand their telephone numbers or post codes, which irritated them as well.

To find the best value of L we carried out additional experiments employing the posterior probabilities obtained for each emotion category and the different lengths of dialogue act sequences, $L = \{1, 2, \ldots, 10\}$. Fig. 3 sets out the results obtained.

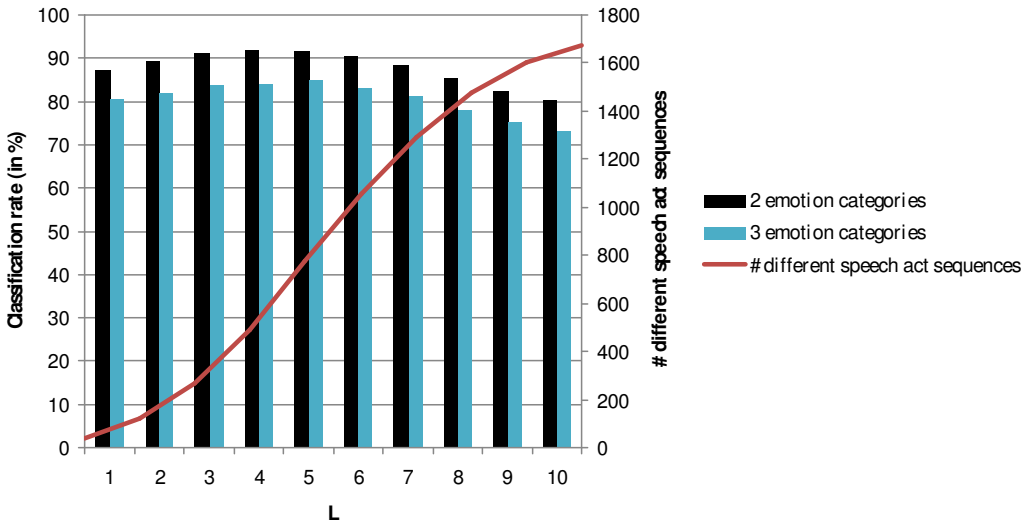


Fig. 3. Performance of the dialogue acts classifier.

It can be observed that the classification rate increases with L until it reaches a threshold ($L = 4$) that yields the best results: 91.77% and 84.45% for the 2 and 3-emotion experiments respectively. This happens because for small values of L ($L < 4$) the sequence of dialogue acts does not provide enough information to correctly predict the emotion category. In particular, when $L = 1$ the emotion prediction relies only on the previous prompt type, which does not take into account whether the current prompt has already been answered by the user in previous dialogue turns. On the contrary, for large values of L ($L > 4$) the number of different sequences of dialogue acts is much larger (e.g. 1,674 for $L = 10$) given that the sequences are longer. Consequently, the probabilities of these sequences are lower, which decreases the correlation with the emotion category and thus leads to worse classification results. Given the results for the different values of L, we have selected 4 as the value for L to be used in the testing.

Table 8 sets out the confusion matrices obtained from the analysis of the test utterances. It can be observed that it is relatively easy for the classifier to detect whether, according to the sequence of previous system turns, the user is likely to be in a 'Non-negative' or 'Negative' emotional state.

Table 8. Confusion matrices of the dialogue acts classifier (results in %).

2 emotion categories		
	Non-negative	Negative
Non-negative	**95.70**	4.29
Negative	25.13	**74.86**
Correctly classified: 91.77		

3 emotion categories			
	Neutral	Tired	Angry
Neutral	**95.70**	2.36	1.92
Tired	22.87	**36.17**	40.95
Angry	19.89	44.08	**36.02**
Correctly classified: 84.45			

The difficulty lays in detecting the type of negative state (either 'Tired' or 'Angry'). As commented above, this difficulty arises because the users of the Saplen system expressed their irritation due to system malfunctions in different ways: in some cases, their emotional state changed from 'Neutral' to 'Tired' whilst in others, from 'Neutral' to 'Angry'.

4.4.5 Performance of the fusion modules

This section examines the performance of Fusion-0 in combining the predictions generated by the classifiers, and the performance of Fusion-1 in combining the predictions of Fusion-0.

4.4.5.1 Performance of Fusion-0

Table 9 sets out the results obtained for Fusion-0 considering several combinations of the classifiers and employing the three fusion methods discussed in Section 3. The results (in %) are reported in terms of accuracy (Acc), unweighted average precision (avg. Prec) and unweighted average recall (avg. Rec) for the emotion categories considered, and *F*-measure of avg. Prec and avg. Rec. As can be observed, MP is the best fusion method. The best classification rates (92.23% and 90.61%) are obtained employing the four classifiers, which proves that they all are sufficiently independent from each other (as discussed in Section 4.1). For the best configuration, Fusion-0 outperforms the baseline (81%) by 11.23% and 9.61% absolute for the 2 and 3-emotion experiments, respectively.

Table 9. Performance of Fusion-0.

Fusion method	Classifiers	2 emotion categories				3 emotion categories			
		Acc	avg. Prec	avg. Rec	*F*	Acc	avg. Prec	avg. Rec	*F*
AP	Aco+Pro	84.15	75.71	84.90	80.04	82.48	65.38	79.41	71.72
	Lex+Pro	85.06	76.09	82.48	79.16	82.69	63.87	72.72	68.01
	DA+Pro	90.51	83.65	87.59	85.57	87.48	70.36	75.33	72.76
	Aco+Lex+Pro	89.20	81.62	86.47	83.97	86.17	68.44	74.31	71.25
	Aco+DA+Pro	90.26	83.25	87.33	85.24	88.54	74.07	79.70	76.78
	DA+Lex+Pro	90.01	82.87	86.97	84.37	88.04	72.61	78.39	75.39
	Aco+DA+Lex+Pro	90.51	83.65	87.59	85.57	88.84	74.78	80.14	77.37
MP	Aco+Pro	84.15	75.71	84.90	80.04	82.84	66.17	80.66	72.70
	Lex+Pro	85.16	76.21	82.54	79.25	83.70	66.29	75.97	70.80
	DA+Pro	91.47	85.25	88.38	86.79	89.80	76.32	80.85	78.52
	Aco+Lex+Pro	89.15	81.54	86.44	83.92	87.88	73.49	80.53	76.85
	Aco+DA+Pro	91.32	84.98	88.29	86.60	89.25	74.90	79.36	77.07
	DA+Lex+Pro	90.06	82.94	87.10	84.97	87.83	72.02	77.68	74.74
	Aco+DA+Lex+Pro	**92.23**	**86.54**	**89.16**	**87.83**	**90.61**	**78.27**	**81.81**	**80.00**
UV	Aco+Pro	88.64	80.84	89.41	84.91	85.21	66.91	77.70	71.90
	Lex+Pro	86.42	77.80	83.83	80.70	82.99	62.96	70.33	66.44
	DA+Pro	88.19	80.17	85.23	82.62	84.91	65.79	72.06	68.78
	Aco+Lex+Pro	88.74	80.96	85.98	83.40	85.56	67.21	73.28	70.11
	Aco+DA+Pro	88.89	81.19	85.97	83.51	85.87	67.75	73.72	70.61
	DA+Lex+Pro	88.49	80.60	85.62	83.04	85.61	67.44	73.93	70.53
	Aco+DA+Lex+Pro	89.05	81.42	86.07	83.68	87.58	72.66	79.30	75.84

Analysis of Fusion-0 using MP (see Table 10) shows that for the 2-emotion experiments, Fusion-0 performs very well at predicting the Non-negative emotion category (94.08%) and works slightly better than the baseline at predicting the Negative category (84.22%). Regarding the 3-emotion experiments, we observe that Fusion-0 performs very well at predicting the 'Neutral' category (94.08%), but clearly worse for the 'Tired' (75.53%) and 'Angry' (75.80%).

Table 10. Confusion matrices of Fusion-0 employing MP (results in %).

2 emotion categories		
	Non-negative	Negative
Non-negative	**94.08**	5.91
Negative	15.77	**84.22**
Correctly classified: 92.23		

3 emotion categories			
	Neutral	Tired	Angry
Neutral	**94.08**	2.55	3.36
Tired	12.76	**75.53**	11.70
Angry	18.81	5.37	**75.80**
Correctly classified: 90.61			

4.4.5.2 Performance of Fusion-1

Table 11 shows the results obtained when Fusion-1 is used to combine the predictions of Fusion-0. 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.

Table 11. Performance of Fusion-1.

Fusion-1 method	Fusion-0 method	2 emotion categories				3 emotion categories			
		Acc	avg. Prec	avg. Rec	*F*	Acc	avg. Prec	avg. Rec	*F*
AP	AP+ PM	93.69	89.33	90.37	89.84	91.77	80.50	81.81	81.15
	AP+UV	83.19	88.33	89.95	89.13	91.67	80.69	82.87	81.77
	MP+UV	93.34	88.64	90.05	89.34	91.27	79.19	80.98	80.07
	AP+MP+UV	93.24	88.43	89.98	89.20	91.57	80.42	82.36	81.38
MP	AP+ PM	**94.50**	**90.92**	**91.17**	**91.05**	**93.99**	**87.31**	**87.61**	**87.46**
	AP+UV	93.24	88.43	89.98	89.20	93.13	85.58	87.89	86.72
	MP+UV	94.40	90.69	91.11	90.90	93.99	87.41	87.92	87.66
	AP+MP+UV	94.35	90.64	90.98	90.81	93.94	87.21	87.74	87.48
UV	AP+ PM	93.54	89.01	90.27	89.64	90.96	77.93	79.43	78.67
	AP+UV	93.19	88.33	89.95	89.13	90.11	75.46	77.35	76.40
	MP+UV	93.19	88.38	89.95	89.11	89.50	73.44	75.05	74.24
	AP+MP+UV	93.19	88.33	89.95	89.13	89.05	71.99	73.61	72.79

Comparison of Table 9 and Table 11 shows that Fusion-1 outperforms Fusion-0. When MP is used in Fusion-1 to combine the predictions generated by Fusion-0 employing AP and MP, the classification rates increase by 2.27% absolute (from 92.23% to 94.50%) and 3.38% absolute (from 90.61% to 93.99%) for the 2 and 3-emotion experiments, respectively. Table 12 sets out the confusion table for this configuration. Comparing this table with Table 10, it can be observed that Fusion-1 slightly enhances the classification rate of Fusion-0 for the 'Non-negative' (94.08% vs. 96.51%) and the 'Negative' (84.22% vs. 85.82%) category. Overall, the best performance of Fusion-1 employing MP (94.50%) outdoes that of Fusion-0 employing AP (90.51%) and MP (92.23%) (see Table 9).

Table 12. Confusion matrices of Fusion-1 employing MP (results in %).

2 emotion categories		
	Non-negative	Negative
Non-negative	**96.51**	3.48
Negative	14.17	**85.82**
Correctly classified: 94.50		

3 emotion categories			
	Neutral	Tired	Angry
Neutral	**96.51**	1.49	1.99
Tired	11.70	**83.51**	4.78
Angry	16.66	0.53	**82.79**
Correctly classified: 93.99			

Comparing again Table 12 with Table 10, we can see that for the 3-emotion experiments, Fusion-1 slightly increments the classification rate of the 'Neutral' category obtained by Fusion-0 (94.08% vs. 96.51%), and that it raises the rate of the 'Tired' category (75.53% vs. 83.51%), and the 'Angry' category (75.80% vs. 82.79%). Overall, the performance of Fusion-1 employing MP (93.99%) outdoes that of Fusion-0 employing AP (88.84%) and MP (90.61%).

4.4.6 Test of statistical significance

In order to show the statistical significance of our results, we have carried out a paired *t*-test in which we have compared the performance of the proposed technique (Fusion-0 plus Fusion-1) with that of the standard fusion (Fusion-0). To do this we have considered, on the one hand, the standalone performance of Fusion-0 employing the three fusion methods (AP, MP and UV), whilst on the other, the performance of Fusion-0 plus Fusion-1 with the latter employing different fusion methods to combine emotion predictions of Fusion-0.

Table 13 shows the significance levels (critical *p*-values) obtained, where "FxYY" denotes "Fusion at level x employing method YY"; for example, F0AP+F0MP+F1UV means that Fusion-1 combines using UV the result of Fusion-0 employing AP and the result of Fusion-0 using MP. In all cases, the input to Fusion-0 is the output of the four classifiers. To obtain the scores in the table we have employed a partition comprised of 100 dialogues taken at random from the test corpus. We have analysed each dialogue in this partition twice, firstly using Fusion-0 only, employing all of the three fusion methods (see rows in Table 13), and secondly employing Fusion-0 plus Fusion-1, employing several combinations of the fusion methods (see columns in the table). The result has been 100 pairs of average emotion recognition accuracy per dialogue, which have been employed to compute the corresponding score in the table. It can be observed that all scores are lower than a threshold for statistical significance $\alpha = 0.05$, which means that our results are statistically significant.

Table 13. *t*-test significance results.

	F0AP +	F0AP +	F0AP +	F0AP +	F0AP +	F0AP +	F0AP +	F0AP +	F0AP +

	F0MP + F1AP	F0MP + F1MP	F0MP + F1UV	F0UV + F1AP	F0UV + F1MP	F0UV + F1UV	F0MP + F0UV + F1AP	F0MP + F0UV + F1MP	F0MP + F0UV + F1UV
F0AP	0.00071	1.6E-08	8.0E-11	0.00889	0.00889	7.2E-09	0.4149	2.8E-08	0.00111
F0MP	2.1E-05	0.00388	0.00422	6,7E-06	6.7E-06	1.3E-14	1.8E-05	0.00019	1.6E-05
F0UV	1.6E-10	4.1E-17	8.1E-19	2.0E-08	2.0E-08	0.00084	1.4E-09	4.5E-17	3.0E-12

(a) 2 emotion categories.

	F0AP + F0MP + F1AP	F0AP + F0MP + F1MP	F0AP + F0MP + F1UV	F0AP + F0UV + F1AP	F0AP + F0UV + F1MP	F0AP + F0UV + F1UV	F0AP + F0MP + F0UV + F1AP	F0AP + F0MP + F0UV + F1MP	F0AP + F0MP + F0UV + F1UV
F0AP	0.00160	0.00439	0.00488	2.4E-05	9.3E-06	5.3E-11	9.3E-06	0.00499	0.00438
F0MP	0.00241	0.04212	0.26297	0.00073	0,00047	2.1E-10	0.00047	0.00158	0.00347
F0UV	9.5E-11	2.6E-10	1.0E-11	1.6E-11	5.5E-11	0.08696	5.5E-11	4.3E-11	2.7E-12

(b) 3 emotion categories.

4.4.7 Comparison of results with previous studies available in the literature

In this section we compare the results of the proposed technique with others reported in the literature. As making such a comparison is a difficult task, to simplify the problem we looked for studies considering the same number of emotion categories as used in our study. However, it should be noted that the number of categories is not the only aspect that matters; for example, another important factor is the difficulty in differencing between the categories. As a result of our search, we found a number of studies employing two emotion categories, and just one considering three categories.

The comparison of these studies is illustrated in Table 14, where the last two rows show classification accuracies[5]. It must be noted that the conclusions drawn from the comparison should be considered relative given that previous studies were performed under a diversity of factors, for example, information sources (e.g., acoustic, prosodic, linguistic, dialogue-related) and emotional speech databases (e.g., types of emotion categories, balance in the number of categories, quality of the recordings, labelling reliability, etc.). For example, Lee et al. (2001)

[5] The result of Morrison et al. (2007) refers to that obtained with their NATURAL database (2 emotion categories), whereas that of Litman and Forbes-Riley (2006) refers to their study with human-computer dialogues (3 emotion categories).

carried out experiments employing a balanced database comprised of 142 utterances that were selected when the two labellers completely agreed in the tagging. In order to work as well with a balanced database, Batliner et al. (2003) employed a WOZ scenario to control the number of utterances in each emotion category. On the contrary, our experiments have been carried out employing an unbalanced database collected from user interactions with a real spoken dialogue system, without controlling ourselves the behaviour of the system in order to elicit specific user emotions.

Regarding the emotion categories, Litman and Forbes-Riley (2006) worked with three categories (Negative, Neutral and Positive). Therefore, in principle these categories are more easily distinguishable that the ones we have considered, given that we have worked with two negative categories (Tired and Angry) instead of one.

It must be mentioned as well that Litman and Forbes-Riley (2006) employed high-quality audio recording with head-mounted microphone, whereas we have employed recordings of telephone calls made in a call centre. Thus, the lower signal quality may affect the reliability of the labelling of our corpus. To address this potential problem, Morrison et al. (2007) employed nine judges to re-label an initial corpus and finally employed for the experiments a corpus comprised of 388 utterances (155 angry, 233 neutral), with a mean labelling agreement of 81.95%. On the contrary, we have used the whole corpus and employed four judges, attaining a value of Kappa that suggests just moderate agreement.

Table 14. Comparison of results with previous studies.

	Lee et al. (2001)	Ang et al. (2002)	Batliner et al. (2003)	Litman & Forbes-Riley (2006)	Morrison et al. (2007)	Our technique
Emotional speech database	142 utterances with call centre (balanced). Spontaneous emotions	837 dialogues with call centre. 21,899 utterances. Spontaneous emotions	Actor (1,336 utter.) Read (1,900 utter.) WOZ (24 dialogues; 2,863 utterances)	15 dialogues with computer tutor. 333 utterances. Spontaneous emotions	388 utterances with call centre (after re-judgement for better objective labelling). Spontaneous emotions	440 dialogues with call centre. 7,923 utterances. Spontaneous emotions
Application domain	Not stated	Travel arrangement	Appointment scheduling	Tutoring	Electricity billing	Fast food ordering
Information sources for emotion classification	Acoustic	Prosodic POS (Part-of-speech) Dialogue acts Speaking style	Prosodic Lexical Dialogue acts POS (Part-of-speech)	Acoustic Prosodic Speaker gender Subject ID Problem ID	Prosodic	Acoustic Prosodic Speaker gender ASR results Dialogue history
Classification algorithm	KNN	Decision trees Class-based trigram model	MLP	Decision trees	SVM MLP KNN K* RF	Based on GMMs, emotional salience and dialogue act n-gram
Emotion categories	Negative/Non-negative	Frustration/Other	Emotional/Neutral	Negative/Positive/Neutral	Anger/Neutral	Negative/Non-negative; Neutral/Tired/Angry
Combination of information sources (fusion)	No	Language model features added to prosodic decision trees	Conversational labels added to prosodic features	User and problem ID, utterance transcriptions and ASR results added to acoustic-prosodic features	Stacked generalisation Unweighted vote	Average of prob. Multiplic. of prob. Unweighted vote
Fusion type	-	Feature level	Feature level	Feature level	Decision level	Decision level
Classification accuracy (2 emotion categ.)	**75.81%** for male speakers **80%** for female speakers (Baseline: **50%**)	**91.80%** using true words **86.70%** using ASR words (Baseline: **84%**)	**95.70%** (Actor) **79.60%** (Read) **73.70%** (WOZ) (Baseline: **55%**)	-	**79.43%** (Baseline: **60%**)	**94.48%** (Baseline: **81%**)
Classification accuracy (3 emotion categ.)	-	-	-	**66%** (Baseline: **49%**)	-	**94%** (Baseline: **81%**)

5 Conclusions and future work

In this paper we have presented a technique to enhance emotion detection in spoken dialogue systems taking into account two modules (Fusion-0 and Fusion-1) that combine different information sources. Fusion-0 combines emotion predictions generated by a set of classifiers, which is a method commonly employed in previous studies to enhance classification rates. Fusion-1 represents the novelty of the technique, which is the combination of emotion predictions generated by Fusion-0. One advantage of the technique is that it can be applied as a posterior processing stage to any other methods that combine information from different information sources at the decision level. This is so because the technique works on the predictions (outputs) of the methods, without interfering in the procedure used to obtain these predictions. Another advantage is that the technique can be implemented as a modular architecture, which facilitates the setting up within a spoken dialogue system as well as the deduction of the emotional state of the user in real time.

We have carried out experiments employing a speech database collected from real users interacting with an experimental dialogue system, which was labelled considering three emotion categories ('Neutral', 'Tired' and 'Angry'). We have employed classifiers to deal with prosodic, acoustic, lexical and dialogue acts information. Three fusion methods (average of probabilities, multiplication of probabilities and unweighted vote) have been employed to combine in Fusion-0 the predictions made by the classifiers, and in Fusion-1 the predictions generated by Fusion-0. The results obtained show that the proposed technique is useful to improve the classification rates of the standard fusion. Comparing results in Table 9 and Table 11 we can observe that for the 2-emotion experiments, Fusion-1 enhances Fusion-0 by 2.27% absolute (from 92.23% to 94.50%), while for the 3-emotion experiments, the improvement is 3.38% absolute (from 90.61% to 93.99%). These improvements are obtained 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 the improvements presented in this paper is that, in accordance with our experiments, the two stage fusion process makes it possible to get more benefit from the advantages of using different methods to combine information. According to our results, these 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 context). MP provides better results when the data contain small errors, which occurs when the data provided by the classifiers is processed by Fusion-0 and the output of this module is the input to Fusion-1.

We think that the proposed technique could be applied to spoken dialogue systems designed for any domain, for example, transactional tasks, games or tutoring. This is so because it employs a general, domain-independent procedure based on using emotion predictions generated at the decision level by a set of classifiers, which are combined by two fusion modules. In application domains different from our experimental field perhaps we would need to employ other information sources and/or fusion methods in order to attain the optimal results, but the basic principles would remain. However, we have experimented with one domain only and thus have not provided any experimental evidence in this direction.

Therefore, more empirical work is needed to explore whether the proposed technique is useful in other application domains, where lexicon, user utterances and emotional reactions can be more complex, for example, tutoring systems. This study would be very important, given that our goal in detecting emotion categories is the adaptation of the performance of spoken dialogue systems used in automated call centres. In this setting, the basic motivation for a negative emotional state of the user is typically a consequence of malfunction of the system due to speech recognition or understanding errors. However, negative emotional states in other application types, e.g. tutoring systems, may also deal with other factors, such as temporal issues, psychological aspects or speaker personality types. This fact may affect the

independence of information sources and thus the selection of the more useful sources for emotion prediction. Moreover, it may affect the classification rates of some classifiers. For example, it would be reasonable to expect lower accuracy in our dialogue acts classifier if there are other reasons for negative emotional states in addition to system errors, e.g. personality types. Therefore, it would be very interesting to find out to what extent the proposed technique provides enhanced classification rates in other application domains.

Future work will include as well 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) have been commonly employed in previous studies. However, there are also studies in the literature that suggest using other information sources, such as speaking style, subject and problem identification, and non-verbal cues. Another possibility for 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.

We also plan to investigate the use of weights in the fusion processes. In the experiments we have assumed that all the classifiers and fusion methods have the same significance when the combination of the predictions takes place. However, it would be possible to measure the accuracy of classifiers and fusion modules and apply weights accordingly.

Acknowledgements

This research has been funded by the Spanish project HADA TIN2007-64718, the Czech Grant Agency project no. 102/08/0707 and the Student Grant Scheme (SGS) at the Technical University of Liberec. The authors would like to thank the reviewers and the guest

editors for their comments, suggestions and corrections that significantly improved the quality of this paper.

6 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.

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

Barra-Chicote, R., Fernández, F., Lufti, S., Lucas-Cuesta, J. M., Macías-Guarasa, J., Montero, J. M., San-Segundo, R., Pardo, J. M. 2009. Acoustic emotion recognition using dynamic Bayesian networks and multi-space distributions. Proc. of Interspeech, pp. 336-339.

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

Bozkurt, E., Erzin, E., Erdem, C. E., Erdem, A. T. 2009. Improving automatic emotion recognition from speech signals. Proc. of Interspeech, pp. 324-327.

Carletta, J. 1996. Assessing agreement on classification tasks: The kappa statistic. Computational Linguistics, 22(2), pp. 249-254.

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

Cover, T., Thomas, J. 1991. Elements of Information Theory. New York, Wiley.

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.

Hastie, H. W., Prasad, R., Walker, M. A. 2002. What's the trouble: Automatically identifying problematic dialogs in DARPA Communicator dialog system. Proc. of ACL, pp. 384-391.

Huber, R., Batliner, A., Buckow, J., Nöth, E., Warnke, V., Niemann, H. 2000. Recognition of emotion in a realistic dialogue scenario. Proc. of ICSLP, pp. 665-668.

Kanda, T., Hirano, T., Eaton, D., Ishiguro, H. 2004. Interactive robots as social partners and peer tutors for children: A field Trial. Human-Computer Interaction, 19(1), pp. 61-84.

Kittler, J., Hatef, M., Duin, R. P. W., Matas, J. 1998. On combining classifiers. IEEE Trans. on Pattern Analysis and Machine Intelligence, 20(3), pp. 226-240.

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.

Kuncheva, L., Bezdek, J. Duin, R. 2001. Decision templates for multiple classifier fusion: An experimental comparison. Pattern Recognition, 34(2), pp. 299-314.

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

Le, C. A., Huynh, V.-N., Shimazu, A. 2005. Combining classifiers with multi-representation of context in word sense disambiguation. LNAI 3518, pp. 262-268.

Lee, C. M., Narayanan, S. S., Pieraccini, R. 2001. Recognition of negative emotions from the speech signal. Proc. of ASRU, pp. 240-243.

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

Lee, C. M., Narayanan, S. 2003. Emotion recognition using a data-driven fuzzy inference system. Proc. of Eurospeech, pp. 157-160.

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

Lee, C.-C., Mower, E., Busso, C., Lee, S., Narayanan, S. 2009. Emotion recognition using a hierarchical binary decision tree approach. Proc. of Interspeech, pp. 320-323.

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

Litman, D. J., Forbes-Riley, K. 2006. Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. Speech Communication, 48, pp. 559-590.

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.

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

Luengo, I., Navas, E., Hernáez, I. 2009. Combining spectral and prosodic information for emotion recognition in the Interspeech 2009 Emotion Challenge. Proc. of Interspeech, pp. 332-335.

Lugger, M., Yang, B. 2009. On the relevance of high-level features for speaker independent emotion recognition of spontaneous speech. Proc. of Interspeech, pp. 1995-1998.

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

Möller, S. 2004. Quality of Telephone-Based Spoken Dialogue Systems. Springer, 2004.

Nakatsu, R., Hicholson, J., Tosa, N. 1999. Emotion recognition and its application to computer agents with spontaneous interactive capabilities. Proc. of the International Conference on Multimedia Computing and Systems, Florence, Italy.

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

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

Ortony, A., Clore, G., Collins, A. 1990. The cognitive structure of emotions. Cambridge, U.K., Cambridge Univ. Press.

Petrushin, V. 2000. Emotion recognition in speech signal, experimental study, development, and application. Proc. of ICSLP, Beijing, China.

Plutchik, R. 1994. The psychology and biology of emotions. New York: HaperCollins College.

Polzehl, T., Sundaram, S., Ketabdar, H., Wagner, M., Metze, F. 2009. Emotion classification in children's speech using fusion of acoustic and linguistic features. Proc. of Interspeech, pp. 340-343.

Roli, F., Kittler, J., Windeatt, T. (Eds.) 2004. Multiple classifier systems. Proc. of 5th International Workshop MSC 2004, LNCS, Vol. 3077. Springer.

Scheirer, J., Fernandez, R., Klein, J., Picard, R. W. 2002. Frustrating the user on purpose: A step toward building an affective computer. Interacting with Computers, 14(2), pp. 93-118.

Schuller, B., Steidl, S., Batliner, A. 2009. The INTERSPEECH 2009 Emotion Challenge. Proc. of Interspeech, pp. 312-315.

Steidl, S. 2009. Automatic classification of emotion-related user states in spontaneous children's speech. Logos Verlag, Berlin.

Stolcke, A., Ries, K., Coccaro, H., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Meteer, M., Van Ess-Dykema, C. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. Computational Linguistics, 26(3), pp. 339-373.

Taylor, P., King, S., Isard, S., Wright, H. 1998. Intonation and dialog context as constraints for speech recognition. Language and Speech, 41(3-4), pp. 489-508.

Xu, L., Xu, M., Yang, D. 2009. ANN based decision fusion for speech emotion recognition. Proc. of Interspeech, pp. 2035-2038.

Yacoub, S., Simske, S., Lin, X. Burns, J. 2003. Recognition of Emotions in Interactive Voice Response Systems. Proc. of Interspeech, pp. 729-732.