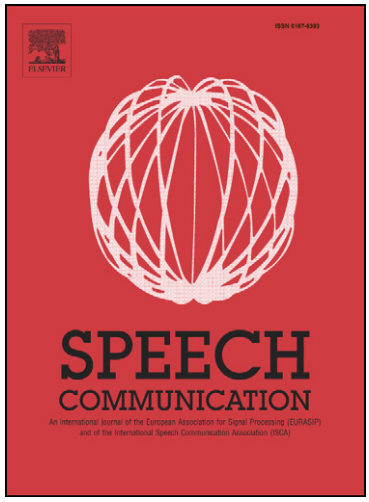
Influence of contextual information in emotion annotation for spoken dialogue systems

Zoraida Callejas, Ramòn Lòpez-Còzar



Influence of contextual information in emotion annotation for spoken dialogue systems

Zoraida Callejas, Ramòn Lòpez-Còzar

PII: S0167-6393(08)00003-4
DOI: 10.1016/j.specom.2008.01.001
Reference: SPECOM 1684

To appear in: *Speech Communication*

Received Date: 2 April 2007
Revised Date: 30 October 2007
Accepted Date: 6 January 2008

Please cite this article as: Callejas, Z., Lòpez-Còzar, R., Influence of contextual information in emotion annotation for spoken dialogue systems, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.01.001

Influence of contextual information in emotion annotation for spoken dialogue systems

Zoraida Callejas * Ramón López-Cózar

Dept. of Languages and Computer Systems

Faculty of Computer Science and Telecommunications

University of Granada

18071 Granada Spain

Abstract

In this paper we study the impact of considering context information for the annotation of emotions. Concretely, we propose the inclusion of the history of user-system interaction and the neutral speaking style of users. A new method to automatically include both sources of information has been developed making use of novel techniques for acoustic normalization and dialogue context annotation. We have carried out experiments with a corpus extracted from real human interactions with a spoken dialogue system. Results show that the performance of non-expert human annotators and machine-learned classifications are both affected by contextual information. The proposed method allows the annotation of more non-neutral emotions and yields values closer to maximum agreement rates for non-expert human annotation. Moreover, automatic classification accuracy improves by 29.57% compared to the classical approach based only on acoustic features.

Key words: emotion annotation, emotion recognition, emotional speech, spoken dialogue system, dialogue context, acoustic context, affective computing.

1 Introduction

One of the main research objectives of dialogue systems is to achieve human-like communication between people and machines. This eliminates the need for keyboard and mouse in favour of more intuitive ways of interaction, such as natural language, thus leading to a new paradigm in which technologies can be accessed by non-expert users or handicapped people.

However, multimodal human-computer interaction is still not comparable to human dialogue. One of the reasons for this is that human interaction involves exchanging not only explicit content, but also implicit information about the affective state of the interlocutor. Systems that make use of such information are described as incorporating "affective computing" or "emotion intelligence", which covers the areas of emotion recognition, interpretation, management and generation.

Due to its benefits and huge variety of applications, affective computing has become an outstanding research topic in the field of HCI, and numerous important international and interdisciplinary related projects have appeared. Some of the latest are MEGA (Camurri et al., 2004), NECA (Gebhard et al., 2004), VICTEC (Hall et al., 2005), NICE (Corradini et al., 2005), HUMAINE (Camurri et al., 2005) and COMPANIONS (Wilks, 2006), to mention just a few.

Accurate annotation is a first step towards optimized detection and management of emotions, which is a very important task in order to avoid significant

* Corresponding author.
Email addresses: zoraida@ugr.es (Zoraida Callejas), rlopezc@ugr.es (Ramón López-Cózar).

problems in communication, such as misunderstandings and user dissatisfaction, that lead to very low task completion rates. Despite its benefits, the annotation of emotions in spoken dialogue systems encounters restrictions as a result of certain important problems. Firstly, as shown in Section 3.4, the percentage of neutral vs. emotive speech is usually very unbalanced (Forbes-Riley and Litman, 2004; Morrison et al., 2007). Secondly, all information must be gathered through the oral modality and in some systems where the dialogue is less flexible, the length of the user utterances can be insufficient to enable other knowledge sources like linguistic information to be employed.

To solve these problems, we propose to use contextual information for the annotation of user emotions in spoken dialogue systems. We are interested specifically in recognizing negative emotions as some studies, like for example (Riccardi and Hakkani-Tür, 2005), have shown that once the user is in a negative emotional state, it is difficult to guide him out. Furthermore, these bad experiences can also discourage users from employing the system again. Concretely, we take into account three negative emotions. The first is *doubtful*, which is useful to identify when the user is uncertain about what to do next. A user is in this emotional state when he has doubts about what to say in that turn. The second and third emotions are *angry* and *bored*, two negative emotional states that must be recognized before the user gets too frustrated because of system malfunctions. In the activation-evaluation space (Russell, 1980; Scherer, 2005), *angry* corresponds to an active negative emotion, whereas *bored* and *doubtful* to passive negative emotions.

Different approaches are presented in order to include contextual information in both human annotation (as discussed in Section 3) and machine learned classification (Section 4). In human annotation, non-expert annotators were

provided with contextual information by giving them the utterances to be annotated along with the dialogues where these were produced. In this way, annotators had information about the user speaking style and the moment of the dialogue at which each sentence was uttered. For machine-learned classification, we introduce a novel method in two steps which enhances negative emotion classification with automatically generated context information. The first step calculates users' neutral speaking style, which we use to classify emotions into *angry* and *doubtfulORbored*; whereas the second step introduces dialogue context and allows the distinction between *bored* and *doubtful* categories. One of the main advantages of the proposed method is that it does not require including additional manually annotated data. Hence, it permits a straightforward automatic integration of context information within emotion recognizers for spoken dialogue systems.

To evaluate the benefits of our proposals, different experiments have been performed over a corpus of real emotions extracted from the interaction of 60 different users with our spoken dialogue system (Section 3.1). Our objective is to demonstrate that the proposed contextual information influences human as well as machine recognition, and that better results can be obtained when context is included using the proposed methods, if compared to recognition based on traditional acoustic features or the baseline classification methods.

The paper is structured as follows: in Section 2 there is an overview of the related work done in the area and the points in which we make our main contributions. Section 3 presents the human annotation procedure and discusses the corpus facts and the results in terms of emotions annotated and agreement between annotators. In Section 4 there is a description of the automatic classification of emotions, and a discussion of the experimental results

obtained for it. Finally, in Section 5 we evaluate the benefits of the proposed methods, present conclusions extracted from them, and suggest some future work guidelines.

2 Related work

Emotional information has been used in Human Computer Interaction (HCI) systems for several purposes. In some application domains it is necessary to recognize the affective state of the user to adapt the systems to it or even change it. For example, in emergency services (Bickmore and Giorgino, 2004) or intelligent tutors (Ai et al., 2006), it is necessary to know the users' emotional state to calm them down, or to encourage them in learning activities. However, there are also some applications in which emotion management is not a central aspect, but contributes to the better functioning of the system as a whole. In these systems emotion management can be used to resolve stages of the dialogue that cause negative emotional states, as well as to avoid them and foster positive ones in future interactions. For example, Burkhardt et al. (2005) use an anger detector to avoid user frustration during the interaction with their voice portal. Furthermore, emotions are of interest not just for their own sake, but also because they affect the explicit message conveyed during the interaction: they change peoples' voices, facial expressions, gestures, speed of speech, etc. This is usually called "emotional colouring" and can be of great importance for the interpretation of user input. For example, Streit et al. (2006) use emotional colouring in the context of the SmartKom system to detect sarcasm and thus tackle false positive sentences.

In the area of emotion recognition the great majority of studies [1] focus on studying the appropriateness of different machine learning classifiers (Shafran and Mohri, 2005), such as K-nearest neighbours (Lee and Narayanan, 2005), Hidden Markov Models (Ververidis and Kotropoulos, 2006; Pitterman and Pitterman, 2006), Support Vector Machines (Morrison et al., 2007), Neural Networks (Morrison et al., 2007) or Boosting Algorithms (Liscombe et al., 2005; Forbes-Riley and Litman, 2004). In addition, important research has been directed towards finding the best features to be used for classification. These features can be categorized at different levels. The lowest level deals with physiological features, which are usually measured with intrusive methods. Some examples are galvanic skin response (Lee et al., 2005), facial muscle movements (Mahlke, 2006) or brain images (Critchley et al., 2005). Acoustic and linguistic levels are more widespread and features like articulation changes (Cowie et al., 2001), statistical measures of acoustic features (Ververidis and Kotropoulos, 2006) or word emotional salience (Lee and Narayanan, 2005) are frequently found in the literature. In addition, visual features like facial expression, body posture and movements of hands have recently been adopted, especially in multimodal systems (Picard and Daily, 2005; Zeng et al., 2006). More recently, some authors like Boehner et al. (2007) have proposed cultural information as an additional source of information for detecting emotional states.

However, less attention is being paid to the training process of the algorithms in which automatic emotion classification is based, and for which emotional annotated corpora are needed. A good annotation scheme is essential as it affects the rest of the stages in the learning process. Besides, manual an-

[1] See Ververidis and Kotropoulos (2006) for a good review

notation of corpora is very difficult, time-consuming and expensive, and thus must be carefully designed. Authors that study emotional corpora are mainly interested in how it is gathered, especially comparing acted vs. real emotions acquisition (Morrison et al., 2007), but less work has been done in how the annotation of such a corpus must be achieved. Among others, Devillers et al. (2005) have proposed guidelines to design and develop successful annotation schemes in terms of labels, segmentation rules and validation processes. Gut and Bayerl (2004) have also worked on reliability measures of human annotations, whereas Craggs and Wood (2003) have proposed several layers of emotion annotation.

In this paper we go a step further and study how to add contextual information to the corpus annotation process, and suggest the inclusion of two new context sources: users' neutral speaking style and dialogue history. The former provides information about how users talk when they are not conveying any emotion, which can lead to a better recognition of users' non-neutral emotional states (Section 4.2). The latter involves using information about the current dialogue state in terms of dialogue length and number of confirmations and repetitions (as we will discuss in Section 4.3), which gives a reliable indication of the users' emotional state at each moment. For example, the user is likely to be angry if he has to repeat the same piece of information in numerous consecutive turns in the dialogue.

In the literature we can find three main approaches for collecting emotional speech corpora: recording spontaneous emotional speech, recording induced emotions, and using actors to simulate the emotions. As shown in Figure 1, in these approaches there is a compromise between naturalness of the emotions and control over the collected data: the more control over the generated data,

the less spontaneity and naturalness of the expressed emotion, and vice versa. Therefore, spontaneous emotional speech, which reflects completely natural emotional speech production in the application domain of the emotion recognizer, is the most realistic approach. However, a lot of effort is necessary for the annotation of the corpus, as it requires an interpretation of which emotion is being expressed in each recording. Sometimes, the corpus is recorded from human-to-human interaction in the application context (Forbes-Riley and Litman, 2004). In these cases, the result is also natural but it is not directly applicable to the case in which humans interact with a machine. In the other extreme, acted emotional speech is easier to manipulate and avoids the need for annotation, as emotions conveyed in each recording are known beforehand. The results obtained with acted speech are highly dependent on the skills of the actors, therefore the best results are obtained with actors with good drama preparation. When non-expert actors are used, another phase is necessary to discard the recordings that fail to reproduce the required emotion appropriately. In a middle point are the induced emotions, which can be more natural, like the ones elicited when playing computer games (Johnstone, 1996), or easier to manipulate like the ones induced by making people read texts that relate to specific emotions (Stibbard, 2000).

As some authors have indicated, e.g. Douglas-Cowie et al. (2003), the relationship between acted data and spontaneous emotional speech is not exactly known. But, as stated by Johnstone (1996), even professionally acted speech loses realism as there are some effects that cannot be controlled consciously. Thus, different studies have shown that it is not appropriate to use acted data to recognize naturally occurring emotions (Vogt and André, 2005; Wilting et al., 2006).

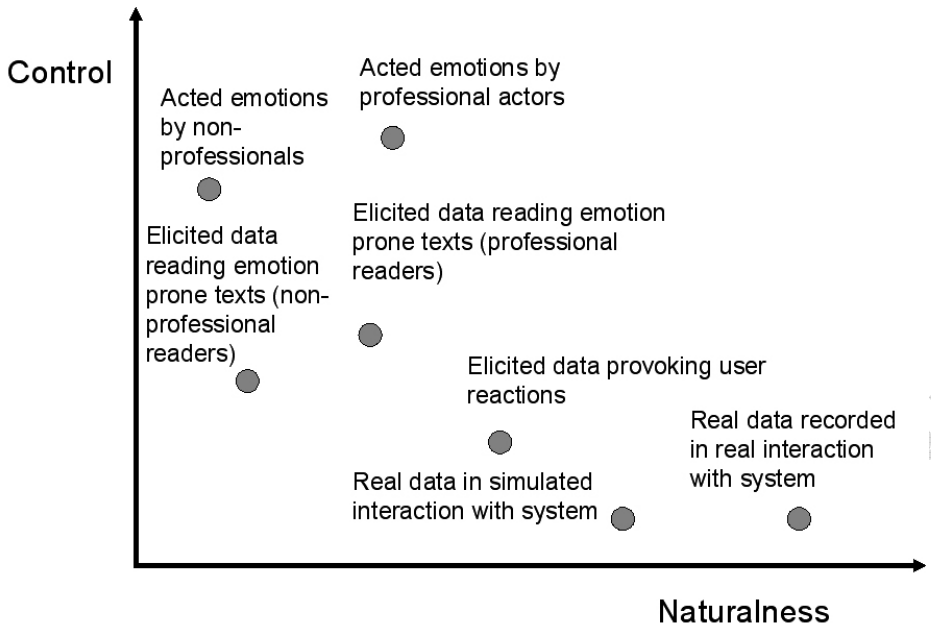


Fig. 1. Naturalness vs. control in the main emotional corpora generation approaches

As our objective is to build an emotion recognizer for the UAH (Universidad Al Habla - University on the Line) dialogue system (Section 3.1), and it would have to work with natural emotions occurring in real time, we used for the experiments an utterance corpus collected from real users interacting with the system. This is an important contribution to the state of the art as real non-elicited emotions are difficult to find in Spanish corpora. For example, out of the 70 corpora studied by Douglas-Cowie et al. (2003) and Ververidis and Kotropoulos (2006), only three are in Spanish: González (1999), Montero et al. (1999) and Iriondo et al. (2000). As shown in Table 1, two of these are used for emotion synthesis instead of recognition. The table also sets out Spanish corpora employed for emotion feature studies (Adell et al., 2005) and general purpose studies (Hozjan et al., 2002). None of these corpora were collected from real user interactions and the maximum number of actors used was 8, whereas our corpus was collected from 60 users and real emotional speech.

Reference	Subjects	Purpose	Kind
González (1999)	-	Recognition	Elicited
Montero et al. (1999)	1 actor	Synthesis	Simulated
Iriondo et al. (2000)	8 actors	Synthesis	Simulated
Hozjan et al. (2002)	2 actors	Study, synthesis and recognition	Simulated
Adell et al. (2005)	1 actress, 1 professional reader, 1 member of Spanish Parliament	Emotion features study	2 simulated, 1 natural
UAH corpus	60 UAH system users	Recognition	Natural

Table 1
Spanish emotional speech corpora

3 Human annotation of the UAH corpus

3.1 The UAH emotional speech corpus

The UAH dialogue system was developed in our laboratory to provide telephone-based spoken access to the information available in our Department's web page (Callejas and López-Cózar, 2005). The corpus used for the experiments described in this paper is comprised of 85 dialogues of 60 different users interacting with UAH. It contains 422 user turns, with an average of 5 user turns per dialogue. The recorded material has a duration of 150 minutes. The size of the corpus is similar to that of other real emotional speech corpora like those used by Forbes-Riley and Litman (2004) (10 dialogues, 453 turns) or Morrison et al. (2007) (391 user turns). For illustration purposes, we show below a sample dialogue from our corpus (translated from Spanish to English), where S = system, U = user:

S1> Welcome to the UAH system.

U1> I want information about a subject.

S2> Please say the name of the subject.

U2> Compilers.

S3> Compilers is a subject in third year of the Computer Science degree. What information do you need about it?

U3> The timetable for the classes.

S4> Compilers is imparted on Mondays, Tuesdays and Thursdays from 10 to 11 and on Fridays from 11 to 12. Do you need any other information about the subject?

U4> No, thanks.

S5> Thank you for using the UAH system. Have a nice day.

3.2 The UAH corpus human annotation procedure

The annotation of emotions is a highly subjective task, given that for the same utterance, different annotators may perceive different emotions. The most reliable way to obtain rigorous annotations is to recruit specialized annotators, for example psychologists who are trained to recognize human emotions. Unfortunately, in most cases expert annotators are difficult to find and thus the annotation must be done by non-experts. In our case, all annotators were non-expert as they had not received any specific training on emotion recognition.

To get the best possible annotation employing non-expert annotators, the labelling process must be rigorously designed. Vidrascu and Devillers (2005) suggest several phases to decide the list of labels and annotation scheme, seg-

mentation rules, number of annotators, validation procedures, and consistency study.

The first step is to decide the labels to be used for annotation. Our main interest is to study negative emotional states of the users, mainly to detect frustration because of system malfunctions. In this paper, we discern between the three major negative emotions encountered in the UAH corpus, namely *angry*, *bored* and *doubtful*. For the human annotation of the corpus we have also used a fourth category: *neutral*, which represents a non-negative emotional state (i.e. positive emotions such as happiness are also treated as *neutral*). The neutral category was used only for the human annotation of the corpus. For the rest of the experiments we will focus exclusively on the distinction between the negative emotions considered.

We decided to use an odd, high number of annotators - nine, which is more than is typically reported in previous studies, e.g. Forbes-Riley and Litman (2004) and Lee and Narayanan (2005). Regarding the "segment length", in our study this is the whole utterance because it was not useful to employ smaller segmentation units (i.e. words). The reason is that our goal was to analyze the emotion as a whole response to a system prompt, without considering the possible emotional changes within an utterance.

In our annotation procedure the corpus was annotated twice by every annotator, firstly in an ordered style and secondly in an unordered style. In the first mode the annotators had information about the dialogue context and the users' speaking style. In the second case the annotators did not have this information, so their annotations were based only on the acoustic information of the current utterance.

The final emotion assigned to each utterance in the ordered and unordered schemes was the one annotated by a majority of annotators in each of them. Gold standard emotions for the whole corpus were then computed from the results of each of the schemes. In situations where there was no majority of an emotion above the others (e.g. 4 *neutral*, 4 *bored* and 1 *doubtful*), priority was given to the non-neutral ones (in the last example *bored*). If this conflict was between two non-neutral emotions (e.g. 4 *doubtful*, 4 *bored* and 1 *neutral*), the results were compared between both annotation schemes to choose the emotion annotated by majority among the 18 annotations (the 9 of the ordered and the 9 of the unordered scheme).

3.3 Calculation of the level of agreement between annotators

Four Kappa coefficients [2] were used to study the degree of inter-annotator agreement for the ordered and unordered case (Figure 2).

With these coefficients, we studied two main issues: i) the impact of annotator bias, that is, given a fixed number of agreements, the effect that the distribution of disagreements between categories has in the Kappa value; and ii) the level of importance of all possible disagreements in our task, i.e. disagreements between emotions which are easily distinguishable should have a more negative impact in the Kappa coefficient than disagreements in more

[2] Kappa coefficients are based on the idea of rating the proportion of pairs of annotators in agreement (P_o) with the expected proportion of pairs of annotators that agree by chance (P_c). Thus obtaining a proportion of the agreement actually achieved beyond chance ($P_o - P_c$) with all possible agreements that are not by chance ($1 - P_c$):

$$\kappa = \frac{P_o - P_c}{1 - P_c} \quad (1)$$

similar categories.

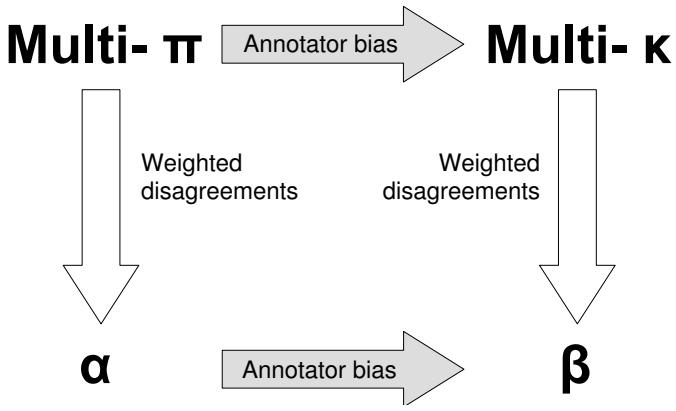


Fig. 2. Kappa coefficients used in the experiments

In order to avoid inconsistencies we follow Artstein and Poesio (2005) notation for all the Kappa coefficients employed. The simplest Kappa coefficient used was proposed by Fleiss (1971), as a generalization for multiple annotators of the two-coders Scott's π (Scott, 1955). We have noted Fleiss' Kappa as multi-π.

The calculation of multi-π assumes that each annotator follows the same overall distribution of utterances into emotions, this means that the probability that an annotator classifies an utterance 'u' with a particular emotion 'e', can be computed as the overall probability of annotating 'u' as 'e'. However, such a simplification may not be plausible in all domains due to the effect of the so-called *annotator bias* in the Kappa value. In our experiments, the annotator bias can be defined as the extent to which annotators disagree on the proportion of emotions, given a particular number of agreements. With the rest of the parameters fixed, the Kappa value increases as the bias value gets higher, that is, when disagreement proportions are not equal for all emotions and there is a high skew among them. This is the so-called *Kappa second paradox*. Different studies of its impact can be found in the literature, e.g. Feinstein and

Cicchetti (1990), Cicchetti and Feinstein (1990), Lantz and Nebenzahl (1996), and Artstein and Poesio (2005).

To study whether inclusion of the different annotating behaviours could improve the Kappa values, we calculated Davies and Fleiss (1982) Kappa, which we have noted as multi-κ, following the study of Artstein and Poesio (2005). Multi-κ includes a separate distribution for each annotator. Thus, in this case the probability that an annotator 'a' classifies an utterance 'u' with emotion 'e' is computed with the observed number of utterances assigned to emotion 'e' by that annotator, divided by the total number of utterances.

Despite accounting for differences between annotators, multi-κ gives all disagreements the same importance. In practice, all disagreements are not equally probable and do not have the same impact on the quality of the annotation results. For example, in our experiments, a disagreement between *neutral* and *angry* is stronger than between *neutral* and *doubtful*, because the first two categories are more easily distinguishable.

To take all this information into account we have used weighted Kappa coefficients (Cohen, 1968; Fleiss and Cohen, 1973), which put the emphasis on disagreements instead of agreements [3]. The computation of the weighted coefficients implies employing distance metrics between the four emotions used for annotation (*neutral*, *angry*, *bored* and *doubtful*). To do so, we have arranged

[3] Their calculation is based on Equation 2 (equivalent to Equation 1):

$$\kappa_w = 1 - \frac{\overline{P}_o}{\overline{P}_c} \tag{2}$$

where $\overline{P}_o$ indicates observed disagreement, and $\overline{P}_c$ disagreement by chance. For all the coefficients used, the observed disagreement has been calculated as the number of times each utterance 'u' was annotated with two different emotions by every pair of annotators, weighted by the distance between the emotions.

our discrete list of emotions within a continuous space, using the bidimensional activation-evaluation space (Russell, 1980). In its horizontal axis, evaluation deals with the "valence" of emotions, that is, positive or negative evaluations of people, things or events. In the vertical axis, activation measures the user disposition to take some action rather than none. Emotions form a circular pattern in this space. This is why other authors proposed a representation based on angles and distance to the centre. For example, some tools like FEELTRACE (Cowie et al., 2000) have been implemented to give a visual representation of the dynamic progress of emotions inside this circle.

Taking advantage of this circular disposition, we have used angular distances between our emotions for the calculation of the weighted Kappa coefficients. Instead of establishing our own placement of the emotions in the space, we employed some already established angular disposition to avoid introducing measurement errors. With this purpose, we used the list of 40 emotions with their respective angles proposed by Plutchik (1980), which has been widely accepted and used by the scientific community. In this list, *bored* (136.0º) and *angry* (212.0º) were explicitly contemplated, but this was not the case for *doubtful*. The most similar emotions found were "uncertain", "bewildered" and "confused", which only differentiated in 2º in the circle. We chose "uncertain" (139.3º) which was the one that better reflected the emotion we wanted to annotate. However, other authors like Scherer (2005) have explicitly considered *doubtful* as an emotional state. Plutchik (1980) did not reflect neutral in his list as it really is not an emotion but the absence of emotion. Instead, he used a state called "accepting" as the starting point of the circle (0º), which we used as *neutral* in our experiments.

With the angle that each of the four emotions forms in the circle we cal-

culated the distance between them in degrees. We chose always the smallest angle between the emotions being considered (x or 360-x). This way, the distance between every two angles was always between 0 and 180 degrees. For the calculation of the Kappa coefficients, distances were converted into weights with values between 0 and 1. A 0 weight (which corresponds to 0º distance in our approach) implies annotating the same emotion, and thus having no disagreement. On the contrary, weight=1 (180º distance) corresponds to completely opposite annotations and thus maximum disagreement. The resulting distances and weights are listed in Table 2.

Angle/ Weight	Neutral	Angry	Bored	Doubtful
Neutral	0.00º / 0.00	148.00º / 0.82	136.00º / 0.75	139.30º / 0.77
Angry	148.00º / 0.82	0.00º / 0.00	76.00º / 0.42	72.70º / 0.40
Bored	136.00º / 0.75	76.00º / 0.42	0.00º / 0.00	3.30º / 0.02
Doubtful	139.30º / 0.77	72.70º / 0.40	3.30º / 0.02	0º / 0.00

Table 2
Distance between emotions

There is not a consensus in the scientific community about the properties of the distance measures. However, Artstein and Poesio (2005) have proposed some constraints: the distance between a category and itself should be minimal and the distance between two categories should not depend on the order (i.e. distance from A to B should be equal to distance from B to A). As can be observed by the symmetry of the table, our distance measures and weights follow these restrictions.

As can be observed in the table, the highest distances were between non-neutrals and neutral. Thus, when calculating weighted Kappa coefficients, disagreements in which an annotator judged an utterance as neutral and the

other as non-neutral were given more importance than for example an *angry* vs. *bored* disagreement.

We calculated two weighted Kappa coefficients: Krippendorff's α (Krippendorff, 2003), and the β coefficient proposed by Artstein and Poesio (2005). Both share the same calculation for the observed agreement, but whereas α [4] does not take into account annotator bias, they are considered in the calculation of agreement by chance in the β coefficient.

The results for each described coefficient are listed in Table 3 and discussed in the next section.

Coefficient	Unordered	Ordered
multi-π	0.3256	0.3241
multi-κ	0.3355	0.3256
α	0.3382	0.3220
β	0.3393	0.3237

Table 3
Values of the Kappa coefficients for unordered and ordered annotation schemes

3.4 Discussion of human annotation results

As previously commented, one of the difficulties of emotion recognition in spoken dialogue systems is that in most application domains the corpora obtained are very unbalanced, because there is usually a higher proportion of neutral than emotional utterances (Forbes-Riley and Litman, 2004; Morrison et al., 2007). This is in accordance with our experimental results since, on average among the nine annotators, more than 85.00% of utterances were an-

[4] Results obtained for the alternative formulation of α presented in (Artstein and Poesio, 2005) where very similar to Krippendorff's α (see Table 3): 0.3381 in the unordered, and 0.3218 in the ordered case.

notated as *neutral*. We have also observed that this proportion is affected in 3.40% of the cases by the annotation style. Concretely, for the ordered annotation, 87.28% were tagged as *neutral*, whereas for the unordered annotation the corpus was even more unbalanced: 90.68% of the utterances were annotated as *neutral*. Figure 3 shows the proportion of non-neutral emotions tagged by the 9 annotators. As can be observed, the ordered annotation style yielded a greater percentage for the *bored* category: 39.00% more than in the unordered style. The figure also shows that the *angry* category is substantially affected by the annotation style (i.e. ordered vs. unordered): 70.58% more *angry* annotations were found in the ordered annotation style. On the contrary, the *doubtful* category is virtually independent of the annotation style: only 2.75% more doubts were found in the unordered annotation.

A plausible reason for these results is that the incorporation of context in the ordered case influences the annotators in assigning the utterances belonging to the same dialogues into the same emotional categories. This way, there are no very noticeable transitions between consecutive utterances. For example, if anger is detected in one utterance then the next one is probably also annotated as *angry*. Besides, the context allows the annotators to have information about user's speaking style and the interaction history. In contrast, in the unordered case the annotators only have information about the current utterance. Hence, sometimes they cannot tell whether the user is either angry or he normally speaks loudly and fast.

Thus, it is an important fact to be taken into account when annotation is carried out by non-expert annotators, which is the most common, cheapest and least time consuming method. In addition, when listening to the corpus in order, the annotators had information about the position of the current user

turn within the whole dialogue, which also gives a reliable clue to the user's state. For example, a user is more likely to get bored after a long dialogue, or to become angry after many confirmation prompts generated by the system.

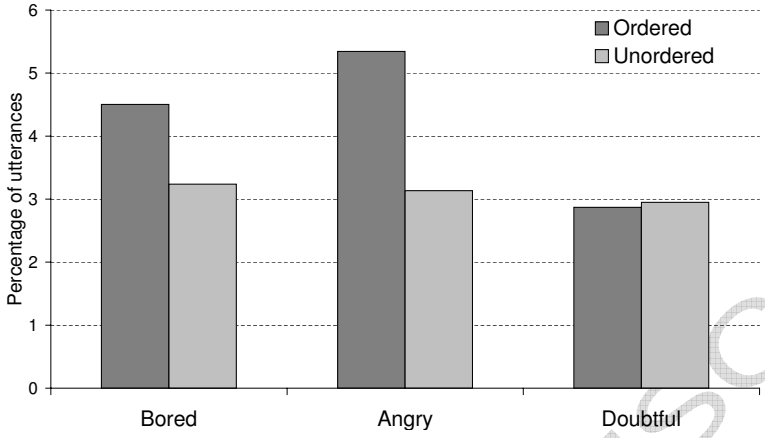


Fig. 3. Proportion of non-neutral annotated utterances

As can be observed in Table 3, the values of the different Kappa coefficients also vary slightly depending on the annotating scheme used. In the unordered case, both taking into account annotator bias (multi-κ vs. multi-π, and β vs. α), and weighting disagreements (β and α vs. multi-κ) improves the agreement values. However, in the ordered case only taking into account annotator bias enhances the agreement values, whereas weighting the disagreements reduces Kappa. This is a consequence of the increment of non-neutral annotations already discussed. Taking into account that the great majority of agreements occur when annotators tag the same utterance as neutral (as can be observed in Figure 4), an increment in the number of emotions annotated as non-neutral provokes more discrepancies among the annotators and thus reduces the Kappa value.

Furthermore, as can be observed in Figure 5 most of the disagreements occur between neutral and non-neutral categories, which are the emotions with higher distances according to our weighting scheme (Table 2), thus provoking

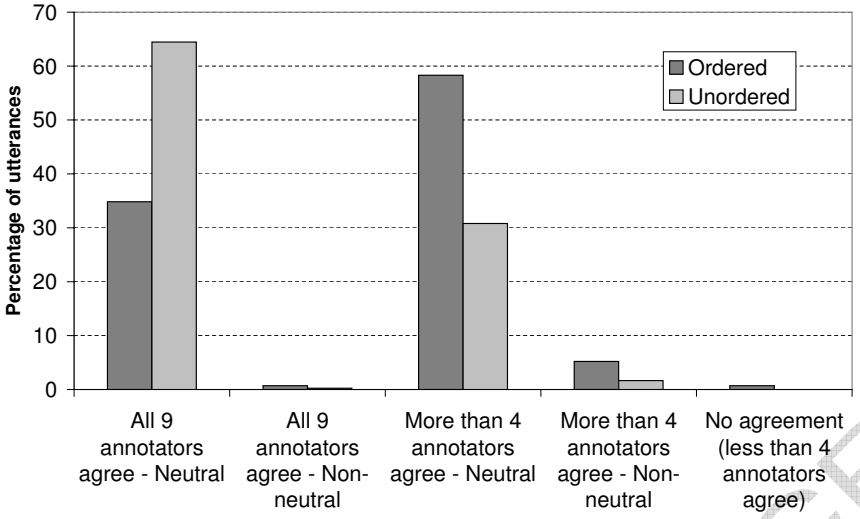


Fig. 4. Percentage of utterances in which annotators agree

weighted agreements to be lower in the case of the ordered scheme.

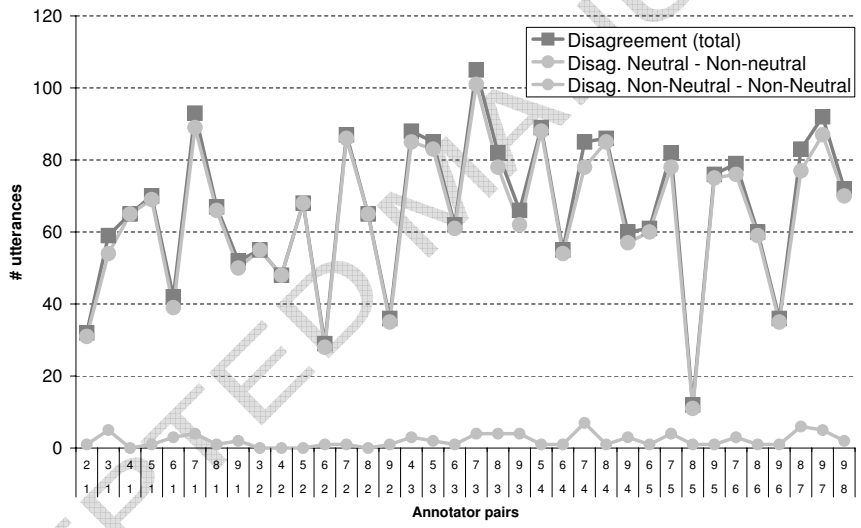


Fig. 5. Pair wise disagreement between annotators following the ordered scheme

To study the effect of annotator bias, we measured pair wise agreement between all annotators. As can be observed in Figure 5 there were no annotators who had a significantly poor agreement with the rest. However, when we examined the annotation results, we found that there were remarkable differences between those annotators who were used to the Andalusian dialect [5] (in

[5] Spanish spoken in Southern Spain.

which the utterances were pronounced) and those who were not so accustomed. As explained in Section 3.1, the corpus was recorded from user interactions with the UAH system. The users were mainly students and professors at the University of Granada, which is in south eastern Spain. The way these users express themselves is influenced by the Eastern Andalusian dialect (Gerfen, 2002; O'Neill, 2005), which although similar to Spanish Castilian has several differences like the aspiration of the final 's' in words, a relaxed pronunciation of 'd' in some terminations (like -ado or -ido), faster rhythm and lower expiratory strength. In our group of annotators, 6 were used to the Andalusian dialect (annotators 1, 2, 3, 4, 6 and 9 in Figure 5) and 3 were not (annotators 5, 7 and 8 in the figure).

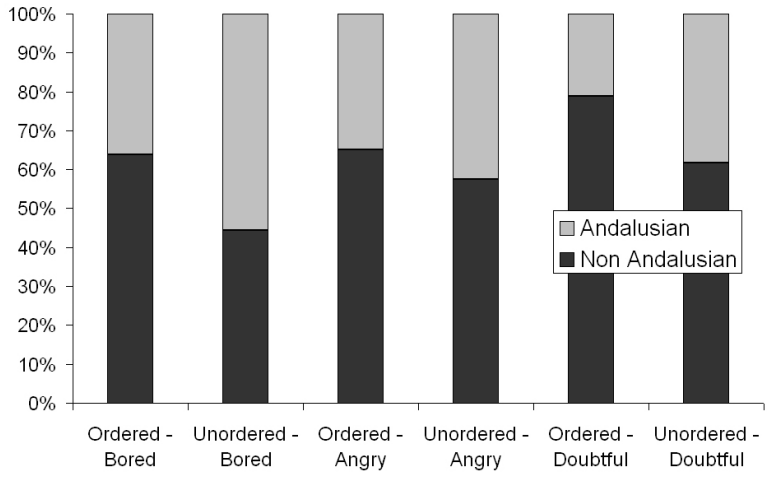


Fig. 6. Proportion of annotated emotions depending on dialect

Figure 6 shows, for the total number of annotations made in each category, which percentage corresponds to each type of annotators. As can be observed, in all the cases but one (especially in those obtained employing the ordered scheme), the annotators not used to the Andalusian dialect marked around 50% of the emotions encountered for the emotional category. This is caused by the confusion of characteristics of the dialect with emotional cues, for example

confusing the Andalusian fast rhythm with an indication of anger. We studied the effect on the annotation schemes for both kinds of annotators and obtained the results shown in Table 4.

	Andalusian annotators		Non-andalusian annotators	
	Unordered	Ordered	Unordered	Ordered
multi-π	0.3608	0.3234	0.3734	0.5593
multi-κ	0.3621	0.3275	0.3746	0.5598
α	0.3595	0.3248	0.3644	0.5691
β	0.3607	0.3265	0.3703	0.5697

Table 4
Kappa values for the different annotator types

As can be observed in the table, the annotators used to the Andalusian dialect obtained Kappa values for both annotation schemes which were more similar (ranging between 0.3234 to 0.3621). For these annotators, the Kappa values were smaller for the ordered scheme because there were fewer utterances annotated as neutral.

On the contrary, annotators not used to the Andalusian dialect had very different Kappa values depending on the annotating scheme used: in the ordered case values ranged from 0.5593 to 0.5697 whereas in the unordered these were between 0.3639 and 0.3746. This is due to a big decrement of the chance agreement. As shown in Figure 7, the observed agreement was more or less constant, whereas the chance agreement drastically decreased in the ordered scheme.

The most likely reason for this is the decrement in the number of neutrals annotated by annotators not used to Andalusian. This happens for both annotation schemes, but the number of neutrals annotated is higher in the unordered one, and that is why results are more similar to those obtained by An-

dalusian annotators with the unordered annotation scheme. Even though the number of non-neutral annotations increased proportionally with the decrement of neutrals, the unbalancement of the corpus made the probability of agreeing by chance in the neutral emotion more important in the computation of the overall agreement by chance. For example, in the case of multi-κ, agreement by chance (P_c) was calculated as the sum of agreeing by chance in each emotion ($P_c = P_c^{neutral} + P_c^{bored} + P_c^{angry} + P_c^{doubtful}$). The values for agreeing by chance when annotators not used to Andalusian used the ordered scheme were $P_c^{neutral} = 0.6645$, $P_c^{bored} = 0.0052$, $P_c^{angry} = 0.0069$ and $P_c^{doubtful} = 0.0008$. For the rest of annotators these values were: $P_c^{neutral} = 0.8137$, $P_c^{bored} = 0.0010$, $P_c^{angry} = 0.0014$ and $P_c^{doubtful} = 0.0008$. Thus, $P_c^{neutral}$ was the determining factor in obtaining the global P_c.

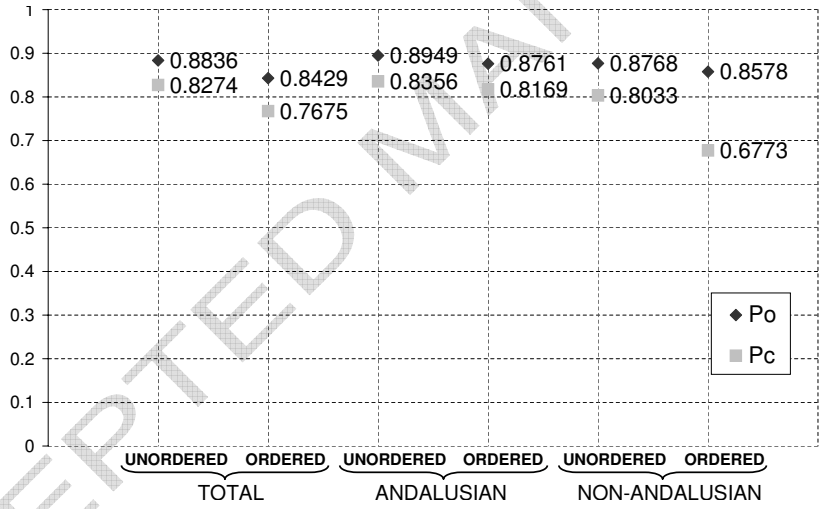


Fig. 7. Relative values of chance and observed agreements for multi-κ

The situation in which although having an almost identical number of agreements, the distribution of these across the different annotation categories deeply affects Kappa, is typically known as the *first Kappa paradox*. This phenomenon establishes that other things being equal, Kappa increases with more symmetrical distributions of agreement. That is, if the prevalence of a category compared to the others is very high, then the agreement by chance

(P_c) is also high and the Kappa is considerably decremented (Feinstein and Cicchetti, 1990; Cicchetti and Feinstein, 1990).

As already reported by other authors, e.g. Feinstein and Cicchetti (1990), the first Kappa paradox can drastically affect Kappa values and thus must be considered in its interpretation. There is not an unique and generally accepted interpretation of the Kappa values. One of the most widely used is the one presented by Landis and Koch (1977), which makes a correspondence between intervals for Kappa values and interpretations of agreement. Following this approach, our experimental results indicate fair agreement for both annotating schemes and with the four different Kappa coefficients. Alternatively, Krippendorff (2003) established 0.65 as a threshold for acceptability of agreement results. Hence, considering this value our 0.3393 highest Kappa would not be acceptable. However, most authors seem to agree in that using a fixed benchmark of Kappa intervals does not provide enough information to make a justified interpretation of acceptability of the agreement results. In order to provide a more complete framework, some authors like Dunn (1989), propose to place Kappa into perspective by reporting *maximum*, *minimum* and *normal* values of Kappa which can be calculated from the observed agreement (P_o) as follows (Lantz and Nebenzahl, 1996):

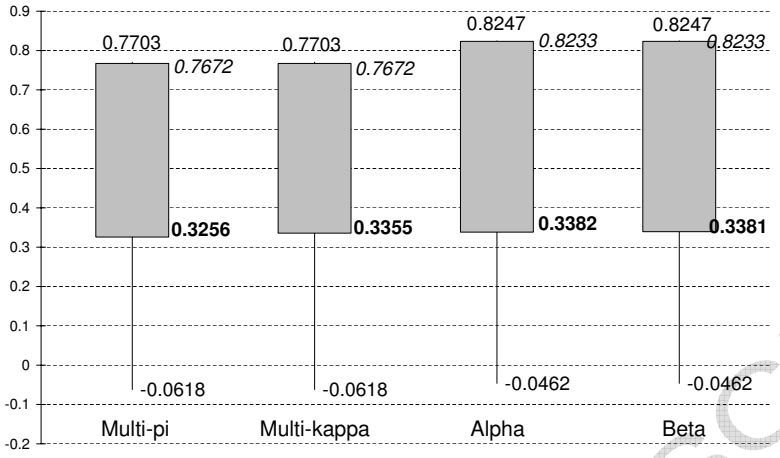
$$kappa_{max} = \frac{P_o^2}{(1 - P_o)^2 + 1} \tag{3}$$

$$kappa_{min} = \frac{P_o - 1}{P_o + 1} \tag{4}$$

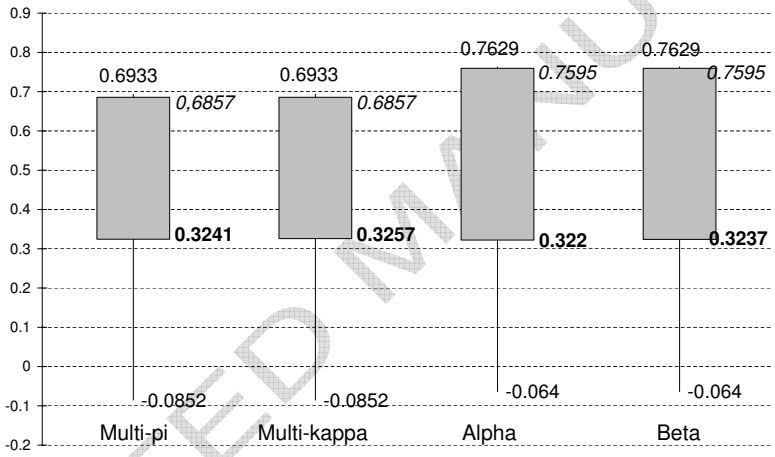
$$kappa_{nor} = 2P_o - 1 \tag{5}$$

We compared the obtained Kappa values (Table 3) with their $kappa_{max}$, $kappa_{min}$ and $kappa_{nor}$ values and obtained the results shown in Figure 8,

where *normal* values are marked in italics and the actual values obtained in bold.



(a) Unordered scheme



(b) Ordered scheme

Fig. 8. Kappa *maximum*, *minimum*, *normal* (italics) and observed (bold) values

As can be observed in the figure, for the same observed agreement, the possible values of Kappa can deeply vary from $kappa_{min}$ to $kappa_{max}$ depending on the balancement of the corpus. $Kappa_{max}$ is obtained when maximally skewing disagreements while maintaining balanced agreements, whereas $kappa_{min}$ is obtained when agreements are skewed and disagreements balanced. $Kappa_{nor}$ does not correspond to an ideal value of Kappa, but rather to symmetrical distributions of both agreements and disagreements. It can be observed in the

figure how the displacement between actual and normal values is smaller in the ordered scheme (Figure 8(b)). Thus, this scheme does not only allow recognizing more non-neutral emotions, but also obtaining Kappa values which, although smaller than in the unordered scheme in absolute value, are much closer to the *normal* and *maximum* agreement values attainable and further from the *minimum*.

As stated in (Lantz and Nebenzahl, 1996), departures from the $kappa_{nor}$ value indicate asymmetry in agreements or disagreements depending in if they are closer to the minimum or maximum value respectively. In Figure 8, the shift between the observed and the normal Kappa values is represented with a grey box. The results corroborate that presenting Kappa values is more informative when they are put into context, as we obtain a valuable indicative of possible unbalancements that has to be considered to reach appropriate conclusions about reliability of the annotations. For example, in our case there were significant departures from $kappa_{nor}$ in all cases (grey boxes), which corroborates that there was a big asymmetry in the categories. This is due to the prevalence phenomena previously discussed (first Kappa paradox).

As discussed before, prevalence appeared as an unavoidable consequence of the natural unbalancement of non-acted emotional corpora, where the neutral category is clearly predominant. Thus, approaches based uniquely on already established values of acceptability (Landis and Koch, 1977; Krippendorff, 2003) are not suitable for our application domain. Some authors have already reported additional measures to complement the information provided with the Kappa coefficients. For example, Forbes-Riley and Litman (2004) report on both observed agreement and Kappa, whereas Lee and Narayanan (2005) report on Kappa along with an hypothesis test.

Although reported Kappa values in emotion recognition employing unbalanced corpora are usually low, e.g. from 0.32 to 0.42 in Shafran et al. (2003) and below 0.48 in Ang et al. (2002) and Lee and Narayanan (2005), there is not a deep discussion about the problematic of Kappa values in the area, not even in papers explicitly devoted to challenges in emotion annotation, e.g. Devillers et al. (2005). Furthermore, even when other agreement measures are reported along with Kappa, e.g. Forbes-Riley and Litman (2004) and Lee and Narayanan (2005), there is only one Kappa coefficient calculated (usually multi-π) and no discussion about why there is such a big difference between the Kappa values and the other measures reported. Thus, we believe that our study may be one of the first reports about different Kappa values and the issues related to their use and interpretation in annotation of real emotions.

Finally, to obtain a more approximate idea about the real level of agreement reached by the nine annotators, we report the values of observed agreement in Table 5, which has been used along with Kappa by other authors in different areas of study (Ang et al., 2002; Forbes-Riley and Litman, 2004). As can be observed in the table, in all the cases the observed agreement was above 0.85. This measure does not contemplate the effect of prevalence (see Figure 7), and thus values were not higher for the annotators not used to the Andalusian dialect in the ordered case.

From all the previous results we can conclude that employing the ordered scheme allowed the annotation of more non-neutral emotions. Unfortunately, this translates in lower Kappa coefficients as most of the agreements occur for neutrals. These low Kappas indicate that multiple annotators should be used for annotating natural emotions to obtain reliable emotional corpora. One possible way to overcome the problem of high chance agreements, consists in

		Observed agreement	Weighted observed agreement
Unordered	Total	0.8836	0.9117
	Andalusian	0.8950	0.9197
	Non-andalusian	0.8767	0.9050
Ordered	Total	0.8429	0.8800
	Andalusian	0.8761	0.9049
	Non-andalusian	0.8578	0.895

Table 5
Observed agreement for all annotation schemes and annotator types

maximizing the observed agreement. For example, Litman and Forbes-Riley (2006) propose the usage of "consensus labelling", i.e. to reach a consensus between annotators until a 100% observed agreement is obtained.

In our case, the computed Kappa values were useful to compare the two annotation schemes. As shown in Figure 8, although the Kappa value and observed agreement percentages are lower in the ordered case, we found that it can be useful to obtain results which are closer to the maximum achievable. We can also deduce from our study that evaluating the reliability of an emotion annotation process where agreements are so highly skewed, can lead to very low Kappas (Table 3) that are far from the high agreement values observed (Table 5). Hence, it was necessary to include other sources of information like observed agreement and minimal, maximal and normal values along with the values obtained for the different Kappa coefficients in order to make meaningful interpretations. Besides, as shown in Table 5, giving a weight to the different disagreement types can considerably increment the observed agreement between annotators. We have presented a method to compute distances between such disagreements.

4 Automatic classification of the UAH corpus

As shown by the experimental results described in Section 3, contextual information is very important for human annotators. Therefore, in this section we examine whether discrimination between emotions in machine-learned classification is also affected by this factor. We are specifically interested in distinguishing between emotions and thus we only use non-neutral emotions as input for the learning algorithms. The reason for doing this is not to reduce the effect of the corpus unbalancement, but to carry out a deeper study on the differences between the negative emotions considered. Thus, our main future work guideline will be to add a neutral vs. non-neutral emotion recognizer to build an emotion recognizer that copes with this natural skewness (Section 5). The experiments in this section can be classified into two types: speech-related and dialogue-related. For the first group we have applied machine learning to distinguish the three emotions of interest (*angry*, *bored* and *doubtful*) and have measured the benefits of using our novel approach for acoustic normalization to improve classification. For the second group we have considered in addition knowledge about the context of the interaction. The number of non-neutral emotions reached after the complete annotation procedure (Section 3.2) was 63, of which 16% were *bored*, 32% *doubtful* and 52% *angry*.

For comparison purposes, the first approach used is a baseline that always annotates user utterances with the same emotion regardless of the input. In our corpus the most frequent emotion category is *angry*, therefore the baseline annotated each utterance with this label. The second algorithm used is a Multilayer Perceptron (MLP) (Rumelhart et al., 1986; Bishop, 2006). We used a topology with a hidden layer with $\frac{number\ of\ features\ +\ number\ of\ emotions}{2}$ nodes.

The learning rate, which determines the speed of the search convergence, was set at 0.3 to prevent it being too large (in which case it might miss minimums or oscillate abruptly) or too small (thus provoking slow convergence). To prevent the MLP from over fitting, the passes through the data (epochs) were restricted to 500. In addition, a validation threshold of 0.2 was set to determine the consecutive times that the validation set error could deteriorate before the training was stopped. To improve the performance, we also introduced a momentum of 0.2 for the learning of the weights that configured the MLP. To train the MLPs and carry out the experimentation we employed the WEKA toolkit (Witten and Frank, 2005), an open source collection of machine learning algorithms for data pre-processing, classification, regression, clustering and visualization.

For training and testing our learning algorithms we have used a 5-fold cross-validation technique. Therefore, the experiments consisted of five trials where the corpus was randomly split into five approximately equal subsets (20% of the corpus each). Every trial used each of the partitions in turn for testing, and the remainder (80% of the corpus) for training, so that after the 5 trials every instance had been used exactly once for testing. Additionally, a tuning partition (20%) was extracted from each training partition in order to make the feature selection. Thus, the evaluation was carried using two phases. In the first one, the learning algorithms were trained with the 60% of training utterances and evaluated with the 20% employed for tuning. In the second phase, the complete training partition was used for training the MLP, and the test part (20%) for evaluation. For comparison purposes, this second step was carried out, on the one hand, employing all the features of the 80% training utterances, whilst on the other, employing only the features selected in the

first step.

Finally, for all the experiments described in this section, the significance of the results was checked using the corrected paired t-tester available in the analysis tool of the Weka 3.5.4 Experimenter [6] (Witten and Frank, 2005), with 0.05 significance.

4.1 Automatic classification based on standard acoustic features

For these experiments we used 60 features that incorporate those typically used in previous studies (Devillers et al., 2005; Lee and Narayanan, 2005; Morrison et al., 2007). These are utterance-level statistics corresponding to the four groups set out in Table 6.

The first group is comprised of pitch features. Pitch depends on the tension of the vocal folds and the sub glottal air pressure (Ververidis and Kotropoulos, 2006), and can be used to obtain information about emotions in speech. As noted by Hansen (1996), mean pitch values may be employed as significant indicators for emotional speech when compared to neutral conditions. We have computed all the pitch features in the voiced portion of speech. Specifically, we focused on the minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation coefficient, slope, and error of the linear regression that describes the line that fits the pitch contour. All the duration parameters (e.g. slope)

[6] t-statistic is calculated in Weka as:

$$t = \frac{\overline{d}}{\sqrt{\left(\frac{1}{k} + \frac{n_2}{n_1}\right)\sigma_d^2}}$$

In our case, with 5-fold cross-validation repeated 5 times: $k = 25$, $\frac{n_2}{n_1} = \frac{0.2}{0.8}$ and σ_d^2 is the variance on 25 differences.

Category	Features
Fundamental frequency (F0)	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
F1, F2, B1, B2	Min, max, range, mean, median value at first voiced segment, value at last voiced segment
Energy	Min, max, range, mean, median, standard deviation, slope, correlation coef., regression error, value at first voiced segment, value at last voiced segment
Rhythm	Rate, voiced duration, unvoiced duration, value at first voiced, number of unvoiced segments

Table 6
Acoustic features used for classification

were normalized by the utterance duration to obtain comparable results for all the utterances in the corpus. To extract the pitch we have used the modified autocorrelation algorithm (Boersma, 1993).

The second group is comprised of features related to the first two formant frequencies (F1 and F2) and their bandwidths (B1 and B2). We have used only the first two formants because it has been empirically demonstrated that adding information about a third frequency does not introduce any informative features, neither in real nor in acted emotional corpora (Morrison et al., 2007). These frequencies are a representation of the vocal tract resonances. Speakers change the configuration of the vocal tract to distinguish the phonemes that they wish to utter, thus resulting in shifts of formant frequencies. Different speaking styles produce variations of the typical positions of formants. In the particular case of emotional speech, the vocal tract is modified by the emotional state. As pointed out by Hansen (1996), in stressed or depressed

states speakers do not articulate voiced sounds with the same effort as in neutral emotional states. The features that we have used for categories F1, F2, B1 and B2 are minimum value, maximum value, range, mean, median, standard deviation and value in the first and last voiced segments of each utterance.

Energy is considered in the third group of features. As stated by Ververidis and Kotropoulos (2006), this feature can be exploited for emotion recognition because it is related to the arousal level of emotions. The variation of energy of words or utterances can be used as a significant indicator for various speech styles, as the vocal effort and ratio (duration) of voiced/unvoiced parts of speech change. For example, Hansen (1996) demonstrated that loud and angry emotions significantly increase intensity, i.e. energy. For these features, we have used non-zero values of energy only, similarly as for pitch, obtaining minimum value, maximum value, mean, median, standard deviation, value in the first voiced segment, value in the last voiced segment, correlation, slope, and error of the energy linear regression.

The fourth group is comprised of rhythm features. These are based on the duration of voiced and unvoiced segments. A segment is considered to be unvoiced if its fundamental frequency is zero. The reason for this is that F0 equals the fundamental frequency of the glottal pulses, which are only generated in the presence of speech. Rhythm and duration features can be good emotion indicators, as shown by previous studies. For example, Boersma (1993) noted that the duration variance decreases for most domains under fast stress conditions.

We have calculated five rhythm features, namely speech rate, duration of

voiced segments, duration of unvoiced segments, duration of longest voiced segment and number of unvoiced segments. All these features were normalized by the overall duration of the utterance. We have computed the speech rate as the number of syllables normalized by the utterance duration. To compute the utterance duration we have multiplied the number of frames used by the frame step. Using the 60 acoustic features described above and the 5-fold cross-validation strategy, we obtained for the MLP an emotion recognition rate of 35.42%, whereas for the majority-class baseline it was 51.67%. The significance studies using a t-test with 0.05 significance showed that this difference is significant.

Not all the features employed for classification are necessarily very informative. Unnecessary features make the learning process slower and increase the dimensionality of the problem. Therefore, we have carried out a feature selection process (Guyon and Elisseeff, 2003) employing three methods. Firstly, a forward selection algorithm like that used by Lee and Narayanan (2005), which selects the B1 value at the last voiced segment and the maximum energy. Secondly, a genetic search method that starts with no attributes and uses a population size of 20, 20 generations, 0.6 crossover probability and 0.033 mutation probability. The selected features for this method were the following: F1 maximum, F1 median, B1 minimum, B1 range, B1 median, B1 in the last voiced segment, B2 minimum, B2 maximum, B2 median, energy maximum, energy range, and energy in the last voiced segment. Thirdly, we have ranked the attributes using information gain as a ranking filter. The results employing this method were: energy maximum ranked with 0.50, B1 in the last voiced segment with 0.46, and other features were evaluated with 0. Taking into account the three approaches, the optimal subset was composed of

B1 in the last voiced segment and energy maximum. When we classified with the selected features only, we obtained no improvements in the experiment, as the percentage of correctly classified utterances was 49.00%, which is worse than obtained with the baseline. However, this difference was not significant in the t-test, which indicates that the results for both the MLP and the baseline are equivalent after feature selection.

4.2 Automatic classification based on normalized acoustic features

To reproduce the user's speaking style information that the annotators had when they labelled the corpus in the ordered case, we propose a new approach in which acoustic features are normalized around the neutral voice of the user. For example, let us say that user 'A' always speaks very fast and loudly, and user 'B' always speaks in a very relaxed way. Then, some acoustic features may be the same for 'A' neutral as for 'B' angry, which would make the automatic classification fail for one of the users. This is what happened to the annotators who were not used to the Andalusian dialect (Section 3.4), as they were confused by the fast rhythm and loud speech of the speakers who recorded the corpus.

In order to carry out the proposed normalization we obtain the user's neutral voice features in each dialogue, and subtract them from the feature values obtained in the rest of the user's utterances. To calculate the neutral voice, we could have used the average value of all utterances of that user in the corpus labelled as *neutral* by the annotators. However, our intention for future work is to integrate our emotion recognizer in the UAH system, so that it can be adaptive to user's emotions. It is impossible to carry out this computation in

execution time as this would require to have all the user turns in the dialogue in advance. Therefore, we have considered that the first utterance of the user is neutral, assuming that he is initially in a non-negative emotional state. Besides, we are only interested in emotions caused by the interaction with the system, assuming that the user is in a neutral emotional state when he starts the interaction with the dialogue system. This assumption is possible in domains not directly related to highly affective situations, such as bookings or information extraction, which are the typical application domains of spoken dialogue systems. For these application domains, dialogues in which the user is already in a negative emotional state are negligible.

The accuracy obtained with the Multilayer Perceptron using the normalized features was 53.17%. Thus, introducing acoustic context enables the MLP to improve over the results obtained by the baseline, but the improvement was not significant. Employing the features selected in Section 4.1 (B1 in the last voiced segment and energy maximum) we obtained 69.33% correctly classified utterances, which showed to be a significant improvement following the t-test. In the non-normalized case the feature selection did not yield any improvement. Thus, using normalized acoustic features yielded an improvement of 17.66% (69.33% recognition rate) over the baseline, which was also the best case in the non-normalized classification (Figure 9).

Thus, the normalization of traditional acoustic features yields a noticeable increase in the percentage of correctly recognized emotions with respect to the baseline. This is a very important result as, due to the natural skewness of non-acted emotional corpora, high accuracies can be obtained when directly assigning the most frequent category to all the prompts. In our case, the baseline yielded an accuracy higher than 50.00%. Only with normalization

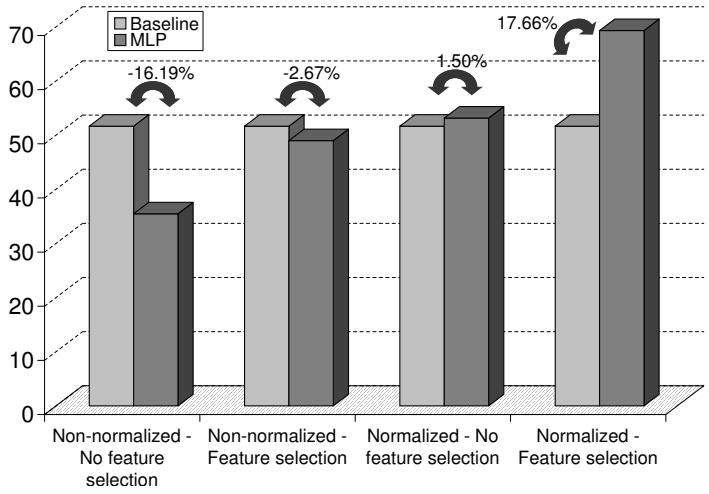


Fig. 9. Recognition accuracy for *angry*, *bored* and *doubtful* considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection

the MLP could obtain better results than the baseline, which were improved by a 17.66% when using acoustic features selection.

A study of the confusion matrices in the normalized and non-normalized classification showed that the *doubtful* category was often confused with the *angry* or *bored* categories, with confusion percentages above 20% in most cases. A similar result was obtained for human annotation given that the ordered scheme did not improve the annotation of the doubtful emotion (as can be observed in Figure 3). These results show that contextual information affects automatic speech recognition using these classification methods, similarly as it affects human annotation.

Thus, in order to distinguish between *doubtful* and the other negative emotions, additional sources of contextual information must be added. We propose to automatically recognize the three emotions using a two-step method. In the first one, acoustic information and contextual information about the user's neutral voice are used to distinguish between *angry* and *doubtfulORbored*. In the second step, dialogue context is used to discern between *doubtful* and *bored*.

For the first step, we used the previously described normalization procedure to recognize *angry* vs. *doubtfulORbored*.

To optimize the results, we carried out another feature selection in which the optimal features are those that best discriminate between *angry* and *doubtfulORbored*. Using the same feature selection algorithms, we obtained a subset comprised of three features: F2 median, energy maximum and energy mean. The results obtained are shown in Figure 10. All of them proved to be significantly better than the baseline using the t-test, except for the first case (non-normalized and no feature selection), where the results were the same order for the MLP and the baseline. The best result for *angry* and *doubtfulORbored* was achieved with feature selection in the ordered scheme, where an 80.00% accuracy was obtained.

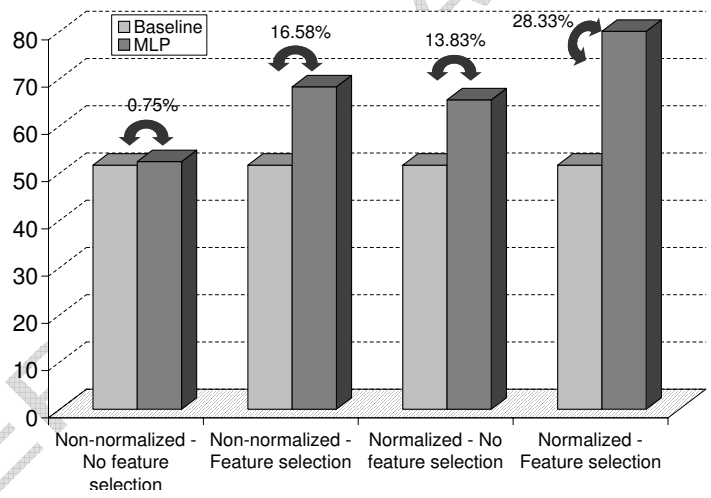


Fig. 10. Recognition accuracy for *angry* and *doubtfulORbored* considering non-normalized vs. normalized acoustic features, and no feature selection vs. feature selection

With these experiments we have shown that acoustic features normalized with neutral style of the users are preferable to the non-normalized ones, as these yield 17.66% improvement (69.33% vs. 51.67% success rate) when recognizing between the three negative emotions, as shown in Figure 9. Moreover,

if acoustic information is used in the first recognition step in which we only distinguish between *angry* and *doubtfulORbored*, accuracy of 80.00% can be achieved, which represents an improvement of 28.33% over the baseline, and of 11.75% over the case when no context information about the user's neutral voice is used (Figure 10). In the next section, a second step is described so that dialogue context can be added to distinguish between *doubtful* and *bored*.

4.3 Automatic classification based on dialogue context

As discussed in Section 3.2, dialogue context was provided to human annotators by giving them the ordered sequence of utterances in each dialogue. To represent context information for automatic recognition we have employed two labels: *depth* and *width*. The first of these indicates the total number of dialogue turns, whereas the second denotes the number of additional user turns necessary to obtain a particular piece of information (e.g. a person's surname). Other authors have already studied the use of discourse structure in similar ways for other areas. For example, Rotaru and Litman (2006) studied how specific locations in the discourse structure are more prone to speech recognition errors in dialogue systems. To do so, they quantified the position of the user turns employing two similar features: "depth" and "transition". However, their approach is based on intentional models of dialogue, which consider it divided in different goals or intentions that have to be satisfied to complete a task. Thus, they define "depth" of a user's turn as the number of different intention subdialogues (or subgoals) in the previous history. This way in their approach several user turns can have the same "depth". Similarly, their "transition" feature captures the position in the discourse structure relative to

the previous turn, describing it using different labels for the type of transition between the last and the current turn (e.g. if the new turn introduces a new intention or continues with the same goal).

Although also based in vertical and horizontal positions of user turns in the dialogue structure, our approach employs numerical representations, which allows the automatic calculation of dialogue context from the dialogue history, without introducing new manual annotations like Rotaru and Litman's "transition" tags. Thus, we propose a completely new calculation of the dialogue context, in which the different values of the vertical component (*depth*) are not shared by different turns. Only our *width* label is concerned with the subdialogue in which the user is engaged. However, we are only interested in subdialogue types that could cause a negative emotion, which we consider to be those in which the user can not provide any new information and has to repeat, correct or confirm data.

The computation of the dialogue context in our approach is as follows:

- *Depth* is initialized to 1, given that 0 would mean that the user hangs up the telephone before he says anything. *Depth* is incremented by 1 for each new user turn and each time the interaction goes backwards (e.g. to the main menu).
- *Width* is initialized to 0, which indicates that the user does not need to repeat any information to the system initially. It is incremented by 1 for each user turn generated to confirm, repeat data or ask the system for help. It is also incremented by 1 if the user has to disambiguate his own input, even if providing new information. Then *width* is restablished to 0 when the user introduces a new piece of information.

For illustration purposes, we show below a sample dialogue from our corpus (translated from Spanish to English), where S = system, U = user:

S1> Welcome to the UAH system.

U1> I need information about professors.

S2> Please say the name of the professor.

U2> López.

S3> I have found two professors named López: Pedro López and Sara López. Who do you mean?

U3> Sara López.

S4> What information do you need about Prof. Sara López?

U4> I want her telephone number.

S5> Sorry, I did not understand. What information do you need?

U5> Telephone.

S6>The telephone number of Prof. Sara López is 958123456. Do you need any other information about professors?

U6> No thanks.

By applying the annotation scheme to this sample dialogue we obtain the following values for *Depth* and *Width*:

U1> Depth=1, Width=0

U2> Depth=2, Width=0

U3> Depth=3, Width=1

U4> Depth=4, Width=0

U5> Depth=5, Width=1

U6> Depth=6, Width=0

It can be observed in this example that the user needed to employ two turns (U2 and U3) to make the system understand the professors' name. In turn U5 he rephrased what he said in turn U4, which solved the system misunderstanding. This is the reason why *width* is 1 for these two user turns.

This scheme is implemented automatically in our system using the dialogue history, which is stored in log files. The *depth* and *width* values of a user turn are computed checking the type of the previous system prompt. For example, as shown in Figure 11, *width* would only be 0 after a system prompt of type "subject_name" if the current prompt type were "subject_information". If the current prompt type were "subject_disambiguation", *width* would be incremented by 1 because an additional user turn would be needed to provide the desired subject to the system.

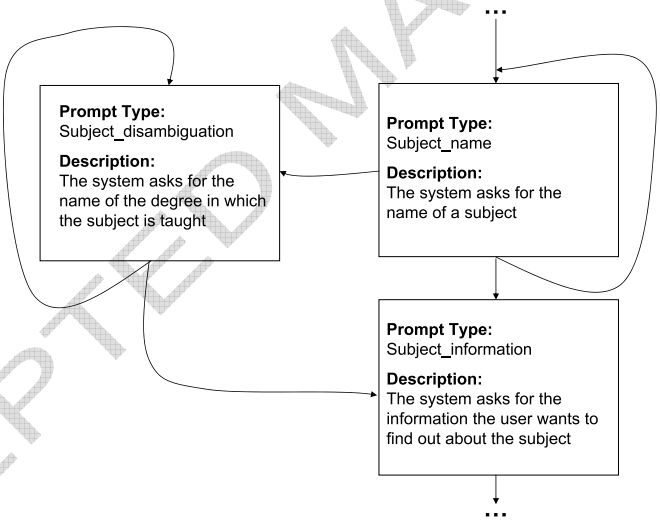


Fig. 11. Example of transitions between system prompts in the UAH system

An exhaustive study of our corpus showed that different users react with different emotions to the same dialogue state, in a less predictable way than we initially expected. Employing one threshold for *depth* and another for *width* to distinguish independently between emotions was found to be inefficient as emotions are influenced by a mixture of the two. Furthermore, the study of our

corpus revealed that it is not sufficient to compute *width* considering only the previous turn or current subdialogue, but it is necessary to take into account the whole dialogue history. This differs from the results reported by Rotaru and Litman (2006), which annotate their horizontal variable ("transition") only with information about the previous system prompt. For example, in the following dialogue:

(...)

S1> Please say the name of the professor.

U1> Martín.

S2> Sorry, I did not understand. Please repeat the name.

U2> Luis Martín.

S3> Did you say Luis Marín?

U3> No, Luis Martín.

S4> What information do you need about Prof. Martín?

U4> His email address.

(...)

width would be 0 in U1, 1 in U2, 2 in U3 and 0 again in U4 because the dialogue starts to deal with a new piece of information. A high *width* value in U2 indicates a higher probability of the user being angry because of the misunderstandings and repetitions needed to make the system understand the name of the professor referred to. However, in turn U4 the user may still be angry but it has *width*=0.

This is why we have defined another measure called *accumulated width*. Whereas *width* is a measure of the extra turns necessary to obtain a particular piece of information, *accumulated width* denotes the total number of extra

turns employed in the whole dialogue up to the current utterance. *Accumulated width* is initialized to 0 and it is increased by 1 each time *width* is incremented. Thus, in the previous example, in U3 $width = 2$, which indicates that it was necessary to repeat or confirm the information of the professor's name twice. Note that in turn U4 $width = 0$ again because the system is gathering different data, namely the type of information about the professor that the user wants. Hence, accumulated width is more representative than width because it takes into account all the problematic points in the previous dialogue. For example, in U4 *accumulated width* $= 2$, which lets us know that there were 2 problematic turns before the current prompt.

The algorithm employed to classify the emotions based on the dialogue context information is as follows:

if Any of the 2 previous turns has been tagged as *ANGRY* **then**
  ANGRY
else if $(D \leq 4)$ AND $(A \leq 1)$ **then**
  DOUBTFUL
else if $(\frac{A}{D} \geq 0.5)$ OR $((D > 4)$ AND $(\frac{A}{D} < 0.2))$ **then**
  BORED
else
  ANGRY
end if

where 'D' denotes *depth* and 'A' the *accumulated width*. In our approach, the user utterances are considered as *doubtful* when the dialogues are short and have no more than one error, as in the first stages of the dialogue is more probable that the users are unsure about how to interact with the system. An

utterance is recognized as *bored* when most of the dialogue has been employed in repeating or confirming information to the system. The user can also be *bored* when the number of errors is low but the dialogue has been long. Finally, an utterance is recognized as *angry* when the user was considered to be angry in at least one of the two previous turns in the dialogue (as described at the beginning of Section 3.4 with human annotation), or the utterance is not in any of the previous situations (i.e. the percentage of the full dialogue length comprised by the confirmations and/or repetitions is between 20% and 50%).

If we consider a baseline that always classifies utterances with the most frequent emotion, which in our case is *angry* (same baseline as in Section 4.2), we obtain 51.67% accuracy in distinguishing between the three emotions. This rate is improved by 13.61% employing our algorithm, which attains an accuracy of 65.28%.

4.4 Automatic classification based on normalized acoustic features and dialogue context

We propose a two-step method that integrates both contextual sources: users' neutral speaking style (Section 4.2) and dialogue context (Section 4.3). The acoustic features normalized with the users' neutral speaking style are used to discriminate whether each utterance is *angry* or *doubtfulORbored*. Then, if an utterance is classified as *doubtfulORbored*, dialogue context information is used to distinguish between *doubtful* and *bored*. Additionally, dialogue context is used to classify utterances as *angry* if they were misrecognized in the first step. The results obtained by the two-step method are shown in Figure 12, and proved to be significant following the t-test.

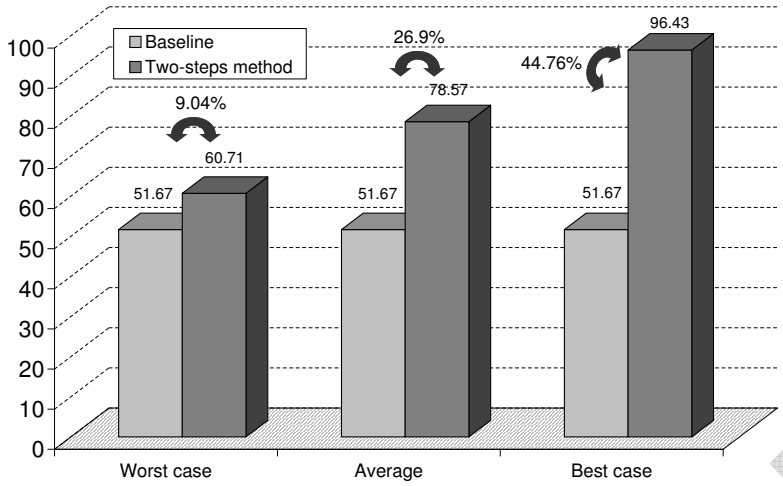


Fig. 12. Emotion recognition accuracy using both acoustic and dialogue context information

Obviously, the result of the two-step method depends on the result of the first one, given that the *angry* vs. *doubtfulORbored* step can fail in the distinction of the two categories and the second step may have to categorize as *doubtful* or *bored* an utterance that belongs to the *angry* category. In the worst case, the first step can fail to recognize all the *angry* utterances, so that all the utterances are recognized as *doubtfulORbored* and passed to the second step. In this case, the recognition accuracy is 60.71%, as a mechanism to detect possible angry utterances has been incorporated to the second step (see Section 4.3). In an ideal best case, the first step would have 100% accuracy, and thus would correctly classify all the utterances as *angry* or *doubtfulORbored*. Thus, the second step would only have to classify the *doubtful* and *bored* utterances. The recognition rate in this case is 96.46%. However, as it was discussed in Section 4.2, with our corpus the first step obtains a maximum 80.00% accuracy, which means that 20.00% of the *angry* utterances may be misrecognized. Employing the two-step method the recognition rate was again 96.43%. Thus, the misrecognized *angry* utterances could be correctly classified in the second step, obtaining a recognition rate for our best case in practice, which is

identical to the ideal best case.

On average between the worst and best case, the two-step method obtains a 78.57% accuracy (as observed in Figure 13), which outperforms the baseline by 26.90%. The improvement over the baseline is 44.76% in the best case, i.e. when the first step does not fail. The average improvement over the recognition based only on neutral acoustic context is 9.24% (27.10% in the best case). If the recognition is based on dialogue context only, the average improvement is 14.29% (32.15% in the best case), and if it is based on the traditional approaches considering non-normalized acoustic features, the average improvement is 29.57% (47.43% in the best case).

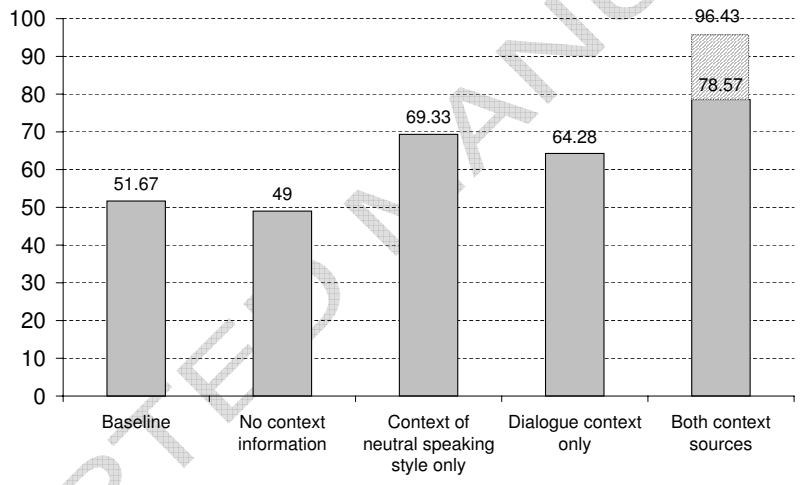


Fig. 13. Recognition accuracy using both acoustic and dialogue context information

Thus, using only one context source (neutral voice or dialogue context), improves over both the baseline and the traditional approach where no context information is used. Besides, combining the two context sources in the proposed two-step method considerably outperforms the baseline, the traditional approach based on acoustic features without additional context sources, and the approach considering only one context source either the neutral voice of the user or the dialogue context.

5 Conclusions and future work

In this paper we have carried out experiments to study the annotation of human emotions in a corpus collected from real interactions with the UAH dialogue system. The experiments considered both manual annotation by 9 non-expert human annotators, as well as automatic classification employing MLPs and different feature selection techniques. We found that human annotators marked 3.40% more non-neutral emotions when they had contextual knowledge. A plausible reason for this result is that context information makes it possible to identify non-trivial emotional speech (e.g. detecting emotions expressed more subtly). On the contrary, when the traditional non-normalized acoustic features were used, only very easily distinguishable emotions were annotated. Additionally, we have discussed the problems that the nature of non-acted emotional corpora impose in evaluating reliability of human annotations. Although deeply affected by the so-called paradoxes of Kappa coefficients, we have studied how the inclusion of context information during annotation helps to obtain values closer to the maximum agreement rates obtainable when compared to not using any additional information.

For machine-learned classification methods, the experimental results show that, due to the natural unbalancement of the corpus, it is difficult to improve the baseline. This makes traditional recognition based on acoustic features yield results very similar to the majority-class baseline. However, as with human annotators, the emotion classification process is substantially improved when adding information about the user's neutral voice and the dialogue history. Just introducing the user's neutral acoustic context gives an improvement of 17.66%. Similarly, employing dialogue context information

improved the baseline results in 12.67%. In this paper we have described a method in two steps to integrate both sources of contextual information. In this method, the dialogue context is useful to distinguish between *angry* and *doubtfulORbored* categories with a 80.00% success rate. Once an utterance is classified as *doubtfulORbored*, the normalized acoustic features enable us to distinguish between *doubtful* and *bored*. When the first step attains maximum accuracy, the two-step method obtains 96.43% accuracy. In the average case, the proposed method obtains 78.57% accuracy, which is 29.57% better than not using contextual information, 47.43% better in the best case (when the first step reaches its maximum performance).

In addition, the proposed methods can be employed during the running of a dialogue system as the contextual information sources can be obtained automatically and at execution time. Plans for future work include the design and implementation of an emotion recognizer for the UAH system to adapt its behaviour automatically, considering negative emotional states of the user. The implementation of the recognizer will involve the use of the approaches proposed in this paper, as well as a classifier to distinguish between neutral and non-neutral emotions. With the results obtained from the evaluation of this recognizer, we will measure both objectively (e.g. in terms of task success) and subjectively (considering users' opinions) the benefits of adding emotion intelligence mechanisms to the dialogue system.

Acknowledgements

The authors would like to thank the reviewers for their knowledgeable comments and helpful suggestions to improve this paper.

References

Adell, J., Bonafonte, A., Escudero, D., 2005. Analysis of prosodic features towards modelling of emotional and pragmatic attributes of speech. Procesamiento de Lenguaje Natural 35, 277–284.

Ai, H., Litman, D., Forbes-Riley, K., Rotaru, M., Tetreault, J., Purandare, A., 2006. Using systems and user performance features to improve emotion detection in spoken tutoring dialogs. In: Proceedings of the 9th International Conference on Spoken Language Processing (Interspeech 2006 — ICSLP). Pittsburgh PA, USA, 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. In: Proceedings of the 7th International Conference on Spoken Language Processing (Interspeech 2002 - ICSLP). Denver, USA, pp. 2037–2040.

Artstein, R., Poesio, M., 2005. $kappa_3$ = alpha (or beta). Tech. rep., University of Essex.

Bickmore, T., Giorgino, T., 2004. Some novel aspects of health communication from a dialogue systems perspective. In: Proceedings of AAAI Fall Symposium on Dialogue Systems for Health Communication. Washington DC, USA, pp. 275–291.

Bishop, C. M., 2006. Pattern Recognition and Machine Learning. Springer.

Boehner, K., DePaula, R., Dourish, P., , Sengers, P., 2007. How emotion is made and measured. International Journal of Human-Computer Studies, Special Issue on Evaluating Affective Interactions 65 (4), 275–291.

Boersma, P., 1993. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In: Proceedings of Institute of Phonetic Sciences. Vol. 17. University of Amnsterdam, pp. 97–

110.

Burkhardt, F., van Ballegooy, M., Englert, R., Huber, R., 2005. An emotion-aware voice portal. In: Proceedings of Electronic Speech Signal Processing. Prague, Czech Republic, pp. 123—-131.

Callejas, Z., López-Cózar, R., 2005. Implementing modular dialogue systems: a case study. In: Proceedings of Applied Spoken Language Interaction in Distributed Environments (ASIDE 2005).

Camurri, A., Mazzarino, B., Volpe, G., 2004. Expressive interfaces. Cognition, Technology and Work 6 (1), 15–22.

Camurri, A., Mazzarino, B., Volpe, G., 2005. Piecing together the emotion jigsaw. Lecture Notes on Computer Science 3361/2005, 305–317.

Cicchetti, D. V., Feinstein, A. R., 1990. High agreement but low Kappa: II. Resolving the paradoxes. Journal of Clinical Epidemiology 43 (6), 551–558.

Cohen, J., 1968. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychological Bulletin 70 (4), 213–220.

Corradini, A., Mehta, M., Bernsen, N. O., Charfuelán, M., 2005. Animating an interactive conversational character for an educational game system. In: Proceedings of the 2005 International Conference on Intelligent User Interfaces. San Diego, CA, USA, p. 183–190.

Cowie, R., Douglas-Cowie, E., Savvidou, S., McMahon, E., Sawey, M., Schröder, M., 2000. 'feeltrace': An instrument for recording perceived emotion in real time. In: Proceedings of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Northern Ireland, pp. 19–24.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., Taylor, J., 2001. Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine 18 (1), 32—-80.

Craggs, R., Wood, M. M., 2003. Annotating emotion in dialogue. In: Proceedings of the 4th SIGdial Workshop on Discourse and Dialogue. Sapporo, Japan, pp. 218–225.

Critchley, H. D., Rotshtein, P., Nagai, Y., O'Doherty, J., Mathias, C. J., Dolana, R. J., 2005. Activity in the human brain predicting differential heart rate responses to emotional facial expressions. NeuroImage 24, 751–762.

Davies, M., Fleiss, J. L., 1982. Measuring agreement for multinomial data. Biometrics 38 (4), 1047–1051.

Devillers, L., Vidrascu, L., Lamel, L., 2005. Challenges in real-life emotion annotation and machine learning based detection. Neural Networks 18, 407–422.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: towards a new generation of databases. Speech Communication 40, 33–60.

Dunn, G., 1989. Design and analysis of reliability studies: the statistical evaluation of measurement errors. Edward Arnold.

Feinstein, A. R., Cicchetti, D. V., 1990. High agreement but low Kappa: I. The problems of two paradoxes. Journal of Clinical Epidemiology 43 (6), 543–549.

Fleiss, J. L., 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76 (5), 378–382.

Fleiss, J. L., Cohen, J., 1973. The equivalence of weighted kappa and the interclass correlation coefficient as measures of reliability. Educational and Psychological Measurement 33, 613–619.

Forbes-Riley, K., Litman, D. J., 2004. Predicting emotion in spoken dialogue from multiple knowledge sources. In: Proceedings of the Human Language

Technology Conference - North American chapter of the Association for Computational Linguistics annual meeting (HLT-NAACL 2004). pp. 201–208.

Gebhard, P., Klesen, M., Rist, T., 2004. Coloring multi-character conversations through the expression of emotions. In: Proceedings of Tutorial and Research Workshop on Affective Dialogue Systems. Kloster Irsee, Germany, pp. 128–141.

Gerfen, C., 2002. Andalusian codas. Probus 14, 247–277.

González, G. M., 1999. Bilingual computer-assisted psychological assessment: An innovative approach for screening depression in chicanos/latinos. Tech. Rep. 39, University of Michigan.

Gut, U., Bayerl, P. S., 2004. Measuring the reliability of manual annotations of speech corpora. In: Proceedings of the Second International Conference on Speech Prosody (SP2004). Nara, Japan, pp. 565–568.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of Machine Learning Research 3, 1157–1182.

Hall, L., Woods, S., Aylett, R., Paiva, A., Newall, L., 2005. Achieving empathic engagement through affective interaction with synthetic characters. In: Proceedings of the 1st International Conference on Affective Computing and Intelligent Interaction. Beijing, China, pp. 731–738.

Hansen, J. H. L., 1996. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. Speech communication 20 (2), 151–170.

Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., Nogueiras, A., 2002. Interface databases: Design and collection of a multilingual emotional speech database. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC 2002). Las Palmas de Gran Canaria,

Spain, pp. 2019–2023.

Iriondo, I., Guaus, R., Rodríguez, A., Lázaro, P., Montoya, N., Blanco, J. M., Bernadas, D., Oliver, J. M., Tena, D., Longhi, L., 2000. Validation of an acoustical modelling of emotional expression in spanish using speech synthesis techniques. In: Proceedings of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Northern Ireland, pp. 161–166.

Johnstone, T., 1996. Emotional speech elicited using computer games. In: Proceedings of the 4th International Conference on Spoken Language Processing (ICSLP 1996). Vol. 3. Philadelphia, PA, pp. 1985–1988.

Krippendorff, K., 2003. Content Analysis: An Introduction to its Methodology. Sage Publications, Inc.

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

Lantz, C. A., Nebenzahl, E., 1996. Behavior and interpretation of the κ statistic: Resolution of the two paradoxes. Journal of Clinical Epidemiology 49 (4), 431–434.

Lee, C., Yoo, S. K., Park, Y. J., Kim, N. H., Jeong, K. S., Lee, B. C., 2005. Using neural network to recognize human emotions from heart rate variability and skin resistance. In: Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference. Shanghai, China, pp. 5523–5525.

Lee, C. M., Narayanan, S. S., 2005. Toward detecting emotions in spoken dialogs. IEEE transactions on speech and audio processing 13 (2), 293–303.

Liscombe, J., Riccardi, G., ür, D. H.-T., 2005. Using context to improve emotion detection in spoken dialog systems. In: Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech 2005 -

Eurospeech). Lisbon, Portugal, 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 (5), 559–590.

Mahlke, S., 2006. Emotions and EMG measures of facial muscles in interactive contexts. In: Proceedings of the 2006 Conference on Human Factors in Computer Systems (CHI 2006). Montréal, Canada.

Montero, J. M., Gutiérrez-Arriola, J., Enríquez, E., Pardo, J. M., 1999. Analysis and modelling of emotional speech in spanish. In: Proceedings of the 14th International Conference of Phonetic. San Francisco, USA, p. 957–960.

Morrison, D., Wang, R., Silva, L. C. D., 2007. Ensemble methods for spoken emotion recognition in call-centers. Speech communication 49, 98–112.

O'Neill, P., 2005. Utterance final /s/ in Andalusian Spanish. The phonetic neutralization of a phonological contrast. Language Design 7, 151–166.

Picard, R. W., Daily, S. B., 2005. Evaluating affective interactions: Alternatives to asking what users feel. In: Proceedings of the 2005 Conference on Human Factors in Computer Systems (CHI 2005), Workshop: Evaluating Affective Interfaces—Innovative Approaches, In CHI 2005. Portland, Oregon, USA.

Pitterman, J., Pitterman, A., 2006. Integrating emotion recognition into an adaptive spoken language dialogue system. In: Proceedings of the 2nd IEEE International Conference on Intelligent Environments. Athens, Greece, pp. 197–202.

Plutchik, R., 1980. EMOTION: A psychoevolutionary synthesis. Harper and Row publishers.

Riccardi, G., Hakkani-Tür, D., 2005. Grounding emotions in human-machine conversational systems. Lecture Notes in Computer Science, 144–154.

Rotaru, M., Litman, D. J., 2006. Discourse structure and speech recognition problems. In: Proceedings of Interspeech 2006. Pittsburgh PA, USA, pp. 53—-56.

Rumelhart, D. E., Hinton, G. E., Williams, R. J., 1986. Learning internal representations by error propagation. MIT Press.

Russell, J. A., 1980. A circumplex model of affect. Journal of Personality and Social Psychology 39, 1161–1178.

Scherer, K. R., 2005. What are emotions? and how can they be measured? Social Science Information 44 (4), 694–729.

Scott, W., 1955. Reliability of content analysis: the case of nominal scale coding. Public opinion quarterly 19 (3), 321–325.

Shafran, I., Mohri, M., 2005. A comparison of classifiers for detecting emotion from speech. In: Proceedings of IEEE International Conference on Acoustic Signal and Speech Processing 2005 (ICASSP 05). pp. 341–344.

Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. In: Proceedings of IEEE Automatic Speech Recognition and Understanding Workhop. pp. 31–36.

Stibbard, R., 2000. Automated extraction of tobi annotation data from the reading/leeds emotional speech corpus. In: Proceedings of the International Speech Communication Association (ISCA) Workshop on Speech and Emotion. Newcastle, Northern Ireland, UK, pp. 60–65.

Streit, M., Batliner, A., Portele, T., 2006. Emotion analysis and emotion-handling subdialogues. In: Wahlster, W. (Ed.), SmartKom - Foundations of Multimodal Dialogue Systems. Springer Verlag, pp. 317–332.

Ververidis, D., Kotropoulos, C., 2006. Emotional speech recognition: resources, features and methods. Speech communication 48, 1162–1181.

Vidrascu, L., Devillers, L., 2005. Real-life emotion representation and detec-

tion in call centers data. Lecture Notes on Computer Science 3784, 739–746.

Vogt, T., André, E., 2005. Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition. In: Proceedings of IEEE International Conference on Multimedia and Expo. pp. 474–477.

Wilks, Y., 2006. Artificial companions as a new kind of interface to the future internet. Tech. Rep. 13, Oxford Internet Institute.

Wilting, J., Krahmer, E., Swerts, M., 2006. Real vs. acted emotional speech. In: Proceedings of Interspeech 2006. Pittsburgh PA, USA, pp. 805–808.

Witten, I. H., Frank, E., 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann.

Zeng, Z., Hu, Y., Fu, Y., Huang, T. S., Roisman, G. I., Wen, Z., 2006. Audio-visual emotion recognition in adult attachment interview. In: Proceedings of the 8th International Conference on Multimodal interfaces. Banff, Alberta, Canada, pp. 828–831.