Annotation and Detection of Emotion in a Task-oriented Human-Human Dialog Corpus *

Laurence Devillers, Ioana Vasilescu, Lori Lamel*
LIMSI-CNRS, BP 133 - 91403 Orsay cedex, France
* ENST-TSI, 46, rue Barrault - 75634 Paris cedex 13, France
{devil, lamel}@limsi.fr, vasilesc@tsi.enst.fr

Abstract

Automatic emotion detection is potentially important for customer care in the context of call center services. A major difficulty is that the expression of emotion is quite complex in the context of agent-client dialogs. Due to unstated rules of politeness it is common to have shaded emotions, in which the interlocutors attempt to limit their internal attitude (frustration, anger, disappointment). Furthermore, the telephonic quality speech complicates the tasks of detecting perceptual cues and of extracting relevant measures from the signal. Detecting real emotions can be helpful to follow the evolution of the human-computer dialogs, enabling dynamic modification of the dialog strategies and influencing the final outcome. This paper reports on three studies: the first concerns the use of multi-level annotations including emotion tags for diagnosis of the dialog state; the second investigates automatic emotion detection using linguistic information; and the third reports on two perceptual tests for identifying emotions as well as the prosodic and textual cues which signal them. Two experimental conditions were considered – with and without the capability of listening the audio signal. Finally, we propose a new set of emotions/attitudes tags suggested by the previous experiments and discuss perspectives.

1 Introduction

This paper addresses the analysis, annotation and automatic detection of emotions in a spontaneous speech corpus comprised of real human-human dialogs. Different school of thoughts, such as psychology, cognitive science, sociology and philosophy, have developed independent theories about personality and emotions. In recent years there has been growing interest in the study of emotions to improve the capabilities of current speech technologies (speech synthesis, automatic speech recognition, and dialog systems). Despite progress made in speech and language technologies, which has led to the development and experimentation of spoken dialog systems as well as a variety of commercial applications, automatic system performance remains substantially inferior to human performance in comparable tasks. Therefore, researchers are still looking for ways to improve the current systems both in terms of efficiency and conviviality. Several of the linguistic areas being actively explored are the study of speech and dialog acts, and markers for prosody and emotion. As a consequence, the definition, categorization and automatic recognition of emotions and attitudes in speech is a research topic with increasing popularity. If the purpose of studying emotions differs for each of the above mentioned domains, all are confronted with the complexity of the domain of emotions and of their means of expression which is multimodal, combining verbal, gestural, prosodic and nonverbal markers such as laughter, throat clearing, hesitations, etc.

Current research can be classify according to three methodological choices: the type of corpus employed, the classes of emotion detected and the emotional markers studied.

Actors, WOz and real-life corpus

In order to control the experimental conditions, many of the reported studies make use of artificial corpora in which human actors read prepared texts simulating a set of archetypal emotions (Dellaert et al., 1996). Other studies use Wizard of Oz (WOz) systems in order to acquire more natural data (Batliner et al., 2000).

*This work was partially financed by the European Commission under the IST-2000-25033 AMITIES project.

Both methods (and their results) are still far from the complex issues found in real spontaneous conversations. Spontaneous corpora obtained from real tasks provide a very large variability in terms of the different parameters carrying emotion manifestations. This variability concerns both the acoustic and linguistic levels, as well as the dialogic strategies and/or the politeness markers. Generally speaking, the study of the emotional behavior of naive users in real-life settings is relatively infrequent in the speech research community, such as (Lee et al., 2002) which use real human-computer call center dialogs.

1.1 Emotion, attitude and stress classes

For traditional linguists, this domain is particularly "slippery", because it implies fuzzy categories and polymorphic notions (Kerbrat-Orecchioni, 2000). As evidence, current researchers are still far from a consensual opinion concerning distinction between the definition of emotion, attitude, personality, mood and even human behavior in general. One of the traditional ways of studying emotions from both the linguistic and psychological perspectives, is to explore the verbal material used in different cultural contexts to express and communicate emotions (Galati et al., 2000; Plutchik, 1994). Moreover, trans-cultural emotional patterns can be detected with a limited number of archetypal emotions. Plutchik (1994) identified eight primary emotions: fear, anger, joy, sadness, acceptance, disgust, anticipation, surprise. Four fundamental terms describing the emotional behavior (i.e. anger, fear, joy and sadness) are widely accepted in the literature as primary emotions. Another point of view is to focus on acoustic parameters, allowing to the correlation of some basic emotions (Morlec et al., 2001; Mozziconacci, 2001) or psychologic stress (Sherer et al., 2002) with prosodic and voice quality markers. For research in the field of human-machine modelization, the study of emotions has generally tried to provide evidence for questions such as: How are mood and speaker intention correlated with dialog quality and success? Which user attitudes and emotional factors may affect human-computer interaction? Generally the goal of this research is to automatically extract mood features in order to dynamically adapt the dialog strategy of the automatic system or for the more critical phases, to pass the communication over to a human operator. Most of the studies focus on a minimal set of emotions or attitudes such as positive/negative emotions (Lee et al., 2002) or emotion/neutral state (Batliner et al., 2000) or stressed/non stressed speech (Fernandez et al., 2002).

1.2 Prosodic, textual and nonverbal markers

In the speech processing area, three main directions for automatic recognition of emotions have been explored. The first direction can be considered acoustic, as it concerns the extraction of prosodic features from the speech signal (i.e., fundamental frequency, energy, speaking rate, etc.) allowing the automatic detection of different emotions (Dellaert et al., 1996; Lee et al., 2001). The second direction can be defined as linguistic (lexical, semantic, dialogic...). It concerns the extraction of linguistic cues identifying emotions. While, this direction has been exploited in traditional linguistics, research in automatic modeling typically uses linguistic cues combined with other information. Studies have also aimed at validating a set of archetypal emotions (anger, fear, sadness, happiness, neutral attitude) via perceptual tests, enabling the extraction of representative lexical items associated with the emotions. For example, in (Petrushin, 1999), perceptual tests were carried out in which naive listeners identified the emotions in simulated utterances produced by the actors. The same corpus was also independently used to study automatic detection of emotions based on a set of acoustic features extracted from the audio signal (pitch, first and second formants, energy and speaking rate). The reported emotion detection performance is about 70%. The third direction consists of combining acoustic information with language information in the spoken utterances (Batliner et al., 2000; Lee et al., 2002). For these purposes the linguistic information has been evaluated in terms of "emotional salience" (i.e. the salience of a word in emotion recognition can be defined as the mutual information between a specific word and an emotion category). This combined approach enables a signification error reduction compared to the use of acoustic or linguistic information separately. Although nonverbal events such as laughter, pauses and throat clearing are considered as significant emotion markers emotion, there has been little evidence of the best way to model this information. Recent developments with the three approaches highlights a real need for integrating several parameters, since the manifestation of emotion is particularly complex and concerns several levels of communication. In addition, the respective weight of the cues characterizing each level is not easily observable. Ultimately, despite the number of parameters employed to automatically distinguish emotions, automatic detection scores are corpus-dependent and thus far from being generalizable to other real-world situations.

1.3 Our approach

This present study is being carried out within the framework of the IST Amities (Automated Multilingual Interaction with Information and Services) project (Amities, 2001-2004) and makes use of a corpus of real Human-Human dialogs recorded in a Stock

Exchange Service Center. The aim of the project is to develop automatic speech dialog systems which are able to communicate with the human user in a natural and efficient way. In call-center applications, an automated system must be able to determine if a critical phase of the dialog is reached and decide if the call should be passed over to a human operator or if the system should initialize a clarification dialog strategy. Customer's vocal expression will carry a number of mixed parameters such as prosodic (i.e breathy voice, increasing energy), lexical (i.e. the use of swear words, and the repetition of words or sub-dialogs) and nonverbal (laughter, screaming) signaling the critical phase.

Most reported studies on emotion detection have explored and validated a predefined number of emotions in a artificially built corpus. Our goal is different. Given the real-life data, we aim to identify salient markers indicating the presence of emotion in the corpus. A major challenge is to overcome the complexity of the expression of emotion in the context of the agent-client call center dialogs. Due to unstated rules of politeness it is common to encounter shaded emotions, in which the interlocutors attempt to limit their internal attitude (frustration, anger, disappointment). In addition, natural dialogs (compared to actors or WOz interaction) allow the expression of emotions in different ways by using a wide range of emotional marks. In this work, we favor the notion of application dependent emotions, and thus focus on a reduced space of emotions, in the context of developing algorithms for conversational interfaces.

In the following sections, we present the corpus and the adopted emotion annotation methodology. We then explore how to correlate emotion labels with dialog quality, progression, and success. Preliminary results on automatic emotion detection using the lexical information are presented in Section 4. Finally, we report on two perceptual tests aimed at validating the selected set of emotions and comparing the respective roles of linguistic (i.e. lexical, syntactic) and acoustic emotion indicators. Finally, we propose a new set of emotion/attitude tags as were validated in the previous experiments and discuss perspectives.

2 Corpus and emotion annotation

The dialogs are real agent-client recordings from a Web-based Stock Exchange Customer Service center. The service center can be reached via an Internet connection or by directly calling an agent. While many of the calls are concerned with problems in using the Web to carry out transactions (general information, complicated requests, transactions, confirmations, connection failures), some of the callers simply seem to prefer interacting with a human agent. A corpus of 100 agent-client dialogs (4 different agents) in French has been orthographically transcribed and annotated at multiple levels (Devillers et al., 2002). The dialogs cover a range of investment related topics such as information requests (services, commission fees, stock quotations), orders (buy, sell, status), account management (open, close, transfer, credit, debit) and Web questions and problems. There are about 5229 speaker turns after excluding overlaps which are known to be frequent phenomena in spontaneous speech. In this work, we make use of a corpus of 5012 sentences corresponding to the in-task exchanges. The corpus contains a total of 44.1k words, of which 3k are distinct. The mean average dialog length is 50 utterances (min 5, max 227), the mean average sentence length is 9 words (min 1, max 128).

2.1 Choice of emotion/attitude labels

A task-dependent annotation scheme has been developed with the assumption that the basic affective disposition towards a computer is either trust or irritation. Emotive communication is not necessarily related to the speaker's "real" inner affective state. Our studies are based on only the external behavioral expression of this state. A speaker's emotive mark should be considered as marking intention in the context of the ongoing dialog. In this specific task, customers are observed to use (consciously or unconsciously) emotive marks in order to persuade the agent to do something for them.

Two of the four classical emotions are retained: *anger* and *fear*. In this Web-based stock exchange context, joy and sadness are uncommon emotions and have been excluded from the emotion set. In addition, we have considered some of the agent and customer behaviors directly associated with the task, which are useful for capturing some of the dialog dynamics. For this purpose, *satisfaction* and *excuse* (embarrassment) are included as emotion labels. These correspond to a particular class of expressive speech acts described in the classical pragmatic theory (Searle, 1985). The anger tag applies to emotions ranging from nervosity to aggressivity, whereas fear ranges from doubt to fear. Finally the "neutral attitude" label corresponds to a default state in the dialog which is progressing normally. The satisfaction label could also be associated with this neutral state, as a particular manifestation of the normal dialog progression.

2.2 Annotation strategy

Two annotators independently listened to the 100 dialogs, labeling each sentence (agent and customer) with one of the five emotions. Sentences with ambiguous labels were judged by a third independent listener in order to decide the final label. Table 1 gives the proportion of sentences in the corpus for each emotion label. Based on the auditory classification, sen-

Emotion Annotation				
A	*F*	*S*	*E*	*N*
5.0%	3.8%	3.4%	1.0%	86.8%

Table 1: Proportion of each emotion label in the dialog corpus determined by listening to the audio signal. (*A: anger, F: fear, S: satisfaction, E: excuse, N: neutral*)

tences with non-neutral labels (F, A, S, E) comprise about 13% (665 sentences) of the corpus. Ambiguities occurred on 138 of the 5012 in-task sentences (2.7% of the corpus) and most often involved indecision between neutral state and other emotions: anger (26/138), fear (25/138), and satisfaction (14/138).

3 Multi-level annotations for diagnosis of the dialog state and factor analysis

3.1 Study objectives

The aim of the study is to explore how to correlate the emotion labels with the dialog quality, progression, and success. The final goal is to exploit the relationships between dialogic annotations and emotion labels in order to determine features which can be automatically extracted and to dynamically adapt the dialog strategy of the spoken language dialog system accordingly.

3.2 Multi-level annotations

With the goal of relating dialog annotations (lexical, pragmatic and dialogic) with emotion annotations, the 100 agent-client dialogs were annotated with DAMSL-like dialogic labels and dialog progression axis labels (Devillers et al., 2002). The dialogic annotations (Hardy et al., 2002) were adapted from the DAMSL standard dialog acts (Allen et al., 1997). The turn based annotations were entered using the XDML tool provided by the State University of New York, Albany, a partner in the AMITIES project (Amities, 2001-2004). For the purposes of this study the DAMSL-like annotations were limited to the dialogic level: no semantic labels were annotated. The selected dialogic labels are applied at three main levels: **Information**, **Forward Looking function** and **Backward Looking function**. The dialog progression labels are represented on two axes to take into account the dynamic aspect of dialog. A progression axis represents the "good" progression of the dialog whereas an accidental axis represents the accidents that occur, corresponding to misunderstandings between the agent and the user (Rosset and Lamel, 2002). The emotion, dialog and progression annotations were carried out independently.

3.3 Choice of parameters

The three annotation types (emotion, dialog, progression) are used to determine a set of factors with which predictive models are estimated. For this experiment, 14 parameters were extracted from the annotated corpus. These parameters primarily denote negative factors in the three annotation types (Non-Understanding, Accident, Anger,...) which can be expected to have an influence on the dialog quality. Five parameters are taken from the dialogic annotations (Devillers et al., 2002): at the `Statement` level, Reassert (REA); at the `Agreement` level, Reject (REJ) and I-Don't-Know (IDK); and at the `Understanding` level, Non-Understanding (NUN) and correct (COR). Three parameters concern the dialog progression axis: Residual Error (RER), Accident (ACC) and Progression (PRO) (Rosset and Lamel, 2002). The five emotion labels are: Fear (FEA), Anger (ANG), Neutral state (NEU), Excuse (EXC) and Satisfaction (SAT). The last parameter is the dialog duration in number of turns (LEN). Some of these parameters can be categorized as utterance-level features (emotion and dialogic labels), and some others are per-turn features (dialog progression axis parameters). As a first measure of the dialog quality a global predictive feature vector is extracted from each dialog. This vector is formed by summing and normalizing all the occurrences of each of the selected 14 parameters.

3.4 Methodology, results and analysis

Table 2 shows the correlations between the 14 parameters. Correlations higher than 0.4 are shown in bold. There are very high correlations between dialog length and dialog progression with neutral state, which is to be expected since over 86% of the sentences have this label. Another notable correlation is between Residual Error and Accident, which is also expected.

We used classical multiple linear regression techniques to find which combination of factors are able to predict parameters such as Accident and Residual Error or Emotion (Anger and Fear) in a dialog. Different multiple regression models were assessed by adding and dropping terms as appropriate using ANOVA.

Table 3 shows some prediction models for detecting dialogs with problems, in particular for Accidents and Residual Errors. A correct prediction for the parameter ACC is obtained with the predictive factors: ERR, ANG, EXC, FEA, COR and REJ (first entry). Taken together these factors explain 81.6% of the variance of accidents, with the highest contribution from RER. The next three models remove the RER factor, which is highly correlated with accidents and may mask the contributions of other factors. The second entry explains 65.5% of the variance of the accidents. Comparing the 3rd and 4th entries, the Emotion factors

	ACC	RER	PRO	FEA	ANG	SAT	EXC	NEU	IDK	COR	NUN	REA	REJ	LEN
ACC	1.0													
RER	**0.81**	1.0												
PRO	**0.52**	0.31	1.0											
FEA	**0.46**	**0.41**	**0.47**	1.0										
ANG	**0.63**	**0.54**	**0.41**	0.30	1.0									
SAT	0.15	0.07	0.3	**0.42**	0.12	1.0								
EXC	**0.51**	0.35	0.21	0.10	0.32	0.04	1.0							
NEU	0.35	0.17	**0.84**	0.29	0.30	0.21	0.10	1.0						
IDK	0.38	0.38	0.38	0.38	0.30	0.1	0.08	0.11	1.0					
COR	0.07	0.01	0.17	-0.06	-0.15	0.02	-0.05	0.04	-0.03	1.0				
NUN	0.27	0.28	0.17	0.15	0.21	0.01	0.12	0.08	0.07	0.10	1.0			
REA	**0.56**	**0.43**	**0.56**	0.34	**0.50**	0.19	0.27	0.36	0.21	0.01	0.29	1.0		
REJ	**0.54**	0.35	0.39	**0.40**	**0.45**	0.27	0.25	0.23	0.16	-0.03	0.05	**0.59**	1.0	
LEN	0.33	0.16	**0.82**	0.29	0.30	0.23	0.1	**0.99**	0.11	0.04	0.09	0.34	0.22	1.0

Table 2: Correlations among the 14 selected factors. ACC: accident, RER: residual error, PRO: progression, FEA: fear, ANG: anger, SAT: satisfaction, EXC: excuse, NEU: neutral, COR: correct, NUN: non-understanding, REA: reassert, REJ: reject, and LEN: dialog duration. (After (Devillers et al., 2002)).

Variable	*Main Predictors*						*Explanation*
ACC	.55 · RER	.22 · EXC	.18 · REJ	.17 · ANG	.12 · COR	.10 · FEA	81.6%
ACC	.34 · ANG	.33 · EXC	.22 · FEA	.20 · REJ	.17 · COR	.12 · IDK	65.5%
ACC	.42 · ANG	.35 · EXC	.32 · FEA				58.8%
ACC	.34 · REJ	.27 · IDK	.25 · REA	.16 · NUN			47.6%
RER	.28 · ANG	.20 · FEA	.18 · EXC	.14 · IDK	.13 · NUN	.10 · REA	44.6%
RER	.38 · ANG	.29 · FEA	.19 · EXC				39.9%
RER	.29 · IDK	.19 · REA	.19 · NUN	.17 · REJ			31.9%
ANG	.33 · ACC	.21 · REA	–.18 · COR	.14 · IDK	.07 · RER		48.6%
FEA	.24 · IDK	.22 · ACC	.21 · REJ				30.6%

Table 3: Prediction models for **ACC**, **RER**, **ANG**, **FEA**. The weighted main factors predict the variable with the percentage given in the Explanation column. ACC: accident, RER: residual error, PRO: progression, FEA: fear, ANG: anger, SAT: satisfaction, EXC: excuse, NEU: neutral, COR: correct, NUN: non-understanding, REA: reassert, REJ: reject, and LEN: dialog duration. (After (Devillers et al., 2002)).

Emotion Annotation (no dialog context)				
A	*F*	*S*	*E*	*N*
2.3%	1.8%	5.8%	1.2%	88.9%

Table 4: Proportion of each emotion label in the dialog corpus based only on lexical information (without dialog context). (*A: anger, F: fear, S: satisfaction, E: excuse, N: neutral*)

EXC, FEA and ANG seem to be better predictors of accidents (58.8%) than the dialogic factors (47.6%) retained here. It can be inferred that the Emotion factors account for most of the explanation of the 2nd model.

Models were also built to predict the RER at the end of the dialog, which is an important indication of the overall dialog success. The first model is able to explain 44.6% of the variance of the residual dialog progression errors with a p_value of 4.496e-10. Anger is also seen to be correlated with error at the end of the dialog and is a good predictor of dialog problems.

Finally, we tried to predict emotions such as Anger and Fear. Client anger can be partially explained with dialog axis Accidents, and the dialogic labels (reassertion, correction), but Fear is unable to be predicted with better than 30% using any combination of the 14 parameters. Client anger is to some degree correlated with the need to repeat information, but the negative weight of correlation seems to imply that correcting errors is not a big deal. Problems arise when the one of the interlocutors is unable to correct an error.

3.5 Discussion

Using standard multiple linear regression techniques, a predictive function of dialog problems was derived, estimating the relative contributions of various factors extracted from dialogic, progression and emotion annotations. These measures are able to explain about 80% of the dialog accidents. The observed correlations between DAMSL-like dialogic labels and the annotations for emotion and dialog progression axes provide evidence that these latter annotation types are relevant.

4 Emotion Detection Model

4.1 Study objectives

Our goal is to analyze the emotional behaviors observed in the linguistic material of the human-human interactions present in the dialog corpus in order to detect which kinds of lexical information are particularly salient to characterize each emotion.

4.2 Dialog context-independent Emotion annotation

A second type of emotion annotation based only on lexical information was carried out at the sentence level without listening to the audio signal and without the dialog context. Each sentence transcription (the sentences were randomized in order to avoid using the dialog context in making a judgment), was labeled as to the presence or absence of the five non-neutral emotions (anger, fear, satisfaction, excuse, neutral attitude). When uncertain, the annotator could indicate that the lexical cue was ambiguous. The lexically-based emotion labels were made by one annotator who had never listened to the dialogs, thereby avoiding any subjective influence from the audio signal. The percentage of sentences with each emotion label is shown in the Table 4 (*Lexical*). Based on only lexical information, the non-neutral emotion labels (F, A, S, E) are seen to apply to 11% of the corpus (554 sentences).

4.3 Emotion detection

Our emotion detection system uses the same basic unigram model as is used in the LIMSI Topic Detection and Tracking system (Lo and Gauvain, 2001). The similarity between an utterance and an emotion is the normalized log likelihood ratio between an emotion model and a general task-specific model.

Five emotion models were trained, one for each annotated emotion (A, F, S, E, N). For each emotion a unigram model is constructed from the set of on-emotion training utterances (without using the off-emotion training utterances). Due to the sparseness of the on-emotion training data (about only 11% of the corpus), the probability of the sentence given the emotion is obtained by interpolating its maximum likelihood unigram estimate with the general task-specific model probability. The general model was estimated on the entire training corpus. An interpolation coefficient of $\lambda = 0.75$ was found to optimize the results. The emotion of an unknown sentence is determined by the model yielding the highest score for the utterance u given the 5 emotion models E:

$$logP(u|E) = \frac{1}{L_u} \sum_{w \in u} tf(w,u) \log \frac{\lambda P(w|E) + (1-\lambda)P(w)}{P(w)}$$

where $P(w|E)$ is the maximum likelihood estimate of the probability of word w given the emotion model, $P(w)$ is the general task-specific probability of w in the training corpus, $tf(w,u)$ are the term frequencies in the incoming utterance u, and L_u is the utterance length in words.

4.4 Stemming, Stopping and Compounding

Stemming and stopping are commonly used procedures in information retrieval tasks for removing very frequent words in order to increase the likelihood that the resulting terms are relevant. We have adopted these techniques for emotion detection. In order to reduce the number of lexical items for a given word sense, an automatic part of speech tagger (Toussaint et al., 1998) was used to derive the word stems. Stopping is a standard filtering procedure which removes high frequency words which are assumed to be uncorrelated with the task. Experiments were carried out using different stop lists (containing from 60 to 200 entries). Our stop-lists differed from standard lists in that some frequent words that can be meaningful for the emotion detection such as *no, yes, okay, thanks* were not filtered.

A list of about 20 compound words was constructed to compensate for the limited span of a unigram model. Compound words are needed to account for negative expressions which can be important indicators for emotion detection. For example, *pas_marcher (doesn't_work)* and *pas_normal (not_normal)* can suggest that the person is upset about the situation.

4.5 Experiments

Emotion detection experiments using lexical cues were carried out with two sets of 125 test sentences (25 sentences per emotion class) extracted from the annotated corpus. The remaining sentences (about 5k) were used for the training. For these experiments the lexically based annotations are used as the reference emotions.

Table 5 summarizes the emotion detection results for the baseline unigram system, and the improvements due to the normalization procedures. Since the normalization procedures change the lexical forms, the number of words in the lexicon are also given for each condition. The results are given for the complete test set and for the anger subset. Using the baseline system, emotion can be detected with about 62% precision. Stemming and compounding are both seen to improve the detection rate. Despite trying multiple stop-lists, stopping did not improve the detection rate. Compounding seems to be helpful for detecting anger, due to the inability of the unigram model to account for the word context. (Given the limited size of our corpus, it is not possible to build larger span models than unigram models.)

For the remaining tests, the training and test data were normalized by the stemming and compounding procedures. Table 6 gives the detection scores for each of the five emotion classes averaged across two test sets of 125 sentences. The results show that some emotions are better detected than others, the best being satisfaction and the worst fear. The high detection of satisfaction can be attributed to strong lexical markers which are very specific to this emotion (*thanks, I agree*). On the contrary, the expression of fear is more syntactic than lexical, i.e., word repetitions, restarts, etc. For example: *ou alors je vends des ou alors je je je vends je ne sais pas encore (or so I sell the so I I I sell I don't know yet)*. Examples of the more specific lexical items for each class are shown in Table 7. Some words such as *problem, bizarre* are shared by the anger and fear classes. Other potential detection cues such as idiomatic expressions are language dependent (*je laisse courir (forget about it), je m'assois dessus (I don't care), on m'a ôté une épine du pied (s/he solved the problem for me)*) are too difficult to model with a unigram lexical model. Although quite meaningful to a human, such expressions are relatively infrequent and quite varied, making them difficult to capture in a statistical model.

In order to refine the list of lexical cues, it would be

Condition	*Test (125)*	*Anger (25)*	*#words*
Baseline	61.6%	48.0%	2732
Stem	64.8 %	48.0%	1911
Stem+Comp	67.2%	56.0%	1927

Table 5: Emotion detection performance of the baseline system; the baseline system with stemming; and the baseline system with stemming and compounding. Results are shown for the complete test set (125 utterances) and for the anger emotion subset (25 sentences). The lexicon sizes (#words) are also shown.

Detected Emotion (%)					
Total	*A*	*F*	*S*	*E*	*N*
68	56	38	88	68	88

Table 6: Average emotion detection scores on two 125-sentence test sets. (*A: anger, F: fear, S: satisfaction, E: excuse, N: neutral*

A	*F*	*S*	*E*
abnormal	worry	agree	mistake
"swear words"	fear	thanks	error
irritating	panic	perfect	sorry
embarrassing	afraid	excellent	excuse
bothering	disastrous	excellent	pardon

Table 7: Examples of specific lexical items for each emotion class which from a statistical point of view are unambiguous. (*A: anger, F: fear, S: satisfaction, E: excuse)*

necessary to consider the role of non-linguistic events in the dialog (nonverbal markers, filler words, hesitations, backchannel information, interruptions) for mood detection. In this corpus of 100 dialogs, there were too few nonverbal markers (laughter and throat clearing) in the transcripts to find a correlation with the emotion labels. Similarly, although hesitations (*euh*) are relatively frequent in spontaneous speech, they do not appear to be relevant for distinguishing different emotions.

Preliminary results on the automatic detection of a set of 5 task-dependent emotions resulted in a detection rate of around 70% using a simple lexical unigram model trained on the lexically annotated training corpus.

5 Perceptual tests

Subjective tests have been carried out on a subset of the dialog corpus for two experimental conditions, with and without the capability of listening to the audio signal. The [-signal] condition requires emotion detection using only linguistic information (i.e. using only the orthographic transcriptions of the utterances extracted from a dialog as stimuli). The [+signal] condition provides both linguistic and acoustic information and highlights the role of both sources of information. Different sets of native French speakers participated in the one of the two conditions. The experimental tests consisted of naming the emotion present in each stimulus and of describing the linguistic cues ([-signal] condition) and the linguistic and acoustic cues ([+signal] condition) used to make the judgment.

5.1 Corpus and experimental protocol

The test stimuli consist of 45 sentences extracted from Stock Exchange dialog corpus. The sentences were selected with respect to the 5 categories of emotions previously annotated in context (see Section 4). There are 9 sentences for each emotion class. In order to allow more liberty in the tests, the set of emotion labels proposed to the subjects was enlarged to cover shaded emotions. Five of the sentences (one sentence per annotated emotion) were used in the training phase of the perceptual experiments. The test corpus is balanced with respect to the agent and customer sentences. The 20 agent sentences are equally distributed between the four agents (three men and one woman). For the customer sentences, the majority are spoken by men (18 male, 2 female utterances).

Forty native French subjects have participated in the one the two tests: 20 for each condition.

The instructions for the subjects using the test interface were:

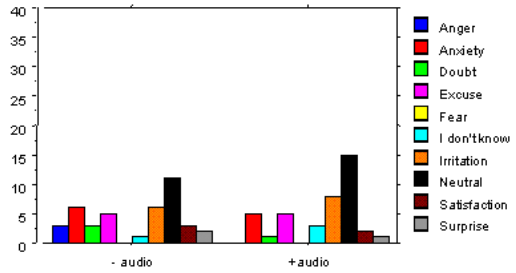

Figure 1: Majority votes for the 10 emotion classes for both experimental conditions [-audio, +audio]. On average the majority of subjects agreed over 55% of the time.

1. to name the emotion (open-choice) present in each stimulus without the dialog context

2. to describe the linguistic cues ([-signal] condition) or the linguistic and acoustic cues ([+signal] condition)

3. to choose among the emotion labels: *anger, fear, irritation, anxiety, doubt, surprise, satisfaction, excuse, neutral attitude* and *I don't know*.

All elements to be completed were presented at the same time on the screen, for each stimulus individually. Consequently, subjects could complete the information in any order they wanted.

5.2 Results

Open choice emotion

Subjects were asked to classify each sentence in terms of the perceived emotions. The results suggest two major strategies: (1) subjects use one or more of the labels of the forced choice present in the test window. The labels are employed alone or combined with other emotion and attitude markers. (2) subjects use another emotion label or complex syntactic structure describing the agent/client behavior. Concerning this second strategy, embarrassment and disappointment were the most frequent labels.

Results for emotion classes

Figure 1 summarizes the identification results for the perceptual experiments. 55% of the sentences are identically labelled with and without listening to the audio signal. These results highlight the importance of linguistic information in telephone-based applications and the influence of commonly accepted politeness rules. The majority of the customer sentences

were judged to be in the anxiety and irritation emotion classes. As expected for this sort of application the agent sentences were usually judged as neutral attitude or excuse emotion. There are only few exceptions such as marks of anxiety "ouh la c'est pas normal tout ça" (*oh oh that is suspect*).

A surprising result is that the proportion of non-neutral emotions is lower when subjects were able to listen to the signal than when they were not (55% compared to the 70%). Some possible explanations are first, the rules of politeness encourage callers to control the expression of the underlying emotion, and second, the subject may associate voice quality rather than emotion with the audio characteristics. Some examples in the context-independent perceptual tests indicated the significance of the voice quality. There are sentences that were clearly judged by subjects as in the class "I don't know" or "neutral" with audio listening because the voice tone did not correspond to the semantic meaning of the sentence. For example, the sentence "d'accord très bien je vous remercie" (ok, very good, thanks) was always judged as "satisfaction" without listening the sentence, but when subjects were able to listen the majority selected the "I don't know" label suggesting that the lexical and prosodic cues were contradictory. The audio condition provides a number of complex acoustic parameters which carry the emotion including the voice quality, the environment quality (recording, noise) which results from a mixture of emotional state of speaker and interaction-dependent emotion, etc. In addition, the acoustic correlates of emotion in the human voice are subject to large individual differences. Prosody adds complementary information to the linguistic content of the message or can be contradictory and even modify, as shown in the previous example, the literal meaning of the sentence. In real-life dialogs, the context helps to correctly evaluate the interaction-dependent emotion. For example, the context-dependent audio-based annotation provides the satisfaction label for the previous sentence suggesting that the context can solve the contradiction between the acoustic cues and linguistic information. This example highlights the importance of integrating dialog information to improve emotion recognition.

Prosodic and textual cues

Here we report on the perceptual prosodic and textual cues detected by the subjects. We will present the major choices in terms of perceptual cues analysis. In order to compare the role of the two types of cues, we focus on the [+audio] condition experiment. As mentioned in experimental protocol, subjects were ask to describe the linguistic and prosodic cues used in making their judgement. The linguistic cues concern emotionally charged words and particular syntactic phenomena. Generally speaking, the acoustic and prosodic parameters accepted and described in the literature are speech rate, intensity and F0 variation. For the perceptual test, we supposed that providing acoustic prosodic cues is not intuitive for a naive subject, so we guided them with a number of forced choices for describing the selected prosodic parameters. The choices for the speech rate were: slow, normal and fast; for intensity: normal and high; and for F0 variation: flat or variable.

The majority of subjects judged the speech rate as fast for irritation and satisfaction, whereas the F0 variation allowed subjects to distinguish neutral state and excuse (flat) from other emotional states (variable). Finally, there was no noted perceptual difference in intensity across the stimuli. Two possible explanations of these results are: (i) there is no objective perceived acoustic variation among the stimuli of the test; (ii) the telephonic speech does not allow subjects to perceive this variation. In addition, pitch and energy extraction for telephone speech is an especially difficult problem, due to the fact that the fundamental is often weak or missing, and the signal to noise quality is usually low. Concerning the first explanation, in contrast to the perceptual cues found to be relevant using simulated emotions produced by actors, in the WOz experiments (Batliner et al., 2000) and real agent-client dialogs the acoustic and prosodic cues are much less easily identifiable. A reason for this difference is that in acted speech, emotions are expressed with more prosodic clues than in realistic speech data where speakers express their emotions multiple linguistic strategies (i.e lexical, dialogic...).

Concerning the emotionally charged keywords, the subjects' answers can be grouped into a few main classes: words denoting emotion (*nervous* for irritation, *I am afraid* for anxiety, *thanks so much* for satisfaction...), swear words ('4-letter' words for irritation), exclamations, negation, etc. Concerning syntactic structure, the responses point out a number of characteristics of spontaneous speech (hesitation, repetition, reformulation...) but only a few are explicitly correlated with a particular emotion (such as spluttering for anxiety). We are currently looking to correlate the perceptual cues with objective measures carried out on the test corpus. The next step will be to validate these measures on a larger corpus.

6 Discussion

Three main emotional behaviors emerge from the perceptual tests in both conditions: irritation, anxiety and neutral attitude (the default normal state). In addition, the excuse attitude was identified as an agent behavior directly associated to the task. After exploring a

range of possibilities for annotating emotions with the perspective of carrying out automatic detection experiments, we decided to refine the set of emotion/attitude tags. These tags reflect complex mixed and shaded emotions and attitudes found in the agent-client call center dialogs. The six selected tags are the shaded emotion tags of anxiety and irritation, the neutral attitude and three attitudes dependent on the interaction in the perceptive tests presented above: embarrassment, disappointment and satisfaction.

7 Perspectives

This work has been carried out in the context of the AMITIES project which aims to explore novel technologies for adaptable multilingual spoken dialog systems. Central to the project is the study and modelization of large corpora of human-human and human-computer dialogs which serve as the basis for system development. Part of our ongoing work is the multilevel annotation of agent-client dialogs from another financial call center. Concerning the emotion labels, 130 dialogs have already been annotated with the new tag set and will serve as a larger real-life corpus to better evaluate the role of different linguistic parameters. Further work will be to explore the combination of emotion information conveyed by the textual information and the contextual dialogic information with the prosodic features.

8 Acknowledgments

The authors thank Catherine Mathon for helping with perceptual tests and Isabelle Wilhem for helping with the corpus annotation.

References

AMITIES, http://www.dcs.shef.ac.uk/nlp/amities

J. Allen and M. Core, "Draft of DAMSL: Dialog Act Markup in Several Layers," October 1997. http://www.cs.rochester.edu/research/trains/annotation

A. Batliner, K. Fischer, R. Huber, J. Spliker, E. Nöth, "Desperately seeking emotions or: actors, wizards, and human beings", *ISCA Workshop on Speech and Emotion*, 2000.

F. Dellaert, T. Polzin, A. Waibel,"Recognizing Emotion In Speech," *ICSLP*, 1996.

L. Devillers, S. Rosset, H. Maynard, L. Lamel, "Annotations for Dynamic Diagnosis of the Dialog State," *LREC'02*.

R. Fernandez, R. Picard, "Modeling Drivers' Speech Under Stress," *Speech Communication*, 2002.

D. Galati, B. Sini, "Les structures sémantique du lexique français des émotions", in Les émotions dans les interactions, C.Plantin, M.Doury, V.Traverso (eds.), PUL 2000. ch 3, p75-87.

H. Hardy, K. Baker, L. Devillers, L. Lamel, S. Rosset, T. Strzalkowski, C. Ursu and N. Webb, "Multilayer Dialogue Annotation for Automated Multilingual Customer Service", *ISLE Workshop on dialogue tagging for multi-modal Human-Computer Interaction*, Edinburgh, Dec 2002.

C. Kerbrat-Orecchioni, "Les émotions dans la linguistique", in *Les émotions dans les interactions*, C.Plantin, M.Doury, V.Traverso (eds.), PUL 2000. ch2, p33-74.

C.M. Lee, S. Narayanan, R. Pieraccini, "Recognition of Negative Emotions from the Speech Signal", *ASRU*, 2001.

C.M. Lee, S. Narayanan, R. Pieraccini, "Combining acoustic and language information for emotion recognition", *ICSLP*, 2002.

Y.Y. Lo, J.L. Gauvain, "The Limsi Topic tracking system for TDT2001," *DARPA Topic Detection and Tracking Workshop*, Gaithersburg, Nov. 2001.

Y. Morlec, G. Bailly, V. Aubergé, "Generating prosodic attitudes in French: data, model and evaluation", *Speech Communication*, March 2001.

S. Mozziconacci "Emotion and attitude conveyed in speech by means of prosody," *2nd Workshop on Attitude, Personality and Emotions in User-Adapted Interaction*, Sonthofen, Germany, July 2001.

V. Petrushin, "Emotion in speech: recognition and application to call centers," *Artificial Neural Network ANNIE'99*.

R. Plutchik, *The psychology and Biology of Emotion*, HarperCollins College, New York, 1994.

S. Rosset, L. Lamel,"Representing Dialog Progression for Dynamic State Assessment," *ISLE Workshop on dialogue tagging for multi-modal Human-Computer Interaction*, Edinburgh, Dec 2002.

JR. Searle, D. Vanderveken, "Fondations of Illocutionary Logic," Cambridge: CUP, 1985.

K. Sherer, D. Frandjean, T. Johnstone, G. Klasmeyer, T. Bänziger, "Acoustic correlates of task load and stress," *ICSLP*, 2002.

Y. Toussaint, F. Namer, C. Jacquemin, B. Daille, J. Royauté, N. Hathout, "Une approche linguistique et statistique pour l'analyse de l'information en corpus," *TALN'98*.