EMOTION DETECTION IN TASK-ORIENTED SPOKEN DIALOGS[1]

Laurence Devillers, Lori Lamel

LIMSI-CNRS, TLP
BP 133 - 91403 Orsay cedex, France
{devil, lamel}@limsi.fr

Ioana Vasilescu

ENST-CNRS, TSI
46, rue Barrault - 75634 Paris cedex 13, France
vasilesc@tsi.enst.fr

ABSTRACT

Detecting emotions in the context of automated call center services can be helpful for following the evolution of the human-computer dialogs, enabling dynamic modification of the dialog strategies and influencing the final outcome. The emotion detection work reported here is a part of larger study aiming to model user behavior in real interactions. We make use of a corpus of real agent-client spoken dialogs in which the manifestation of emotion is quite complex, and it is common to have shaded emotions since the interlocutors attempt to control the expression of their internal attitude. Our aims are to define appropriate emotions for call center services, to annotate the dialogs and to validate the presence of emotions via perceptual tests and to find robust cues for emotion detection. In contrast to research carried out with artificial data with simulated emotions, for real-life corpora the set of appropriate emotion labels must be determined. Two studies are reported: the first investigates automatic emotion detection using linguistic information, whereas the second concerns perceptual tests for identifying emotions as well as the prosodic and textual cues which signal them. About 11% of the utterances are annotated with non-neutral emotion labels. Preliminary experiments using lexical cues detect about 70% of these labels.

1. INTRODUCTION

In recent years there has been growing interest in the study of emotions [1, 2, 7] to improve the capabilities of current speech technologies (speech synthesis, speech recognition, and dialog systems). While different school of thoughts, such as psychology, cognitive science, sociology and philosophy, have developed independent theories about personality and emotions [5, 10, 11], 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. In the context of human-machine interaction, the study of emotion has generally been aimed at the automatic extraction of mood features in order to be able 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.

Despite the lack of consensus describing human behavior (emotion, attitude, mood, etc.), four primary emotions are widely accepted in the literature: fear, anger, joy, sadness. These emotions are not necessary well adapted to human-machine interaction, where studies have focused on a minimal set of emotions/attitudes such as positive/negative emotions [9] or emotion/neutral state [1] or stressed/non-stressed speech [4]. With real-life data, the emotions are often considered as application-dependent [9]. Three main directions for emotion detection have been explored. The acoustic direction concerns the extraction of prosodic features from the speech signal (i.e., fundamental frequency, energy, speaking rate, etc.) which allow automatic detection of different emotions [2, 7]. The linguistic direction concerns the extraction of lexical cues identifying emotions. While, this direction has been exploited in traditional linguistics research in automatic modeling typically combines lexical cues with prosodic information [1, 9]. These recent developments highlight the need for integrating several parameters, since the manifestation of emotion is particularly complex and concerns several levels of communication. Although nonverbal events (laughter, pauses, throat clearing) are considered as significant emotion markers, there has been little evidence of the best way to model this information.

This present study is carried out within the framework of the IST Amities (Automated Multilingual Interaction with Information and Services) project, and makes use of a corpus of real agent-client dialogs recorded (for independent purposes) at a Stock Exchange Customer Service Center.

2. CORPUS AND EMOTION ANNOTATION

A corpus of 100 agent-client dialogs (4 agents) in French has been orthographically transcribed and annotated at multiple levels: semantic, topic, and dialogic [3, 6]. The Customer Service center can be reached via an Internet connection or by directly calling an agent. The dialogs cover a

[1]This work was partially financed by the European Commission under the IST-2000-25033 AMITIES project `http://www.dcs.shef.ac.uk/nlp/amities`.

range of investment related topics such as information requests (services, commission fees, stock quotations), orders (buy, sell, status), account management and Web questions and problems. There are about 5229 speaker turns after excluding overlaps. This work uses 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 average dialog contains 50 utterances (min 5, max 227), the average sentence length is 9 words (min 1, max 128).

An annotation scheme has been developed with the assumption that the basic affective disposition towards a computer is either trust or irritation. Although our annotation scheme can be considered to be domain-dependent, the annotations cover a large range of services such as financial, lost luggage etc. For the initial set of labels, 2 of the 4 classical emotions are retained: *anger* and *fear*.[1] In addition, some of the agent and customer behaviors directly associated with the domain which can be useful for capturing the dialog dynamics have also been considered. As a result, *satisfaction* and *excuse* (embarrassment) are included as emotion labels. 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 the default state in which a dialog progresses normally.

Two types of emotion annotations were carried out. For the first type, two annotators independently listened to the 100 dialogs, labeling each sentence (agent and customer) with one of the five emotions (anger, fear, satisfaction, excuse, neutral attitude). The second type of 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 by two annotators with one of the five emotions. Sentences with ambiguous labels ($\sim$ 3%) for both annotations were judged by a third independent annotator. The percentage of sentences with each emotion label for the *Auditory* and *Lexical* conditions is shown in the Table 1. Based on only lexical information, non-neutral emotion labels *(fear, anger, satisfaction, excuse)* apply to 11% of the corpus, compared to 13% of the corpus when the audio and dialogic context are available. This comparison highlights the importance of lexical cues and dialog context in emotion detection.

3. EMOTION DETECTION MODEL

Our goal is to analyze the emotional behaviors observed in the dialog corpus in order to detect what, if any, lexical information is particularly salient to characterize each emotion. Our emotion detection system is based on a unigram model, as is used in the LIMSI Topic Detection and Tracking system [8]. 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 unigram emotion models were trained, one for each annotated emotion, using the set of on-emotion training utterances. Due to the sparseness of the on-emotion training data, 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. Stemming and stopping are commonly used procedures in information retrieval tasks for normalizing and removing 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 was used to derive the word stems. Experiments were carried out using different stop lists (containing from 60 to 200 entries) of high frequency words assumed to be uncorrelated with the task. 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 (abnormal)* can suggest that the person is upset about the situation. Since the corpus is quite limited, emotion balanced test sets were ramdomly selected using the lexically based reference annotations (5 sentences per emotion) following a jackknifing procedure. The remaining (about 5k) sentences were used for the training. Table 2 shows the emotion detection results for the baseline unigram system, and with the normalization procedures. Since

	Emotion Annotation				
Cond.	*Anger*	*Fear*	*Sat*	*Excuse*	*Neutral*
Auditory	5.0%	3.8%	3.4%	1.0%	86.8%
Lexical	2.3%	1.8%	5.8%	1.2%	88.9%

Table 1. Proportion of utterances in the corpus with each emotion label. *Auditory*: annotated by listening to the audio signal in the dialog context, *Lexical*: annotated using only the lexical information in the sentence.

[1] In the context of automatic call centers, joy and sadness are uncommon emotions, and have thus been excluded from the emotion set.

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

Table 2. Emotion detection performance of the baseline system and the baseline system with stemming and compounding. Results are given for the complete test set and the anger emotion subset.

Detected Emotion (%)					
Total	*Anger*	*Fear*	*Sat*	*Excuse*	*Neutral*
68	56	38	88	68	88

Table 3. Average emotion detection scores on two 125-sentence test sets.

the normalization procedures change the lexical forms, the number of words in the lexicon are also given. 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 seen to 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. Despite trying multiple stop-lists, stopping did not improve the detection rate.

Table 3 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*). In contrast, 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 4. 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)*) are too difficult to model with a unigram lexical model. Although quite meaningful humans, such expressions are relatively infrequent and quite varied, making them difficult to capture in a statistical model. Preliminary results using the simple lexical unigram model results in a detection rate of around 70% for a set of 5 task-dependent emotions.

4. PERCEPTUAL TESTS

Systematic and careful evaluations of emotion tagsets are generally lacking. In order to validate our emotion labels and to identify perceptual cues, perceptual tests were carried out using 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., the stimuli are the orthographic transcriptions of the extracted utterances). The [+signal] condition provides both linguistic and acoustic information and highlights the role of both sources of information. The tests consisted of naming the emotion present in each stimulus and of describing the linguistic cues ([-/+signal] conditions) and the acoustic cues ([+signal] condition) used to make the judgment.

Anger	*Fear*	*Sat*	*Excuse*
abnormal	worry	agree	mistake
"swear words"	fear	thanks	error
irritating	panic	perfect	sorry
embarrassing	afraid	excellent	excuse

Table 4. Examples of unambiguous lexical items for each emotion class.

4.1. Corpus and experimental protocol

The test stimuli consist of 40 sentences extracted from the dialog corpus, with 8 sentences for each of the 5 emotion classes. Five additional sentences (one per annotated emotion) were used in the training phase of the perceptual experiments. Forty native French subjects participated in one of the two tests: 20 for each condition. Two conditions for emotion annotation were used: one with free-choice, another with forced choice. The set of forced-choice emotion labels was enlarged with shaded emotions such as irritation and anxiety, and potential complex emotions such as surprise and doubt were added. In addition, the I-don't-know label allowed subjects to differentiate neutral attitude from ambiguous emotions. The same user interface was used for both tests, allowing a free choice for the linguistic cues and a forced choice for the prosodic ones.

4.2. Results

The free choice alternative resulted in two major strategies: (1) one label of forced choice or (2) a combination of forced choice and/or other label choices. The emerging forced-choice labels are irritation, anxiety and satisfaction, and the most frequent new labels are embarrassment and disappointment. Figure 1 summarizes the identification results for the perceptual experiments. On average the majority vote is obtained with an agreement of 55% for the ten classes. 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. 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%). One possible explanation is that politeness rules encourage callers to control the expression of the underlying emotion. Another possibility is that subjects may have associated voice quality rather than emotion with the audio

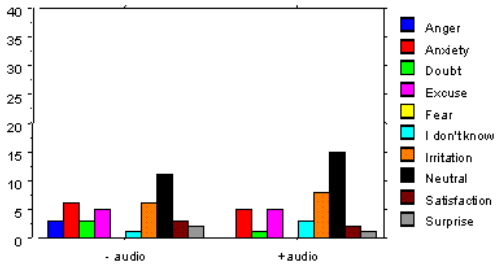

Fig. 1. Majority votes for the 10 emotion classes for both experimental conditions [-audio, +audio]. On average the majority of subjects agreed 55% of the time.

characteristics. 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. The audio condition provides a number of complex acoustic parameters which can influence the listener, including voice quality and the environmental conditions. In addition, the acoustic correlates of emotion in the human voice are subject to large individual differences. Prosody can add complementary information to the linguistic content of the message or can be contradictory and even modify the literal meaning of the sentence. In real-life dialogs, the context helps to correctly evaluate the interaction-dependent emotion, thus suggesting the importance of integrating dialog information to improve emotion recognition.

For the prosodic cues, the choices for the speech rate were: slow, normal and fast; for intensity: normal and high; and for pitch variation: flat or variable. The majority of subjects judged the speech rate as fast for irritation and satisfaction, whereas the pitch variation allowed subjects to distinguish neutral state and excuse (flat) from other emotional states (variable). 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 (which are often expressed with more prosodic clues than in realistic speech data), in the WOz experiments [1] and real agent-client dialogs the acoustic and prosodic cues are much less easily identifiable as callers may use multiple linguistic strategies. 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 made on the test corpus.

5. CONCLUSION AND PERSPECTIVES

As a result of the perceptual tests, we have selected 6 new shaded emotion tags: anxiety and irritation, embarrassment, disappointment, satisfaction and neutral attitude. Part of our ongoing work is the multilevel annotation of agent-client dialogs from another financial call center. An additional 250 dialogs have been annotated with the new tag set (10% of the utterances have non-neutral labels). This corpus will serve as a larger real-life corpus for training models. We are also comparing several algorithms for automatic extraction of prosodic features.

Despite the complexity involved, in our opinion it is crucial to work on real-life data. Emotion detection requires first identifying and validating emotion labels. Our studies indicate that for accurate emotion detection, lexical, prosodic, voice quality and contextual dialogic information need to be combined.

6. REFERENCES

[1] A. Batliner et al., "Desperately seeking emotions or: actors, wizards, and human beings", *ISCA ITRW Speech and Emotion*, 2000.

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

[3] L. Devillers et al., "Annotations for Dynamic Diagnosis of the Dialog State," *LREC'02*.

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

[5] D. Galati, B. Sini, "Les structures sémantique du lexique français des émotions", *Les émotions dans les interactions*, C.Plantin, M.Doury, V.Traverso (eds.), PUL 2000, ch 3.

[6] H. Hardy et al, "Multi-layer Dialogue Annotation for Automated Multilingual Customer Service", *ISLE Workshop on dialogue tagging*, Edinburgh, Dec 2002.

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

[8] Y.Y. Lo, J.L. Gauvain, "The Limsi Topic tracking system for TDT2001," *DARPA TDT'01* Nov. 2001.

[9] C.M. Lee et al., "Combining acoustic and language information for emotion recognition", *ICSLP*, 2002.

[10] R. Plutchik, *The psychology and Biology of Emotion*, Harper-Collins College, New York, 1994.

[11] K. Sherer et al., "Acoustic correlates of task load and stress," *ICSLP*, 2002.