Prosodic cues for perceptual emotion detection in task-oriented Human-Human corpus

Laurence Devillers[◇] and **Ioana Vasilescu**[♠] and **Catherine Mathon**[◇♣]

[◇] *LIMSI-CNRS 0rsay, France*
[♠] *ENST-CNRS, TSI, Paris, France*
[♣] *UFR Linguistique, Université Paris VII, France.*

ABSTRACT

This paper addresses the question of perceptual detection and prosodic cues analysis of emotional behavior in a spontaneous speech corpus of real Human-Human dialogs. Detecting real emotions should be a clear focus for research on modeling human dialog, as it could help with analyzing the evolution of the dialog. Our aims are to define appropriate emotions for call center services, to validate the presence of emotions via perceptual tests and to find robust cues for emotion detection. Most research has focused mainly on artificial data in which predefined-emotions were simulated by actors. For real-life corpora a set of appropriate emotion labels must be determined. To this purpose, we conducted a perceptual test exploring 2 experimental conditions: with and without the capacity of listening the audio-signal. Perceived emotions reflect the presence of shaded and mixed emotions/attitudes. We report correlations between objective values and both perceived prosodic parameters and emotion labels.

1 INTRODUCTION

In recent years there has been growing interest in the study of emotions [1, 3, 7] to improve the capabilities of current speech technologies (speech synthesis, speech recognition, and dialog systems). While different schools of thoughts, such as psychology, cognitive science, sociology and philosophy, have developed independent theories about personality and emotions [6, 9, 10], 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 [8] or emotion/-neutral state [1] or stressed/non-stressed speech [5]. With real-life data, the emotions are often considered as application-dependent [8]. Three main directions for emotion detection have been explored. The acoustic direction concerns the extraction of acoustic features from the speech signal (i.e., fundamental frequency, energy, speaking rate, etc.) which allow automatic detection of different emotions [3, 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, 8]. 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.

The 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. 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). Around 13% of the corpus when the audio and dialogic context are available, are annotated with marked emotion. Sentences with ambiguous labels ($\sim$ 3%) for those annotations were judged by a third independent annotator.

In a previous work [4], we have analyzed the emotional behaviors observed in this dialog corpus (around 5K speaker turns) in order to detect the type of lexical information particularly salient to characterize each

ISBN 1-876346-48-5

emotion. For this particular experiment, re-annotated sentences without considering the context and the audio signal were employed. Preliminary results using the simple lexical unigram model results in a detection rate of around 70% for a set of 5 task-dependent emotions. 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 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)*. From this initial dialog corpus, we selected 45 sentences representing the 5 emotion classes as material for the perceptual tests described in this paper.

The following sections describe perceptual tests conducted under two experimental conditions: with and without the capacity of listening the audio-signal. We present the choice of test corpus, experimental protocol and subjects participating in the experiment (section 2). In section 3, we focus on the results provided by subjects, i.e. perceived emotion labels and perceived linguistic (lexical, syntactic) and prosodic cues. Section 4 considers the correlation between perceived emotion labels and our initial annotation. In section 5, we compare the objective prosodic cues with emotion labels and perceived prosodic cues. Conclusion and further research are discussed in section 6.

2 PERCEPTUAL TEST

2.1 Experimental Protocol

Systematic and careful evaluations of emotion tag-sets are generally lacking. In order to validate an appropriate set of emotion labels and to identify perceptual cues, two tests were carried out using 45 sentences extracted of the global dialog corpus. 8 sentences for each of the 5 emotion classes (*Anger, Excuse, Fear, Neutral and Satisfaction*) have thus been extracted. Five additional sentences (one per annotated emotion) were used in the training phase of the perceptual experiments. The corpus covers a large range of possible spontaneous realizations in terms of topics, sentence lengths and types (interrogative, assertive etc.) and speaker characteristics (voice quality, sex etc.).

Two experimental conditions have been considered : 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 prosodic information in order to highlight 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 prosodic cues ([+signal] condition) used to make the judgment.

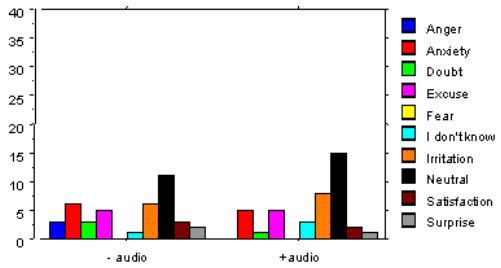

Figure 1: Majority votes for the 10 emotion classes for both experimental conditions [-audio, +audio]. On average the majority of subjects agreed 55% of the time.

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. Finally, 10 classes have been obtained (*Anger, Anxiety, Doubt, Excuse, Fear, I don't know, Irritation, Neutral, Satisfaction and Surprise*).

2.2 Subjects

40 native French subjects participated in one of the two tests: 20 for each condition. 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.

3 RESULTS

3.1 Perceived emotion labels

The free choice alternative resulted in categories with 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 labeled 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 expres-

Perceived prosodic cues [+ audio condition]		
Rate	Value	Emotion
	Slow	Doubt
	Normal	Anxiety
		Neutral
		Surprise
	Fast	**Irritation**
		Satisfaction
		Excuse
δF0	Value	Emotion
	Flat	**Neutral**
		Excuse
	Variable	Other emotions
E	Value	Emotion
	Normal, Low, High	All emotions

Table 1: Perceptual classification of main prosodic features.

Confusion Matrix					
Labels	*Anger*	*Fear*	*Sat.*	*Exc.*	*Neutral*
Nb sent.	*8*	*8*	*8*	*8*	*8*
Anger	-	-	-	-	-
Irritation	6	-	-	1	-
Surprise	1	-	-	-	-
Fear	-	-	-	-	-
Anxiety	-	6	-	-	-
Doubt	-	1	-	-	-
Excuse	-	-	-	5	-
Neutral	-	1	6	2	6
Satisfaction	-	-	1	-	1
I don't know	1	-	1	-	1

Table 2: Confusion Matrix between the emotional perceived labels (majority votes) versus initial emotion annotations for the 40 sentences of the test with auditory condition.

sion of the underlying emotion. Another possibility is that subjects may have associated voice quality rather than emotion with the audio 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.

3.2 Prosodic cues

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 (see Table 1). 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 objectively perceived prosodic 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 prosodic cues are much less easily identifiable as callers may use multiple linguistic strategies.

3.3 Linguistic cues (lexico-semantic, syntactic)

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).

4 PERCEIVED LABELS VS INITIAL EMOTION ANNOTATIONS

When comparing perceptual choices (10 classes) of subjects (see Table 2) with the previous contextual annotated labels (5 classes), we observe that *anger* and *fear* are identified as *irritation* and *anxiety*. This distribution is possibly due to the politeness rules and social conventions avoiding the extreme manifestation of emotions in ecological dialogs. 75% of negative shaded emotion (*irritation* and *anxiety*) are thus correctly perceived. There is no change in identifying the excuse (perceived choice vs. initial annotation). An interesting result is that the majority of the satisfaction in initial annotated corpus (5 classes) is found as neutral by subjects. We can explain this finding by satisfaction marks which generally indicate a normal dialog progression.

5 OBJECTIVE PROSODIC CUES

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.

The Praat software [2] has been employed for acoustic features detection. It is based on a robust algorithm for periodicity detection, working in the lag (auto correlation) domain. This algorithm is particularly adapted

ISBN 1-876346-48-5

Initial emotion annotations					
Labels	*Ang*	*Fea*	*Sat*	*Neu*	*Exc*
mean range F0	++	+	=	=	−
max δF0	++	=	=	=	−

Table 3: Trends for emotion effects on selected prosodic parameters correlated with initial annotations (5 classes). Symbols: ++: very high, +: high, =: medium, -: low. Ang= Anger, Fea= Fear, Sat= Satisfaction, Neu= Neutral, Exc= Excuse.

Perceived emotions					
Lab.	*Sur*	*Idk*	*Dou*	*Sat*	*Irr*
mean range F0	++	++	++	++	+
max δF0	++	++	++	+	+
Lab.	*Neu*	*Anx*	*Exc*	*Ang*	*Fea*
mean range F0	=	=	−	No	No
max δF0	=	=	−	No	No

Table 4: Trends for emotion effects on selected prosodic parameters correlated with perceived emotion (10 classes). Symbols: same as Table 3, No: No data. Sur= Surprise, Idk= I don't know, Dou= Doubt, Sa= Satisfaction, Irr= Irritation,Neu= Neutral, Anx= Anxiety, Exc= Excuse, Ang= Anger, Fea= Fear.

for noise condition (telephonic speech).

All studies point to the pitch as the main prosodic cues for emotion recognition. The other classical variables contributing to vocal emotion detection are vocal energy, formants, temporal feature such as speech rate and pausing. For our study, we estimated: pitch (F0) (min, max, mean, range, standard deviation), mean energy and speaking rate (number of syllables per second). For pitch calculation, only voiced regions were taken into account. We also calculated the maximum cross-variation of pitch between two adjoining segments of voicement. The Table 3 and 4 give some results for F0 variation.

Table 3 and 4 allow to observe the main trends in selected prosodic features variation according to different emotions. Table 3 focuses on the prosodic features variation correlated with initial emotion annotation (5 classes). We can notice a strong difference between values in negative emotion manifestation vs *neutral, satisfaction* and *excuse*, thus comforting the findings provided by the literature. When correlating objective values and perceived emotion labels (10 classes), two main groups oppose (*surprise, doubt, satisfaction, irritation, I-don't-know*) to (*neutral, excuse, anxiety*). The difference between neutral vs non neutral emotion classes is less important because of the redistribution of the stimuli among some of the new classes (*doubt, surprise, I-don't-know*). For example, a stimulus predicted as anger has been perceived as ambiguous even if showing high variation of F0 values. In addition, we can notice a correlation between perceived cues (table 1) and prosodic objective values (table 4) for the F0 variation.

Concerning the other parameters such as speech rate, *satisfaction* and *irritation*, they show higher values, whereas for energy there is no objective difference.

Obviously, given the large variability of prosodic features in a real-life corpus recorded in noise conditions (telephonic speech), more data is needed to validate trends presented and obtain robust cues.

6 CONCLUSION

Emotion detection requires first identifying and validating task-dependent emotion labels. Two main emotional behaviors emerge from the perceptual tests in both conditions: *irritation* and *anxiety*. In addition, the *excuse attitude* was identified as an agent behavior directly associated to the task. Correlations between F0 variation and emotion labels (annotated and perceived) have been found validating the prosodic cues provided by the literature for negative vs positive/neutral emotions. However emotions have complex manifestations integrating several linguistic levels and/or non linguistic markers. Our ongoing work focuses on validating the perceptual findings on a larger corpus. Further work will be to explore the combination of emotion information conveyed by the textual information and the contextual dialogic information with prosodic features.

REFERENCES

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

[2] P. Boersma, "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", '*IFA Proceedings*, 1993, p 97-110.

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

[4] L. Devillers, I. Vasilescu, L. Lamel, "Annotation and Detection of Emotion in a Task-oriented Human-Human Dialog Corpus", *ISLE Workshop on dialogue tagging*, Edinburgh, Dec 2002.

[5] R. Fernandez, R. Picard, "Modeling Drivers' Speech Under Stress," *Speech Communication*, 2003.

[6] 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.

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

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

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

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