Reliability of Lexical and Prosodic Cues in two Real-life Spoken Dialog Corpora[1]

L. Devillers [(1)], I. Vasilescu[(2)]

[(1)] LIMSI-CNRS , BP133, 91 403 Orsay Cedex, France, [(2)] LTCI-ENST, 46, rue Barrault, 75013 Paris, France
devil@limsi.fr, vasilescu@tsi.enst.fr

Abstract

The present research focuses on analyzing and detecting emotions in speech as revealed by task-dependent spoken dialogs corpora. Previously, we have conducted several experiments on a real-life corpus in order to develop a reliable annotation method and to detect lexical and prosodic cues correlated to the main emotion class. In this paper we evaluate both the robustness of the annotation scheme and of the lexical and prosodic cues by testing them on a new corpus. This work is carried out in the context of the Amities project in which spoken dialog systems for call center services are being developed.

Introduction

Emotion manifestations in real-life spoken corpora are particularly complex and dependent of the subject of the interactions. As a result, the literature on emotions shows that annotations strategies and detection cues are dependent on the corpus specificities (Sherer, 2003, Douglas-Cowie & al, 2003). Our aim is to define a generic method for emotion annotation and cues detection.

Our research focuses on analyzing and detecting emotions in speech as revealed by task-dependent spoken dialogs corpora. We have previously conducted several experiments on a real-life corpus (Devillers & al, 2003a, Devillers & Vasilescu, 2003). Emotion detection in spontaneous speech is a part of a larger study aiming at modeling user behavior in human-machine interactions. Detecting emotions in the context of automated call center services can be helpful for the management of the human-computer dialogs, enabling dynamic modification of the dialog strategy according to the user behavior and influencing the final outcome. We make use of two corpora of Agent-Client spoken dialogs recorded in French call centers. The recordings were made for purposes independent of this study, and have been made available for use in developing an automated call routing service within the context of the Amities project[1]. In both corpora, the manifestation of emotion is complex, i.e. shaded emotions occur since the interlocutors attempt to control the expression of their internal attitude.

In this paper we study different emotion annotation schemes and discuss about a generic annotation scheme founded on abstract dimensions. We also evaluate the robustness of lexical and prosodic cues identified in a first corpus, by testing them on a second corpus.

The following section describes the two corpora employed in the study. In the third, we describe the annotation strategies and describe the emotion labels adopted, the inter-annotators agreement measures and the perceptual validation of the labels. In the fourth section, we focus on the lexical cues detection, followed by the description of the prosodic cues. We compare the two types of cues on CORPUS 1 vs. CORPUS 2. Finally we discuss on this study.

Real-life call centers corpora

The first corpus (CORPUS 1) analyzed in previous experiments has been recorded at a 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 concern problems in using the Web interface 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. 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 related questions and problems. 100 dialogs have been annotated with emotion tags by two annotators (5K speaker turns). Overlaps representing additional 1,1K speaker turns have been eliminated.

The second corpus (CORPUS 2) of real agent-client dialogs has been recorded in Capital Bank Service Center. The dialogs cover personal accounts management topics such as opening/closure of an account, loans, transaction fees etc. 250 dialogs have been annotated with emotion tags by an annotator. Among the 250 dialogs, 5K speaker turns have been extracted for lexical and prosodic parameters estimation. As for CORPUS 1, overlaps (1,5K speaker turns of the global corpus of 250 dialogs) have been eliminated. Given the important amount of time needed to annotate the entire corpus, for this study uniquely 1K speaker turns randomly extracted from the CORPUS 2 are annotated by two raters.

Dialogs are generally longer in CORPUS 1 than in CORPUS 2. Both corpora have been orthographically transcribed and annotated at multiple levels: semantic, topic, dialogic and emotional behaviors in Amities project.

Globally speaking, there is a similarity in terms of general task between the two corpora, i.e. financial telephonic transactions. As the emotions are task-dependent, we expect that similar emotion labels will be able to define CORPUS 1 and CORPUS 2, and the analogous selected lexical and prosodic cues will characterize them.

[1] This work was partially financed by the European Commission under the IST-200-25033 AMITIES project http://www.dcs.shef.ac.uk/nlp/amities.

Consequently, the comparison will allow estimating the general reliability of the parameters. In addition and with the same purpose, we evaluate on CORPUS 1 the emotion labels by using an annotation scheme previously confirmed on two other corpora (Cowie & al, 2001; Craggs & Wood, 2004).

Annotation strategies

Corpora annotation

A task-dependent annotation scheme was initially developed for emotion annotation in CORPUS 1, keeping in mind that the basic affective disposition towards a computer is generally either trust or irritation. Two negative of the four classical emotions (Anger, Fear, Joy and Sadness) are retained as appropriated for both corpora: Anger and Fear. However, in the call center dialogic contexts, most of Anger and Fear manifestations are shaded emotions such as irritation for Anger, and worry or anxiety for Fear. Among the labels, we also considered Agents' and Clients' behaviors directly associated with the task in order to capture some of the dialog dynamics. For this purpose, Satisfaction (greetings) and Excuse (embarrassment) were included as emotion labels. Both labels correspond to a particular class of the speech acts (expressive acts) as described in the classical version of pragmatic theory. However, Excuse class is poorly represented and will not be considered in this study. The Neutral state, i.e. the normal progression of the dialog, is also considered. Finally, about 12% of the utterances are annotated with non-neutral emotion labels (13% for the first corpus, 11% for the second one). In this paper we focus mainly on negative emotions Anger and Fear. We consider them in opposition to Positive class of emotions containing Satisfaction and Neutral.

Inter-annotators agreement

We intended at standardising the annotation by calculating the Kappa coefficient which shows the inter-annotators reliability (Carletta, 1996). Kappa measures the agreement between a number of annotators by comparing the number of times they agree upon a label for an object (speaker turns for our corpora) against the number of disagreements. The results is a value between 0 to 1, where zero equates to the level of agreement that might be expected if annotators behaved randomly and one represents perfect agreement.
The Kappa coefficients obtained for our corpora are 0.8 for CORPUS 1 and 0.5 for CORPUS 2. We can explain this difference by the subject of the dialogs which allow more emotion manifestations in CORPUS 1 than in CORPUS 2. Indeed, in CORPUS 1 the interaction Clients/Agents focusing on stock exchange aspects causes longer dialogs with stronger emotion effects in Clients and Agents turns (Devillers & Vasilescu, 2004). In the new corpus (CORPUS 2) the subjects of the dialogs concern financial transactions in which Agents show globally neutral behaviours and Clients are thus more moderate. Accordingly, the annotators encountered more problems in annotating the second corpus due to the slighter variability in terms of emotion effects (Amities Project, 2002).

Perceptual validation

A first perceptual experiment has validated the emotion labels employed in CORPUS 1 and encouraged at adopting the same strategy for the annotation of CORPUS 2 (Devillers & al, 2003b).

Furthermore, we aimed at finding a more general annotation strategy. In this purpose, we make use of an experimental method adopted by researchers on emotion detection in spoken dialogs which employs a 2-dimensional annotation scheme.
This scheme is based on two abstract dimensions, Activation-Evaluation, suggested as salient for emotion categorization (Cowie & al., 2001). It represents an adaptation of the original theory developed by (Osgood & al., 1975). According to Osgood, the communication of affect is conceptualized following three major dimensions of connotative meaning: arousal (Activation), pleasure (Evaluation) and power (Control). From the original 3-dimensional scheme, 2-dimensions, Activation (passive vs. active emotions) and Evaluation (positive vs. negative emotions) have been retained as salient for emotion categorization in spoken dialogs (Cowie & al, 2001). We adopt here the scheme as employed by (Craggs & Wood, 2004) in which the Activation is substituted by Intensity. The Intensity is considered a more intuitive dimension for the naive annotators and more appropriated to describe the level of emotions in an utterance or a speaker turn. The Evaluation axis covers discrete values from wholly negative (-3, -2, -1) to wholly positive (1, 2, 3). The Intensity axis provides 5 levels from 0 to 4. Level 0 corresponds to Neutral both for Evaluation and Intensity. The annotation tool and the instructions are described in (Craggs & Wood, 2004).

For this second experiment the same 40 speaker turns (8 per initial emotion class) as for the test described in (Devillers & al, 2003) and extracted from CORPUS 1 has been employed. Additional 5 speaker turns have been provided as training phase and thus not taken into consideration. 10 raters annotated each speaker turn after listening to them as many times as they preferred. Thus, we expect annotators employed lexical and prosodic cues to decide the values on the Evaluation/Intensity axis.

Figure 1 below shows the percentage of Evaluation labels for negative emotions Anger and Fear. As a general observation, stimuli for Anger and Fear are placed in the negative region of the Evaluation axis. The difference concerns the magnitude of the perceived negative values, i.e. stronger for Anger (maximum value: -2) than for Fear (maximum value: -1).

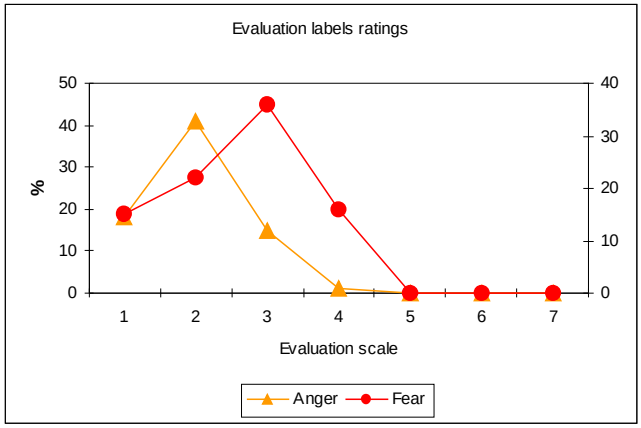

Figure 1: *Percentage for evaluation labels ratings for Anger and Fear. Coding for the initial grid: 1=-3, 2=-2, 3=-1, 4=0, 5=1, 6=2, 7=3.*

The results obtained with the 2-dimensional annotation scheme validate the emotion labels adopted to annotate CORPUS 1 for negative emotions. Indeed, the speaker turns initially labeled as Anger and Fear are perceived as negative emotions on the Evaluation axis. The results on the Intensity axis are not relevant in discriminating Anger from Fear and thus they are not discussed here.
It confirms also the first perceptual experiment and the choice of emotion labels. However, the 2-dimensional scheme does not allow differentiating inside the class of Negative or Positive emotions, but uniquely between the two classes, i.e. Negative vs. Positive emotions. More specifically, it does not allow differentiating between Anger and Fear. Or for our particular application, it is crucial to make a clear distinction between those two negative emotions. Moreover, acoustic and prosodic cues as well as perceptual findings confirm that the distinction between the two emotions is possible (Devillers & Vasilescu, 2004, Devillers & al, 2003b). A possible issue to this question could be the use of the 3d dimension as defined by Osgood, Power (Control). This dimension marks the relation between the speakers producing utterances, the emotion perceived in a given utterance and the stimulus causing the emotion.

Lexical cues detection

Previous emotion detection experiments have been carried out at several levels on CORPUS 1. Initially, we investigated the role of lexical level in automatic emotion detection. The emotion detection system is based on a unigram model, as is used in the LIMSI Topic Detection and Tracking System (Devillers & al, 2003a). The similarity between an utterance and an emotion is the normalized log likelihood ration 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 training corpora for CORPUS 1 and CORPUS 2. An interpolation coefficient 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*.

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

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 procedures are commonly used in information retrieval tasks for normalizing words in order increase the likelihood that the resulting terms are relevant. We have adopted this technique for emotion detection. Since the corpora are quite limited, emotion balanced test sets were randomly selected using the lexically based reference annotations (25 utterances per emotion) following a jackknifing procedure. The remaining sentences were used for the training. We compare emotion detection by using the unigram model on CORPUS 1 and CORPUS 2. We differentiate between Negative (Anger and Fear) and Positive emotions. Models are trained on each corpus.
The results on the test sets (average on 10 test sets) show a better score for the first corpus. Thus, 84% of good detection for Negative vs. Positive emotions is obtained for CORPUS 1, 75% for CORPUS 2. As an additional experiment, we also evaluate the best lexical model (obtained on CORPUS 1) on the test set of CORPUS 2. We notice 78% of correct detection showing the reliability of the lexical cues for these tasks. We can explain this slightly better result by the presence of more robust lexical cues for Negative emotions in CORPUS 1.

Prosodic cues detection

As a second step, we focus on emotion detection using prosodic parameters (F0 features) (Devillers & Vasilescu, 2003 and 2004). PRAAT (Boersma, 1993) has been used to extract F0 features on voiced regions. 1.4% of shorts segments (<40ms) have been considered detection errors and eliminated. These errors are homogenously distributed among the emotion classes in CORPUS 1 and CORPUS 2. The F0 measures are considered for each speaker turn. The z-score normalization method has been used. It is computed by removing the mean obtained over all values of a speaker in a dialog and dividing by the corresponding standard deviation. Five F0 parameters have been estimated for CORPUS 1 and CORPUS 2: MaxDF0, F0Range, MinF0, MaxF0 and MeanF0. All the parameters are not equally salient as shown in tables 2 and 3. We compare the F0 measures on CORPUS 1 and CORPUS 2. We consider uniquely F0 parameters values for Clients' turns which presents emotion manifestations in the two corpora. Agents' turns are not emotionally marked in CORPUS 2.

Emotion	CORPUS 1	CORPUS 2
Anger	216	82
Fear	161	213
Satif.	61	62
Neutral	1921	1832

Table 1: *Clients' turns repartition according to the main emotion classes and the two corpora.*

Previously, we have shown that at the prosodic level two main F0 range parameters differentiate between Negative vs. Positive emotions on the CORPUS 1. The two parameters are the F0 range (sentence level) and the Max delta F0 (MaxDF0), calculated between 2 adjoining voiced segment (segmental level).

Emotion	MaxDF0	F0Range	MinF0	MaxF0	MeanF0
Anger	**133**	**237**	99	**338**	156
Fear	**137**	**249**	88	**348**	153
Satisf.	65	174	107	282	159
Neutral	82	190	104	295	153

Table 2: *Mean values (in Hz) for emotions effects for F0 parameters correlated with Negative and Positive emotion class for CORPUS 1.*

In Table 2, three F0 parameters (MaxDF0, F0Range and MaxF0) divide emotion in two groups, Negative (Anger and Fear) and Positive (Neutral and Satisfaction). MeanF0 and MinF0 do not show difference in emotion distinction.

Emotion	MaxDF0	F0Range	MinF0	MaxF0	MeanF0
Anger	**128**	212	71	284	147
Fear	**144**	**236**	62	300	150
Satisf.	69	164	83	248	151
Neutral	121	216	69	286	151

Table 3: *Mean values (in Hz) for emotions effects for F0 parameters correlated with Negative and Positive emotion class for CORPUS 2.*

In Table 3 MaxDF0 and F0range show same trends as CORPUS 1 for Fear whereas they are rather similar for Neutral and Anger. F0 parameters are globally more emerging for Fear than for Anger for both CORPUS 1 and CORPUS 2. They are also globally lower for CORPUS 2 than for CORPUS 1 compared to Neutral/Positive emotions.

Discussion

The experiments on lexical and prosodic cues conducted for this study show a difference in magnitude between CORPUS 1 and CORPUS 2. More precisely, CORPUS 1 is more marked at all levels. We can explain these findings by the differences in dialogs subjects in the two corpora. In CORPUS 2, the subjects of the dialogs concern personal accounts management. The interactions between Agents and Clients consist mostly in temperate exchanges of information. Anger is thus less present and less marked. Indeed, Clients deal with personal financial resources and when a problem occurs they are generally worried. In addition, Agents and Clients are both polite and obey to social rules which avoid extreme manifestations. On the contrary, in CORPUS 1 Clients and Agents deal with stocks transactions. The negative emotions give an idea about the amount of stress the topic itself produces. The negative emotions are the consequence of the failure in managing this stress by Agents and Clients. As a general behavior, they are less polite and more hurried in their interactions as the topic needs quick reactions. Anger is more present and the negative emotions in general are more marked.

Consequently, we estimate the annotation by emotion labels is not robust enough to reflect the differences in the cues magnitude even if the dialogs topics are quite similar. As mentioned in the literature, a systematic description could be to represent the emotions as coordinates in a space with a small number of dimensions. As a first experiment, annotation on CORPUS 1 is evaluated with a method based on abstract 2-dimensional scheme. The scheme validates the main emotion classes Negative and Positive but does not allow making finer distinction among the emotions of a same class. Further studies on the number of the abstract dimensions are necessary to obtain a more reliable method. Studies will be carried on the two corpora.

The final aim is to define a hierarchy of reliable cues for a general detection model and to elaborate a complex model in which different levels of information are taken into account.

References

Amities project: http://www.dcs.shef.ac.uk/nlp/amities/

Boersma, P, (1993), "Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound", IFA Proc., 1993, pp 97-110.

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

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, Fellens, W, Taylor, J., (2001), "Emotion recognition in human-computer interaction". IEEE Signal Processing Magazine 18.

Craggs, R., McGee Wood, M., (2004), "A two dimensional annotation scheme for emotion in dialogs", AAAI Proc.

Devillers, L., Vasilescu, I., Lamel, L., (2003), "Emotion Detection in a task-oriented Dialog Corpus", IEEE International Confererence on Multimedia ICME 2003, Baltimore (a).

Devillers, L., Vasilescu I., Mathon, C., (2003), "Acoustic cues for perceptual emotion detection in task-oriented Human-Human corpus", 15th International Congress of Phonetic Sciences, Barcelone (b).

Devillers, L., Vasilescu, I., (2003), "Prosodic cues for emotion characterization in real-life spoken dialogs, Eurospeech", Geneva.

Devillers, L., Vasilescu, I., (2004), "Anger versus Fear detection in recorded conversations", Speech Prosody, Nara, Japon.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., (2003), "Emotional speech; Towards a new generation of databases". Speech Communication.

Osgood, C., May, W.H., Miron, M.S., (1975), "Cross-cultural Universals of Affective Meaning". University of Illinois Press, Urbana.

Scherer, K., (2003), "Vocal communication of emotion: A review of research paradigms", Speech Communication.