ANNOTATION AND DETECTION OF BLENDED EMOTIONS IN REAL HUMAN-HUMAN DIALOGS RECORDED IN A CALL CENTER[1]

Laurence Vidrascu, Laurence Devillers

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

ABSTRACT

In the context of call centres, emotion detection is potentially important for customer care. Emotions in natural interaction are often blended. For example, in a Stock Exchange service centre, some customers are angry because they are afraid to lose money. A 100 agent-client dialog corpus has been annotated at the speaker turn level with one label among 5 emotions including Fear and Anger. In this paper, we report on our experiments of automatic emotion detection using acoustic cues with several classifiers. 73% correct detection was achieved in discriminating between Negative and Neutral emotions and 60% between Anger and Fear. An analysis of the confusions led us to question the validity of the initial single valued annotation scheme. It was found that customer emotional states can be a mixture of Anger and Fear. As a result a new annotation scheme is used allowing the selection of two verbal labels per segment.

1. INTRODUCTION

This paper reports experiments on automatic emotion detection using real-life human-human dialogs recorded in a call center. This work is carried out in the context of the IST FP-5 AMITIES Project in which spoken dialog systems for call center services are being developed. We are also involved in the FP-6 HUMAINE NoE "*HUman-MAchine Interaction Network on Emotions*". Recently, there has been a growing interest in the study of real-life emotions [1, 2, 9] to improve the capacities of current speech technologies: recognition, synthesis and dialog. Detecting emotion in call centers can help by orienting evolution of the human-computer interaction via dynamic modification of dialog strategies.

Compared to acted corpora*,* real-life corpora are still under-represented in the research [8] despite the fact that results based on studies with acted stereotypical emotional states transfer poorly to real applications [2]. A major difficulty of using natural corpora is that the expression of emotion is quite complex. Indeed, realistic emotional states are often blended, shaded or masked. Furthermore, they are strongly dependent of contextual and social factors.

One of the main challenges in emotion detection in real-life speech is the categorization and annotation phase. Three types of emotion annotation are generally used: appraisal dimensions, abstract dimensions and most commonly verbal categories.

The appraisal theory [13] provides detailed specification of appraisal dimensions that are assumed to be used in evaluating emotion-antecedent events (pleasantness, novelty, etc). The major methodological problem is that the only reliable way of ensuring a correct annotation is to ask the persons themselves to perform it. Two other types of emotion description are also given by Cowie [4], the speaker emotion based on his internal emotional state named "cause-type" and the effect that he would be likely to have on the listener, named "effect-type". Asking annotators to figure out "how a speaker is feeling" is very subjective and likely to lead to erroneous annotations. Therefore, on speech data recorded in a natural context, we can only annotate the emotion that is expressed in the speech content.

According to Osgood [10], the communication of emotion is conceptualized following three major dimensions of meaning: Activation, Evaluation and Control. So, some researchers define emotions as continuous abstract dimensions instead of naming emotions as discrete categories. The widely employed scheme is based on two perceptive abstract dimensions: Activation-Evaluation. It has been employed for the annotation of several corpora [8]. However, this scheme does not allow for instance to distinguish between Fear and Anger.

Actually, most of the emotion detection studies have only focused on verbal categories using a minimal set of emotions to be tractable [1, 2, 9]. The challenges in the annotation of real-life data is thus to define a pertinent and limited set of labels and an adequate annotation scheme.

In our study we make use of a corpus of real agent-client dialogs recorded in French, for independent purposes, at a Stock Exchange Customer Service Center.

[1] This work was partially financed by the European Commission under the IST-2000-25033 Amities project

0-7803-9332-5/05/$20.00 ©2005 IEEE

The 100 agent-client dialog corpus has been annotated at the speaker turn level with one simple emotional label chosen among Anger, Fear, Excuse, Satisfaction and Neutral labels. In previous studies on this annotated corpus, we have found that during perceptual tests subjects are able to differentiate Anger and Fear with 75% accuracy [7]. We have also analysed correlations between F0 and disfluencies Features (pause and filler-pause) and emotion labels and context (gender and position-type in the dialog) [5]. In those experiments, the acoustic features set selected seems to be sufficient for discriminating between Neutral and Negative emotions but insufficient to separate Fear and Anger.

In this paper, we report on experiments of automatic emotional states detection using acoustic parameters for Negative to Neutral emotion detection and inside the Negative class between Anger and Fear emotion with several classifiers (SVM, decision trees, boosting and bagging techniques). Making assumption that a part of the previously Anger and Fear annotated segments is certainly blended emotions; we have conducted a re-annotation phase providing the possibility to use two emotional labels per segment. A first analysis on these new annotations is reported.

In the following sections, we describe the corpus (section 2), the two emotional annotation schemes (section 3) and the features and classifiers (section 4). Section 5 presents detection results obtained on Neutral/Negative emotion classification, then Fear/Anger emotion detection with the first annotation scheme and the annotations analysis with the second annotation scheme. Conclusions and further research are discussed in section 6.

2. DATA

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. The corpus of 100 agent-client dialogs in French has been orthographically transcribed and annotated. A part of the corpus has been annotated with two different annotation schemes. There are 6241 speaker turns. We considered 5000 sentences after excluding overlaps, which are known to be frequent phenomena in spontaneous speech.

3. EMOTION ANNOTATION

The high subjectivity of human annotation requires the use of rigorous methods of annotation. We illustrate two different strategies for annotation. For both, instead of having many naive annotators (at least 3) and then using a majority voting technique to create reference, two expert annotators have labeled the corpus.

3.1. First annotation scheme

A first task-dependent annotation scheme was developed, adopting, for practical reasons, a selection strategy of a minimal number of labels. Two of the four classical emotions were retained: Anger, Fear. In this application, most of Anger and Fear manifestations are shaded emotions such as irritation for Anger, and anxiety for Fear. We also considered some of the agents' and clients' behaviours directly associated with the task: Satisfaction and Excuse. The Neutral attitude is also added.

Two annotators independently listened to the 100 dialogs, labelling each of the sentences with one of the 5 classes. In order to assess the consistency of the selected labels, we conducted perceptive tests and calculated the inter-annotation agreement (Kappa value is 0.8). Ambiguous labels concern 2.7% of the global corpus and most often involved indecision between neutral state and other emotion. Sentences with ambiguous labels (19% of the sentences labelled with non-neutral emotion labels) were judged by a third independent listener in order to decide on the final label. Based on the auditory classification, sentences with non-neutral labels comprise about 13.2% of the entire corpus (Table 1). For detection experiments, only Fear, Anger and Neutral classes will be considered.

	Anger	Fear	Sat	Excuse	Neutral
Client	9.9	6.7	2.6	0.1	80.7
Agent	0.7	1.3	4.0	1.8	92.1

Table 1 *Proportion of each emotion label in the dialog corpus labelled by listening to the audio signal.*

3.2. New annotation scheme

In order to deal with blended emotions, annotators have the possibility of selecting a second label, called the "minor" if he perceives two different emotions for an utterance such as Anger and Fear, Anxiety and Relief. The labels were chosen from a list of 50 emotions (resulting from studies of the NoE HUMAINE). Five people appraised their relevance in the context of call centres and a list of 20 labels was derived using majority voting between 5 persons. Those labels can be grouped in 8 broad classes: Neutral, Sadness, Fear, Anger, Embarrassment, other Negative, Empathy, other Positive.

To evaluate this new protocol, all utterances previously annotated as Negative emotion, Fear or Anger, were re-annotated in context by two expert annotators, different from those of the first study. For this task, the annotators used mostly labels from the classes Neutral, Fear, Anger, Embarrassment, Surprise and Positive. The two annotators chose the same major emotion in 64% of the cases and 13% utterances were ambiguous (no common label between the 2 annotators). As a first experiment, we convert the decisions of the two labelers into an emotion vector. The vector size is the number of Emotional classes per annotator (Table 2). Different weights were given to the major emotion (w_M) and the minor emotion (w_m) and the two vectors were added. W is the sum of the weights.

Segment annotation: **Labeler 1** *Major* Anger, *Minor* Fear **Labeler 2:** *Major* Anger, *Minor* Other Neg **Conversion into an emotion vector:** (w_M/W **Anger**, w_m/W **Fear**, w_m/W **OtherNeg)**

Table 2 *Conversion of the decisions of two labelers into an emotion vector. For w_M=2, w_m=1, W=6, (0. 66 Anger, 0. 165 Fear, 0.165 OtherNeg).*

4. FEATURES AND CLASSIFIERS

Because emotions in real-life interaction have complex manifestations, we have focused on the extraction of several types of parameters. Previous work was done on emotion detection based on lexical cues [6]. For acoustic cues extraction, we have used the Praat program [3]. It is based on a robust algorithm for periodicity detection, working in the lag auto-correlation domain. Praat has been used to extract F0 features on voiced regions. As a lot of errors are made on F0 feature detection, we have used a filter to try to eliminate some errors. 1.4% of short segments (< 40 ms) have been considered detection errors and eliminated. These errors are homogeneously distributed among all the 5 classes. In our current experiments we have a big and redundant feature set and use a classifier to select the relevant features.

4.1. Features

The set of features includes:

F0 features (z-score normalization method): min, max, mean, standard deviation, range on the turn level, slope (mean and max) of the voiced segments, regression coefficient and its mean square error (performed on the voiced parts as well) and also maximum cross-variation of F0 between two adjoining voiced segments.

Energy features (normalization phase): min, max, mean, standard deviation, and range on the turn level.

Duration features: speaking rate: inverse of the average length of the speech voiced parts, number and length of silences (unvoiced parts between 200-800 ms)

Spectral features: formants and their bandwidth.

Disfluency features: We proceeded to an automatic alignment of the orthographic transcription with the acoustic signal in order to extract disfluency cues such as filler and abnormal large silent pauses. 4.6% of utterances have not been automatically aligned. All the utterances labelled with negative emotions have been manually verified in order to avoid alignment errors. The parameters extracted are pauses features per utterance (normalized by the length of the utterance), mean and maximum duration of silence, number of filler pauses ("euh" in French).

4.2. Features set reduction

The aim is to get rid of the noise and reduce the complexity of space of attributes without affecting the performances. Two kinds of methods are very common in data mining: to select the more relevant attributes or to apply linear transformations to reduce the dimension of the data.

4.3. Classifiers

Several classifiers and classification strategies from the machine learning literature tested in our study are listed:

The Decision Tree is a set of rules in a structure of nodes and leaves, with each node representing a test and each leaf a class. The algorithm **C4.5** [12] uses pruning.

Support Vector machines (SVM) [14] algorithms search an optimal hyperplan to separate the data

Voting algorithms (AddTree and AdaBoost) combine the outputs of different models. Boosting methods affect weights to models with better performances.

5. RESULTS

5.1. Neutral/Negative emotion detection

All experiments were done using jack-knifing with 5 subsets (4 subsets are used for training and one for test, the experiment is repeated 5 times with each subset being used for test). This procedure is repeated 6 times with different subsets.

	C4.5	**AdaBoost**	**ADTree**	**SVM**
5att	72.8 (5.2)	71.2 (4.5)	72.3(4.6)	67.2(6.3)
10att	73.0 (5.3)	71.5(4.8)	73.0(5.7)	69.5(5.6)
15att	71.7 (6.4)	71.1(4.7)	71.6(4.9)	70.8(4.9)
20att	71.8 (5.3)	71.3(4.3)	71.8(5.1)	71.0(4.9)
allatt	69.4 (5.6)	71.7(4.3)	71.6(4.8)	69.6(3.5)

Table 3 *Algorithms and attribute selection: comparison of the Neutral/Negative (Fear and Anger) detection performances with the best features; Allatt: all attributes. This table shows the average of correctly classified instances for the 30 runs. The number into parenthesis is the standard deviation.*

The results of our experiments on Neutral/Negative emotion detection, given in Table 3, have been obtained using features set described in section 4 minus disfluency parameters. As shown in the literature [2], the results obtained on Neutral/Negative emotion detection show little differences between these techniques**.**

5.2 Fear/Anger emotion detection

With the first annotation scheme, we obtain better results for Negative/Neutral classification (73%) than for Anger/Fear classification (56%). We obtained around 60% adding disfluencies to the same set of features. That could be explained by the fact that Fear and Anger in this financial task are combined: « Clients are angry because they are afraid of losing money ». When we look at the confusion matrix, there are as many "Fear utterances classified as Anger" as "Anger utterances classified as Fear".

We wanted to see if our second annotation scheme would enable us to account for those ambiguities. Because we focused on Anger and Fear, an emotion vector (Fear, Anger) was computed from the new annotations by giving a weight of 2 to the major emotion and one of 1 to the minor one as shown in section 3.2. 4 classes were deduced from these vectors: Fear (Fear>0; Anger=0), Anger (Fear=0; Anger>0), Blended emotion between Fear and Anger (Fear>0; Anger>0) and Other (0,0) (see figure 1). 40% of the utterances are thus labelled as blended.

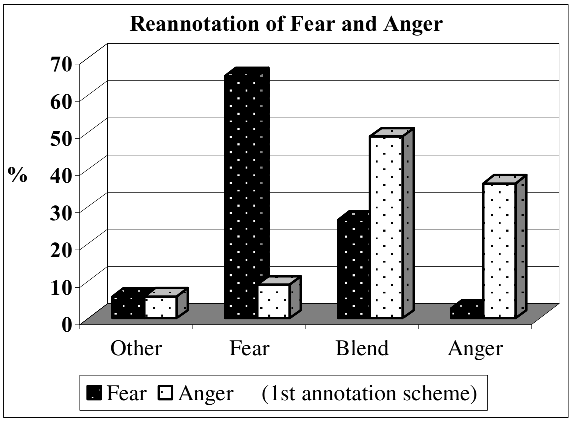

Figure 1 *Repartition of the utterances previously labelled Anger and Fear after the new annotation scheme*

We checked that the second annotation was consistent with the first annotation. For an utterance initially labeled as Fear, we considered that the second annotation was equivalent if the first field of our (Fear, Anger) vector was positive and superior or equal to the second one (idem for Anger). This is the case for 78% of the utterances. As with the classifier's confusion matrix, the re-annotation shows clearly a class of blend emotion Anger and Fear for this task.

6. CONCLUSION

73% correct detection was achieved in discriminating between Negative and Neutral emotions and 60% between Fear and Anger. Analysis of results led us to propose a new annotation scheme for complex natural emotion. This scheme allowing two labels per segment enables the apparition of classes representing complex blended emotions. This annotation scheme is actually used to annotate a 20 hour-corpus recorded in another real call center. Experiments on this larger corpus are under way and will provide complementary results on blend classes that hopefully will improve emotion detection performance for natural data.

7. REFERENCES

[1] Ang, J.; Dhillon, R.; Krupski, A.; Shriberg, E.; Stolck. A.: Prosody-Based Automatic Detection of Annoyance and Frustration in H-M Dialog. ICSLP, (2002).

[2] Batliner, A.; Fisher, K.; Huber, R.; Spilker, J.; Noth, E.: Desperately seeking emotions or: Actors, wizards, and human beings. SpeechEmotion-2000: 195-200 (2000).

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

[4] Cowie, R.: Describing the emotional states expressed in speech. In SpeechEmotion (2000).

[5] Devillers, L., Vasilescu, I., Vidrascu, L.: Anger versus Fear detection in recorded conversations, Speech Prosody 2004, Nara, Japon (2004).

[6] Devillers, L., Vasilescu, I. Lamel, L.: Emotion detection in task-oriented dialogs corpus, IEEE ICME Baltimore, (2003).

[7] Devillers, L.; Vasilescu, I.; Mathon, C.: Acoustic cues for perceptual emotion detection in task-oriented human-human corpus. 15th ICPhS, Barcelona Spain (2003)

[8] Douglas-Cowie, E.; Campbell, N.; Cowie, R.; Roach, P.: Emotional speech; towards a new generation of databases, Speech Communication 40 (2003).

[9] Lee, C.M.; Narayanan, S.; Pieraccini, R.: Combining acoustic and language information for emotion recognition, ICSLP, (2002).

[10] Osgood, C.; May, W.H.; Miron, M.S.: Cross-cultural Universals of Affective Meaning, Urbana, (1975).

[11] Plutchik, R.: The psychology and Biology of Emotion, HarperCollins College, New York, (1994).

[12] Quinlan, JR.: C4.5: Programs for Machine Learning, Morgan Kaufman, (1993).

[13] Sherer, K.: Emotion, In M. Hewstone & W. Stroebe (Eds.), Introduction to Social Psychology (2000).

[14] Vapnik, V.N.: The Nature of Statistical Learning Theory, Springer, (1995).