Real-life emotions detection with lexical and paralinguistic cues on Human-Human call center dialogs

Laurence Devillers and Laurence Vidrascu

LIMSI-CNRS – BP 133, 91 403 Orsay cedex, France

{devil, vidrascu}@limsi.fr

Abstract

The emotion detection work reported here is part of a larger study aiming to model user behavior in real interactions. We already studied emotions in a real-life corpus with human-human dialogs on a financial task. We now make use of another corpus of real agent-caller spoken dialogs from a medical emergency call center in which emotion manifestations are much more complex, and extreme emotions are common. Our global aims are to define appropriate verbal labels for annotating real-life emotions, to annotate the dialogs, to validate the presence of emotions via perceptual tests and to find robust cues for emotion detection. Annotations have been done by two experts with twenty verbal classes organized in eight macro-classes. We retained for experiments in this paper four macro classes: Relief, Anger, Fear and Sadness. The relevant cues for detecting natural emotions are paralinguistic and linguistic. Two studies are reported in this paper: the first investigates automatic emotion detection using linguistic information, whereas the second investigates emotion detection with paralinguistic cues. On the medical corpus, preliminary experiments using lexical cues detect about 78% of the four labels showing very good detection for Relief (about 90%) and Fear (about 86%) emotions. Experiments using paralinguistic cues show about 60% of good detection, Fear being best detected.

Index Terms: emotion detection, real-life emotion, lexical and paralinguistic cues

1. Introduction

The emotion detection work reported here is part of a larger study aiming to model user behaviour in real interactions. We have already worked on other real life data: financial call centers [1] and EmoTV clips [2]. In this paper, we make use of a corpus of real agent-client spoken dialogs in which the manifestation of emotion is stronger [1, 3]. The context of emergency gives a larger palette of complex and mixed emotions. About 11% of the utterances are annotated with non-neutral emotion labels on financial data compared to 30% in the medical corpus. Emotions are less shaded than in the financial corpus where the interlocutors attempt to control the expression of their internal attitude. In the context of emergency, emotions are not played but really felt in a natural way. In contrast to research carried out with artificial data with simulated emotions or with acted data, for real-life corpora the emotions are linked to internal or external emotional event(s). We might think that natural and complex emotion behaviour could be found in movies data. Yet, emotions are still played and in most cases, except for marvellous actors, they are not really "felt". However, it is also of great interest to study professional movie actors in order to portray a recognisable emotion and to define a scale of naturalness [4]. The difference is mainly due to the context. The context is the set of events that are at the origin of a person's emotions... Different events might trigger different emotions at the same time: for instance a physical internal event as a stomachache triggering pain with an external event as "someone helping the sick person" triggering relief. In the artificial data, this context is "rubbed out" or simulated so that we can expect to have much more simple full-blown affect states which are far away from real affective states.

In contrast to research carried out with artificial data and simulated emotions, for real-life corpora the set of appropriate emotion labels must be determined. There are many reviews on the representation of emotions. For a recent review, the reader is referred to [6]. We have defined in the context of Humaine NoE, an annotation scheme "Multi-level Emotion and Context Annotation Scheme" [1, 2] to represent the complex real-life emotions in audio and audiovisual natural data. It is a hierarchical framework allowing emotion representation at several layers of granularity, with both dominant (Major) and secondary (Minor) labels and also the context representation. This scheme includes verbal, dimensional and appraisal labels. Our aim in this study is to find robust lexical and paralinguistic cues for emotion detection.

One of the challenges when studying real-life speech call center data is to identify relevant cues that can be attributed to an emotional behavior and separate them from those that are simply characteristic of spontaneous conversational speech. A large number of linguistic and paralinguistic features indicating emotional states are present in the speech signal. Among the features mentioned in the literature as relevant for characterizing the manifestations of speech emotions, prosodic features are the most widely employed, because as mentioned above, the first studies on emotion detection were carried out with acted speech where the linguistic content was controlled. At the acoustic level, the different features which have been proposed are prosodic (fundamental frequency, duration, energy), and voice-quality features [6]. Additionally, lexical and dialogic cues can help as well to distinguish between emotion classes [1, 7, 8, 9, 10]. Speech disfluences have also been shown as relevant cues for emotion characterization [11] and can be automatically extracted. Non-verbal speech cues such as laughter or mouth noise are also helpful for emotion detection. The most widely used strategy is to compute as many features as possible. All the features are, more or less,

correlated with each other. Optimization algorithms are then often applied to select the most efficient features and reduce their number, thereby avoiding making hard a priori decisions about the relevant features. Trying to combine the information of different natures, paralinguistic features (prosodic, spectral, disfluences, etc) with linguistic features (lexical, dialogic), to improve emotion detection or prediction is also a research challenge. Due to the difficulty of categorization and annotation, most of the studies [7, 8, 9, 10, 12] have only focused on a minimal set of emotions.

Two studies are reported in this paper: the first investigates automatic emotion detection using linguistic information, whereas the second investigates emotion detection through paralinguistic cues. Sections 2 and 3 describe the corpus and the adopted annotation protocol. Section 4 relates experiments with respectively lexical and paralinguistic features. Finally, in the discussion and conclusion (section 5), the results obtain with lexical and paralinguistic are compared and future research is discussed.

2. The CEMO Corpus

The studies reported in this paper make use of a corpus of naturally-occurring dialogs recorded in a real-life medical call center. The dialog corpus contains real agent-client recordings obtained from a convention between a medical emergency call center and the LIMSI-CNRS. The use of these data carefully respected ethical conventions and agreements ensuring the anonymity of the callers, the privacy of personal information and the non-diffusion of the corpus and annotations. The service center can be reached 24 hours a day, 7 days a week. The aim of this service is to offer medical advice. The agent follows a precise, predefined strategy during the interaction to efficiently acquire important information. The role of the agent is to determine the call topic, the caller location, and to obtain sufficient details about this situation so as to be able to evaluate the call emergency and to take a decision. In the case of emergency calls, the patients often express stress, pain, fear of being sick or even real panic. In many cases, two or three persons speak during a conversation. The caller may be the patient or a third person (a family member, friend, colleague, caregiver, etc.). Table 1 gives the caracteristics of the CEMO corpus.

Table 1. ***CEMO*** *corpus characteristics: 688 agent-client dialogs of around 20 hours (M: male, F: female)*

#agents	7 (3M, 4F)
#clients	688 dialogs (271M, 513F)
#turns/dialog	Average: 48
#distinct words	9.2 k
#total words	262 k

The transcription guidelines are similar to those used for spoken dialogs in previous work [1]. Some additional markers have been added to denote named-entities, breath, silence, intelligible speech, laugh, tears, clearing throat and other noises (mouth noise). The transcribed corpus contains about 20 hours of data. About 10% of speech data is not transcribed since there is heavily overlapping speech.

3. Emotion annotation

Representing complex real-life emotion and computing inter-labeler agreement and annotation label confidences are important issues to address. A soft emotion vector is used to combine the decisions of the two annotators and represent emotion mixtures [1, 2]. This representation allows to obtain a much more reliable and rich annotation and to select the part of the corpus without conflictual blended emotions for training models. Sets of pure emotions or blended emotions can then be used for testing models. In this experiment utterances without emotion mixtures were considered.

The set of labels is hierarchically organized (see Table 2) from coarse-grained to fine-grained labels in order to deal with the lack of occurrences of fine-grained emotions and to allow for different annotator judgments.

Table 2. *Emotion classes hierarchy: multi-level of granularity*

Coarse level (8 classes)	**Fine-grained level (20 classes + Neutral)**
Fear	Fear, Anxiety, Stress, Panic, Embarrassment, Dismay
Anger	Anger, Annoyance, Impatience, ColdAnger, HotAnger
Sadness	Sadness, Disappointment, Resignation, Despair
Hurt	Hurt
Surprise	Surprise
Relief	Relief
Other Positive	Interest, Compassion, Amusement
Neutral	Neutral

The annotation level used to train emotion detection system can be chosen based on the number of segments available. The repartition of fine labels (5 best classes) only using the emotion with the highest coefficient in the vector [1] is given Table 3.

Table 3. *Repartition of fine labels (688 dialogues). Other gives the percentage of the 15 other labels. Neu: Neutral, Anx: Anxiety, Str: Stress, Hur: Hurt, Int: Interest, Com:Compassion, Sur: Surprise, Oth: Other.*

Caller	Neu.	Anx.	Str.	Rel.	Hur.	Oth
10810	67.6%	17,7%	6.5%	2.7%	1.1%	4.5%
Agent	Neu..	Int.	Com..	Ann..	Sur.	Oth
11207	89.2%	6.1%	1.9%	1,7%	0.6%	0.6%

The Kappa coefficient was computed for agents (0.35) and clients (0.57). Most confusion is between a so-called “neutral state” and an emotional set. Because we believe there can be different perceptions for a same utterance, we considered an annotator as coherent if he chooses the same labels for the same utterance at any time. We have thus adopted a self re-annotation procedure of small sets of dialogs at different time (for instance once a month) in order to judge the intra-annotator coherence over time. About 85% of the utterances

are similarly re-annotated [1]. A perception test was carried out [13]. Subjects have detected in a part of the corpus complex mixtures of emotions within different classes both of the same and of different valence. The results validate our annotation protocol, the choice of labels and the use of a soft vector to represent emotions.

4. Classification

Our long-term goal is to analyze the emotional behaviors observed in the linguistic and paralinguistic material of the human-human interactions present in the dialog corpus in order to detect what, if any, lexical information or paralinguistic is particularly salient to characterize each of the four emotions selected. Several classifiers and classification strategies well described in the machine learning literature are used to classify prosodic and lexical.

For this study, four classes at the coarse level have been considered: Anger, Fear, Relief and Sadness (see Table 4). We only selected utterances of callers and non-mixed emotions for this first experiment.

Table 4. *Train and test corpus characteristics*

Corpus	Train	Test
#Speaker turn	1618	640
#Speakers	501(182 M, 319F)	179(60M, 119F)
Anger	179	49
Fear	1084	384
Relief	160	107
Sadness	195	100

4.1. Lexical cues

Our emotion detection system is based on a unigram model, as used in the LIMSI Topic Detection and Tracking system. The lexical model is a unigram model, where the similarity between an utterance and an emotion is the normalized log likelihood ratio between an emotion model and a general task-specific model (eq. 1). Four 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 λ=0.75 was found to optimize the results of CL and RR. The emotion of an unknown sentence is determined by the model yielding the highest score for the utterance u, given the emotion model E.

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

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 to increase the likelihood that the resulting terms are relevant. We have adopted this technique for emotion detection. The training is done on 501 speakers and the test corresponds to 179 other speakers. Table 5 relates experiments with a stemming procedure and without a normalization procedure (the baseline).

Table 5. *Emotion detection with lexical cues*.

	Baseline		Stemming	
Size of lexicon	2856		1305	
lambda	RR	CL	RR	CL
0.65	62.7	47.5	75.9	67.1
0.75	**66.9**	**47.5**	**78.0**	**67.2**
0.85	67.5	44.4	80.3	64.6

Table 6. *Repartition for the 4 classes for stemming condition and lambda = 0.75. Utt: Utterances, A: Anger, F: Fear, R: Relief, S: Sadness. RR: Overall Recognition rate, CL: Class-wise averaged recognition rate*

Stemming	Total	A	F	R	S
#Utt	640	49	384	107	100
% rec.	**78**	59	90	86	34

Table 6 shows the emotion detection results for the baseline unigram system, and with the normalization procedure of stemming. Since the normalization procedures change the lexical forms, the number of words in the lexicon is also given.

Results are given for the complete test set and for different λ. Using the baseline system, emotion can be detected with about 67% precision. Stemming is seen to improve the detection rate, we obtained around 78% of recognition rate (67.2 for class-wise averaged recognition rate). The results in Table 6 show that some emotions are better detected than others, the best being the Fear class and the worst Sadness. Anxiety is the main emotion for the callers. The high detection of Relief can be attributed to strong lexical markers which are very specific to this emotion ("thanks", "I agree"). In contrast, the expression of Sadness is more prosodic or syntactic than lexical in this corpus. The main confusions are between Fear and Sadness, and Fear and Anger.

4.2. Paralinguistic cues

A crucial problem for all emotion recognition systems is the selection of the set of relevant features to be used with the most efficient machine learning algorithm. In the experiments reported in this paper, we have focused on the extraction of prosodic, spectral, disfluency and non-verbal events cues, The Praat program [14] was used for prosodic (F0 and energy) and spectral cue extraction. About a hundred features are input to a classifier which selects the most relevant ones:

- **F0 and Spectral features** (Log-normalized per speaker): *min*, median, first and third quartile, *max*, *mean*, *standard*

deviation, *range* at the turn level, *slope* (mean and max) in the voiced segments, *regression coefficient* and its *mean square error* (performed on the voiced parts as well), *maximum cross-variation of F0* between two adjoining voiced segments (inter-segment) and with each voiced segment(intra-segment), position on the time axis when F0 is maximum (resp. minimum), *ratio* of the number of voiced and non-voiced segments, *formants* and their *bandwidth*, difference between third and second formant, difference between second and first formant: min, max, mean, standard deviation, range.

- **Microprosody** : jitter, shimmer, NHR, HNR
- **Energy features** (normalized): *min*, *max*, *mean*, *standard deviation* and *range* at the segment level, position on the time axis when the energy is maximum (resp. minimum), .
- **Duration features**: *speaking rate* (inverse of the average length of the speech voiced parts), number and length of *silences* (unvoiced portions between 200-800 ms).
- **Disfluency features**: number of *pauses and filled pauses* ("euh" in French) per utterance annotated with time-stamps during transcription.
- **Non linguistic event features**: *inspiration*, *expiration*, *mouth noise laughter*, *crying*, and *unintelligible voice*. These features are marked during the transcription phase.

The above set of features are computed for all emotion segments and fed into a classifier. The same train and test are used as for the classifier based on the lexical features. Table 7 shows the emotion detection results using a SVM classifier.

Table 7. *Repartition for the 4 with a SVM classifier. A: Anger, F: Fear, R: Relief, S: Sadness*

	Total	A	F	R	S
#Utterances	640	49	384	107	100
%rec.	**59,8**	39	64	58	57

As for lexical results, the Fear is best detected (64%). The Anger is worst detected (39%) while still above chance. It is mostly confused with Fear (37%). This might be due to the fact that Fear is often in the background.

5. Discussion and Conclusion

We have obtained about 78% and 60% of good detection for respectively lexical and paralinguistic cues on four real-life emotion classes. Both results were better for Fear/Anxiety detection, which is the most frequent emotion in the corpus and occurs with different intensity (anxiety, stress, fear, panic). Because Anger recognition is very low with the paralinguistic model and Sadness is low with the lexical model, we believe there might be a way to combine the two models and yield better results. Thus, future work will be to combine information of different natures: paralinguistic features (prosodic, spectral, disfluences, etc) with linguistic features (lexical), to improve emotion detection or prediction. Comparison with our previous results on lexical, paralinguistic and combined cues on other call center data will be done in a next future.

6. Acknowledgements

The work is conducted in the framework of a convention between the APHP France and the LIMSI-CNRS. The authors would like to thank the Professor P. Carli, the Doctor P. Sauval and their colleagues N. Borgne, A. Rouillard and G. Benezit. This work was partially financed by NoE HUMAINE.

7. References

[1] Devillers L., Vidrascu L. & Lamel L. (2005). Challenges in real-life emotion annotation and machine learning based detection, Journal of Neural Networks 2005, special issue: Emotion and Brain, vol18, Number 4, 407-422.

[2] Devillers, L., Abrilian, S., Martin, J.-C (2005). Representing real life emotions in audiovisual data with non basic emotional patterns and context features, ACII.

[3] Vidrascu L., Devillers, L. (2005). Real-life Emotions Representation and Detection in Call Centers, ACII.

[4] Clavel C., Vasilescu I., Devillers L., Ehrette T., (2004). Fiction database for emotion detection in abnormal situation, ICSLP 2004.

[5] Cowie, R. & Cornelius, R.R (2003). Describing the emotional states expressed in speech, Speech Communication, 40(1-2), 5-32.

[6] Campbell, N. (2004). Accounting for Voice Quality Variation, Speech Prosody 2004, 217-220.

[7] Batliner, A., Fisher, K., Huber, R., Spilker, J. & Noth, E. (2003). How to Find Trouble in Communication. Journal of Speech Communication, 40, 117-143.

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

[9] Forbes-Riley, K. & Litman, D. (2004). Predicting Emotion in Spoken Dialogue from Multiple Knowledge Sources. Proceedings of HLT/NAACL.

[10] Devillers, L., Vasilescu, I. & Lamel, L. (2003). Emotion detection in task-oriented dialog corpus. Proceedings of IEEE International Conference on Multimedia

[11] Devillers, L., Vasilescu, I., & Vidrascu, L. (2004). Anger versus Fear detection in recorded conversations . Proceedings of Speech Prosody. 205-208.

[12] Steidl, S., Levit M., Batliner, A., Nöth, E. & Niemann, E. (2005). Off all things the measure is man Automatic classification of emotions and inter-labeler consistency, Proceeding of the IEEE ICASSP.

[13] Vidrascu, L. Devillers L., (2006). Real-life emotions in naturalistic data recorded in a medical call center, Workshop on Emotion, LREC 2006.

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