Real-life emotions detection on Human-Human spoken dialogs

L. Devillers
LIMSI-CNRS
devil@limsi.fr

L. Vidrascu
LIMSI-CNRS
vidrascu@limsi.fr

Abstract

In the paper we present emotional annotation for a corpus of naturalistic data recorded in a French Medical call center. When studying real-life data, there are few occurrences of full blown emotions but also there are many emotion mixtures. To represent emotion mixtures, an annotation scheme with the possibility to choose two verbal labels per segment was used by 2 expert annotators. A closer study of these mixtures has been carried out, revealing the presence of conflictual valence emotions. Results of the perceptive test show 85% of consensus between expert and naive labellers. When selecting the non-complex part of the annotated corpus, the performances obtained are around 60% of good detection between four emotions for respectively agents and callers.

Keywords: Emotions, real-life spoken interactions, detection system, medical call center

1 Introduction

This decade has seen an upsurge of interest in affective computing. Speech and Language are among the main channels to communicate human affective states. Affective Speech and language processing can be used alone or coupled with other channels in many systems such as call centers, robots, artificial animated agents for telephony, education, medical or games applications. Affective corpora are then fundamental both to developing sound conceptual analyses and to training these 'affective-oriented systems' at all levels - to recognise user affect, to express appropriate affective states, to anticipate how a user in one state might respond to a possible kind of reaction from the machine, etc. Our aim is to study the vocal expression of "emotion" in real-life spoken interactions in order to build emotion detection system.

In the computer science community, the widely used terms of emotion or emotional state are used without distinction from the more generic term affective state which may be viewed as more adequate from the psychological theory point of view. This "affective state" includes the emotions / feelings / attitudes / moods / and the interpersonal stances of a person. There is a significant gap between the affective states observed in artificial data (acted data or contrived data produced in laboratories) and those observed in real-life spontaneous data.

Most of the time, researches are done on a sub-set of the big-six "basic" emotions described by Ekman [1] and on prototypical acted data. In the artificial data, the context is "rubbed out" or "manipulated" so we can expect to have much more simple full-blown affect states which are quite far away from spontaneous affective states.

The affective state of a person at any given time is a mixture of emotion/ attitude/ mood /interpersonal stance with often multi-trigger events (internal or external) occurring at different times: for instance a physical internal event as a stomach-ache triggering pain with an external event as "someone helping the sick person" triggering relief. Thus, far from being as simple as "basic emotion", affective states in spontaneous data are a subtle blend of many more complex and often seemingly contradictory factors that are very relevant to human communication and that are perceived without any conscious effort by any native speaker of the language or member of the same cultural group.

L. Magdalena, M. Ojeda-Aciego, J.L. Verdegay (eds): Proceedings of IPMU'08, pp. 1590–1596
Torremolinos (Málaga), June 22–27, 2008

The first challenge when studying real-life speech data is to find the set of appropriate descriptors attributed to an emotional behaviour. For a recent review of all emotion representation theories, the reader is referred to the Humaine NoE (www.emotion-research.net). Several studies define emotions using continuous abstract dimensions: Activation-Valence or Arousal-Valence-Power. But these three dimensions do not always enable to obtain a precise representation of emotion. For example, it is impossible to distinguish fear and anger. According to the appraisal theory [2], the perception and the cognitive evaluation of an event determine the type of the emotion felt by a person. Finally, the most widely used approach for the annotation of emotion is the discrete representation of emotion using verbal labels enabling to discriminate between different emotions categories. We have defined in the context of Humaine, an annotation scheme "Multi-level Emotion and Context Annotation Scheme" [3, 4] to represent the complex real-life emotions in audio and audiovisual natural data. This scheme is adapted to each different task. We are also involved as expert in the W3C incubator group on emotion representation.

The second challenge 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. The aim is that to extract the main voice characteristics of emotions, together with their deviation which are often present in real spontaneous interaction. 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 [5]. Additionally, lexical and dialogic cues can help as well to distinguish between emotion classes [3, 7, 8, 9]. 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, 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 have only focused on a minimal set of emotions.

In this study, we verify that emotional behaviour is very often complex in real-life data. We also show that by using a large number of different features, we can improve performances obtained with only classical prosodic features for emotion detection. Section 2 describes the corpus of real-life data. Section 3 is devoted to the perceptive test on complex data. Section 4 is the description of the features used. In section 5, the methods for training models are briefly described. Section 6 summarizes our results which are then discussed.

2 Real-life data

In the context of emergency, emotions are not played but really felt in a natural way. The aim of the medical call center 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, and 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.).

he corpus (Table 1) contains 688 agent-client dialogs of around 20 hours (271 males, 513 females, the number of speakers is different of the number of dialogs because several persons can be involved in the same dialogs). The corpus has been transcribed following the LDC transcription guideline.

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.

Table 1: Corpus Description
NB: the number of speakers is different of the number of dialogs because several persons can be involved in the same dialogs

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

Some additional markers (Table 2) have been added to denote named-entities, breath, silence, intelligible speech, laugh, tears, clearing throat and other noises (mouth noise).

Table 2: Number of the main non-speech sounds markings on 20 hours of spontaneous speech.

#laugh	119
#tear	182
# « heu »	7347
#mouth noise	4500
#breath	243

In our experiment, we define one list of emotion labels using a majority voting technique. A first list of labels was selected out of the fusion several lists of emotional labels defined within HUMAINE (European network on emotion http://emotion-research.net/). In a second step, several judges rated each emotion word of this list with respect to how much it sounded relevant for describing emotions present in our corpus.

We have defined an annotation scheme "Multi-level Emotion and Context Annotation Scheme" [3, 4] 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 (Table 3), with both dominant (Major) and secondary (Minor) labels and also the context representation. This scheme includes verbal (from the predefined list), dimensional and appraisal labels. 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 several coders and represent emotion mixtures [3, 4]. This representation allows to obtain a much more reliable and rich annotation and to select the part of the corpus without blended emotions for training models. Sets of "pure" emotions or blended emotions can then be used for testing models. About 30% of the utterances are annotated with non-neutral emotion labels on this medical corpus (Table 4).

Table 3: Emotion classes hierarchy: multi-level of granularity

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

Table 4: Repartition of fine labels (688 dialogues). Other gives the percentage of the 15 other labels. Neu: Neutral, Anx: Anxiety, Ann: Annoyance, Str: Stress, Rel: Relief, 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 (measuring the inter-labeler agreement) was computed for agents (0.35) and callers (0.57). The following experiments have been carried out on the callers' voices for the coarse classes: Fear, Anger, Sadness, Relief and a "Neutral" state.

3 Perceptive test on complex emotions

The main goal of the test was to appraise the presence of emotion mixtures and see if the lack of context would hinder their perception. 41 segments were selected including 14 “simple” segments (annotated with one emotion by both annotators), 27 emotion mixtures (including 13 positive/negative emotion mixtures). A segment is a portion of speech data associated with one label or two labels for blends. The context is the dialog. A typical case of a positive/negative blend for an agent is to feel both annoyance and compassion towards a caller (or try to hide his irritation with a compassionate voice). For the client, recurring examples were to feel worry coupled with relief from knowing help is on its way. The stimuli were given without context but could be replayed indefinitely. The test was run with 43 subjects [12]: 33 French native people (13F/20H) and 10 non native French speakers. The subjects were asked to choose a label for the emotion they perceived in the list of 20 labels + neutral. In the case when a second emotion was perceived, they had to choose it from the same list. Because of the absence of context and of the large number of labels, this task felt very difficult, especially for the non native subjects. Most of them were unable to specify mixtures of labels. Thus we only considered native subjects when studying emotion mixtures.

Evaluation Results per subject

Without the context, every native French subject perceived some occurrences of emotion mixtures and all but 2 subjects among them perceived mixtures of a positive and negative emotion. Table 5 shows for each subset (simple emotion, conflictual (positive/negative) and non conflictual (negative/negative)) the percentage given by naive users.

Table 5: Percentage of simple and complex emotions (non conflictual and conflictual)

Annotated as->	**Simple/ Ambiguous**	**Non conflictual**	**Conflict.**
Simple (14 seg)	87%	7%	6%
non conflictual (11 seg)	76%	19%	5%
Conflictual(13 seg)	71%	10%	18%

For 28% of the conflictual sample, people were able to perceive emotion mixtures (mainly conflictual ones). In the other hand, there were still 13% which were judged as complex when annotated as simple. In this study, women perceived more conflictual mixtures than men. These poor results show the difficulty of the perception of these samples without context.

Evaluation Results per vector

Even when subjects individually chose one label, the vector combining the annotations of all 43 subjects appears to correspond to the vector of the 2 expert labelers. Indeed, when comparing the two highest coefficients of the vectors for expert annotators and naive annotators, there is an agreement of 85% between the two annotations. 70% of the complex emotions were detected (9 segments out of the 11 non conflictual and 9 out of the 13 conflictual have the same 2 coarse emotions). Errors often involve relief that out of context is labeled as fear. The cases where experts and naive annotators disagreed were often accounted for by the context.

The following experiments only used non complex emotions. Our later goal will be to take into account these complex data.

4 Features

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 recent research, a lot of different sets and classifiers have been used. However, the best features set and the most efficient model are still not well established and from published results appear to be data-dependent.

Mainly prosodic features (fundamental frequency (F0) and Energy) are classical features used in a majority of experiments on emotion detection. For accurate emotion detection in natural real-world speech dialogs, not only the prosodic information must be considered.

We use non-verbal speech cues such as speech disfluencies and affect bursts (laugh, tear, etc.) as relevant cues for emotion characterization. For example, we considered the autonomous main French filler pause "euh" as a marker of disfluency. It occurs as independent item and it has to be differentiated from vocalic lengthening. We correlate the filler pause with emotions in [10]. This correlation follows the orthographic

(lexical) transcription of the dialogs and considers the number of occurrences of transcribed "euh" per emotion class. In [10], "euh" was correlated mainly with Fear sentences, followed by Anger sentences and finally the other emotions. In [11], affect bursts such as laughter or mouth noise are shown to be also helpful for emotion detection.

Since there is no common agreement on a top list of features and the feature choice seems to be data-dependent, our usual strategy is to use as many features as possible even if many of the features are redundant, and to optimize the choice of features with attribute selection algorithms. In the experiments reported in this paper, we divided the features into several types with a distinction between those that can be extracted automatically without any human intervention (prosodic, spectral features, microprosody) and the others (duration features after automatic phonemic alignment, features extracted from transcription including disfluencies and affect bursts).

Our set of features includes very local cues (such as for instance the local maximums or inspiration markers) as well as global cues (computed on a segmental unit) [13]. In Table 6, we summarize the different types of features and the number of cues used in our experiments.

We distinguish the following sets of features:

- "Blind": *automatic features extracted only from audio signal including paralinguistic features (prosodic, micro-prosodic, formants)*

The Praat program [14] was used for the extraction of prosodic (F0 and energy), micro-prosody and spectral cues. It is based on a robust algorithm for periodicity detection carried out in the lag auto-correlation domain. Since F0 feature detection is subject to errors, a filter was used to eliminate some of the extreme values that are detected. Energy, spectral cues and formants were only extracted on voice parts (i.e.: parts where Praat detects F0). The paralinguistic features were normalized using Z-norm: zNorm(P) = (P-mean(P))/std(P). The aim is to erase speaker-differences without smoothing variations due to emotional speech.

- *"Trans1": duration features from phonemic alignment*

For the moment we only extracted Duration features from the phonetic transcription, mean and maximum phone duration, phonemic speech rate (#phones/ turn length), length (max and mean) of hesitations.

- *"Trans2"*: *features extracted from the transcription*

Non linguistic event features: inspiration, expiration, mouth noise laughter, crying, number of truncated words and unintelligible voice. These features are marked during the transcription phase.

Table 6: Summary of the feature types

	Feature type		# of cues
Blind	F0 related		25
	Energy		20
	Spectral & formant related	Bandwidths	18
		Formants	30
	Micro-prosody		14
Trans1	Duration features from phonemic alignment		11
Trans2	Speech disfluencies an affect burst from transcription		11

5 Methods

For training the paralinguistic model, we use the Weka machine learning software [16]. Weka is a collection of machine learning algorithms for data mining tests; it contains tools for preprocessing, classification, regression and clustering. The following approaches have been compared for the paralinguistic model: decision trees (C4.5) [17], Support Vector Machine (SVM) [15] and Voting algorithms (AdTree) [18] and (AdaBoost)[19]. The best performances have been obtained with SVMs. In this paper, we only present the results with SVM.

The Support Vector machine algorithm searches an optimal hyperplan to separate the data. The

formulation embodies the Structural Risk Minimisation (SRM) principle, which has been shown to be superior to traditional Empirical Risk Minimization (ERM) principle, employed by conventional neural networks. The SVM used in our experiments are based on Gaussian kernels (Gaussian Radial Basis Functions).

With only blind features and without any knowledge about the speech transcription, we obtained a detection rate of 45% on these 5 emotions. Still, the more emotional classes there are, the more different cues will be needed to achieve good detection rates. By adding knowledge (Fig. 1) derived from the orthographic transcription (disfluencies, affect bursts, phonemic alignment) and after the selection of the best 25 features, we achieved 56% of good detection for the same 5 emotions.

Features from all the types were selected among the 25 features: 15 features in the Blind set, 4 in Trans1 and 6 in Trans2. The experiments described in Fig. 2 focus on a task of discriminating 2 to 5 emotions among Fear, Anger, Sadness, Neutral and Relief.

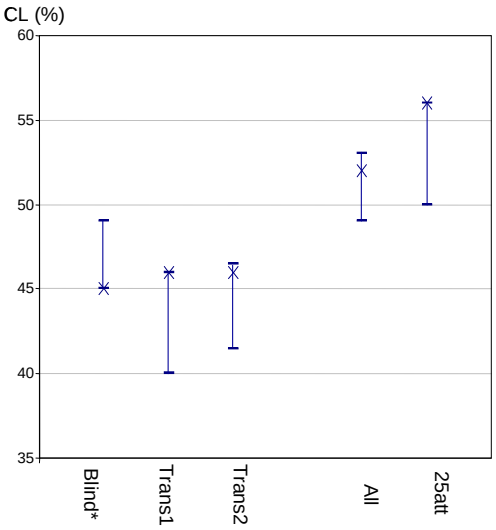

Figure 1*:* CL score for the 5 classes Fear, Anger, Sadness, Relief and Neutral with different set of cues; Blind: all parameters extracted automatically (F0, Formants, Energy, micro-prosody); Trans1: durations from phonemic alignment, Trans2: parameters extracted from the manual transcription, all: everything 25-best : 25 best features

The complexity of the recognition task increases the higher the number of classes and the finest and closest these classes are. For only two emotions (such as Anger/Neutral or Fear/Neutral), we obtained with our best system more than 80% of good detection.

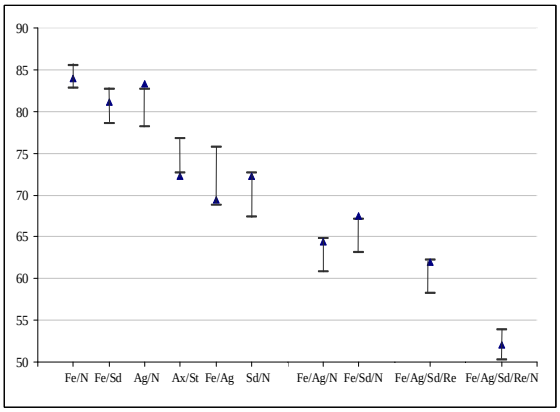

Figure 2: Performances from 2 emotions to 5 emotions (Fe: Fear, N: Neutral state, Ag: Anger, Sd: Sadness, Re: Relief)

6 Conclusion

In conclusion, finding relevant features of various types becomes essential in order to improve the emotion detection performances on real-life spontaneous data. Some of these features such as affect burst or disfluencies could be detected automatically without any speech recognition. Future experiments will be devoted to the automatic detection of such features.

As shown in our perceptive test, there are several emotion mixtures in real-life data. Until now, we have not exploited the emotion soft-vector representation for training emotion detection system. Our perspective is to build a hierarchical structure of several emotion detection system based on different cues to deal with the detection of emotion mixtures.

Acknowledgements

This work was partially financed by several EC projects: FP5-Amities, FP6-CHIL and NoE HUMAINE. The work is conducted in the framework of a convention between a medical call center in France and the LIMSI-CNRS.

References

[1] P. Ekman (1999) "Basic emotions." In *Handbook of Cognition & Emotion*, 301–320. New York: John Wiley.

[2] K. Scherer (1999) Appraisal Theory. In: Dalgleish, T. Power, M. (Eds), *Handbook of Cognition and Emotion*. John Wiley, New York, 637-663.

[3] L. Devillers, L, Vidrascu & L. Lamel (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.

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

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

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

[7] T. Vogt, E. André, (2005) "Comparing Feature Sets for Acted and Spontaneous Speech in View of Automatic Emotion Recognition", *ICME 2005*.

[8] A. Batliner, et al, "Whodunit — Towards the most Important Features Signaling Emotions in Speech: a Case Study on Feature and Extraction Types", *ACII 2007*.

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

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

[11] M. Schröder, "Experimental study of affect bursts", *Proc. ISCA workshop "Speech and Emotion"*, Newcastle, Northern Ireland, 2000, p 132-137.

[12] L. Vidrascu L., Devillers, (2006) Real-life emotions in naturalistic data recorded in a medical call center, *LREC 2006*.

[13] L. Vidrascu L., Devillers, (2007) Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features, *paraling07*.

[14] P. Boersma, (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*, 1993, p 97-110.

[15] V.N. Vapnik, (1995), The Nature of Statistical Learning Theory, *Springer-Verlag*, 1995.

[16] I.H. Witten, et al., (1999) "Weka: Practical machine learning tools and techniques with Java implementations", *Proc ANNES '99* p 192-196.

[17] L. Breiman, (1996), Bagging predictors, *Machine Learning*, 24(2), 123-140.

[18] JR. Quinlan, (1993). C4.5: *Programs for Machine Learnin*g, Morgan Kaufman.

[19] Y. Freund, & R.E. Shapire, (1996). Experiments with a new boosting algorithm. *Proceedings of 19th International Conference on Machine Learning*, pp 148-156