Available online at www.sciencedirect.com

SCIENCE DIRECT®

Neural Networks 18 (2005) 407–422

Neural Networks

www.elsevier.com/locate/neunet

2005 Special Issue

Challenges in real-life emotion annotation and machine learning based detection

Laurence Devillers*, Laurence Vidrascu, Lori Lamel

Department of Human-Machine Communication, LIMSI-CNRS, BP133, 91 403, Orsay Cedex, France

Received 21 March 2005; accepted 23 March 2005

Abstract

Since the early studies of human behavior, emotion has attracted the interest of researchers in many disciplines of Neurosciences and Psychology. More recently, it is a growing field of research in computer science and machine learning. We are exploring how the expression of emotion is perceived by listeners and how to represent and automatically detect a subject's emotional state in speech. In contrast with most previous studies, conducted on artificial data with archetypal emotions, this paper addresses some of the challenges faced when studying real-life non-basic emotions. We present a new annotation scheme allowing the annotation of emotion mixtures. Our studies of real-life spoken dialogs from two call center services reveal the presence of many blended emotions, dependent on the dialog context. Several classification methods (SVM, decision trees) are compared to identify relevant emotional states from prosodic, disfluency and lexical cues extracted from the real-life spoken human-human interactions.

Keywords: Emotion detection; Emotion annotation; Blended emotion; Naturalistic spoken data; Prosodic, disfluency and lexical features; Machine learning; SVM; Decision trees

1. Introduction

The work presented in this paper addresses both experimental and theoretical issues in the study of emotion in natural un-acted spoken data. The brain is the seat of many emotions at the same time, even if at any given moment there is one dominant emotion. We explore issues such as how to annotate real-life non-basic spoken emotions, how to define a typology of blended emotions which tries to take into account the known effects of masking by self-control, and how to detect emotions with machine learning techniques. In this paper, 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.

A large amount of work has been reported in the neurobiology literature on how the human brain recognizes emotional states (e.g. Damasio, 1994; Ledoux, 1989). There are also numerous psychological studies and theories on perception and production models of emotion, among others, the appraisal theory (Scherer, 1999). During the last decade, new results have appeared identifying a close relationship of emotion to knowledge, to brain activity, and to consciousness (e.g. Damasio, 1994; Taylor, 1997). New imaging techniques of the brain are also promising for enabling a new understanding of the neuronal substrate serving complex natural emotional processes. The Handbook of Affective Sciences (Davidson, Scherer, & Goldsmith, 2003) reports on the recent considerable advances in understanding how brain processes shape emotions and how the brain is changed by human emotion using a wide range of inquiry methods, such as neuroimaging techniques or laboratory paradigms designed to assess the cognitive and social constituents of emotion. Cognitive scientists study the nature of affect and emotion from a psychological point of view, mostly building computer models that help to understand memorization, perception and other psychological processes. Within the broad area of cognitive science, the dominant paradigm is the information processing approach. This paradigm has been successfully applied to cognitive processes at different levels such as

* Corresponding author. Tel.: +33 1 698 58 062; fax: +33 1 698 58 088.
E-mail addresses: devil@limsi.fr (L. Devillers), vidrascu@limsi.fr (L. Vidrascu), lamel@limsi.fr (L. Lamel).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.03.007

basic vision, spoken language or higher level thoughts. Most of the previous works on emotion have been conducted on induced or recalled data with archetypal emotions. Everyday emotions in real-life context are still rarely studied. Recently, databases involving people in various natural emotional states in response to real situations have been acquired, which will hopefully allow the development of more sophisticated systems for recognizing the relevant emotional states in such data. Our approach, grounded in findings from cognitive science and neuroscience is to develop tools for representing and modeling real-life emotions in natural contexts. Automatic detection systems based on different types of machine learning architectures, such as localized or distributed connectionist systems, may enable a deeper understanding of the perception of emotion by identifying relevant cues for emotion detection in natural emotional states.

The pluridisciplinarity of the emotion research field is certainly the most powerful means to improve knowledge on the nature of emotions. Emotional behavior has been, since the early days, an important field of research for psychology, philosophy and neuroscience, but more recently it is a growing field of research for computer science. This can be clearly seen by the recent activities in academic laboratories (e.g. Affective Computing Research at MIT (Picard, 1997)), industrial research laboratories (e.g. AT&T, SpeechWorks-Scansoft), and the special interest shown by funding agencies via the promotion of the FP6 trans-European Network of Excellence HUMAINE 'Human–machine interaction network on emotions'(http://www.emotion-research.fr) and the European projects PF-STAR 'Preparing future multisensorial interaction research' (http://pfstar.itc.it/), AMI 'Augmented Multi-party Interaction' (http://amiprotocol.sourceforge.net/) and CHIL 'Computers in the Human Interaction Loop' (http://chil.server.de/serlevts/is/101/) all of which include activities on emotion research, as well as the growing number of workshops and special sessions on emotion (on affective dialog, Embodied Conversational Agents (ECAs), etc.) and articles appearing in the scientific and popular press.

There has been increasing interest in emotion analysis to improve the capabilities of current speech technologies such as speech synthesis, recognition, and spoken dialog systems. In the context of human–machine interaction, the study of emotion has generally been aimed at the generation of Embodied Conversational Agents and at the automatic extraction of emotional behavior related features for dialog systems. Detecting emotion or underlying attitude can help orient the evolution of a human–computer interaction via dynamic modification of the dialog strategy. Emotion is conveyed by several multimodal cues: speech, gesture, face and physiological signals. In this paper, we focus on the speech signal. Emotional speech is more likely to occur in unstructured human-to-human interactions than in restricted contexts. Three types of corpora are generally used for emotion analysis and detection: acted, induced, and natural real-life corpora. People, often even actors, are bad at faking emotions on demand. However acted corpora are the easiest to obtain and to exploit, with only variations at the prosodic level since the linguistic (semantic and lexical) content has been controlled. Induced corpora often are obtained using Wizard of Oz (WOz) techniques and provide much more natural data. In Batliner, Fisher, Huber, Spilker, and Noth (2000), for instance, the authors make use of a WOz corpus to develop a detection model incorporating several layers of information, i.e. prosodic, dialog acts, repetitions, etc. Similar work using both human–human (DHH) and human–computer (DHM) dialogs and focusing on emotion detection was carried out in (Ang, Dhillon, Krupski, Shriberg, & Stolcke, 2002; Lee, Narayanan, & Pieraccini, 2002). However, as shown in Batliner et al. (2000), the closer we get to the real-life context of interaction, the more difficult the detection of reliable emotion markers is. It is quite difficult to obtain natural audio–video material because the camera is rarely invisible. Many researchers have made use of media material such as TV programs as natural corpora, focusing on interviews or reports that exhibit the most spontaneous situations. Others have used meetings or lectures, such as in the CHIL and AMI projects, as audio–visual material for studying emotional behaviors (Burger, Maclaren, & Yue, 2002; Wreded & Shriberg, 2003). Batliner et al. (2004) recorded corpora of children playing with an AIBO robot to incite for emotional reactions. When the focus is on real-life speech signals, call centers can provide interesting solutions for recording people in various natural emotional states since the recordings can be made imperceptivity. Among the natural corpora for emotion detection, we can mention the 'Lifelog' corpus consisting of everyday interactions between a female speaker and her family and friends described in (Campbell, 2004); the 'Interviews corpus' also known as the Belfast database (Douglas-Cowie, Campbell, Cowie, & Roach, 2003); the 'EmoTV' corpus—a set of TV interviews in French recorded in the HUMAINE Noe (Abrilian, Devillers, Buisine, & Martin, 2005); call center data (Lee et al., 2002); and medical dialogs (Craggs & Wood, 2004). Evidently, the types of emotions found in the corpus are heavily dependent on the task and situation/context.

The research presented in this paper focuses on the detection of emotional behavior in real-life speech corpora recorded in call centers. Two corpora of real agent client dialogs recorded in French, one recorded at a Stock Exchange Customer service center and the other at a Medical Emergency call center, are studied. In both corpora, callers manifest real-life emotions whereas the agent plays a predetermined, moderating role. The first corpus was recorded for other purposes within the framework of the IST FP-5 Amities Automated Multilingual Interaction with Information and Services project at a French Stock Exchange Customer Service Centre (http://www.dcs.shef.ac.uk/nlp/amities/). The second corpus is studied in the context of collaboration with a French Medical Emergency call center. Emotion manifestations in the second corpus are much more frequent and intense than in the first one. LIMSI also participates in HUMAINE.

One of the main challenges we address is the categorization and annotation of real-life emotions, requiring the definition of a pertinent and limited set of labels and an appropriated annotation scheme. Some of the problems we face have to do with the dynamic and constantly changing expression of emotions. Emotion manifestations are context-dependent, and are also highly person-dependent. Unambiguous emotions are apparent in only a small portion of any real corpus, therefore the relevant emotion data are too infrequent to provide a basis for consistent annotation and modeling using fine-grained emotion labels. A major difficulty of using natural corpora is that the expression of emotion is much more complex than in acted speech for the above listed reasons. Furthermore, inter-labeler agreement and annotation label confidences are important issues to address.

There are many reviews on the representation of emotions. For a recent review, the reader is referred to Cowie and Cornelius (2003). Here are just briefly described the three types of theories generally used to represent the emotions: appraisal dimensions, abstract dimensions and most commonly verbal categories.

(1) The appraisal theory (Scherer, 1999) provides a detailed specification of appraisal dimensions that are assumed to be used in evaluating emotion-antecedent events (novelty, pleasantness, goal relevance, 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. If done in real-time this can affect the expression of the emotions, and if done a posteriori, relies on the person's recall of the situation.

(2) According to Osgood, May and Miron (1975), the communication of emotion is conceptualized following several dimensions such as Evaluation, Power and Activation. These are measured on scales. The Evaluative scales rank emotions based on evaluation statements such as good–bad. The Power scales measure power and potency of judgmental connotation such as strong–weak and the Activity scales measure judgments such as active–passive. Other subjective dimensions include: intensity, control, tension, etc. As such emotions are defined along continuous abstract dimensions instead of naming emotions as discrete categories. The most widely employed scheme is based on the two perceptive abstract dimensions: Activation–Evaluation and has been employed to annotate several corpora with Feeltrace tool (Cowie et al., 2001). However, other dimensions are necessary, for instance, for distinguishing between Fear and Anger.

(3) Most of the emotion detection studies have only focused on verbal categories using a minimal set of emotions to be tractable. However, it is well-known that linguistic labels are imprecise and capture only a specific aspect of phenomena, i.e. those that are immediately relevant for speakers in a particular context. One question that is asked repeatedly is what are the emotions? There have been many answers to this question (Ortony & Turner, 1990), with the most consensual set comprised of the six primary basic emotions Joy, Sadness, Fear, Anger, Surprise and Disgust (Ekman, 1992). Basic emotions are defined as being inborn and are universal reactions. The distinction between primary vs. secondary or social emotions is widely used. These secondary emotions can also be viewed as combinations or mixtures of emotions. Plutchik's wheel (Plutchik, 1984) consists of eight primary emotions: to the six above are added anticipation and acceptance. Secondary emotions are produced by combinations of primary emotions that are adjacent on the emotion wheel. For example, love is a combination of joy and acceptance, whereas submission is a combination of acceptance and fear. Additional emotions can be classified as representing different degrees of intensity of primary and secondary emotions. For instance, anger can range from annoyance to rage.

In this work we make the assumption that it is possible to perceive and to annotate mixtures of emotions. We adopt a discrete palette theory (Cowie, 2000; Plutchik, 1984) which allows a blending of verbal emotion descriptors like that of a painter mixing colors on a palette, image given by Scherer (1984). There is no clear typology of the different emotion mixtures. Two types of emotions are also described in Cowie (2000), the speaker emotion based on her/his internal emotional state named cause-type and the effect that s/he would be likely to have on the listener, named effect-type. When we perceive a mixture of emotions, are they linked to cause-type or effect-type? Asking annotators to figure out how to assess a speaker's internal feeling is very subjective and likely to lead to erroneous labels. For speech data recorded in a natural context, only the emotion that is expressed in the speech content by lexical and paralinguistic cues can be annotated with (somewhat limited) knowledge of the situational context.

One of the challenges in the study on real-life speech call center data is to identify relevant cues that can be attributed to an emotional behavior 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), spectral (MFCC, cepstral features) and voice-quality (NAQ) (Campbell & Mokhtari, 2003) features. Additionally, lexical and dialogic cues can help as well to distinguish between emotion classes (Batliner, Fisher, Huber, Spilker, & Noth, 2003; Devillers, Vasilescu, & Lamel, 2002; Narayanan, 2002; Forbes-Riley & Litman, 2004). Speech disfluences have also been shown as relevant cues for emotion characterization (Devillers, Vasilescu, & Vidrascu, 2004) and can be automatically extracted. Non-verbal speech cues such as laughter or mouth noise are also helpful for emotion detection (Polzin & Waibel, 1998). 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 parameters and to reduce the number of features, 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 (Batliner et al., 2003; Forbes-Riley & Litman, 2004; Narayanan, 2002). Due to the difficulty of categorization and annotation, most of the studies have only focused on a minimal set of emotions such as positive vs. negative emotions (Lee, Narayanan, & Pieraccini, 2001), emotional vs. neutral state (Batliner et al., 2003). Some researchers consider task-dependent emotional behaviors or attitudes such as stressed vs. non-stressed speech (Fernandez & Picard, 2003; Narayanan, 2002; Petrushin, 1999), frustrated/annoyed vs. neutral/amused attitudes (Ang et al., 2002), or anger, motherese, emphatic and neutral (Steidl, Levit, Batliner, Nöth, & Niemann, 2005). The classification models used for emotion detection include Decision Trees, Support Vector Machines, Multi-Layer Perceptrons, Gaussian Mixture Models, etc. Several papers reporting on automatic emotion detection experiments are listed in Table 1 to give an (non-exhaustive) idea of the current trends in state-of-the art systems.

The remainder of this paper is as follows. The first question we address is how to annotate emotion in natural spoken data. This is the topic of Sections 2 and 3 which describe the corpora and present the Multi-level Emotion

Table 1
Emotion detection studies: author reference, style of corpora (acted, WOZ, DHM, DHH) and number of speakers, corpora size (turns: #speaker turns), emotion labels, type of features (spectral, prosodic (pitch, energy, rate), disfluences, lexical, language (*n*-g), syntax/semantic (part-of-speech-labels, dialogic), classification model (MLB, maximum likelihood Bayes classifier; KR, kernel regression; LDC, linear discriminant classifier; kNN, k nearest-neighbors; SVM, support vector machine; HMM, hidden Markov model; NNs, neural networks, decision trees, adaboost, etc.), detection rate

Author reference	Style of corpora	Corpora size	Emotion labels	Type of features	Machine learning	Detection rate
Dellaert et al. (1996)	Acted	1000 turns (five actors)	Joy, fear, anger, neutral	Prosodic (pitch contour)	MLB, KR, kNN	60–65% (acted, four classes)
Petrushin (1999)	Acted	700 turns	Anger, sadness, fear, happiness neutral	Prosodic spectral (F1, F2, F3)	NNs	70% (acted, five classes)
	Acted (non-professional)	56 calls (15–90 s)	Calm, agitation			77% (acted, two classes)
Batliner et al. (2000)	Acted	96 turns with emotion	Emotion, non emotion	Prosodic spectral part-of-speech dialogic	NNs	95% (acted, two classes)
	Read-speech	50 turns (with emotion), 50 turns (NE)				79% (read, two classes)
	WOZ vermobil	2395 turns (20 dial.)				73% (Woz, two classes)
Ang et al. (2002)	DHM communicator	21 k turns (∼3500 turns with emotion)	Frustration , annoyance, other	Prosodic	CART decision tree	75% (DHM, two classes)
				Language	3-g	60–65% (DHM, two classes)
Lee et al. (2002)	DHM call centers (real-life) Speech-Works	7200 turns	Negative, non negative	Prosodic	LDC kNN	75% (DHM, two classes)
Narayanan (2002)	DHM call centers (real life)	7200 turns	Negative, non negative	Prosodic language dialogic	LDC kNN	80–90% (DHM, two classes)
Shafran, Riley and Mohri (2003)	DHM AT&T how may I help you	5147 turns	Negative, non negative	Spectral (MFCC) F0	HMM	76% (DHM, two classes)
				Lexical	SVM	81% (DHM, two classes)
Forbes-Riley et al. (2004)	DHH machine-mediated	385 Utt. (90 Neg, 15 Pos, 280 Neu)	Positive, negative, neutral	Prosodic lexical dialogic	AdaBoost	84% (DHH-M, three classes)
Steidl et al. (2005)	DHM AIBO	∼6000 words (51 children)	Anger, motherese, emphatic, neutral	Prosodic part-of-speech	NNs	60% (DHM, four classes)

and Context Annotation Scheme (MECAS) and the adopted annotation protocol. MECAS 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. Section 4 describes the annotation experiments. The second question we address is how to detect the emotional state with machine learning, which is the subject of Section 5 on emotion features and detection models. Section 6 reports experiments on automatic emotional state detection using both acoustic and lexical parameters. Conclusions and further research are discussed in Section 7.

2. Real-life corpora

The studies reported in this paper make use of two corpora of naturally-occurring dialogs recorded in real-life call centers. 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 first corpus was recorded on only one channel; therefore there is a significant portion of overlapping speech that is difficult to use (20% of the dialog turns). Even though the second corpus was recorded on two channels; only one of the channels is reasonably clear (agent), the other contains some overlapping portions which were not transcribed (∼10%). These overlapped turns have been excluded in this study, even though they could well correlate with emotional speech. This is due to the difficulty of transcription (it is both very time consuming and error prone), and of automatic extraction of acoustic features for such speech segments. The transcription guidelines are similar to those used for spoken dialogs in Amities project (http://www.dcs.shef.ac.uk/nlp/amities/). Some additional markers have been added to denote named-entities, breath, silence, intelligible speech, laugh, tears, clearing throat and other noises (mouth noise).

The first corpus of dialogs contains real agent–client recordings obtained from a Web-based Stock Exchange Customer Service Center. These recordings were made for purposes independent of this study, and have been made available for use in developing an automated call routing Prototype service within the context of the Amities project (Hardy, Baker, Devillers, Lamel, Rosset, & Strzalkowski, 2002). The service center can be reached via an Internet connection or by directly calling an agent. While many of the calls involve problems in using the Web 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. A corpus of 100 agent–client dialogs (four different agents) in French was orthographically transcribed and annotated. 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 questions and problems. There are about 6200 speaker turns in corpus 1. Our studies are based only on the 5000 speaker turns after overlaps are excluded, which are known to be frequent phenomena in spontaneous speech.

Table 2
Corpora characteristics: Corpus 1: 100 agent–client dialogs of around 3 h (M, male; F, female), 6.2 k speaker turns, Corpus 2: 404 agent–client dialogs of around 10 h (M, male; F, female), 19 k speaker turns

	Corpus 1	Corpus 2
# agents	4 (3 M, 1 F)	6 (2 M, 4 F)
# clients	100 dialogs (91 M, 9 F)	404 dialogs (152 M, 266 F)
# turns/dialog	Average: 50	Average: 47
# distinct words	3 k	6.5 k
# total words	44 k	143 k

The second dialog corpus contains real agent–client recordings obtained from a convention between a medical emergency call center and the LIMSI-CNRS. The transcribed corpus contains about 20 h of data. This study is based on a 10-h subset comprised of 404 agent–client dialogs (six different agents, 404 clients). About 10% of speech data is not transcribed since there is heavily overlapping speech. The service center can be reached 24 h 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. The call topic is classified as emergency situation, medical help, demand for medical information, or finding a doctor. The decision can be to redirect the caller to an emergency doctor or psychologist, to provide immediate help by sending an ambulance, to suggest a medical contact, refer the patient to another medical call center or contact his own primary care physician. In the case of emergency calls, the patients often express stress, pain, fear of being sick or even real panic. The caller may be the patient or a third person (a family member, friend, colleague, caregiver, etc.). Table 2 summarizes the characteristics of both corpora.

3. Multi-level emotion and context annotation scheme

Providing a reference language for emotion description and a reference annotation guide which includes relevant confidence measures is one of the aims of the HUMAINE network. However, at the moment no formal description exists. Our Multi-level Emotion and Context Annotation Scheme (MECAS) has been developed for both multimodal data (Abrilian, Devillers, Buisine, & Martin, 2005) and speech-only data (Vidrascu & Devillers, 2005). The adopted hierarchical scheme allows emotion and context to be represented with several layers of granularity. The global context representation has been adapted for use with speech-only data.

In order to describe emotion, four main problems have to be dealt with: the dynamic aspect of emotions, the possible mixture of emotions, context-dependency, and the highly person-dependent nature of emotion expression. First, the dynamic aspect of emotions can be expressed as a continuous mark in an N-dimensional space or at a coarse level by a sequence of emotionally quasi-stable segments labeled with discrete verbal labels. Second, the mixture of emotions can be described using N-label categories with operators on them (blended, sequential, masking, ambiguous, etc.) (Douglas-Cowie, Devillers, Martin, Cowie et al., 2005) or as a continuous mark in the complex emotion space. Third, some of the context and speaker-dependencies can be annotated as meta-data. After the language and level of representation decisions (context annotation at the dialog level, emotion annotation at the segment level, label definition) are made, an annotation and validation protocol needs to be defined, assuring that the reference emotions can be accurately extracted for use with machine learning.

3.1. Context annotation at the dialog level

At the dialog level, there are many types of information about the context and person that can be annotated. Some of them are task-dependent, and others are not. Corpus 2 was annotated with contextual metadata, such as the call origin (from patient, from medical center), the role in the dialog (direct caller or 3rd person), reason (immediate help, doctor help, medical information, etc.), decision taking, etc. Additional information concerning the acoustic quality of the recording (noise, outside/inside, mobile/fixed/radio phone) and the caller information such as sex, age category (child, middle-age, old), accent (French accents, foreign accents), defaults in pronunciation/voice quality (nasal twang, lisp, slurred, etc.), health/mental state (normal, pathologic, alcohol/drug influenced, hoarsel, groggy, etc.) are also labeled. These data can reveal a lot about the history of the person: health, origin, pain/displeasure, etc.

3.2. Emotion annotation at the segment level

The main difficulty for this representation is to determine the useful levels of description in terms of granularity and temporality. The audio signal was further segmented into emotional segments where the annotators felt it was appropriate, so the temporal-grain can be finer than the speaker turn. The specification of MECAS at the segment level enables annotation of emotion labels and abstract dimensions with one or two emotion labels for segment, selected from fine-grained and coarse-grained labels as well as some local emotional context cues.

In order to minimize the annotation time, only the abstract dimensions which are complementary with verbal categories are labeled. The bi-polar Valence (Negative/Positive) is deducted from the fine-grained verbal labels. The only ambiguity concerns the class 'Surprise' which can be associated with both positive and negative emotions. Activation (passive, normal, active) is often confused with Intensity (low, middle, high) by non-expert annotators. In these annotations, only intensity is rated on a five-level scale. Rough labels for the bi-polar Activation (Passive/Active) were extracted from fine-grained verbal labels. We also added a new dimension named Self-Control (not the Power/Control described in (Osgood, May & Miron, 1975)). The Self-Control dimension is a meta-annotation describing the perception of the self-control of the person (from controlled to uncontrolled on a seven-level scale). All combinations of categories are possible with only one exception; for the verbal label surprise, which has an ambiguous valence, the minor annotation is obligatory and gives the valence.

The set of labels is hierarchically organized, 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. The annotation level used to train emotion models can be chosen based on the number of segments available. A first rough description of mixture emotions has been defined. Mixed emotions in the same coarse-grained label are noted as Ambiguous and a mixture between two coarse-grained labels is called Conflictual if they are not the same valence, Non-conflictual otherwise. A closer study is certainly necessary as well as perceptual tests to obtain a more precise typology.

3.3. How to define labels?

The labels are task-dependent. There are many different possible strategies for finding the best N category labels. Two extreme strategies are the direct selection of a minimal number of labels (typically from 2 to 5) or a free annotation which leads to a high number of verbal labels that must be reduced to be tractable, for instance from 176 (after normalization) to 14 classes in experiments by (Abrilian, Devillers, Buisine, & Martin, 2005). The mapping from fine-grained to coarse-grained emotion labels is not straightforward when free annotations are used. In the previous experiment, the mapping from 176 to 14 was done manually by the same annotators after a consensus was made on a shorter list of 14 emotion labels. An alternative strategy, which seems to us to be more powerful, is to select by majority vote a set of labels before annotating the corpus. However to adopt this strategy, a group of (at least three) persons who have already worked with the corpus, need to select emotions with high appropriateness, appropriateness, moderate appropriateness from a list of reference emotions. Then a majority voting procedure allows a sub-list of verbal categories, the best 20 for instance, to be selected. Several different reference lists can be found in the literature (see http://www.emotion-research.fr).

3.4. Annotation protocol

The high subjectivity of human annotation requires the use of rigorous annotation protocols. After deciding the list of labels and the adopted scheme, precise rules for segmentation must be determined along with the number of annotators and the validation procedures. The emotional units can be at the level of the speaker turn, the segment, or the word. The segments (within a speaker turn) can be defined as a syntactic or semantic group. Concerning label consistency, it is evident that combining the opinions from a larger number of annotators (at least 3), 5 in Steidl et al. (2005), via majority voting, for example, leads to less subjective annotations. Evidently, the larger the size of the corpus, the more difficult it is to obtain multiple annotations. We propose a less time and money-consuming procedure for annotation. First annotations are independently done by two labelers prior to re-annotation of segments labeled as non-neutral by at least one of the two labelers. Since only 13% of turns in Corpus 1 had non-neutral emotion labels, the gain in time is non-negligible. Then the re-annotation phase of the emotionally rich part of the corpora can then be done by more labelers.

We also have to consider inter-labeler consistency and confidence measures. There are different measures of annotation reliability; for instance, the widely used Kappa inter-coder agreement measures (Carletta, 1996) for categorical labels and the Cronbach's alpha measures (Cronbach et al., 1972) for continuous variables. For those measures, one label by segment is normally used. When a mixture of emotions is annotated, a solution is to compare only the Major label or to add some rules to improve inter-labeler agreement such as Major/Minor = Minor/Major. We propose to adopt a self re-annotation procedure of small set of dialogs at different times (for instance once a month) in order to judge the intra-annotator coherence over time.

In emotion recognition from natural speech, for instance, it is unenvisagable to have perfect emotion reference labels, so, as proposed in Steidl et al. (2005), the machine learning results must be compared with the performance of human labelers. They propose an entropy-based method to evaluate the classification results which takes into account systematic labeling errors.

3.5. Labels for machine learning

Since segments were labeled by more than one labeler and also since segments could be assigned one or two labels, it was necessary to create a mapping (i.e. to reduce the multiple labels per segment to one label) for the machine learning experiments. Let us consider each annotation as a vector (Major, Minor). Several choices were considered. The first mapping selects a weighted vote for Major/Minor (or for only Major ignoring the Minor) labels across all annotators. The second mapping combines the N (Major, Minor) vectors (for N annotators) in an emotion soft vector.

Table 3
Representation of the decisions of two labelers for an emotion: for example, for $w_{\mathrm{M}}=2$, $w_{\mathrm{m}}=1$, $W=6$, W is the sum of the weights (0. 67 anger, 0. 165 fear, 0.165 OtherNeg)

Segmentation annotation:
Labeler 1 *major* anger, *minor* fear
Labeler 2 *major* anger, *minor* otherneg
Conversion into an emotion vector:
(w_{M}/W anger, w_{m}/W fear, w_{m}/W OtherNeg)

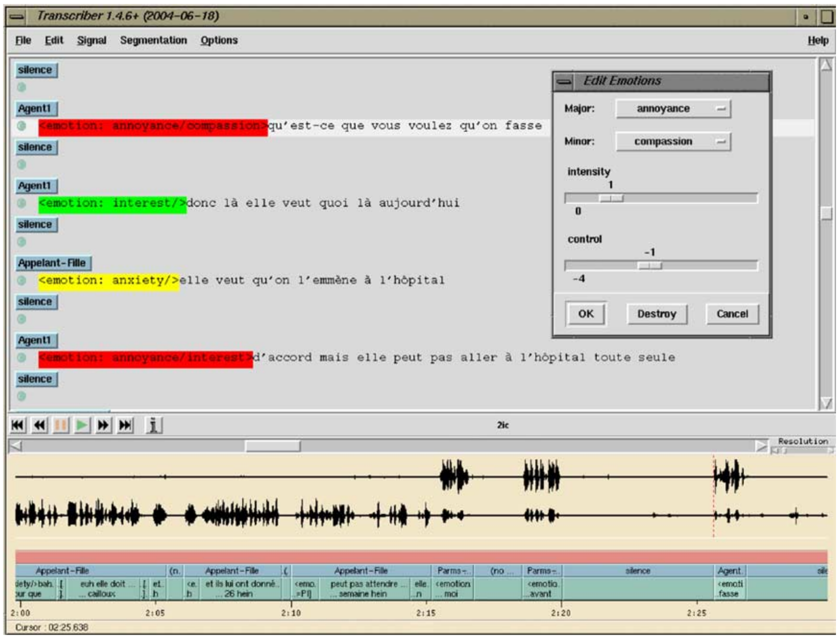
Different weights are given to the emotion annotation, one weight to the Major emotions (w_{M}) and one other to the Minor emotions (w_{m}). If there are two annotators then two vectors were added, W being the sum of the weights. The resulting soft vector size is the number of different emotional classes used by the annotators as shown in Table 3. As for the first mapping, it is possible to select the label which has the major weight as the reference label. Another solution is to combine the first and second highest emotions as a new one-label class (if the class is sufficiently represented). Finally, it is also possible to use a specific training algorithm which uses the reference vector instead of one single reference.

4. Emotion annotation

This section describes our experience in annotating in two Call center corpora, as well as experiments to measure inter-annotator agreement. The initial studies led us to refine the annotation protocol so as to better suit the detection task. Labeling was done using a modified version of the transcriber tool (Barras, Geoffrois, Wu, & Liberman, 2000). An example screen display is shown in Fig. 1.

4.1. Annotation of corpus 1: financial customer service call center dialogs

The initial annotation experiments with this corpus allowed only a single label per segment. A first task-dependent annotation scheme was developed (Devillers, Vasilescu, & Lamel, 2003), adopting, for practical reasons, a limited number of permissible emotion labels. Two of the four classical emotions were retained: Anger and Fear. It should be noted however that in this application most of the Anger and Fear manifestations are shaded, observing irritation for Anger, and anxiety for Fear. Two characteristics of the callers behavior were also included, Satisfaction and Excuse, since these can be directly associated with the task. The Neutral attitude is also included since most of the speaker turns do not exhibit a non-neutral emotion. Two annotators independently listened to the 100 dialogs, labeling each sentence with one of the five classes. In order to assess the consistency of the selected labels, the inter-annotator agreement was computed (the Kappa value is 0.8). Ambiguous labels concern only 2.7% of the entire



silence

Agent1
<emotion: annoyance/compassion>qu'est-ce que vous voulez qu'on fasse

silence

Agent1
<emotion: interest/>donc là elle veut quoi là aujourd'hui

silence

Appelant-Fille
<emotion: anxiety/>elle veut qu'on l'emmène à l'hòpital

silence

Agent1
<emotion: annoyance/interest>d'accord mais elle peut pas aller à l'hòpital toute seule

silence

Major: annoyance
Minor: compassion
intensity 1 (0)
control -1 (-4)

Appelant-Fille | (n. | Appelant-Fille | (| Appelant-Fille | Parms-.. | (no ... | Parms-.. | silence | Agent. | sile
iety/>bah. pur que | euh elle doit ... cailloux | et. h | <e. b | et ils lui ont donné. ... 26 hein | <emo. =P!] | peut pas attendre ... semaine hein | elle. ..n | <emotion ... moi | <emotio. ..avant | <emoti fasse

Fig. 1. Screen copy of the modified Transcriber tool, showing the emotion annotation of an excerpt taken from the end of a 3 min call.

corpus and most often involved indecision between the neutral state and another emotion. Sentences with ambiguous labels (19% of the sentences labeled 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% of the entire corpus. For the detection experiments presented in Section 6, only Fear, Anger and Neutral classes are considered.

The proportion of turns for each emotion label is shown in Table 4 based on the non-overlapping speaker turns. Although the same emotion labels are used for the agent and the client, they do not have the same meaning. Anxiety on the part of the agent is less marked than for the client, it concerns problems carrying out his work (for instance problems with WEB services), whereas for the client, his money is at stake.

Since most of the emotions observed in the corpus were shaded, a second annotation phase was carried to better deal with blended emotions. In this second phase, labelers were given the possibility of selecting a second label, called a 'Minor' label when two different emotions for a segment were perceived, such as Anger and Fear. To evaluate this new protocol, all utterances previously annotated with Negative emotions in the first corpus, Fear or Anger, were re-annotated in dialog context by two expert annotators, different from those who carried out the first annotations. For this task, the annotators used a list of 21 emotions (detailed in Section 4.2) from the seven coarse classes: Fear, Anger, Sadness, Hurt, Surprise, Positive and Neutral. The two expert annotators chose the same Major emotion in 64% of the cases, and only 13% utterances were ambiguous or contractictory having no common label between the two annotators.

4.2. Annotation of corpus 2: medical emergency call center dialogs

Based on our experience labeling corpus 1, the first labeling experiments with the second corpus were conducted with the possibility of annotating a mixture of labels per segment. The emotion and emotion-related list (55 terms) proposed by Rodie Cowie at the HUMAINE Summer school in Belfast (http://emotion-research.net/ws/summer-school1/emotion%20words) was used as a reference list. Eighteen fine-grained classes were selected from this list after a majority voting procedure with five labelers. The list of fine-grained labels is Anxiety, Stress, Fear, Panic, Annoyance, Impatience, ColdAnger, HotAnger, Disappointment, Sadness, Despair, Hurt, Embarrassment, Relief, Interest, Amusement, Surprise and Neutral. Three other verbal labels have been proposed by the annotators and added for this task: Dismay, Resignation and Compassion making a list of 21 emotions including Neutral. During the

Table 4
Proportion of each emotion label in the dialog Corpus 1 labeled listening to audio signal in the dialog context for 5000 non-overlapping speaker turns

	Anger (%)	Fear (%)	Satisfaction (%)	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 5
Emotion classes hierarchy: multi-level of granularity

Valence-level	Coarse level (7 classes)	Fine-grained level (21 classes including Neutral)
Negative	Fear	Fear, anxiety, stress, panic, embarrassment, dismay
	Anger	Anger, annoyance, impatience, coldAnger, hotAnger
	Sadness	Sadness, disappointment, resignation, despair
	Hurt	Hurt
Negative or positive	Surprise	Surprise
Positive	Positive	Interest, compassion, amusement
Neutral	Neutral	Neutral

Table 6
Labeler inter-reliability in terms of % agreement between two annotations by the same labeler at different times

	Client		Agent	
	Dec–Feb	Jan–Feb	Dec–Feb	Jan–Feb
Labeler 1	76.4 (369 seg.)	82.9 (287 seg.)	73.9 (356 seg.)	83.9 (255 seg.)
Labeler 2	66.5 (369 seg.)	80.8 (279 seg.)	78.5 (350 seg.)	76.5 (254 seg.)

Dec–Feb means first annotation in December, re-annotation in February (14 dialogs), Jan–Feb First annotation in January, re-annotation in February (11 dialogs).

annotation phase, labelers were also given the possibility to choose between Negative, Positive and Unknown labels for cases where it was difficult to find the right fine-grained class.

The labelers were asked to create emotional sub-segments and to annotate them using the list of 21 fine-grained labels plus the three labels (positive, negative, unknown). The fine-grained labels were grouped into seven coarse-grained emotion label families: Fear, Anger, Sadness, Hurt, Positive, Surprise and Neutral. This hierarchy (see Table 5) was empirically decided.

An excerpt of an annotated dialog is given Fig. 1. This dialog is between an agent and the daughter of a patient, who is calling for the second time in a few days. The previous time an ambulance was sent to take her to the hospital, but her condition was considered non-critical and she was sent home. The interaction is a bit long and it is hard for the agent to determine exactly what the caller wants. The agent, although showing compassion, is a bit annoyed with the situation. We can also see an example of two emotional segments for one of the agent turn in this excerpt. For each segment, the 'Edit Emotions' box allows to annotate the Major emotion and Minor emotion if necessary, the Control (Self-control) and Intensity dimensions. Here is the translation of the Fig. 1 excerpt:

Agent: what is it that you want us to do so, now, what does she want now, today
Daughter: she wants to be brought to the hospital
Agent: ok, but can't she go to the hospital by herself

In order to verify the coherence of annotations for each labeler, a re-annotation phase has been carried out at two different times, for a randomly selected set of dialogs (14 and 11 dialogs, respectively). This procedure will continue to be carried out as the remaining part of the corpus is labeled in order to ensure annotation stability.

Comparing the coarse-grained labels (Table 6), a global improvement of the annotation coherence between two re-annotations phases can be seen. Most of the re-annotation confusions were made between Neutral and Positive classes on the agents' data and between Neutral and Fear classes for the clients' segments.

Table 7 gives the proportion of the most frequent emotion labels for both the agent and the client in the 404 annotated dialogs. This corpus has a larger proportion of emotional behavior than was found in the financial call center data. The percentage reflects the number of segments where two annotators agreed, i.e. having assigned the same Major label. In total there are 19,045 emotion segments, with roughly 9% more segments than speaker turns in the transcribed corpus, since speaker turns were subdivided when more than one emotion was manifest during the turn. Globally speaking, 67% (6109/9204) of the segments have common emotional labels for Agent, 57% (5643/9841) for the Client.

In order to assess the consistency of the selected labels, the inter-annotator agreement was calculated. The Kappa value is 0.61 for clients and 0.35 for agents when only considering the Major annotation. The Kappa values are slightly better (0.65 and 0.37, respectively) if a rule allowing common annotation in one of the two annotations Major and Minor is used. The annotation is seen to be much more reliable for the clients speech than for the agents, which may be due to their respective roles as the agents has the job of controlling the dialog to extract the needed information from the client. Ambiguous labels concern 38% of the global corpus and most often involved indecision between neutral state and other emotions. We envision two possible ways of dealing with ambiguous segments. The first is to make a judgment by a third independent listener, and the second is to use

Table 7
Repartition of the fine-grained emotion labels (five top classes) with common major annotations (404 dialogs)

Client	Neutral	Anxiety	Stress	Relief	Annoyance	Others
5643 seg.	64.5%	20%	8.7%	3.3%	0.6%	3.2%
Agent	Neutral	Interest	Compassion	Annoyance	Surprise	Others
6109 seg.	89%	4.3%	2.5%	2.3%	0.4%	2.5%

the information that the segment is inherently ambiguous by combining the labels in a vector (as illustrated in Table 3). The experiments presented in this paper are carried out only on the part of the corpus where the two labelers assigned the same coarse-grained Major labels.

4.2.1. Blended emotion annotation

Blended emotions can be considered as the presence of two or more emotions at the same time, of which one may be considered Major and the others Minor. The Major emotion can be related to the dominant emotion in the brain, with the other perceived emotions in a mixture being secondary. Because of the high level of subjectivity, we decided to annotate only one Minor per mixture, the most intense one. The majority of the perceived emotions were assigned only a single emotion label. Labeler 1 gave 37% of the non neutral segments a Minor label, whereas labeler 2 assigned secondary emotions for only 16% of the non neutral segments (see Fig. 2).

A coarse typology of the blended emotions can be identified. The first type of mixture is labeled Ambiguous, and corresponds to Minor and Major labels belonging to the same emotion family (or same coarse-grained label). The second type is named the Confictual emotions, where the Major and Minor labels come from different valence emotion families (Fear/Positive, Anger/Positive, etc.). The third type is the Unconflictual mixture of emotions such as a mixture of Anger and Fear, or Fear and Sadness which shares the same valence value.

The proportions of the Conflictual, Unconflictual and Ambiguous blended emotions in Agent and Client data for both labelers are given Fig. 3. The Ambiguous blended emotions are a means for labelers to give more accurate labels and levels of intensity. This strategy was frequently used by the first labeler who annotated 1860 segments in Client data with Ambiguous mixed emotions. Evidently, the Conflictual and Unconflictual emotions mixtures are the most interesting data to study. It is interesting to note that they appear in different positions in the dialog (i.e. for Agent and Client) and that both annotators have perceived mixtures in those classes and more particularly in the conflictual class.

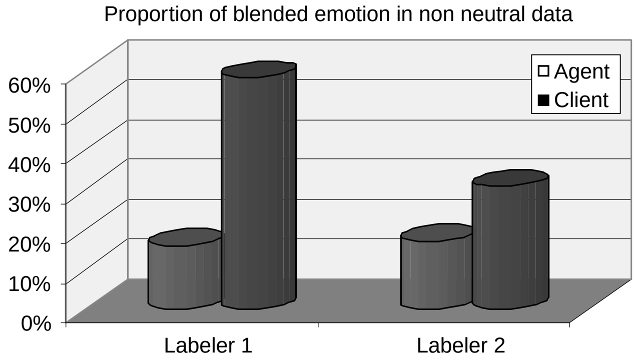


Fig. 2. Repartition of mixed emotions for the Clients' and the Agents' non neutral segments.

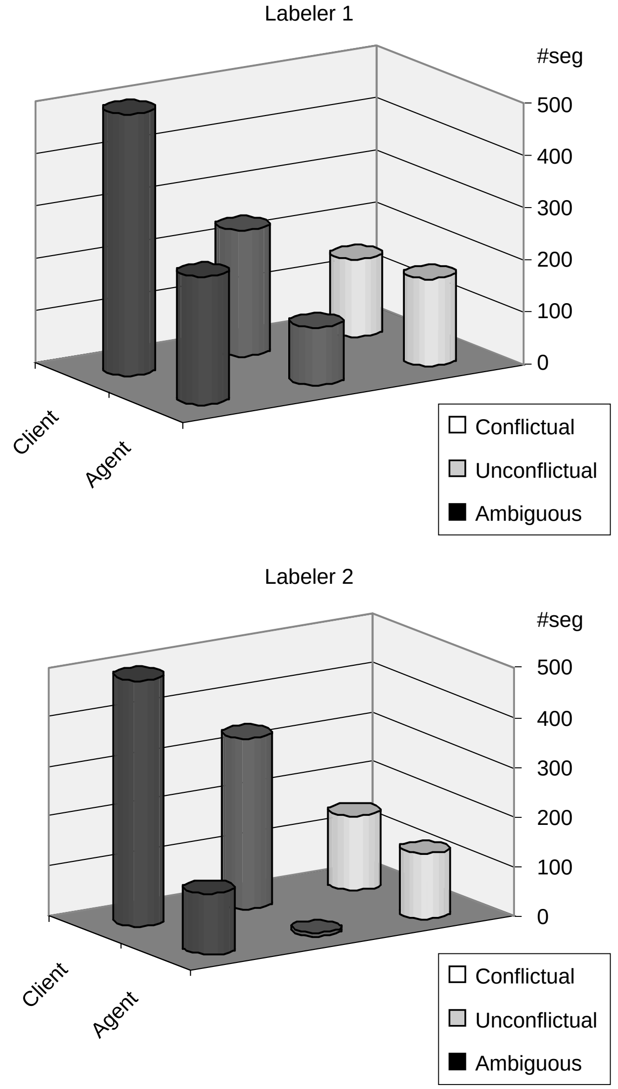


Fig. 3. Proportion of the conflictual, unconflictual and ambiguous blended emotion annotated in client and agent data for labelers 1 and 2.

5. Features and classification with machine learning

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.

Prosodic features (mainly Pitch and Energy) are classical features used in a majority of applications and research systems. For accurate emotion detection in natural real-life speech dialogs, lexical, prosodic, disfluency and contextual cues should be considered and not only the prosodic information. Many other measures related to spectral features: zero-crossing, LPC, MFCC, etc. have been also studied. Spectral parameters are not irrelevant but have not yet led to significant improvements. Since there is no common agreement on a top list of features and the feature choice seems to be data-dependent, our 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 the classification algorithm.

In the experiments reported in this paper, we have focused on the extraction of lexical, prosodic, spectral, disfluency and non-verbal events cues. For prosodic (F0 and energy) and spectral cue extraction, the Praat program (Boersma, 1993) has been used. It is based on a robust algorithm for periodicity detection carried out in the lag auto-correlation domain. Praat has been used to extract F0 features on voiced regions. Since F0 feature detection is subject to errors, a filter was used to eliminate some of the extreme values that are detected. About 50 features are input to a classifier which selects the most relevant ones. This 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):

(1) F0 features (Log-normalized per speaker): min, 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) and maximum cross-variation of F0 between two adjoining voiced segments (inter-segment) and with each voiced segment (intra-segment).
(2) Energy features (normalized): min, max, mean, standard deviation and range at the segment level.
(3) Duration features: speaking rate (inverse of the average length of the speech voiced parts), number and length of silences (unvoiced portions between 200 and 800 ms).
(4) Spectral features: formants and their bandwidth (first and second): min, max, mean, standard deviation, range.
(5) Disfluency features: from time-aligned orthographic reference transcriptions, disfluency cues such as filler and silence pauses can be identified (Devillers et al., 2004). For corpus 1, the time-alignments were carried out automatically, whereas for corpus 2 a limited number of disfluences markers were annotated with time-stamps during transcription. The disfluency pause features are the number of pauses per utterance (normalized by the length of the utterance), the mean and maximum pause duration, and the number of filled pauses ('euh' in French).

On the corpus 1, an automatic alignment of the orthographic transcription with the acoustic signal was used in order to extract disfluency cues such as filler and abnormal large silent pauses. The orthographic transcriptions are aligned with the signal using existing models already developed at LIMSI for telephonic conversations. The alignment system uses continuous density Hidden Markov Models (CDHMMs) with Gaussian mixtures for acoustic modeling. The vocabulary contains 3022 words with a phonetic transcription based on 37 phones, details are given in Devillers et al. (2004). All the utterances labeled with negative emotions have been manually verified in order to avoid alignment errors. The features extracted are pauses features per utterance (normalized by the length of the utterance), mean and maximum duration of silence, number of filler pauses.

On the corpus 2, we have not used automatic alignment but disfluency markers were annotated and aligned during the transcription: silence (number and size) and number of filler pauses.

(6) 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 aim is to get rid of the noise and reduce the complexity of feature space without affecting the performance. Two kinds of methods are very common in data mining: selecting the most relevant attributes or to apply linear transformations to reduce the dimension of the data. Feature selection and reduction methods such as PFS (Promising First Selection), FS (Forward Selection) and PCA (Principal Composant Analysis) have been shown to provide improved performance compared to base features as shown in (Dellaert, Polzin, & Waibel, 1996; Lee et al., 2001; Petrushin, 1999).

Several classifiers and classification strategies well described in the machine learning literature are used to classify prosodic and lexical and to combine of both models. 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)). 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 increase the likelihood that the resulting terms are relevant. We have adopted this technique for emotion detection.

For the paralinguistic model, we use the Weka machine learning software (Witten et al., 1999). 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:

- The Decision Tree is a set of rules in a structure of nodes and leaves, with each node represents a test and each leaf a class. The algorithm C4.5 is a well-known system for training decision-trees (Quinlan, 1993).
- Support Vector machine (SVM) (Vapnik, 1995) algorithms search 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).
- Voting algorithms (AdTree) (Breiman, 1996) and (AdaBoost) (Freund & Shapire, 1996) combine the outputs of different models. Boosting methods assign weights to models with better performances. In our experiments, they are applied to C4.5. Both, boosting (AdaBoost) and bagging (AdTree) manipulate the training data in order to generate different classifiers. Bagging produces replicate training sets by sampling with replacement from the training instances. Boosting uses all instances at each repetition, but maintains a weight for each instance in the training set that reflects its importance, different weights lead to different classifiers. In both cases, the multiple classifiers are then combined by voting to form the final classifier.

For combining lexical with paralinguistic (prosodic, spectral, disfluency, non-events) cues, a linear combination of results from the lexical and paralinguistic models is used.

6. Experiments

Emotion detection experiments are reported using the two corpora described earlier. Three sets of experiments are reported using the first corpus: a lexical emotion detection system based on a unigram model, acoustic models for emotion detection, and the combination of the two. Experiments with second corpus are carried out separately for the Agents and Clients, since different emotions are expressed for the two roles. All of the experiments make use of a jack-knifing procedure in order to avoid biases due to selected subsets of the data. The number of features specified in the tables is the top-N most relevant features selected after automatic preprocessing data.

6.1. Corpus 1

A lexical emotion detection system was developed using a unigram model. The similarity between an utterance and an emotion is the normalized log likelihood ratio between the emotion model and a general task-specific model. Two unigram emotion models were trained, one for each emotion (Neutral and Negative), using a set of on-emotion training utterances. The general model was estimated on the entire training corpus. An interpolation coefficient of lambda = 0.75 was found to optimize the results. The emotion of an unknown sentence is determined by the model yielding the highest score for the segment, given the two emotion models. Stemming and stopping, commonly used procedures in information retrieval tasks for normalizing and removing frequent words in order to increase the likelihood that the resulting terms are relevant have been incorporated. Named entities (dates, numbers) were also used order to generalize the vocabulary. In order to reduce the number of lexical items for a given word sense, an automatic part of speech tagger was used to derive the word stems. Experiments were carried out using different stop lists (containing from 60 to 200 entries) of high frequency words assumed to be uncorrelated with the task. Since the corpus is quite limited, emotion-balanced test sets were randomly selected using the lexically based reference annotations following a jack-knifing procedure. The results are obtained on the average of 10 runs. The average detection rate is 68% using the named entity generalization (NORM) and 71% when stemming is also used (LEM = NORM + stemming). Despite trying multiple stop-lists, stopping did not improve the detection rate. Results on positive–negative emotion detection reported in (Devillers & Vasilescu, 2004) are somewhat better, around 78% on this corpus.

6.1.1. Paralinguistic models: neutral/negative emotion detection

Experiments using paralinguistic (prosodic, spectral) models were carried out to distinguish between Neutral and Negative emotions. The jack-knifing experiments used five data subsets (four subsets are used for training and one for test, repeating the experiment five times with each subset being used for test). This procedure was carried out six times with different randomly selected subsets.

The results summarized in Table 8, were obtained using the feature set described in Section 5, excluding the disfluency features which have not been verified on the neutral part of the corpus and non-linguistic events (tears, laugh, etc.) features which occur rarely in this corpus. The list shows about the same recognition rate for sets of features from 5 to 40 showing the redundancy of the features. The best rate obtained is 73%: ADTree and 10 features. As shown in Petrushin (1999), very few features (here five) yield high level of detection (71%). Petrushin achieved ∼77% classification accuracy of two emotion states, ‘agitation’ and ‘calm’ for eight features chosen by a feature selection for a smaller acted corpus (56 calls from 15 to 90 s recorded by non professional actors).

Then, we also tried to see if our set of cues is relevant for Anger/Fear discrimination. With the first annotation (one label per segment), better results are obtained for

Table 8
Algorithms and feature selection: comparison of the neutral/negative (fear and anger) detection performances with the best N features on Corpus 1

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

This table shows the average of correctly classified instances for the 30 runs. The number into parenthesis is the standard deviation.

Negative/Neutral classification (73%) than for Anger/Fear classification (56%). The detection rate of Anger/Fear improves to 60% by adding disfluencies (using automatic alignment) to the same set of features. This may 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 therefore explored if the mixture annotations would enable us to account for these ambiguities. Focusing on Anger and Fear, an emotion vector (Fear, Anger) was computed from the new annotations by giving a weight of two to the Major emotion and a weight of one to the Minor one as shown in Table 3. Four classes were deduced from these vectors: Fear (Fear > 0; Anger = 0), Anger (Fear = 0; Anger > 0), Blended emotions of Fear and Anger (Fear > 0; Anger > 0) and Other (0, 0). As can be seen in Fig. 4, 40% of the utterances are labeled as blended.

The consistency of the two annotations was verified. For an utterance initially labeled as Fear, the second annotation was considered equivalent if the first field of the (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 blended emotion Anger and Fear for this task.

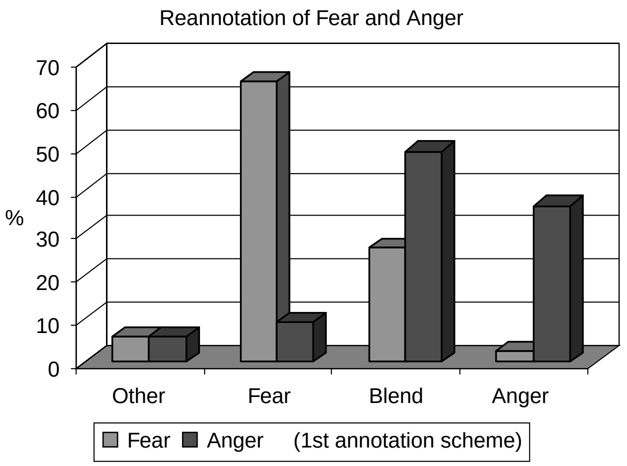


Fig. 4. Repartition of the utterances previously labeled Anger and Fear after the second mixture annotation.

6.1.2. Combining lexical and paralinguistic scores

Experiments combining lexical and paralinguistic output scores were carried out for Neutral/Negative emotion detection. Once again a jack-knifing procedure was used, averaged over 10 runs where each test contains 50 randomly chosen segments. In all tests, better results were obtained by using a linear combination of the lexical and paralinguistic scores.

The lexical scores (average score 71%) were obtained using the unigram model described above. The paralinguistic model selected is based on ADTree and has an average global score 71.6%. The average of combination score is 76.6%, about a 5% gain in performance. The emotion identification curves for lexical, paralinguistic and combination are shown in Fig. 5 for the 10 jack-knifing runs. As in (Forbes-Riley & Litman, 2004; Narayanan, 2002), we show that the lexical content is meaningful for emotion detection.

A perceptive test was carried out on these data for two conditions: with and without listening the audio signal (Devillers, Vasilescu, & Mathon, 2003). When subjects were able to listen to the signal, the speaker turns annotated with negative emotion were correctly perceived with 75% of accuracy by the subjects of the test (20 subjects) validating the initial labels. The model combining paralinguistic and lexical cues yields the same performance as the human perception.

6.2. Corpus 2

During the analysis of the annotations of the Medical Emergency Call Center, we observed that the emotions associated with the agents are quite different from those of the clients. Therefore in these detection studies, separate models and results are given for each of these. The results of

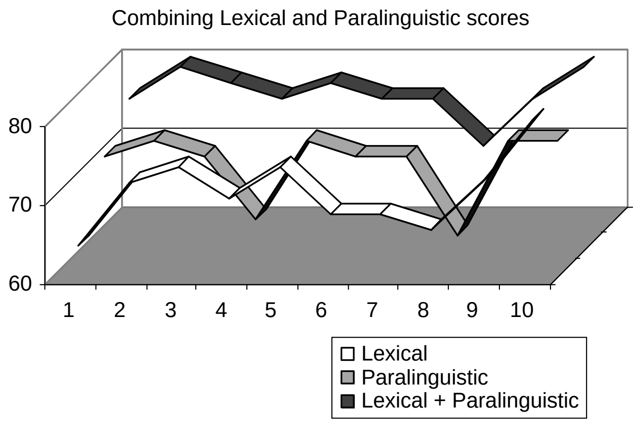


Fig. 5. Emotion identification for lexical, paralinguistic and combination of lexical and paralinguistic models for 10 jack-knifing runs for Corpus 1.

Table 9
Algorithms and feature selection: comparison of the neutral/negative (fear and anger) detection performances for agents and clients with the best N features on Corpus 2

Role in the dialog	Emotion classification	# Top-N features	ADTree	SVM
Agent	Neutral/anger	16	66.7 (4.1)	73.0 (4.1)
	Neutral/anger	21	67.4 (4.6)	72.6 (3.9)
	Neutral/negative	16	66.8 (4.2)	72.8 (3.3)
Client	Neutral/negative	16	82.6 (1.1)	82.9 (1.3)
	Neutral/negative	21	82.8 (1.0)	83.2 (1.2)
	Neutral/fear	16	82.9 (1.3)	83.1 (1.2)

This table shows the average of correctly classified instances for the 30 runs. The number into parenthesis is the standard deviation.

the emotion detection experiments carried out on Corpus 2 are summarized in Table 9.

These experiments made use of all the prosodic and spectral features described in Section 5, including the disfluency and the non-linguistic event (respiration, crying, etc.) features extracted from the manual annotations. All experiments were done using a jack-knifing procedure with five subsets (four subsets are used for training and one for test, the experiment is repeated five times with each subset being used for test). This procedure is repeated six times with different divisions of the data. We again tried different sets of the N-top features. For the agent, the best detection score is obtained with SVM model: around 72% of correct detection between Neutral and Negative emotions, the latter being comprised of the coarse-grained labels Anger and Fear. For the client, the best scores are obtained using the SVM and ADTree models, with correct detection rate of about 83% between Neutral and Negative, where the latter mainly corresponds to the coarse-grained label Fear. We obtain high detection scores on this natural data for a z-way distinction.

A first experiment was also conducted to distinguish four coarse-grained label classes (Anger, Fear, Positive and Neutral) for client segments, using the best SVM model with all features. A global score of 61% correct detection was obtained, which is a very encouraging first result.

7. Discussion and perspectives

This paper has explored some of the issues that are faced when studying real-life non-basic emotions. These include how to annotate the data, how to define a typology for blended emotions and how to detect emotions from the speech with machine learning techniques. We have adopted a new annotation scheme allowing multi-layers of annotation with different granularity (MECAS). This scheme allows emotion to be annotated along with its context. Its originality is that it is possible to use two emotion labels per segment—one label for the dominant emotion perceived called the Major emotion, and second label if another emotion is perceived in background, called the Minor emotion. This annotation scheme is motivated by the fact that the brain is the seat of many emotions at the same time, even if at any given moment there is one dominant emotion.

In certain states of mind, it is possible to exhibit more than one emotion; for instance, when trying to mask a feeling about something, when suffering, or when there are conflicting intentions, etc. This study has found the manifestation of naturally-occurring mixed emotions in telephone dialogs recorded in two call centers. For the corpus recorded at the financial call center, mixed emotions were observed for the clients combining Fear and Anger (or more appropriately anxiety and annoyance). Many clients show annoyance when they are fearful of losing money. This emotion mixture is never seen in the agents' data.

In the second corpus, comprised of dialogs recorded in a Medical Help call center, specific emotion mixtures were found in different parts of the dialog. Agents showed impatience/anxiety mixtures when they identified a high level of emergency and experienced difficulties in dialoguing with the caller (difficulty of understanding non-native persons, social differences, physical condition, etc.). For the callers, the most frequent mixtures involved relief/anxiety, positive/stress which at the first view seem impossible to obtain. Such conflicting emotions are often observed near the end of the dialog, when the person knows that help is coming, but still remains fearful about his condition. Evidence suggests that such a perception is possible, because the two emotions are expressed at different levels, one linguistic and contextual and the other paralinguistic. A study to verify this assumption is in progress on these types of mixtures.

We have separated the emotion mixtures into three categories: ambiguous (two labels of different intensity in the same emotion family or coarse-grained label, for instance Annoyance/Anger); conflictual (two conflicting labels, one positive and one negative, for instance Relief/Anxiety); and non-conflictual (two labels in different emotion family with the same valence, for instance Fear/anger). These annotations are highly dependent of the context of the dialog. We have only started to exploit the various meta-data (call subject, age, etc.) associated with the calls, which can be correlated with the emotion annotations. We have not yet used intensity and self-control values which can be also correlated with the emotion annotations.

The perception of emotion is very subjective, for instance, some persons are more compassionate (or receptive) than others. How does this affect the reliability of the annotations? Can annotators be wrong in their perceptions? Are there good and bad annotators? In our opinion, a good labeler is coherent over time and is able to explain his/her decisions. Just as the expression of emotion is highly personal, so is its perception. Our philosophy is to exploit these differences by combining the labels from multiple annotators in a soft emotion vector. How to then use this vector effectively in machine learning is one of our future objectives.

Automatic systems can also lead to a deeper understanding of the perception of emotion by identifying the relevant cues to emotion detection in natural emotional states in the context of social interaction. In the reported experiments, we only made use of segments with one-label annotations and a high level of reliability. To the mainly classical prosodic (F0, duration and energy) and also spectral (1st and second formant) features, we added some non verbal event markers and disfluency features. For Corpus 1, a correct detection rate of 71% was obtained between Neutral and Negative (Anger and Fear) using either a lexical unigram model and or a prosodic feature model alone. The choice of classifier (SVM, Decision Tree) did not have a large affect on the results. Combining the lexical and prosodic scores increased the detection performance by 5%. This corpus mostly contains low-intensity emotions due to the social aspect of the interaction. The separation between anger and fear classes is not straightforward, the adding of disfluences parameters such as pauses and filler pauses 'euh' yields better distinction between the two classes. It was noted that fear provokes more disfluencies than anger. Listening experiments have shown the presence of a hybrid class between Anger and Fear, explaining the difficulty to distinguish between them.

Experiments with corpus two used the same prosodic and spectral features. Disfluencies (hesitation, pauses) and some non-verbal events such as laughter, inspiration, expiration, which were time-stamped during the transcription phase, were added to the paralinguistic models and appear to improve the capability of the models to discriminate between emotions. A more careful study of the relevance of disfluences is under way. The correct detection rate between Neutral and Negative emotions on this corpus is about 80%. For a four-class separation (Anger, Fear, Positive and Neutral) for the clients, a correct detection rate of 60% is obtained, which is comparable to other state-of-the-art performance on emotion detection. We have obtained similar performance on real-life complex data to other reported state-of-the-art results obtained with WOz or DHM systems.

The MECAS annotation scheme is currently being used to annotate the remaining 10 h of the corpus. Experiments on this larger corpus will provide complementary results on blended emotion classes that hopefully will improve emotion detection performance for natural data.

Acknowledgements

This work was partially financed by several EC projects: FP5-Amities, FP6-CHIL and NoE HUMAINE.

The authors would like to thank I. Vasilescu, for her collaboration on first studies using Corpus 1 and I. Wilhem, C. Mathon, M. Lesprit and J. Martel for their help with data collection and annotation. The work on Corpus 2 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.

References

Abrilian, S., Devillers, L., Buisine, S., & Martin, J.-C. (2005). EmoTV1: Annotation of Real-life Emotions for the Specification of Multimodal Affective Interfaces. *Proceedings of Human-Computer Interaction International*, Las Vegas, USA, July 2005.

Ang, J., Dhillon, R., Krupski, A., Shriberg, E., & Stolcke, A. (2002). Prosody-based automatic detection of annoyance and frustration in human–computer dialog. *Proceedings of International Conference on Spoken Language Processing, Denver*, *3*, 2037–2040.

Barras, C., Geoffrois, E., Wu, Z., & Liberman, M. (2000). Transcriber: Development and use of a tool assisting speech corpora production. *Speech Communication*, *33*(1), 5–22.

Batliner, A., Fisher, K., Huber, R., Spilker, J., & Noth, E. (2000). Desperately seeking emotions or: Actors, wizards, and human beings. *Proceedings of the International Speech Communication Association Workshop on Speech and Emotion* , 195–200.

Batliner, A., Fisher, K., Huber, R., Spilker, J., & Noth, E. (2003). How to find trouble in communication. *Speech Communication*, *40*, 117–143.

Batliner, A., Hacker, Ch., Steidl, S., Nöth, E., D'Arcy, S., Russell, M., et al. (2004). 'You stupid ting box'-children interacting with the AIBO robot: A cross-linguistic emotional speech corpus. *Proceedings of fourth Internacional Conference on Language Resources and Evaluation*, 171–174.

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 (http//www.fon.hum.uva.nl/praat/).

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Burger, S., Maclaren, V., & Yue, H. (2002). The ISL meeting corpus: The impact of meeting type on speech style. *International Conference on Spoken Language Processing*, 301–304.

Campbell, N. (2004). Accounting for voice quality variation. *Proceedings of Speech Prosody*, 217–220.

Campbell, N., & Mokhtari, P. (2003). Voice quality: The 4th prosodic dimension. *15th international congress of phonetic sciences*, 2417–2420.

Carletta, J. C. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics* , *22*(2), 249–254.

Cowie, R. (2000). Emotional states expressed in speech. In describing the emotional states expressed in speech. *Proceedings of ISCA ITRW on speech and emotion: Developing a conceptual framework for research*, 224–231.

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

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human–computer interaction. *IEEE Signal Processing Magazine*, *18*(1), 32–80.

Craggs, R., & Wood, M. M. (2004). *A 2-dimensional annotation scheme for emotion in dialogue Proceedings of AAAI spring symposium on exploring attitude and affect in text: Theories and applications*. Stanford University.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratman, N. (1972). The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley.

Damasio A. (1994). Descartes' Error: emotion, reason and the human brain. New York: Grosset/Putnam.

Davidson, J., Scherer, K., & Goldsmith, H. (2003). *Handbook of affective sciences*. Oxford University Press.

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion, in speech. *Proceedings of fourth international conference on spoken language processing*, *3*, 1970–1973.

Devillers, L., & Vasilescu, I. (2004). Reliability of lexical and prosodic cues in two real-life spoken dialog corpora. *Fourth international Conference on Language Resources and Evaluation*, *4*, 1423–1426.

Devillers, L., Vasilescu, I., & Lamel, L. (2002). Annotation and detection of emotion in a task oriented human–human dialog corpus. *International Standards for Language Engineering, Edinburgh*.

Devillers, L., Vasilescu, I., Lamel, L. (2003). Emotion detection in task-oriented dialog corpus. *In proceedings of the IEEE International conference on multimedia.*

Devillers, L., Vasilescu, I., & Mathon, C. (2003). Acoustic cues for perceptual emotion detection in task oriented human–human corpus. *Proceedings of 15th International Congress of Phonetic Sciences*, 1505–1508.

Devillers, L., Vasilescu, I., & Vidrascu, L. (2004). Anger versus fear detection in recorded conversations. *Proceedings of speech prosody*, 205–208.

Douglas-Cowie, E., Campbell, N., Cowie, R., & Roach, P. (2003). Emotional speech: Towards a new generation of databases. *Speech Communication*, *40*, 33–66.

Douglas-Cowie, E., Devillers, L., Martin, J-C., Cowie, R., Savvidou, S., Abrilian, S., & Cox, C. (2005). Multimodal Databases of Everyday Emotion: Facing up to Complexity, *Proceedings of Interspeech 2005.*

Ekman, P. (1992). An argument for basic emotions. *Cognition and Emotion*, *6*, 169–200.

Fernandez, R., & Picard, R. W. (2003). Modeling drivers' speech under stress. speech. *Speech Communication*, *40*, 145–159.

Forbes-Riley, K., & Litman, D. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. *Proceedings of human language technology conference of the north american chapter of the association for computational linguistics (HLT/NAACL).*

Freund, Y., Shapire R. E. (1996). Experiments with a new boosting algorithm. In proceedings of the of 19th *International Conference on Machine Learning*, pp 148–156.

Hardy, H., Baker, K., Devillers, L., Lamel, L., Rosset, S., Strzalkowski, T., et al. (2002). Multi-layer dialogue annotation for automated multilingual customer service. *International Standards for Language Engineering Workshop*.

Ledoux, J. E., et al. (1989). *Journal of Cognitives Neurosciences*, *1*, 238–243.

Lee, C.M., Narayanan, S., Pieraccini, R. (2001). Recognition of negative emotions from the speech signal. *In proceeding of the IEEE Automatic Speech Recognition and Understanding*.

Lee, C. M., Narayanan, S., & Pieraccini, R. (2002). Combining acoustic and language information for emotion recognition. *Proceedings of international conference on spoken language processing*, 873–876.

Narayanan, S. (2002). Towards modeling user behavior in human–machine interactions: Effect of errors and emotions. *International Standards for Language Engineering workshop, Edinburgh.*

Ortony, A., & Turner, T. J. (1990). What's basic about basic emotion? *Psychological Review*, *97*, 315–331.

Osgood, C., May, W. H., & Miron, M. S. (1975). *Cross-cultural universals of affective meaning*. Urbana: University of Illinois Press.

Petrushin, V. (1999). Emotion in speech: Recognition and application to call centers. *Artificial Neural Network Intelligence Engineering*, 7–10.

Picard, W. R. (1997). *Affective computing*. The MIT Press.

Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. In K. R. Scherer, & P. Ekman (Eds.), *Approaches to emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum, 293–317.

Polzin T., Waibel A. (1998). Detecting emotions in speech. *Cooperative Multimodal Communication* 1998. Tilburg Netherlands.

Quinlan, J. R. (1993). C4.5: Programs for machine learning. *Morgan Kaufman*.

Scherer, K. R. (1984). On the nature and function of emotion: A component process approach. In K. R. Scherer, & P. Ekman (Eds.), *Approaches to emotion* (pp. 293–317). Hillsdale, NJ: Erlbaum, 293–317.

Scherer, K. R. (1999). Appraisal theory. In T. Dalgleish, & M. Power (Eds.), *Handbook of cognition and emotion* (pp. 637–663). New York: John Wiley, 637–663.

Shafran, I., Riley, M., & Mohri, M. (2003). Voice signatures. *Proceedings of IEEE automatic speech recognition and understanding workshop*, 31–36.

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 international conference on acoustics, speech, and signal processing*.

Taylor, J. (1997). In L. Landau, & J. G. Taylor (Eds.), *Neural networks and the mind in concepts for neural networks* (pp. 243–268). London: Springer, 243–268 chapter 9.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. Springer-Verlag.

Vidrascu, L., & Devillers, L. (2005). Annotation and Detection of Blended Emotions in Real Human-Human Dialogs recorded in a Call Center. *Proceedings of IEEE International Conference on Multimedia.*

Witten, I. H., Franck, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical machine learning tools and techniques with Java implementations. *Proceedings of ANNES'99 International Workshop on emerging Engineering and Connectionnist-based Information Systems*, 192–196.

Wreded, E., & Shriberg, E. (2003). Spotting 'Hots Spots' in meetings: Human judgments and prosodic cues. *Proceeding of Eurospeech*, *3*, 1677–1680.