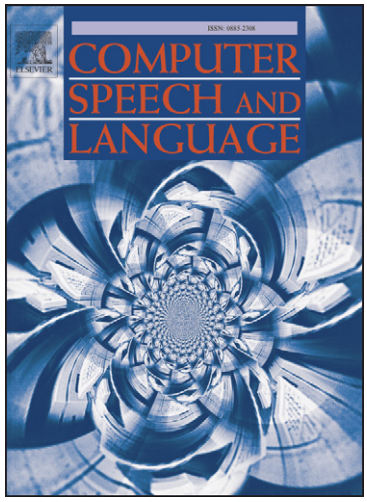
Automatic Inference of Complex Affective States

Tal Sobol-Shikler



Automatic Inference of Complex Affective States

Tal Sobol-Shikler

PII: S0885-2308(09)00076-X
DOI: 10.1016/j.csl.2009.12.005
Reference: YCSLA 440

To appear in: *Computer Speech and Language*

Received Date: 5 May 2008
Revised Date: 23 March 2009
Accepted Date: 24 December 2009

Please cite this article as: Sobol-Shikler, T., Automatic Inference of Complex Affective States, *Computer Speech and Language* (2009), doi: 10.1016/j.csl.2009.12.005

Automatic Inference of Complex Affective States

Tal Sobol-Shikler [a,b]

[a] *Computer Laboratory, University of Cambridge, UK*

[b] *Present address: Department of Industrial Engineering and Management, Ben-Gurion University of the Negev, Israel*

Abstract

Affective states and their non-verbal expressions are an important aspect of human reasoning, communication and social life. Automated recognition of affective states can be integrated into a wide variety of applications for various fields. Therefore, it is of interest to design systems that can infer the affective states of speakers from the non-verbal expressions in speech, occurring in real scenarios. This paper presents such a system and the framework for its design and validation. The framework defines a representation method that comprises a set of affective-state groups or archetypes that often appear in everyday life. The inference system is designed to infer combinations of affective states that can occur simultaneously and whose level of expression can change over time. The framework considers also the validation and generalisation of the system. The system was built of 36 independent pair-wise comparison machines, with average accuracy (tenfold cross validation) of 75%. The accumulated inference system yielded total accuracy of 83% and recognised combinations for different nuances within the affective-state groups. In addition to the ability to recognise these affective-state groups, the inference system was applied to characterisation of a very large variety of affective state concepts (549 concepts) as combinations of the affective-state groups. The system was also applied to annotation of affective states that were naturally evoked during sustained human-computer interactions and multi-modal analysis of the interactions, to new speakers and to a different language, with no additional training. The system provides a powerful tool for recognition, characterisation, annotation (interpretation) and analysis of affective states. In addition, the results inferred from speech in both English and Hebrew, indicate that the vocal expressions of complex affective states such as thinking, certainty and interest transcend language boundaries.

Key words: Affective computing, affect recognition, cognition, emotions, human perception, intelligent systems, machine learning, multi-label inference, multi-modal analysis, multi-modal database, speech analysis, speech corpora.
PACS:

1 Introduction

Affective states and their non-verbal expressions are important aspects of human reasoning, decision-making and communication. Recognition of affective states can be integrated into fields such as human-computer interfaces and interactions (HCI), human-robot interactions (HRI) and speech technologies. The recognition can enhance such systems and user performance and has many potential applications [1–3]. Recognition results can be used for analysis of user reactions in order to predict intentions and to generate appropriate response. It can also be used for annotation of speech corpora for synthesis of affective speech. In order to achieve that, the systems designed should be able to infer affective states occurring in real scenarios. The development of such systems entails collection and labelling of speech corpora [4], development of signal processing and analysis techniques, as well as consolidation of psychological and linguistic analyses of affective states [5].

In this paper the term *affective states* refers to emotions, attitudes, beliefs, intents, desires, pretending, knowledge and moods. Their expression reveals additional information regarding the identity, personality and physiological state of the speaker, in addition to context-related cues and cultural display rules. This wide definition of the term *affective states* draws on a comprehensive approach to the role and origin of emotions [5,6]: affective states and their expressions are part of social behaviour [7,8], with relation to physiological and brain processes [9,10]. They comprise both conscious [11] and unconscious reactions [12,13,10], and have cause and effect relations with cognitive processes such as decision making [14,10]. A number of affective states can co-occur simultaneously [15–17], and change dynamically over time. A similar definition of the concept affective states is given by Höök [18] who describes affect as human, rich, complex and ill-defined *experience*.

The affective states are inferred from their non-verbal expressions. The term *expression* refers here to the outward representation of the affective states. This is the observable behaviour (conscious or unconscious) that people perceive and interpret. It can be affected by factors such as context and cultural display rules.

There are three main approaches to the design of affect recognition systems. These approaches are used for inference from expressions in speech and in other behavioural cues such as facial expressions. The most commonly used approach [19,20,55] is to infer a small set of *basic emotions* [21,22], such as *happy*, *sad*, *angry*, *afraid*, *disgusted* and *surprised*. The term refers to qualitatively distinct states that are held to be universal at least in essence, i.e. recognisable by most people from most backgrounds, and associated with brain systems that evolved to cope with various situations. These affective states

have relatively clear definitions, although even within this set the need for finer definitions has been addressed, for example distinguishing between *cold anger* and *warm anger* [5,23,6,24]. Stereotypical expressions of these affective states are perceived as easier to act and to recognise, and therefore useful for both quick acquisitions of data-sets, and as a starting point for an emerging research field. However, these affective states do not encompass the entire range of human affective states and (in most applications) do not relate to nuances of expression. If the small set is used only as a starting point, it is an open question whether the same behavioural cues are used for both extreme emotions and subtle expressions of complex mental states. The inference is of a single emotion for each analysed sentence so the systems encompass a small set of affective states. This approach is a limited version of a broader perspective called the *categorical* representation method, in which additional affective states are defined, either as a blend of basic emotions, or in conjunction with other cognitive processes [25].

The second approach is to detect the existence of a selected affective state in real situations, such as drivers' stress, attempts at insurance fraud or post-natal depression [26–28]. This method is not used in order to recognise which affective states are expressed in the speech, but rather to detect whether a certain affective state exists or not. It does not refer to other co-occurring affective states or to different levels of the expression.

The third approach, which has recently become more widespread, is the *dimensional* approach, in which several expressions are identified each on a one-dimensional (1-D), two (2-D) or three dimensional (3-D) space, with dimensions such as passive-active, positive-negative and low-high arousal levels [5,29–33,38]. The dimensional approach provides in theory a more continuous scale for interpretation but the research usually refers to recognition of the edges or areas, for example: positive and low arousal or negative and high arousal level. These descriptions are often correlated to physiological processes such as changes in heart rate or skin conductivity [34], but they do not reflect the large variety of affective states nor the different levels of their experience. Furthermore, although people can annotate the affective states they perceive of a sequence of sentences on a dimensional space [35,31] and radial representations of affective states on dimensional space exist in research, in everyday situations radius and angle are not commonly used to describe affective states, and relatively coarse descriptions such as positive and negative are rare. Therefore, the intelligibility of the inference results is limited and can affect the applicability of systems that use this method. Various combinations of the categorical approach and the dimensional approach have been offered [36,37].

These approaches refer to affective states as single entities, although co-occurrences of affective states are common, for example, a *happy* person can *think* and

show *interest* at the same time. Mixtures of affective states, such as aversion-attraction, like some feel to snakes, also appear. These approaches do not refer to different level of experience of the affective states. The number of affective states or dimensions that can be recognised is limited and does not represent the range of affective states and their definitions as people use and express in everyday life (except for the recognition of a single state existence).

There is a growing effort to use real recorded data for recognition [24,38]. However, using corpora of real (not acted) recorded data for training often cannot overcome the limitations posed by the manner of representation. Using real data may further limit the scope of the system because annotation of real data is complicated [4,39], which in practice limits the developers to labelling few affective states (or dimensions).

For all these reasons, a framework should be developed for the design of systems that can recognise and represent affective states in real scenarios in a manner that is meaningful.

In order to be applicable to naturally evoked affective states in real scenarios, a recognition system should be able to handle a large variety of affective states and their expressions; recognise affective states that often occur in everyday life (rather than full-blown emotions that are rarely experienced or seen); handle various affective states that occur simultaneously [16,17]; handle dynamic variations of affective states over time and define the vocal-features that distinguish different expressions [5,55,20]. Another challenge is generalising to new speakers without additional training. A related question for further generalisation is whether the vocal expressions of affective states, beyond the basic emotions, prevail over different languages and cultures so that one system can infer mental states in different languages. A system that can do all that can also be used for annotation and analysis of speech corpora for various applications.

This paper presents an innovative framework for the design and validation of a system that infers complex affective states from their non-verbal expressions in speech. Complex affective states refer to affective states, such as moods, emotions, mental states, attitudes and the like, that occur in everyday life (beyond the set of basic emotions), including co-occurring affective states, dynamic variations and nuances of affective states. The framework includes the choice and definition of two complementing data-sets for training, testing and generalisation; the choice of affective states to be recognised; the design of a system to infer different levels of recognition for combinations of co-occurring affective states for each sentence or utterance and the validation and generalisation of the system. At the basis of the design is the observation that different features distinguish different pairs of affective states [40,36], instead of finding a single set of features to distinguish all expressions [20,41,24]. The validation

and generalisation of the system include generalisation to a very large set of affective state concepts and to automated annotation and analysis of naturally evoked affective states during sustained human-computer interactions, with validation by multi-modal analysis. The system was tested on data in two different languages.

The paper presents the framework, its implementation and results. It aims to present an overall approach for the design and validation of the system. More detailed description of the implementation of each stage and its validation may be found elsewhere [42,43].

2 Methodology

A system was designed to infer a ranked list of co-occurring affective states from aural expressions in an utterance or a sentence. The methodology includes the choice of representation method, i.e. how affective states are represented and what the system should recognise. It also includes the choice and definition of data-sets for training, testing and validation. The data-sets are chosen according to the stated goal of the system, and the underlying theory. Their structure defines the scope and capabilities of the system. The architecture of the inference system is also a part of the methodology.

2.1 Affective states

There are large numbers of lexical definitions of affective states in most languages. For example, four thousand definitions for describing affective states were identified in English [31]. Research in this area of affective computing aims to clarify the states most likely to matter to emotion-oriented computing, and adapting ideas from psychology such as soft coding, dimensional representation, and appraisal theory to provide representations that are more tractable than lists of irreducible categories [44,24].

However, the *prototype* approach has not been widely used for the design of systems that infer affective states from speech. The *prototype* approach consists of both the contents of individual categories and the hierarchical structures among them. It is a combination of the *categorical* approach and a *hierarchical* approach. It can represent a wide range of affective states and uses terms that are intelligible and reflect knowledge. An example for this approach is the Mind Reading taxonomy [45]. This taxonomy describes several hundred affective states, divided into 24 groups, such as *happy*, *fond*, *kind*, *liked*, *romantic*, *thinking*, *interested*, *bothered*, *interested*, *angry*, *unfriendly* and *afraid*. Each

of these groups includes many different emotions or affective state concepts that share a common meaning and knowledge. For example, the *unfriendly* group includes 120 affective state concepts, such as *argumentative*, *cold* and *discouraging*. In addition to the meaningful groups of the taxonomy Baron *et al.* [45,46] refer to different levels of affective states according to the typical age at which people start to recognise and understand them (including affective state concepts that are commonly understood only over the age of 18).

The differences between the *categorical*, *dimensional* and *prototype* representation approaches lie in their scope and intelligibility. The scope is the number of emotions and mental states that can be described by each method. (It is smaller in the first method, while reaching hundreds and thousands of affective-state labels in the other two methods). The intelligibility of the description method is derived from the similarity between the terms used by this method and those used by people in everyday situations, as well as the ability of these terms to cover the large extent of knowledge and meanings of the different affective-state concepts. In this respect the prototype approach is the most intelligible. However, taking as a reference to this approach the Mind Reading taxonomy, a large variety of affective states that belong to the same general meaning group often represent opposite characteristics regarding descriptors such as active-passive (*acknowledging* and *acknowledged*) and positive-negative (*romantic* and *lecherous*, *confident* and *vain*). Therefore, the affective-state groups for automatic systems should be carefully chosen. In addition, like the other methods, taxonomies such as the Mind Reading usually consider affective states as mutually exclusive entities, each belonging to a single meaning group.

In order to represent affective state that occur in real settings, a representation method that draws on the *prototype* approach was chosen. The representation consists of a small set of nine affective-state groups or archetypes. The affective-state groups are: *joyful*, *thinking*, *absorbed* or *concentrating*, *stressed*, *excited*, *opposed* or *disagree*, *interested*, *confident* or *sure*, and *unsure*. Each of these affective state groups comprises several affective states that belong to the wider concept or affective-state group, for example, the *excited* group comprises affective states such as *alert*, *lively* and *inspired*. The chosen affective-state groups are based, to a large extent, on the definitions and groups of the Mind Reading taxonomy and database [45]. Table 1 lists the concepts that were used for training.

This set of affective-state groups was chosen because it represents behavioural patterns and terms that are common in everyday situations. Some of these expressions have been observed in manual labelling of human-computer interactions [47]. El-Kaliouby used quite a similar approach for recognition of six affective-state groups from facial expressions and head gestures [48]. Afzal's survey of affect in intelligent tutoring reveals that different fields that approach

Table 1
The nine affective-state groups that constitute the expression groups recognised by the inference machine (left), and the corresponding concepts that were used for training the machine (right).

Expression	**Concepts**
joyful	tickled,carefree, amusing,overjoyed, festive, merry, enjoying, glad, happy, joking, joyful, triumphant
absorbed	absorbed, engaged, committed, concentrating, focused, thorough
sure	adamant, assertive, confident, sure, convinced, decided, determined, resolved
stressed	bothered, hurried, hampered, overwrought, overrun, pressured, rushed, stressed, flustered, impatient, tense, worried
excited	alert, dynamic, lively, excited, inspired, invigorated, adventurous
opposed	argumentative, confrontational, contradictory, contrary, disagreeing, disapproving, disinclined
interested	asking, fascinated, probing, questioning, scrutinising, interested
unsure	confused, clueless, unsure, undecided, insecure, ambivalent, puzzled, baffled, considering, debating
thinking	fantasising, thinking, choosing, deciding, wool-gathering, comprehending, realising

this subject refer to another subset of these groups or to the individual concepts within them [49]. Vidrascu *et al.* present a similar approach for manual annotation (labelling) of speech samples, using eight groups of basic emotions that contain 20 definitions of fine-grained affective-state concepts [24].

Combinations (co-occurrences) of the nine affective-state groups represent a large variety of affective state concepts. This set is simple enough to be used by both people and systems. Furthermore, the chosen affective-state groups appear in many cultures and languages, although the single affective states that are the constituents of these groups can be different [50,51]. Therefore, the inference results can be used by people from different backgrounds. The representation approach enables the system to recognise different nuances of the affective-state groups. Using affective-state groups, each comprises several affective-state concepts, also compensates for the relatively small number of samples of each of the fine-grained affective-state concepts (six samples in this case) [24].

However, the affective state concepts that were chosen to represent each affective-state group for training and testing are not the only affective states that belong to the same affective-state groups in meaning, nor to the related groups in the Mind Reading taxonomy and database. The remaining affective states and their recorded expressions were later used for further testing.

2.2 Corpora of affective speech

The choice of requirements and representation method affects the definition and choice of data-sets and the manner of data acquisition. Two data-sets were used. The Mind Reading database was used for training [52,45]. This is a commercial product [45] that aims to teach children diagnosed with Autism to recognise the behavioural cues of a large variety of affective states (emotions) from vocal expressions, from video recordings of head gestures and facial expressions and from video recordings of body language in dialogues. The experimental version that was used contains over 700 affective state concepts, arranged into the 24 groups defined by the Mind Reading taxonomy. Six sentences represent each concept. These sentences are uttered by different actors with different (neutral) textual content. In total, the database includes 4400 recorded sentences, by ten UK English speakers, of both genders and of different age groups, including children. According to its publishers, the expressions were induced [53] and the database was labelled by ten people. The database is acted, but its original purpose (teaching children to recognise affective states from their expressions in the children's daily lives) and the large number of affective states that it represents, make it a suitable choice for training a machine to recognise affective states and for validation on a large variety of affective states (although children need fewer samples for training).

The Hebrew database, Doors, was defined and recorded as part of this research [47]. Doors is a multi-modal database of recorded sustained human-computer interactions. The participants were engaged in a computer game designed to evoke emotions and expressions, based on the Iowa Gambling Test (IGT) [10]. Each interaction lasted approximately 15 minutes. The game comprised a series of 100 events in which the participant had to choose and open one of four doors, each with a hidden profit or loss in points and a different gain expectation, unknown to the player. The goal of the player was to maximise the total accumulated profit. The speech part consists of two repeated sentences (*petah delet zo*, open this door, and *segor delet*, close door, in Hebrew), forming a corpus of 200 sentences for each participant, in addition to speech sessions with un-controlled text and uttered sounds of various lengths, freely evoked during the game and during an intervening interview. In addition to speech, the database comprises video recordings of facial expressions and head gestures, game events, (including participants' choices), mouse movement rate, reaction

delay between events and physiological measurements, including: galvanic skin response (GSR), echo-cardiograms (ECG) and blood-volume at the periphery (BVP). Due to the Doors design, each sentence could be associated with specific game events that occurred before and after it was uttered (the 100 door openings). The physiological cues were also synchronised to the openings of doors in the game. The participants were Hebrew speaking graduate students and academic staff, of both genders in the age range of 24-55, mostly from engineering background, whose Hebrew is the first or second language. Segmentation of the continuous recorded speech signal into sentences was done automatically for most of the speakers, using an adaptation for an entropy based algorithm [54].

The two databases complement each other. Mind Reading provides *labelled* samples of *a very large variety* of *complex affective state concepts* (beyond the set of basic emotions) for training and testing. It provides recordings of speakers of different age groups (although few). It also provides information about the relations between different affective states. Doors represents *naturally-evoked* expressions during *sustained human-computer interactions*, i.e. *dynamic changes over time* and *nuances of expressions*. This database provides *controlled text* for extraction of vocal-features (and temporal metrics) that are related only to affective states. The various modalities supply *cross-modalities information* for annotation and for verification of the annotation.

Doors also provides a controlled environment, which means a single speaker, same time and location, identical text, neutral content, multiple repetitions and no influencing parameters outside the interaction. Under these terms the only behavioural differences between utterances were related to the affective state of the speaker which changed only due to the events of the interaction, and therefore nuances of expressions could be compared and statistical analysis of the vocal expressions in comparison to the other modalities was enabled.

These additional measures of control and cross validation are important because the two databases represent two different languages and different cultures. In addition, the choice of data-sets provides the means to check if the vocal correlates of complex affective-state groups, beyond the set of basic emotions, transcend language boundaries. The *non-controlled text* in the Doors database further extends the scope of the system toward universality.

Over 7000 sentences by different speakers in two different languages were used, including 4400 acted and labelled sentences in English, over 2700 text repetitions with naturally evoked expressions in Hebrew and around 100 utterances with un-controlled text, uttered by 25 speakers of both genders and different age groups, actors and non-actors.

2.3 Architecture

The inference system was designed to recognise and rank the level of recognition of the nine chosen affective-state groups, so that several affective states could co-occur simultaneously, and different features could distinguish different affective state groups [40,36]. As described in Figure 1, the first stage of the system consisted of a speech processing stage in which vocal-features were extracted from the speech signal and processed to be the input to the classification system. The next stage consisted of classification or comparisons between every two affective-state groups followed by consolidation of the classification results from all the 36 pair-wise comparisons into a single list of the affective-state groups. In the list (from hereon referred to as the 'ranked list'), a number between 0-8 represented the number of comparisons in which each of the affective states was chosen as result. This process is described in the next sections.

3 Speech processing

The speech processing consisted of 3 stages: the first stage was the extraction of vocal-features from the speech signal (signal processing), the second stage was calculation of statistical and temporal metrics from the extracted vocal-features for each analysed sentence and the third stage was normalisation of the metrics. The normalised metrics were then entered into the classification system. Feature extraction algorithms and the definitions of temporal metrics were developed and tested using both databases, Mind Reading and Doors. The features were derived from research into the fields of affective speech recognition [55,20], speech recognition, linguistics and pragmatics [5,56], psychology, musicology [57–62], acoustics, hearing [63–65], communication disorders [66], brain research and neurology [67–71]. Models of speech production [72–74] and speech perception were examined, in addition to various tools and algorithms for each feature and metric.

The extracted vocal-features included the fundamental frequency, f_0, the vibration rate of the vocal chords, which depends on the size and tension of the vocal fold at any given time. It changes up and down in response to factors relating to stress, emotions and intonations [74–76]. Multiple extraction algorithms of the fundamental frequency were examined [77,28,74,75,78,79]. The chosen algorithm was based on Boersma's algorithm [79,80] with modifications that extended the continuity considerations in time and frequency, and required no manual intervention; energy of the speech signal, smoothed using average over a time frame and a window; spectral content based on Bark scale up to 9 KHz [81,82]; and harmonic properties [43], such as consonance and

Table 2
Distribution of features and metrics

Types	Metrics	# of metrics
f_0	Speech rate, voiced/unvoiced durations, f_0, up/down slopes, properties of peak values	34
Energy	Amplitude, max energy, durations and lapses between peak values	19
'Tempo'	Relative durations of speech parts shape of energy peaks	17
Harmonic Properties	Number and duration of harmonic intervals	19
Spectral Content	Central frequencies: 101,204,309,417,531, 651,781,922,1079,1255,1456,1691,1968, 2302,2711,3212,3833,4554,5412,6414,7617	84

dissonance. Two tones are perceived as pleasant (consonance) when the ear can separate them clearly and when they are in unison for all harmonics. Relatively small frequency intervals (relative to the fundamental frequency), are not well-distinguished and therefore perceived as 'roughness' (dissonance) [58–61,43]. All these vocal-features were calculated for overlapping time frames, with duration of 50 msec (except for very low fundamental frequency) and overlap of 40 msec. All the vocal-features were extracted automatically, with no manual intervention.

A set of metrics was defined from these basic vocal-features. It provided the means to automatically analyse the vocal-features and their temporal characteristics for the whole utterance or sentence. The metrics included statistical properties of each vocal feature, such as median, range, maximum value, mean and standard deviation. The temporal metrics drew on definitions from linguistics and musicology, such as durations and time lapses between occurrences of speech parts, for example, duration of speech or silence, voiced or unvoiced durations - places in which there are and there are no vibrations of the vocal chords, respectively, as well as metrics that drew on the definitions of tempo and melody that take into account the relations in duration and intensity between different speech parts. As these speech parts have different values along an utterance, their statistical properties were also considered. In total, the set of secondary metrics used for inference consisted of 173 parameters for each speech signal (a sentence). This set comprised of all the observed characteristics from both datasets. A summary of the metrics for which statistical measures were used appears in Table 2.

The different metrics greatly differed in their ranges (over several orders of magnitude). This caused a bias. Therefore, the metrics were normalised. Each metric was normalised separately for every speaker (and not for all the speakers). Each speaker has individual characteristics that derive from the speaker's identity, including parameters such as gender, body structure, personality, spo-

ken language and accent, or from the recording conditions. The normalisation compensated for the inter-speaker variability. As a result, the characteristics of an expression in comparison to other expressions could be compared between speakers. Another advantage of this method was that there was no need for a 'neutral' expression (which is not definable in the case of subtle expressions).

4 Classification

The input to the classification system was the values of the normalised metrics calculated for the current sentence or utterance (from hereon referred to as *features*). The output was a ranked list that rated the level of recognition of each of the affective-state groups in this sentence. The classification system consisted of a set or a series of pair-wise decision machines, each resolving a dichotomy. Each machine decided with which of the two affective-state groups, the utterance-to-be-analyzed should be associated. This method has been used for classification of affective states [24]. However, here each machine was built independently in terms of the feature sets and classification algorithms.

Observations show that different sets of features and metrics distinguish different affective states [40]. The implication for design is that classification machines that distinguish between different affective-state groups should use different metrics, and there is no need to find one small set of metrics to distinguish between all the affective states. The different machines use different metrics, therefore they are independent and the classification algorithms can also be optimised for each machine separately. The training process entailed finding the best combination of features and classification algorithm for each pair of affective-state groups. The chosen classification algorithms (after comparison to various classification and clustering algorithms [43]) were Linear Support Vector Machine and C4.5/C5.0 decision tree [83,84]. All the training was done with the data mining tool, Weka [85]. The vocal-feature extraction, the metric calculations, the implementation of the classification machines and the testing were done using Matlab. The chosen pair-wise comparison algorithms were compared to other classification methods [43] and were found at least as good while easier to implement. Incidentally, the chosen methods define a border or a threshold between classes which agrees with previous observations [40] that the distinction between expressions of complex affective states in some cases relates to thresholds rather than to cluster centres. An affective-state group was recognised by the differences or borders between its samples and the samples of each of the other affective-state groups, rather than characterised by the centre of the majority of its samples. Training was done independently for each pair-wise machine. In total 380 sentences were used for training. Using tenfold cross-validation and making sure that the two classes were recognised at similar rates. The average number of features in the

pair-wise machines was 10, which is very low compared to the full set of 173 features. However, nearly the entire set of features was required for the classification and inference of the nine chosen affective states (166 features). The features that did not appear in any of the machines were 4 of the harmonic properties, duration of f_0 down slopes and properties of two of the spectral-bands. Tenfold cross validation results for the 36 machines were on average 76% As can be seen in Table 3(classification over 80% in 11 machines, between 70-80% in 19 machines, and between 66-70% in 6 machines).

It is quite impossible to compare these results to other classification reports because the affective states, the classification methods, the features used and the speech corpora are different, though classification results of affect in speech in general could be seen in many of the references (for example [24,36,26,86,20,87]). For example, Devillers *et al.* [39] review ten sets of pair-classification results, in the range of 60%-90% (median 76%). They mostly refer to classification between well-distinct affective states or dimensions, such as positive-negative, negative vs. non-negative, emotion vs. non-emotion, frustration vs. others, and the like. When pair-wise comparisons are used, as reported for example by Vidrascu and Devillers [24], who used pair-wise comparisons because they preferred to use SVM-based classification, the results are usually reported for the overall machine that distinguishes between all the affective states. The results presented here refer to 36 pair-wise machines of more subtle and more intricate (less distinctive) affective-states and show that such affective states can be classified with similar accuracy rates. The classification results are lower for pairs of affective-state groups that are not necessarily mutually exclusive in meaning, such as joyful and excited (60%). Other relatively low results could signify similarity in the behavioural or vocal expressions of the two examined affective-state groups, or (automatically selected) sub-optimal set of metrics.

The metric sets were chosen mostly automatically, using various methods of feature selection. The observation that different features and metrics distinguish different expressions of affective states was tested. Sets of metrics that yielded near optimal classification between certain pairs of affective-state groups A-B(tenfold cross-validation over 80%) were used for classification between one of these affective-state groups and a third affective-state group A-C, the results in these tests were often close to random probability, which means that these metrics do not distinguish between these affective-state groups (A-C).

The comparison results were consolidated into a single ranked list, in which each of the recognisable affective-state groups was ranked according to the number of comparisons in which it was chosen, in the range of 0-8. This list was used in different manners for different applications.

This architecture was compared to others architectures, including a single ma-

chine for all the expressions, i.e. a machine that chose one of the nine affective states, implemented with decision-tree, neural network, polynomial SVM and Gaussian SVMs, as well as to a pair-wise architecture of only SVM machines. In these methods the performance was not good (close to random) in terms of true-positive values and tenfold cross-validation. Of these methods, only the pair-wise comparison method allows inference of more than one mental state for a single sample. Finding an optimised algorithm for each machine in a pair-wise system improved the results in comparison to a single arbitrary algorithm (such as the SVM only pair-wise system). The same applies to a single sub-set of features. Because the training was done for each pair of affective states, the machine training was relatively simple and did not take long to implement. The integration of multiple classification results from different machines further improved the reliability. The trained machines and the consolidation of the results were implemented in Matlab.

As each pair of affective states has its own comparison machine, the inference machine can be extended to accommodate additional expressions, based on different training data. In this case, only the new machines require training and they can be added to or subtracted from the existing machine without re-training the pair-wise machines that remain relevant.

5 Validation and generalisation

Validation and generalisation were performed in several stages. Each stage evaluated another capability of the inference system, extended and expanded its scope.

5.1 Inference of a single affective state

The Mind Reading database was used for training and testing. The first stage of validation was to infer one affective state for each sentence so that it could be compared to the label of the sentence in the database. For this evaluation, the Condorcet voting method was used [88,89], with the two round runoff method, a second round of pair-wise comparisons between the candidates with the maximum number of votes, in order to ensure the selection of a single leading candidate. For nine expressions of mental-states, the probability of randomly choosing an affective state is 11%, all the affective states were recognised with a higher score, as can be seen in Table 4. True-positive results appear along the diagonal. The testing was done on the full set of 549 sentences and the total detection rate was 79%. Here again the higher error rates can be related to affective-state groups that are not mutually exclusive, such as opposed and

sure.

The inference relates to complex affective states that have not been considered in other studies. Although the speech corpus is acted, and not built of samples that were carefully chosen from real data, the speech corpus comprises sentences of many nuances of affective states rather than stereotypical expressions of extreme emotions. It would be difficult to compare the detection rates to those from other studies. For example, Vidrascue and Devillers [24] report detection rate of 45% for 5 affective-state groups with a corpus from the Berlin database from call centres. They refer to other studies that achieved 77% of good detection on acted speech with 7 classes, 28% with 7 classes and 39% with 5 classes on the same data. Xiao *et al.* [36] report success rate of 78.6% for inferring 6 basic emotions from single words, and refers to other studies that examined detection of 3-6 affective states and did not reach similar detection rates. However, the methodology of testing, the data-sets, the representation manner, the number of classes are all different and this comparison is not very meaningful and can only serve as a general indicator.

5.2 Inferring co-occurring affective states

Several of the affective states can co-exist in an utterance. Therefore, instead of finding a single winner (chosen in most comparisons) it is better to look for the leading candidates. Affective-state groups that were chosen by several machines, for the same utterance, were selected. The threshold for selection was set at *over* one standard deviation above the mean number of machines, i.e. at least six machines. With this method, the average accuracy of the recognition was over 81% (random probability in this case is 14%). This method is more accurate in the sense that the label of the examined expression is more likely to be included in the inference results. A confusion matrix is presented in Table 5. It is not exactly a confusion matrix by definition because multiple affective states were chosen intentionally (the sum is greater than 100%).

5.3 Characterising affective state concepts

In the threshold method several affective states can be recognised, rather than a single affective state. The third stage of validation was to check the inference results of the co-occurring affective states beyond the given labels. The inference results were compared to the lexical meaning of the affective state concepts or to the expected behavioural characteristics. Although these criteria are highly subjective, the inferred combinations often agreed with the definitions of the concepts in dictionaries [90] and thesaurus engines. In addi-

tion, complex affective state concepts that have similar meaning had a similar or an identical inferred combination, for example fantasizing and dreamy.

Table 6 shows an example of the inference results for each sentence with the label *choosing* from the *thinking* group. It shows that in all the sentences, the expression *thinking* is the most dominant. The ranking of 8 means that an affective-state group was recognised in all the pair-wise comparisons as the most probable candidate for the expressions in the examined speech utterance. Dominant affective states appear also with ranking of 7 and 6. The affective-state groups *stressed*, *opposed* and *unsure* were recognised with high rates in the sentences labelled as *choosing*. These recognised affective states are expected and accepted behavioural expressions of the affective state concept *choosing* in different contexts. The inference can be considered correct for all the inferred combinations. However, the expression is subject to the context, and different sentences convey different nuances of the affective-state concept (different combinations of the recognised affective-state groups). A more reliable test is to accumulate the inference results of all the sentences that belong to a certain label or concept and check if they can characterise the concept, finding the affective-state groups that were recognised in most of the sentences that represent and affective-state concept, i.e. ranking by concept. Therefore, a double-threshold was applied - only affective states chosen by six or more machines (over one standard deviation above the mean), in four of the six sentences (over one standard deviation above the mean number of sentences, >66%) were considered *dominant* for the concept. Affective states chosen with the same high rate in three of these sentences, i.e. 50% of all the sentences and 50%-75% of the sentences in which a *dominant* affective state was found, were considered as *possibly influential*. In the example of *choosing*, the affective states *thinking* and *unsure* were dominant, chosen by six or more machines in more than four of the six sentences, as summarised in the two rows at the bottom of the table. The combinations of *thinking*, *stressed* and *opposed* appeared only in a small number of sentences and therefore could not be considered significant or dominant for the affective state concept in general. The result of *thinking* and *unsure* was closer to the meaning of the term *choosing* and characterised its meaning and the expected behaviour related to it. This example demonstrates the cause for differences between the results in Table 4 and Table 5. It also justifies the choice of database and architecture, because nuances of affective states are distinguishable, using different combinations of the affective state groups.

In the same manner the inference results of all the 93 concepts in the testing set were evaluated and found to be similar to the lexical definitions and meanings of the examined concepts and to the behavioural expressions associated with them. All the thresholds were calculated and summarised automatically.

5.4 Characterising a large variety of affective state concepts

An experiment was performed in order to extend the scope of the examination to the entire voice part of the Mind Reading database. For validation, the characteristics of affective state concepts were examined rather than the inference results for a single sentence. The double-threshold procedure was applied as before and 459 affective state concepts were characterised. At least 85% of the characterisation results are comparable to the related lexical definitions or signify the expected behaviour associated with the concept. These definitions are not precise, and six samples per affective state are not a large set, but characterising 459 affective state concepts comprising at least 1700 sentences, using the double threshold procedure, is a meaningful result (far from random). The testing at this stage included concepts from all the 24 groups of the Mind Reading database.

This experiment includes also concepts that could belong to the chosen affective-state groups but for this additional validation were left out of the initial training and testing sets. For example, the affective states *dreamy* and *considering* from the *thinking* group in the Mind Reading taxonomy, for which the affective state *thinking* was recognised with nearly 100% accuracy, in combination with other affective states.

The inference was done automatically for all the sentences and concepts. The inference results of most of the 4400 sentences of the database reveal recognised affective-state groups (first threshold) from the accumulated results of the 36 machines. Only 36% of the affective-state concepts could not be characterised, using the 2nd threshold criteria.

5.5 Comparison to human performance

The distinguishing capabilities of the system were compared to human performance on the CAM Battery Test [46], in which the inference results of a sentence that belongs to a certain affective state concept were compared to the inference of three different affective state concepts, for 50 sentences. The results were compared to those of a test in which humans had to choose the correct label for this sentence, given its label and the labels of the three other affective state concepts. All the affective states that were chosen for this test can generally be recognised and understood only by humans over the age of 15. In this test, the machine inferred different combinations for the sentence and the other concepts in 49 of the 50 sentences, which outperforms the human results, with average of 46 of the 50 sentences, as reported by Golan *et al.* [46].

5.6 Annotation and analysis of sustained interactions

The inference system was used for automatic annotation of six sustained interactions by six different players from the Doors database (described in Section 2.2).

Thus, a ranked list of the recognition level of the nine affective-state groups was generated for each utterance. Each interaction (game) lasted about 15 minutes, with over 200 sentences. Each of the metrics in the samples of each speaker was normalised, as described in the speech processing stage. No additional training was required. For the analysis of sustained interactions, for each sentence the ranked list was used with all the levels of recognition between 0 and 8. An example of annotation results for a short sequence of sentences from an interaction in the Doors database, can be seen in Table 7. In the table, each line represents a sentence. The annotation results are presented as numbers in the range 0-8 that signify the observed rank of the affective states that head the coloumns. In this case, gradual changes of the observed rank of an affective state can be seen between successive sentences, as well as sudden changes.

The fully controlled environment allowed for statistical correlation to be used. The validation was done through analysis of the interactions and correlation to contextual cues (using t-test, $p<0.01$). Significant correlation was found between the inferred mental states and game events such as total gain, temporary loss and gambling on good or bad doors (doors with positive or negative gain expectation) [42,43].

For example, the *confidence* level of a participant, who could explain the right strategy to maximise the gain in the game, increased after a positive feedback from the experimenters for the explanation in the intervening review (mean before feedback 2.6, mean after feedback 3.3, $p<0.01$). The confidence level was inferred from the speech. In addition, from this stage onward the inference of the affective state *interest* decreased. Analysis of the verbal content revealed that at the same time the participant started complaining of boredom, in sentences of uncontrolled text. The inferred affective states for these sentences revealed the same attitude (mean recognition of *opposed* before feedback 3.3, and after feedback 4.1, $p<0.01$).

Another participant wanted to win (as reported in the interview). The mean of the inferred *stress* level was‘significantly lower after the participant was asked by the experimenters to loose on purpose (mean before 3.5, mean after 2.6, , $p<0.01$).

For a third participant significant differences in the inference of *joyful* and of *stress* were found between events in which the total gain was positive (*joy* was high and *stress* was low) and events in which the gain was negative (*joy* was

low and *stress* was high).

These results, which were statistically significant, showed that the automatic annotation was meaningful. This implied that the inference system could be applied to the Doors database. In order to examine the capabilities of the system to infer dynamic changes over time, dynamic analysis of the affective states was conducted. Research efforts have only started to address the problem of changes in expressions over time [91,92]. For example, Picard successfully recognise certain emotions from physiological reactions of an actor, recorded on different days [91], Fernandez and Picard investigate stress in drivers' speech [26]. The connection between physiological reaction, affect and decision making, though widely used [10,93], is limited to few general affective states. Galvanic skin response (GSR) response-time is limited by the response time of the physiological system (delay of 3-7 sec), the response time of other cues such as heart rate variability are measured in minutes. In general, speech provides a more immediate response, and as demonstrated here, a larger variety of affective states can be inferred. Here, the inferred affective states were compared to events. In addition, preliminary comparisons of the inference results to physiological cues, such as GSR and to other behavioural cues, such as mouse movement rate and the verbal content of the uttered sentences were analysed. The analysis included all the speech utterances, with controlled and uncontrolled text [43]. As can be seen for example in Figure 2, the temporal analysis showed that the level of inferred stress of the participant who was asked during the second interview to loose on purpose, decreased at once with the request to loose. The same change appeared also in the baseline of the GSR measurements recorded at the same time.

During the interview in which the participant was asked to choose the doors with the low expectation, the participant asked questions about the request and thought about it. These events appeared both in the verbal content and in the automatically inferred affective states *thinking* and *interested*.

The participant who got positive feedback during the intermediate interview laughed with the affirmation of success and knowledge, the inferred affective state at this stage was (*joy*. At another stage of the experiment, the same participant complained laughingly about part of the experimental setup (the electrodes for ECG measurements), in this sentence the inference results showed a combination of (*joy* and *stress*).

Correlation was found between the outcome of a door opening and the inferred affective state. For example, a decrease of confidence after a big loss. These results were supported by physiological cues such as a change in the GSR at the same point, and behavioural cues such as statistically significant increased mouse movement rate and longer latent periods before the next decision.

These examples and additional results [42,43] demonstrate the capabilities of the system to infer affective states and to automatically annotate affective states in speech corpora. The examples show that the system can generalise to (at least some) new speakers with no additional training. They also show that the vocal expressions of complex affective states (acted or naturally evoked) transcend language or culture barriers (at least between English and Hebrew, English and Israeli speakers).

6 Summary and conclusions

This paper presents an approach to the inference of affective states from their non-verbal expressions in speech with the goal to develop systems that infer affective states as occur in real scenarios. The framework presents a process of design, implementation and verification that leads towards a general solution, i.e. a system suitable for inference of affect in real scenarios, including a wide range of affective states and speakers. The paper presents an implementation that achieves most of the design goals.

The framework includes a representation method of the domain of affective state concepts using the prototype approach. It uses a small set of affective state groups shared by different cultures and languages, that appear in everyday life and whose combinations represent a large variety of affective states and behavioural patterns. The inference results are presented in a manner that can be both understood by people and used by machines (much of the processing for the analysis was done automatically).

The framework also includes the selection and definition of complementing data-sets that enable training, testing and controlled validation and generalisation. A corpora comprising a very large variety of acted and labelled affective state concepts was chosen for training, testing and characterisation of 459 affective state concepts. A multi-modal database of naturally evoked expressions during sustained interactions, with new speakers and language, was recorded and used for further validation, after verifying that the inference results for this corpora are meaningful. Although the training stage itself was not performed on recorded corpora of naturally evoked affective states, the combination of these data-sets uses their respective advantages for an overall design and validation process that exceeds the capabilities of the current state-of-the-art.

The inference is performed for utterances or sentences, which are meaningful units of speech. A large set of features is used for the classification. Inter-speaker variability was neutralised by normalising the values of each feature for each speaker. It enables the system to be used for new speakers with no

additional training. It also means that the system considers only the changes between different expressions and there is no need to define a 'neutral' expression. Generalisation to new languages is achieved by the combination of normalisation of speech metrics and of the representation method that uses affective-state groups that are common to different cultures.

The representation manner and the choice of affect corpora are fundamental to the design of the system and both are based on the definition of the term *affective state* and the goal. An important guideline in the design of the system is that different sets of features distinguish different expressions.

The architecture consists of pair-wise comparisons between expression groups. Different features distinguish different affective states. Therefore, each pair-wise machine was built with its own sub-set of features and classification algorithm. The pair-wise comparisons are consolidated into a single ranked list that reflects the number of comparisons in which each expression was chosen. The ranked list represents levels of inference of co-occurring affective states.

The validation process is part of the framework. It enables gradually extending the scope of the inference system capabilities by using the respective advantages of both data-sets. Due to the combination of representation method, data-sets and architecture, relatively few affective-state groups sufficed to successfully infer and characterise a very large range of affective states. The system successfully inferred affective states within and beyond the set of affective-state groups that it was trained to infer, including nuances of expressions, and affective states that can co-occur simultaneously and change dynamically over time.

Adding affective-state groups to the system would require adding few pair-wise machines while no re-training of the existing machines is required. Thus the system could be easily adapted to various applications.

The successful inference of affective states in different languages (and cultures) indicates that the vocal cues of complex affective states transcend language and culture barriers.

7 Acknowledgement

The author thanks Peter Robinson, Yael Edan and Yehuda Werner for their help and contribution. The author thanks the AAUW Educational Foundation, Cambridge Overseas Trust and Deutsche-Telekom Labs at Ben-Gurion University of the Negev for their partial support of this research.

References

[1] B. Reeves, C. Nass, The media equation, Cambridge University Press, 1996.

[2] R. W. Picard, Affective Computing, MIT Press, Boston, 1997.

[3] C. Becker, S. Kopp, I. Wachsmuth, Conversational informatics: An engineering approach, T. Nishida (Ed.), Wiley, 2007, Ch. Why emotions should be integrated into conversational agents, pp. 49–68.

[4] E. Douglas-Cowie, N. Campbell, R. Cowie, P. Roach, Emotional speech: towards a new generation of databases, Speech Communication 40 (2003) 33–60.

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

[6] R. Cornelius, Theoretical approach to emotion, in: ISCA Workshop on Speech and Emotion, Belfast, 2000.

[7] A. Whiten, Natural theories of mind, Basil Blackwell, Oxford, 1991.

[8] S. Baron-Cohen, The descent of mind: Psychological perspectives on hominid evolution, M. Corballis, S. Lea (Eds.), Oxford University Press, 1999, Ch. Evolution of a theory of mind?

[9] W. James, What is an emotion?, Mind 19 (1884) 188–205.

[10] A. Bechara, H. Damasio, D. Tranel, A. R. Damasio, Deciding advantageously before knowing the advantageous strategy, Science 275 (1997) 1293–5.

[11] K. R. Scherer, Studying the emotion-antecedent appraisal process: An expert system approach, Cognition and Emotion 7 (1993) 325–355.

[12] R. Zajonc, Feeling and thinking: Preferences need no inferences, American Psychologist 35 (1980) 151–175.

[13] M. V. den Noort, M. P. C. Bosch, K. Hugdahl, Understanding the unconscious brain: Can humans process emotional information in a non-linear way?, in: The International Conference on Cognitive Systems, New Delhi, December, 2005.

[14] D. Kahneman, A. Tversky, Prospect theory: An analysis of decision under risk, Econometrica XVLII (1979) 263–291.

[15] K. R. Scherer, How emotion is expressed in speech and singing, in: Proceedings of the XIIIth International Congress of Phonetic Sciences, ICPhS95, Stockholm, Sweden, 1995, pp. 90–96.

[16] J. D. Haynes, G. Rees, Decoding mental states from brain activity in humans, Nature Reviews Neuroscience 7 (2006) 523–534.

[17] M. Slors, Personal identity, memory, and circularity: An alternative for q-memory, The Journal of Philosophy 98 (4) (2001) 186–214.

[18] K. Höök, From brows to trust: Evaluating embodied conversational agents, Z. Ruttkay, C. Pelachaud (Eds.), Vol. 7, Kluwer, 2004, Ch. User-centred design and evaluation of affective interfaces.

[19] V. Petrushin, Emotion in speech: Recognition and application to call centers, in: ANNIE, 1999.
URL citeseer.ist.psu.edu/petrushin99emotion.html

[20] P. Y. Oudeyer, The production and recognition of emotions in speech: Features and algorithms, International Journal of Human Computer Interaction 59 (1-2) (2003) 157–183.

[21] P. Ekman, Handbook of cognition and emotion, M. Power, T. Dalgleish (Eds.), Wiley, Chihester, UK, 1999, Ch. Basic emotion.

[22] C. Darwin, The Expression of the Emotions in Man and Animals, D. Appleton and Company, New-York, 1898.

[23] R. Cornelius, R. Cowie, Describing the emotional states that are expressed in speech, Speech Communication 59.

[24] L. D. L. Vidrascu, Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features, in: Paraling2007, 2007.

[25] R. Plutchik, The nature of emotions, American Scientist online 89 (4) (2001), URL
http://www.americanscientist.org/articles/01articles/Plutchik.html.

[26] R. Fernandez, R. W. Picard, Modeling drivers' speech under stress, Speech Communication 40 (2003) 145–59.

[27] Nemesysco Ltd.- Voice Analysis Technologies,
URL http://www.nemesysco.com/ (Sept 2006).

[28] C. A. Moore, J. F. Cohn, G. S. Katz, Quantitative description and differentiation of fundamental frequency contours, Computer Speech and Language 8 (4) (1994) 385–404.

[29] M. Schröder, Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis, Tech. rep., The Institute of Phonetics, Saarland University (2004).

[30] K. R. Scherer, Approaches to emotion, K. R. Scherer and P. Ekman, Hillsdale, 1984, Ch. On the nature and function of emotion: a component process approach, pp. 293–317.

[31] C. M. Whissell, Emotion: Theory, Research, and Experience, R. Plutchik and H. Kellerman (Eds.), Academic Press, New York, 1989, Ch. The dictionary of affect in language, pp. 113–131.

[32] J. Kim, Robust Speech Recognition and Understanding, M. Grimm and K. Kroschel (Eds.), I-Tech Education and Publishing, Vienna, 2007, Ch. Bimodal Emotion Recognition using Speech and Physiological Changes.

[33] M. Grimm, K. Kroschel, Robust Speech Recognition and Understanding, M. Grimm and K. Kroschel (Eds.), I-Tech Education and Publishing, Vienna, 2007, Ch. Emotion Estimation in Speech Using a 3D Emotion Space Concept.

[34] J. Wagner, J. Kim, E. Andr, From physiological signals to emotions: Implementing and comparing selected methods for feature extraction and classification, in: IEEE International Conference on Multimedia and Expo (ICME 2005), 2005, pp. 940–943.

[35] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, M. Schröder, 'feeltrace': An instrument for recording perceived emotion in real time, in: ISCA Workshop on Speech and Emotion, Belfast, Northern Ireland, 2000, pp. 19–24.

[36] Z. Xiao, E. Dellandrea, W. Dou, L. Chen, Automatic hierarchical classification of emotional speech, in: Multimedia Workshops, ISMW '07, 2007, pp. 291–296.

[37] Humaine deliverable d5f, HUMAINE Netwrok of Excellence, EU's 6^{th} Framework Project, (2006) URL `http://emotion-research.net/projects/humaine/deliverables`

[38] M. Y. M. Hoque, M. Louwerse, Robust recognition of emotion from speech, in: 6th International Conference on Intelligent Virtual Agents, Marina del Rey, 2006.

[39] L. Devillers, L. Vidrascu, L. Lamel, Challenges in real-life emotion annotation and machine learning based detection, Neural Networks 18 (2005) 407–422.

[40] T. Sobol-Shikler, P. Robinson, Visualizing dynamic features of expressions in speech, in: proceedings of ICSLP, Jeju, Korea, 2004.

[41] F. Dellaert, T. Polzin, A. Waibel, Recognizing emotions in speech, in: ICSLP 96, 1996.

[42] T. Sobol-Shikler, Multi-modal analysis of human computer interaction using automatic inference of aural expressions in speech, in: IEEE International Conference on Systems, Man, and Cybernetics, SMC 2008, Singapore, 2008.

[43] T. Sobol-Shikler, Analysis of affective expressions in speech, Tech. rep., Computer Laboratory, University of Cambridge, US Patent Pending, 2007, (2009).

[44] M. Schröder, R. Cowie, Issues in emotion-oriented computing - towards a shared understanding, in: Workshop on Emotion and Computing at KI 2006, Bremen, Germany, 2006.

[45] S. Baron-Cohen, O. Golan, S. Wheelwright, J. J. Hill, Mindreading: The interactive guide to emotions, Jessica Kingsley Limited,
URL `http://www.jkp.com`, (2004).

[46] O. Golan, S. Baron-Cohen, J. Hill, The Cambridge Mindreading (CAM) face-voice battery: Testing complex emotion recognition in adults with and without Asperger Syndrome, Journal of Autism and Developmental Disorders 23 (2006) 7160–7168.

[47] T. Sobol-Shikler, P. Robinson, Recognizing expressions in speech for human computer interaction, in: Designing a More Inclusive World, S. Keates, J. Clarckson, P. Langdon and P. Robinson (Eds.), Springer-Verlag, 2004.

[48] R. el Kaliouby, P. Robinson, Real-Time Vision for HCI, Spring-Verlag, 2005, Ch. Real-time Inference of Complex Mental States from Facial Expressions and Head Gestures, pp. 181–200.

[49] S. Afzal, P. Robinson, A study of affect in intelligent tutoring, in proceedings of the workshop on modelling and scaffolding affective experiences to impact learning, in: International Conference on Artificial Intelligence in Education, Los Angeles, 2007.

[50] A. Wiezrbicka, Emotion and Culture: Empirical Studies of Mutual Influence, American Psychological Association, S. Kitayama and H. R. Markus (Eds.), Washington, 1994, Ch. Emotion, language and cultural scripts.

[51] A. Wierzbicka, The semantics of human facial expressions, Pragmatics and cognition 8 (1) (2000) 147–183.

[52] S. Baron-Cohen, J. J. Hill, O. Golan, S. Wheelwright, Mindreading made easy., Cambridge Medicine 17 (2002) 28–29.

[53] Z. Xiao, E. Dellandrea, W. Dou, L. Chen, Two-stage classification of emotional speech, in: Digital Telecommunications, ICDT '06, 2006.

[54] S. Jia-Lin, H. Jeih-Weih, L. Lin-Shan, Robust entropy-based endpoint detection for speech recognition in noisy environments, in: International conference on spoken language Processing, Sydney, Australia, 1998.

[55] R. Fernandez, R. W. Picard, Classical and novel discriminant features for affect recognition from speech, in: Interspeech 2005 - Eurospeech 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 2005.

[56] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, J. Hirschberg, Tobi: A standard for labelling english prosody, in: ICSLP92, 1992, pp. 867–870.

[57] I. Johnston, Measured Tones: The interplay of physics and music, Adam Hilger, IOP publishing, New-York, 1989.

[58] P. Gorman, Pythagoras, a life, Routledge and K. Paul, London, 1979.

[59] Galileo, Dialogues Concerning Two New Sciences, Dover Publications Inc., New-York, 1954.

[60] D. A. Schartz, Q. C. Howe, D. Purves, The statistical structure of human speech sounds predicts musical universals, The Journal of Neuroscience 23 (2003) 7160–7168.

[61] M. J. Tramo, P. A. Cariani, B. Delgutte, L. D. Braida, The cognitive neuroscience of music, I. Peretz and R. Zatorre (Eds.), Oxford University Press, New York, 2003, Ch. Neurobiology of harmony perception.

[62] D. M. Green, Form in tonal music: and introduction and analysis 2nd ed, Holt, Rinehart and Winston, New-York, 1979.

[63] S. A. Gelfand, Hearing - an introduction to psychological and physiological acoustics, Marcel Dekker, Inc, 1998.

[64] A. D. Cheveigne, H. Kawahara, Multiple period estimation and pitch perception model, Speech Communication 27 (3-4) (1999) 175–85.

[65] E. D. Scheirer, Music-listening systems, MIT Press, Boston, 2000.

[66] P. G. Stelmachowicz, A. L. Pittman, B. M. Hoover, D. E. Lewis, Effect of stimulus bandwidth on the perception of s in normal- and hearing-impaired children and adults, Journal of the Acoustical Society of America 110 (4) (2001) 2183–90.

[67] D. J. Weiss, A. A. Ghazanfar, C. T. Miller, M. D. Hauser, Specialized processing of primate facial and vocal expressions: Evidence for cerebral asymmetries, in: Cerebral Vertebrate Lateralization, L. Rogers and R. Andrews (eds.), New York, Cambridge University Press, 2002.

[68] A. R. Damasio, D. Tranel, H. Damasio, Faces and the neural substrates of memory, Annual Review of Neuroscience 13 (1990) 89–109.

[69] S. J. L. Mozziconacci, Modeling emotion and attitude in speech by means of perceptually based parameter values, User Modeling and User-Adapted Interaction 11 (4) (2001) 297–326.

[70] I. R. Murray, J. L. Arnott, Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion, Journal of the Acoustical Society of America 93 (2) (1993) 1097–108.

[71] R. Adolphs, D. Tranel, Intact recognition of emotional prosody following amygdale damage, Neuropsychologia 37 (1999) 1285–1292.

[72] G. Fant, Acoustic theory of speech production, Mouton and Co., 1960.

[73] J. Flanagan, Speech analysis synthesis and perception, Springer-Verlag, 1972.

[74] J. Markel, A. H. J. Gary, Linear Prediction of Speech, Springer, Berlin, 1976.

[75] J. R. J. Deller, J. G. Proakis, J. H. L. Hansen, Discrete-Time Processing of Speech Signals, Macmillan Publishing Company, New York, 1993.

[76] L. R. Rabiner, R. W. Schafer, Digital Processing of speech signals, Prentice Hall PTR, 1978.

[77] E. Terhardt, G. Stoll, M. Seewann, Algorithm for extraction of pitch and pitch salience from complex tonal signals, J. Acoust. Sec. Amer. 71 (1982) 679.

[78] W. W. Zhao, T. Ogunfunmi, Formant and pitch detection using time-frequency distribution, International Journal of Speech Technology 3 (1) (1999) 35–49.

[79] P. Boersma, Accurate short-term analysis of the fundamental frequency and the harmonics to-noise ratio of a sampled sound, in: Proceedings of the Institute of Phonetic Sciences, Amsterdam, 1993.

[80] P. Boersma, D. Weenink, Praat: doing phonetics by computer.

[81] E. Zwicker, G. Flottorp, S. S. Stevens, Critical bandwidth in loudness summation., The Journal of the Acoustical Society of America Volume 29 (1961) 548–57.

[82] E. Zwicker, Subdivision of the audible frequency range into critical bands (Frequenzgruppen), The Journal of the Acoustical Society of America Volume 33 (1961) 248.

[83] V. Vapnik, Estimation of Dependences Based on Empirical Data, Springer-Verlag, 1982.

[84] J. R. Quinlan, C4.5: Programs for machine learning, Morgan Kaufmann, San Mateo, CA, 1993.

[85] I. H. Witten, E. Frank, Data mining: Practical machine learning tools with java implementations, in: Morgan Kaufmann, San Francisco, 2000.

[86] R. Fernandez, A computational model for the automatic recognition of affect in speech, Ph.D. thesis, Media Arts and Sciences Lab, Massachusetss Institute of Technology (2004).

[87] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, V. Aharonson, The impact of f0 extraction errors on the classification of prominence and emotion, in: Proceedings of the 16th International Congress of PhoneticSciences, (ICPhS 2007),Saarbrcken, 2007, pp. 2201–2204.

[88] C. M. J. A. N. marquis de Condorcet, Essay on the application of analysis to the probability of majority decisions (1786).

[89] J. Malkevitch, The process of electing a president, AMS, American Mathematical Society (April 2008).

[90] Cambridge dictionaries online, Cambridge University Press, URL `http://dictionary.cambridge.org/` (2008).

[91] R. W. Picard, E. Vyzas, J. Healey, Toward machine emotional intelligence: analysis of affective physiological state, IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (10) (2001) 1175–91.

[92] G. Klasmeyer, An automatic description tool for time-contours and long-term average voice features in large emotional speech databases, in: ISCA workshop (ITRW) on speech and emotion: a conceptual framework for research, Belfast, N. Ireland, 2000.

[93] J. Healey, R. W. Picard, Digital processing of affective signals, in: IEEE Conference on Multimedia, ICASSP 1998, pp. 251–252.

Table 3
Tenfold cross-validation of the 36 pair-wise machines

	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
joyful	82%	83%	61%	60%	71%	77%	75%	72%
absorbed		84%	87%	81%	78%	82%	64%	73%
sure			84%	79%	72%	78%	78%	75%
stressed				73%	84%	66%	68%	72%
excited					74%	71%	64%	79%
opposed						75%	79%	81%
interested							72%	83%
unsure								89%

Table 4
Confusion matrix of the inference machine using the Condorcet method.

		Recognised Expression							
Data Class	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
joyful	0.67	0.05	0.04	0.03	0.05	0.04	0.06	0.03	0.03
absorbed	0.00	0.91	0.07	0.00	0.00	0.00	0.02	0.00	0.00
sure	0.03	0.11	0.75	0.00	0.02	0.05	0.02	0.00	0.02
stressed	0.04	0.06	0.02	0.78	0.02	0.03	0.02	0.03	0.00
excited	0.10	0.07	0.10	0.11	0.60	0.00	0.02	0.00	0.00
opposed	0.05	0.05	0.19	0.05	0.05	0.59	0.00	0.02	0.00
interested	0.02	0.00	0.00	0.07	0.00	0.02	0.89	0.00	0.00
unsure	0.03	0.14	0.01	0.10	0.04	0.04	0.13	0.46	0.05
thinking	0.00	0.05	0.00	0.07	0.00	0.02	0.02	0.05	0.79

Table 5
Results of the inference machine using the threshold method.

	Recognised Expression								
Data Class	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
joyful	0.75	0.09	0.07	0.08	0.16	0.19	0.06	0.11	0.02
absorbed	0.02	0.93	0.10	0.02	0.02	0.02	0.02	0.17	0.17
sure	0.03	0.12	0.87	0.02	0.12	0.28	0.02	0.08	0.02
stressed	0.06	0.08	0.01	0.78	0.11	0.10	0.00	0.20	0.01
excited	0.07	0.10	0.12	0.14	0.81	0.17	0.00	0.02	0.02
opposed	0.12	0.05	0.27	0.02	0.07	0.88	0.00	0.12	0.00
interested	0.02	0.12	0.05	0.10	0.05	0.02	0.95	0.26	0.02
unsure	0.00	0.17	0.01	0.17	0.05	0.05	0.08	0.76	0.18
thinking	0.01	0.16	0.05	0.06	0.05	0.04	0.01	0.25	0.90

Table 6
An example of inference of co-occurring affective states for sentences that are labelled with a single mental state concept *choosing*. The three top rows represent inferred ranked lists for 3 sentences. Grey shades mark expressions chosen by 6-8 machines. The final definition of the concept is at the 2 bottom lines, stating the number of sentences in which an expression was recognised: • recognition in 4-6 sentences

Concept	**joyful**	**absorbed**	**sure**	**stressed**	**excited**	**opposed**	**interested**	**unsure**	**thinking**
choosing1.wav	2	5	2	3	3	4	3	7	7
choosing2.wav	4	0	1	6	5	5	2	6	7
choosing3.wav	4	3	1	4	2	6	3	6	7
choosing4.wav	3	5	2	3	1	6	2	6	8
choosing5.wav	5	5	2	4	3	2	3	4	8
choosing6.wav	5	5	2	4	3	2	3	4	8
Choosing								4	6
Choosing								•	•

Table 7
Automatic annotation results for a sequence of sentences from an interaction in the Doors database.

	joyful	**absorbed**	**sure**	**stressed**	**excited**	**opposed**	**interested**	**unsure**	**thinking**
sentence 1	7	3	2	3	5	7	4	2	3
sentence 2	5	4	2	5	5	7	5	1	2
sentence 3	5	4	6	3	4	5	6	1	2
sentence 4	7	4	5	2	3	5	2	4	4
sentence 5	7	5	4	2	2	5	3	4	4

Fig. 1. Schematic description of the inference machine.

Fig. 2. Recordings of automatic inference results of *stress*, *sure* and *absorbed*, in comparison to gain and to skin conductivity during an interaction. Interludes of interviews are marked by dotted lines.