Classification of complex information: Inference of co-occurring affective states from their expressions in speech

Tal Sobol-Shikler, *Member, IEEE,* and Peter Robinson

***Abstract*—We present a classification algorithm for inferring affective states (emotions, mental states, attitudes and the like) from their non-verbal expressions in speech. It is based on the observations that affective states can occur simultaneously and that different sets of vocal features, such as intonation and speech rate, distinguish between non-verbal expressions of different affective states. The input to the inference system was a large set of vocal features and metrics that were extracted from each utterance. The classification algorithm conducted independent pair-wise comparisons between nine affective-state groups. The classifier used various subsets of metrics of the vocal features and various classification algorithms for different pairs of affective-state groups. Average classification accuracy of the 36 pair-wise machines was 75%, using tenfold cross-validation. The comparison results were consolidated into a single ranked list of the nine affective-state groups. This list was the output of the system and represented the inferred combination of co-occurring affective states for the analysed utterance. The inference accuracy of the combined machine was 83%. The system automatically characterised over 500 affective state concepts from the Mind Reading database. The inference of co-occurring affective states was validated by comparing the inferred combinations to the lexical definitions of the labels of the analysed sentences. The distinguishing capabilities of the system were comparable to human performance.**

***Index Terms*—Affective computing, human perception, cognition, affective states, emotions, speech, machine learning, intelligent systems, multiclass, multi-label.**

I. Introduction

AFFECTIVE states and their behavioural expressions, and in particular their non-verbal expressions in speech, are important aspects of human reasoning, decision-making and communication [1]–[4]. According to the 'Theory of mind' [5], [6], affective states such as beliefs, intents, desires, pretending and knowledge, can be the cause of behaviour and thus can be used to explain and predict others' behaviour. The integration of affective states and their behavioural correlates in fields such as human computer interfaces and interactions (HCI), human-robot interactions (HRI) and speech technologies can enhance the system and user performance and has many potential applications [3], [4], [7]. Therefore, there is an increased interest in detecting, analysing and imitating these cues.

In this paper the term *affective states* refers to emotions, mental states, attitudes, beliefs, intents, desires, pretending, knowledge and moods. Their expressions reveal additional information regarding the identity, personality, psychological 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 [8], [9]: affective states and their expressions are part of social behaviour [10], [11], with relation to physiological and brain processes [2], [12]. They comprise both conscious [13] and unconscious reactions [2], [14], [15], and have cause and effect relations with cognitive processes such as decision making [1], [2]. A number of affective states can occur simultaneously [16]–[18], and change dynamically over time. A similar view of the concept *affective states* is given by Höök [19] who describes affect as human, rich, complex and ill-defined experience.

The term 'expression' refers here to the outward representation of affective states. This is the observable behaviour (conscious or unconscious) that people can perceive and would like to interpret. It can be affected by factors such as context and cultural display rules. This perspective is also reflected in automatic synthesis systems that aim to imitate only the behavioural expressions and not their source (automatic systems do not feel nor think at this stage).

Several affective states often occur simultaneously [17], [18], [20]. The existence of co-occurring affective states can be a product of the different time spans that characterise different affective states, and of the wide range of contexts, personalities and people reactions. Examples of co-occurring affective states and their behavioural expressions include being happy and at the same time showing interest, tiredness etc.; genuine joyful and amused laughter vs. stressed laughter; thoughts resulting in the expression of confidence and excitement, or uncertainty, misunderstanding and stress. Mixtures of conflicting affective states (such as the aversion and attraction that some people feel towards snakes) can also occur. These examples represent only a small part of the repertoire of co-occurring affective states that people express and infer on a daily basis. The issue of co-occurring affective states in speech has been discussed in the literature [17], [18] and annotation (labelling) methods of co-occurring affective states in speech corpora have been presented [20], [21]. For example, Devillers *et al.* [20] present annotation of a major affective state and a secondary affective state for each sentence. However, no automated inference solution has been suggested.

The challenges for the design of systems that infer affective states from their expressions are:

Manuscript received January 2008; revised September 2008, February 2009.
T. Sobol-Shikler is with Ben-Gurion University of the Negev
P. Robinson is with the University of Cambridge

- Create a general framework that can handle a large variety of affective states and their expressions rather than a system that is specific to predefined emotions.
- Recognise affective states that often occur in everyday life (rather than strong expressions of basic emotions that are rarely experienced or seen).
- Handle various affective states that occur simultaneously in a speaker-independent manner.

There are three main approaches to the inference of affective states from their non-verbal expressions. The most commonly used approach [21]–[26] is the *categorical* approach, which entails the inference of a small set of *basic emotions* [27], such as *happy*, *sad*, *angry*, *afraid*, *disgusted* and *surprised*. The term *basic emotions* refers to qualitatively distinct states that are held to be universal at least in essence, i.e. recognisable by most people from most backgrounds. 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 developing inference systems. However, these emotions do not encompass the entire range of human affective states, and do not relate to nuances of affective states and their expressions, although in recent works they are defined as groups comprising several affective states each [25]. Inference of a single emotion for each analysed sentence limits the scope of the inference and its ability to grasp the complexity of the information. Furthermore, if the small set is used only as a starting point, it is an open question whether the same behavioural cues can be used for both extreme emotions and subtle expressions of complex affective states.

The second approach is to detect the existence of a selected affective state in real situations, such as drivers' stress, attempts at insurance fraud and post-natal depression [28]–[30]. This method is basically a bi-polar classification in which a state either exists or not. It does not refer to other co-occurring affective states.

The third approach, which has recently become more widespread, is the *dimensional* approach, in which several expressions are represented in a one, two or three dimensional space, with dimensions such as passive-active, positive-negative and low-high arousal levels [8], [31]–[36]. 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, but they do not reflect the large variety of affective states nor the different levels of their experience. Various combinations of the categorical approach and the dimensional approach have been offered [26], [37]. For example, Xiao *et al.* [26] present two methods for hierarchical inference of six basic emotions. The first comprises classification according to active-passive (dimensional approach) as a first stage, and then classification into the single emotions (categorical approach). The second method comprises first classification of speakers' gender, followed by hierarchical classification into a binary graph. These methods provide a better resolution version of the dimensional approach, with or without a label for each sub-set of emotions. However, there are complex affective states that cannot be distinguished in this manner because the transition between them is gradual and therefore blind clustering (unsupervised learning) is not effective [38].

These three approaches refer to affective states as single entities, although co-occurrences of affective states are common. 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.

Researchers who develop automatic recognition systems of affective states from speech try to define one set of metrics or attributes that are calculated from the vocal or speech features, which are extracted from the speech signal, to distinguish between all the affective states they infer [8], [23], [24], [39]. However, comparison between the results shows no agreement about the role and the significance of each metric. Furthermore, there is no common basis for comparison because they use different databases, affective states, features and metrics. Xiao *et al.* [26] use one set of metrics but refer to different calculated *masses* of metrics to distinguish between different groups of affective states at different levels of a hierarchical classification, and the actual weight of each metric in the classification is not known.

In this paper, we present a different approach, which is to infer co-occurring affective states for each utterance. A classification method whose output for each sample is a set of multiple classes rather than a single class was developed. The classification results reflect shades of affective states and nuances of expressions and not only their detection. This method uses different sets of vocal features and metrics to distinguish between different affective states. This approach has also been adopted by the recently published W3C Emotion Markup Language [40].

The paper first describes the methodology that was used in this research. It then focuses on the classification, the validation of the inference and the system generalisation.

II. Methodology

The design of the classification system was influenced by four main factors. The first was the goal of recognising *co-occurring* affective states. This goal evolved from the need to recognise human behaviour as it occurs in real situations and scenarios. The second factor, which evolved from the first, was the choice of a representation or conceptualisation method to represent the large range of affective states and the relations between them. The third factor was the choice of data-sets for training and testing. The fourth factor was derived from the observation that different vocal features and metrics distinguish the expressions of different affective states [26], [38].

The choice of underlying theory and representation method of knowledge and meanings in the problem domain defines the scope of the system and its limitations. It influences the definitions of the system's input, output and architecture, in addition to the method and scope of training and testing data.

TABLE I
THE 24 AFFECTIVE STATE GROUPS (EMOTIONS) THAT CONSTITUTE THE MIND READING TAXONOMY OF BARON-COHEN *et al*. [44]. BASIC EMOTIONS ARE LISTED IN THE LEFT COLUMN. THE GROUPS THAT ARE ADDRESSED IN THIS PAPER ARE INDICATED WITH A *. TWO GROUPS WERE EXTRACTED FROM THE INTERESTED GROUP: INTERESTED AND ABSORBED.

	afraid	touched	bothered*	unfriendly*	thinking*
	surprised	fond	hurt	sneaky	interested**
	angry	liked	sorry	bored	excited*
	sad	kind	disbelieving	wanting	sure*
	happy*	romantic	unsure*		
	disgusted				

We chose to use the prototype approach [41], [42]. This assumes that language and knowledge shape the way people categorise information. It has both contents of individual categories and the hierarchical structures among them. Therefore, it can represent a large range of affective states in terms that are intelligible and reflect knowledge. The Mind Reading taxonomy is an example of this approach [43]. Table I presents the main group categories of the Mind Reading taxonomy. Each of these groups includes many different affective states that share a common meaning and knowledge. For example, the *unfriendly* group includes 120 affective states, such as *argumentative*, *cold* and *discouraging*. However, the groups or categories in the taxonomy often include affective states that are on opposite sides of dimensions such as passive-active or positive-negative. For example, in the *unfriendly* group there are both *ignoring* and *argumentative*, i.e. not engaged and fully engaged. Therefore, it cannot be taken directly for training a machine.

For the inference system, a set of nine affective-state groups or archetypes was chosen to represent a large variety of affective states and co-occurring affective states. The affective-state groups were: *joyful*, *thinking*, *absorbed* or *concentrating*, *stressed*, *excited*, *opposed* or *disagree*, *interested*, *confident* or *sure*, and *unsure*. Each affective-state group consisted of several affective states that generally represent a dominant common concept. For example, the affective states *absorbed*, *engaged*, *committed*, *concentrating* and *focused* were assigned to the *absorbed* group.

These affective-state groups are often used to describe human behaviour. Several of these affective-state groups can co-occur in everyday situations. Some of them were observed in human-computer interactions [45], [46]. The set and the affective-state groups it comprised were simple enough to be used both by people and systems. Most of the affective-state groups and the single affective states they included drew on definitions and categories from the Mind Reading taxonomy and database [43] with modifications. The chosen set aimed to minimise the duality within category groups, for example in distinguishing between interest and concentration [45]. The choice of several affective-states to represent each affective-state group increased the number of samples for training and testing and the reliability of the system. Vidrascu *et al.* [25] 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. The multiple nuances and subtle affective states within each group compensated, to an extent, for inter-speaker variability (in perception, interpretation and expression) and for some of the limitations posed by acting and labelling. In addition, these affective states are relatively easy to induce and therefore to act in a non-stereotypical manner.

A. Training and testing data

The choice of representation method affects the definition and choice of data-sets and the manner of data acquisition. On the other hand, the data-sets' structure defines the scope and capabilities of the classification, by defining the types and range of available samples and classes for training and testing.

There is a growing effort to use real recorded data for recognition [25], [26], [36]. 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 [20], [47], which in practice limits the developers to labelling few affective states (or dimensions). Most of the researchers use their own hand-picked sets of single words or sentences, chosen from very big data-sets, creating proprietary data-sets.

We used another approach in which the training and initial testing of a machine were conducted on a fully annotated database of acted affective states [43]. The voice part of the Mind Reading database was used for training and testing [43], [48]. The Mind Reading database is classified using a prototypical taxonomy, and is available commercially as a DVD. This can be used to teach children and adults diagnosed with Autism Spectrum Disorder, who have difficulties recognising emotional expression in others, to recognise the behavioural cues of a large variety of affective states in their daily lives. We used an experimental version of the database that consists of over 700 affective states arranged into the 24 groups presented in Table I. Each affective state is represented by six different sentences uttered by six different actors. In total it includes 4400 utterances recorded by ten UK English speakers of both genders and of different age groups, including children. According to its publishers, the acting was induced [44], [50] and the database was labelled by ten different people. (The commercial version includes 412 of these affective states.)

The database is acted, but its original purpose (teaching humans) 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 humans need fewer samples for training).

A set of 380 sentences of 93 affective states from the Mind Reading database was used for training the pair-wise classification machines. A set comprising 253 sentences that belong to the same affective states and to additional similar concepts was used for testing the combined inference machine, i.e. a 60%-40% split between training and testing, respectively. In the affective states that were used both for training and for testing, the ratio was 70%-30% respectively. In total, 633 sentences

were used for training and testing. The remaining affective states from the Mind Reading database and their recorded expressions were later used for further testing (validation). Naturally evoked affective states were examined in a later stage in which the inference results that were obtained by the inference machine were compared to other indicators [49].

B. Attribute set

Research [38], [57] reveals two characteristics of subtle affective states. The first characteristic is that different vocal features and metrics (attributes) distinguish between the expressions of different affective states, i.e. a set of attributes *x* may distinguish between class A and class B, while a different set *y* distinguishes between class A and class C. However, class A and class B may share some attributes that distinguish both of them from class C. The second characteristic is that when one class is compared to another class, a threshold often distinguishes between these classes. It happens when a certain attribute has a continuous range of values between the examined classes. It is especially true when examining consecutive speech samples from sustained interactions in which the change in expression is often gradual, until a change of affective state is observed. The classification algorithms in use and the inference results should reflect these subtle transients.

Several conclusions for the design of a classification system were derived from these observations:

- A large enough set of attributes, based on features and metrics that would characterise the affective states and the differences that were observed between their expressions was defined. The contribution of each attribute was measured as the number of times in which it was used in the classification, i.e. automatically selected by the attribute selection algorithms. Only a few of the 173 attributes were not used at all (listed in Section IV).
- Different sets of attributes were used to distinguish between different affective states, as opposed to one set of attributes that distinguishes between all affective states. This approach was tested and justified also in the training process by the results of using the same sets of attributes for the classification of different pairs of affective-state groups. Often, sets that yielded very good results for one pair of affective-state groups yielded no more than random probability for another pair.
- Although many different algorithms and approaches have been tested, at the end the chosen algorithms were based on thresholds, either as the distance between the samples at the border between the classes (SVM) or on the attribute values that were used to distinguish between classes in the tree-based classification, rather than on the distance between the centres of the classes. Blind clustering techniques (unsupervised learning) were not effective for complex affective states. Several methods of blind clustering were examined unsuccessfully (listed in the Section IV).

C. Dataflow

The data flow in the system consisted of a three-stage pre-processing of the incoming speech signal: extracting vocal features from the input speech signal, extraction of temporal characteristics and metrics calculation, and normalisation.

After the pre-processing stage, the normalised metrics entered as attributes into a two-stage classification system (Figure 1). The classification included a first stage of pair-wise decision machines, i.e. each machine compared between two affective-state groups (one-against-one classification [51], [52]).

The set of attributes and classification algorithm was selected independently for each pair of affective states. Different groups of attributes were used for the different pair-wise decision machines. For a set of 9 affective-state groups, 36 pair-wise machines were required. Each affective-state group appeared in 8 pair-wise machines or comparisons. The second stage was a voting algorithm that consolidated the comparisons into a single ranked list. The decisions of all the pair-wise machines were entered into a voting machine that decides which were the most probable affective states that could be related to the processed speech signal and to what extent. Each of the recognisable affective-state groups was ranked according to the number of comparisons in which it was chosen.

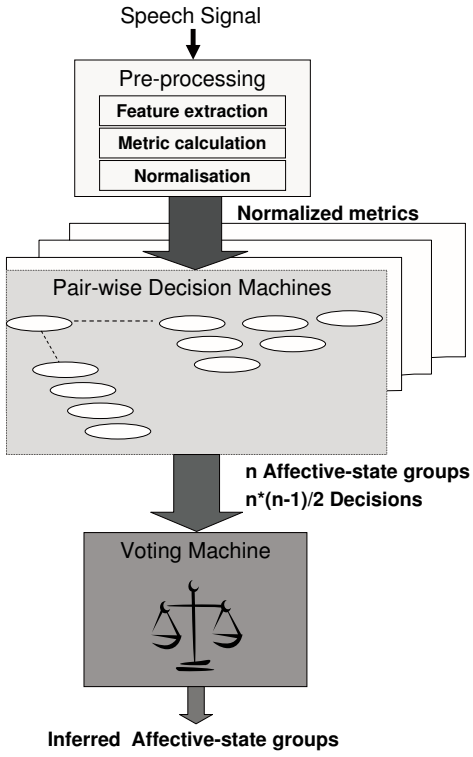


Fig. 1. Schematic description of the dataflow in the inference machine.

D. Classification methodology

Pair-wise machines have been used in research for inference of affective states. For example, Vidrascu *et al.* [25] use them

because they chose to use support vector machines (SVM) as the classification algorithm, with a single feature set. This method may limit the optimisation of each individual pair-wise machine. Furthermore, extending the overall machine to accommodate new affective states requires retraining of the entire machine.

In this paper, two significant guidelines were combined in the implementation of the series of pair-wise classifications. The first stemmed from the observation that expressions of different affective states are characterised by different sets of vocal features and metrics [38]. From this observation evolved the conclusion that there was no requirement for one set of attributes (normalised metrics) to represent all the examined affective-state groups. Therefore, each pair-wise machine had its own sub-set of attributes from the total number of metrics that were extracted from each speech signal (utterance or sentence). The second guideline was that each pair-wise machine could be trained independently, based on a classification method that yielded the best results for the specific affective-state groups and attributes. Optimisation was done for each pair so better results could be achieved for a single pair. The overall machine was flexible because affective-state groups could be added or subtracted according to the requirements of the system, training only the new machines (possibly with samples from new corpora). The attribute selection methods, the chosen attribute sets and the classification algorithms were compared to other sets and methods (described in details in Section IV).

Two voting algorithms for finding one class or affective-state group were examined (using voting of the 36 classification machines). The first was the Condorcet voting method [53]–[55] with a single winner (no co-occurrences). The second was a threshold method that allowed inference of more than one candidate per sentence or utterance, i.e. in each sentence only affective states exceeding the mean number of comparisons by more than one standard deviation were considered. The inference results of these two methods were compared to the labels of the samples, as defined by the Mind Reading database (the expected result).

Training was done with the data mining tool Weka [56]. The extraction of vocal features, metric calculations, the implementation of the classification machines, the voting and the testing were done in Matlab.

E. Validation

In order to examine the inference of co-occurring affective states, the combinations inferred for each sample (sentence or utterance) of the training and testing sets were compared to the lexical definitions of the examined affective states. The inference was then applied to all the 4400 utterances of the Mind Reading database. As a stronger measure, the results for all the samples or sentences that represent an affective state were evaluated. A Friedman test [58], that measures the significance of ranking, was applied to the ranked lists of the sentences that belong to each affective state. In order to find a meaningful interpretation of these results, a double-threshold procedure was applied to each sentence and to the sentences that represent an affective state. The double-threshold procedure automatically inferred the combinations of affective-state groups that characterise each affective state. It found the affective-state groups that were chosen by most of the pair-wise machines (the first threshold) in most of the sentences that represent an affective state (the second threshold). The thresholds were set over one standard deviation above the mean number of machines and sentences, respectively. The results of the double-threshold procedure (combinations of lexical concepts) were compared to the lexical definitions of the examined affective states (Section V).

The distinguishing capabilities of the system were compared to human performance on the Cam Battery Test [44] (SectionV-D). The inference system was also applied (as is) to naturally evoked affective states from the Doors database [38]. Doors is a Hebrew database of affective states naturally evoked during a computer game based on the Iowa gambling test [2]. For each speaker, Doors includes 100 repetition of two sentences uttered throughout the game, in addition to utterances with un-restricted text spontaneously evoked during the game and during intervening interviews. The inference results were compared to other recorded indicators such as event and physiological measurements [49], [59]. Expression changes between successive sentences in a sustained interaction appeared in the distribution of the pair-wise classification results, therefore the full ranked lists were used for the analysis. Significant correlation was found between the inferred affective states and game events, such as gain and speakers' choices. In addition, temporal changes in the inferred affective states were observed simultaneously with events, text and changes in various physiological and behavioural reactions (the details of these measurements are beyond the scope of this paper [49], [59]).

III. Pre-processing: extraction of vocal features and metrics

Before classification, three stages of pre-processing were applied. The first stage was the extraction of vocal features that represent the expressive properties of the speech signal ([8], [23], [24] and references within). These were time series resulting from short-term analysis of the speech signal with a moving window (overlapping time-frames). The vocal features included: the fundamental frequency, the vibration rate of the vocal chords (also referred to as f_0, pitch or intonation), using a completely automated extraction algorithm derived from Boersma's algorithm [60]; smoothed energy curve of the speech signal using average of the energy over a time frame combined with Hamming window; spectral content, the distribution of the energy over the whole frequency range, calculated with a Bark scale based filter bank up to 9 kHz [61], [62]; harmonic properties [59], such as consonance and dissonance, based on findings from physics, musicology and neuro-science that show that people both generate and perceive these properties [59], [63]–[67]. The features were calculated for short and overlapping time frames of 50 msec and overlap of 40 msec for the duration of the utterance. All the vocal features were extracted automatically, with no manual intervention.

TABLE II
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 (Hz): 101,204,309,417,531, 651,781,922,1079,1255,1456,1691,1968, 2302,2711,3212,3833,4554,5412,6414,7617	84

The values of most of the vocal features change during the utterance. An automatic algorithm that divided the duration of the analysed utterance into several parts according to their vocal properties was developed [59]. This rule-based algorithm divided the sentence or the utterance into parts such as silence, voiced (parts in which there are vibrations of the vocal chords and the fundamental frequency is not zero) and unvoiced (where the fundamental frequency is zero), in combination with energy peaks and the like. Temporal characteristics that draw on terms from disciplines such as linguistics and musicology were calculated. For example, units that correspond approximately to linguistic properties such as syllables, consonants and vowels were calculated from combinations of the extracted speech parts, in addition to durations, time and frequency lapses between occurrences of the different speech parts that correspond to terms such as tempo and melody [59].

Secondary metrics were then calculated, including statistical properties, such as the number of occurrences, median, range and maximum or extreme values of each vocal feature and of the temporal metrics [23], [39]. The median was used because it is less sensitive to outliers than the mean. This choice was supported by correlation tests of various metric types. In total, a set of 173 secondary metrics for each speech signal was used. The following list is a rough summary of the metric distribution: 34 pitch related metrics, 19 energy related features, 17 'tempo' related features, 19 harmonic properties and 84 metrics of spectral content in 21 frequency bands [59]. A summary of the metrics for which statistical measures were used appears in Table II. The vocal-features and the secondary metrics were defined and examined by analysis of the two datasets Mind Reading [43] and Doors [57]. Mind Reading supplied a large variety of affective states while Doors provided multiple text repetitions by each speaker, with natural transition between affective states during sustained interactions.

The third stage was normalisation of the values, so that all the metrics were represented on a similar scale. Each metric was normalised separately for every speaker. Each speaker has individual characteristics that derive from the speaker's identity, including parameters such as gender, body structure, personality, spoken language and accent, or from the recording conditions. The normalisation compensated for the inter-speaker variability. As a result, the expressive characteristics of an affective state in comparison to other affective states could be compared between speakers. Therefore, no re-training of the machine was required for new speakers.

IV. CLASSIFICATION

The classification algorithm of co-occurring affective states included two stages. The first stage consisted of pair-wise decision machines (one-against-one classification [51], [52]). The second stage was a voting algorithm that consolidated the comparisons into a single ranked list.

A. Pair-wise machines

The flow of the training process of each pair-wise classification machine can be seen in Figure 2.

For each pair of affective-state groups, the training consisted of finding an optimised combination of both a classification method and a sub-set of attributes (normalised metrics) that yielded the best classification results. Training and attribute selection were jointly conducted. The exploration of both multiple attribute selection methods and classification algorithms follows the experiments conducted by Oudeyer [23]. All the training was conducted using the data-mining tool Weka [56]. Imbalanced training sets cause bias [68], therefore for each pair of affective-state groups the training was conducted on similar-size datasets.

The classification algorithms used were linear SVM [69], [70], a classification algorithm that defines a hyperplane which maximises the distance between the samples of two classes, and C4.5 [71], a decision-tree method constructed through divide-and-conquer strategy, as applied in the J48 package of Weka. These algorithms yielded the best results and their implementation was simple. Additional algorithms were examined, whose performance was not as good and in most cases their implementation was more complicated. For example, polynomial and Gaussian SVMs, Gaussian mixtures, Naive Bayesian, and neural networks were examined, in addition to various rule-based and decision-tree classifiers.

Attribute selection was done as a series of exploratory attribute selection methods. Both scheme-dependent and scheme-independent selections were used [56]. Examples of the examined selection and evaluation methods include: best first selection, forward greedy hill climbing augmented by backtracking, and evaluation of the individual contribution of each attribute [72], principal component analysis (PCA), expectation maximisation (EM), gain ratio evaluation with respect to the class, ranking attributes by their individual evaluation using entropy, information gain attribute evaluation and more, using the algorithms that are implemented in Weka [56]. The attribute selection stage was applied even if the classification algorithm includes an inherent attribute selection, for example decision-trees (C4.5).

The procedure was repeated for different classification algorithms until no further improvement was achieved. No optimal combination of an attribute-selection algorithm and a classification algorithm was found for all the pair-wise machines. When the search was exhausted the best machine and the

best set of attributes were selected. Although the procedure of attribute selection sounds exhaustive and imprecise, in practice for most of the machines the maximal number of sub-sets that were examined before a good enough solution was found is four or five. Good-enough solutions were defined as a combination of all the following criteria simultaneously: tenfold cross-validation over 70-75%, minimisation of the difference between the true-positive values of the two classes (in this case under 5%), Receiver Operator Characteristic (ROC) area close to one (over 0.9), and precision over 95% on the training data. If the automatic methods did not yield good enough results additional manual feature selection was used. One method was to combine the attributes selected by the C4.5 algorithm with attributes selected by attribute-selection algorithms and finding the best sub-set from these attributes. These sets usually comprised less than 10 attributes each, with overlaps. The optimisation using this method was simple and sometimes yielded better results than the original sets. A more radical solution, when the attribute selection algorithms yielded relatively poor results was the selection of attributes or metrics that had been extracted from certain vocal features only, for example, *all the metrics that were calculated from the fundamental frequency*. The training process was concluded when no further improvement was found. Therefore, six machines were accepted with cross-validation values between 60% and 70%.

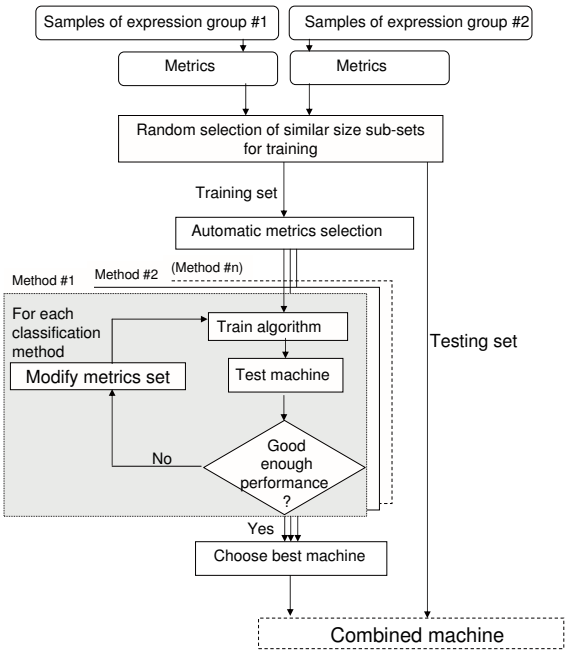


Fig. 2. Flowchart for optimisation of a single pair-wise machine

Table III lists the pair-wise machines, the tenfold cross-validation rate of each machine, the classification algorithm that was chosen, SVM or tree (C4.5), and the number of attributes used in the machine. There were 6 SVM-based machines and 30 tree-based machines, chosen according to the listed objective criteria. A few of the SVM machines presented similar performance to tree-based machines but were not chosen due to an arbitrary decision or due to a larger number of attributes. The cross-validation was usually worse in the SVM machines that were based on the full attribute set though the precision was similar, possibly due to over-fitting. Attribute selection often improved the cross-validation of the SVM machines, and the heuristic method for attribute selection proved better than the more traditional methods in these cases. The tenfold cross-validation was at least 60% in all the machines. The average cross-validation rate for all the 36 machines was 76%. Devillers *et al.* [20] 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. 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 average number of attributes in the pair-wise machines was 10 (the median is 8), which is very low compared to the full set of 173 attributes. Only three machines had more than 20 attributes. These were SVM machines for which fully manual attribute selection was applied, for example *all the metrics that were derived from the spectral content under a certain frequency*, i.e., all the statistical properties of the energy in the relevant filter bands.

Different attributes were automatically found to distinguish different affective-state groups. This result strengthens the initial observations that different vocal-features and metrics distinguish different affective states and justifies the individual attribute-selection step. In some cases, a certain affective-state group shared certain attributes with another (second) affective-state group while the same attributes distinguished between this affective-state group and a different (third) affective-state group. This was confirmed by using the attributes that distinguish between one pair of affective-state groups to compare between a different pair. The result was that a set of attributes that yielded near optimal results for one pair of affective-state groups (precision over 95% on the full set) yielded no more than random probability for another pair of affective-state groups. This test was repeated with similar results for different pairs of affective-state groups, often when one of the affective-state groups remained constant. In addition, if the set of attributes was very large it did not improve the results and usually made them worse, possibly due to over-fitting.

In a few cases, two different sets of attributes yielded similar classification results for the same pair of affective-state groups (either with the same classification algorithm or with different classification algorithms) and an arbitrary decision of which set and algorithm to use was required. There was no other correlation between these attributes in other cases. It may indicate (or confirm) that there is a redundancy in the vocal cues of affective states (or affective-state groups).

The design of such systems require many factors that

may affect their structure, such as the definition of vocal features and precise extraction algorithms, metrics definition and calculation, choice of training samples, and more. The design requires the developers 'to get intimate with the data' [56]. The existence of several good-enough solutions and the occasional redundancy in the data compensate for the possible lack of reproducibility.

At the end of the training process, the number of machines in which each attribute appeared was counted. The attributes that appeared in the largest number of machines were (the normalised values of) the number of different harmonics that appeared in 15 of the 36 pair-wise machines [59]; the median of the fundamental frequency and the standard deviation of the energy in the first filter-band that appeared 11 times each; the minimum durations of unvoiced intervals (speech with no pitch, such as in fricatives) and the range of the energy in the first and second filter bands that appeared in eight machines each. The attributes that did not appear in any of the machines were a few harmonic intervals, the length of the down-slopes of the fundamental frequency and a few properties of the high spectral-bands. These results show that relatively compact and efficient machines can distinguish between pairs of complex affective states, while most of the 173 attributes are required for the classification of all the nine affective-state groups.

TABLE III
DETAILS OF THE 36 PAIR-WISE MACHINES, INCLUDING: TENFOLD CROSS-VALIDATION, MACHINE TYPE: TREE (C4.5) OR SVM, AND THE NUMBER OF ATTRIBUTES THE MACHINE USED.

	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
joyful	82% SVM 19	83% Tree 8	61% Tree 18	60% Tree 8	71% Tree 7	77% SVM 40	75% Tree 8	72% Tree 13
absorbed		84% SVM 12	87% Tree 5	81% Tree 6	78% Tree 6	82% Tree 5	64% Tree 6	73% Tree 10
sure			84% SVM 12	79% Tree 7	72% Tree 8	78% Tree 8	78% Tree 8	75% Tree 7
stressed				73% Tree 9	84% Tree 7	66% Tree 7	68% Tree 15	72% Tree 15
excited					74% Tree 9	71% Tree 9	64% Tree 8	79% Tree 8
opposed						75% Tree 8	79% SVM 6	81% Tree 6
interested							72% Tree 8	83% Tree 8
unsure								89% SVM 22

B. Combination: inferring co-occurring affective states

Distinguishing nine affective-state groups required 36 pair-wise machines, in which each affective-state group was considered by eight machines. These 36 comparisons were then combined to calculate an ordered ranking of the nine affective-state groups. That means that each affective-state group was ranked according to the number of comparisons in which it was chosen in the range 0-8.

Ranked lists of inferred affective-state groups can be used in different manners for different applications, as described in the next sections.

C. Validation

Examination of inference or classification results is mostly done by inferring one candidate so it could be compared to the labels of the testing set. Detection results are given using the CL score (class-wise averaged recognition, i.e. average of the diagonal of the matrix) [25]. Results for 4-7 basic emotions, distinct affective states such as fear, anger, sadness, happiness and 'neutral', are on average 28%-77% ([20], [25], [26] and references within).

In order to select a single leading candidate, Condorcet voting [53] with the two-round runoff method, a second round of pair-wise comparisons between the candidates with the maximum number of votes, was used. For nine affective-state groups, the probability of randomly choosing an affective-state group is 11%. In the single winner method all the affective-state groups were recognised with a much higher rate than that. Most of them, with over 65%, as can be seen in Table IV. The testing yielded an overall recognition accuracy of 70% (true-positive recognition).

TABLE IV
CONFUSION MATRIX OF THE INFERENCE MACHINE USING THE CONDORCET METHOD.

Inferred class / Actual class	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
joyful	75.0	3.6	0.0	0.0	3.6	3.6	7.1	7.1	0.0
absorbed	0.0	69.0	3.4	3.4	0.0	0.0	10.3	6.9	6.9
sure	0.0	14.3	78.6	7.1	0.0	0.0	0.0	0.0	0.0
stressed	0.0	0.0	4.3	73.9	8.7	8.7	4.3	0.0	0.0
excited	11.1	5.6	5.6	11.1	61.1	0.0	5.6	0.0	0.0
opposed	2.6	2.6	17.9	0.0	12.8	61.5	0.0	0.0	2.6
interested	3.6	3.6	3.6	3.6	0.0	0.0	71.4	7.1	7.1
unsure	0.0	3.4	0.0	13.8	0.0	6.9	0.0	65.5	10.3
thinking	0.0	7.1	0.0	3.6	3.6	0.0	7.1	7.1	71.4

However, in the case of affective states it is not necessary to solve conflicts in order to determine a single winning candidate, because several of the candidates can co-exist. Therefore, a second method was tested, selecting affective-state groups that were chosen by several machines. The threshold for selection was set over one standard deviation above the mean number of machines, which means that at least six machines preferred an affective-state group.

For example, Table V represents the inference results (the ranked list) of one sentence of the affective state *choosing* from the affective-state group *thinking*. Each cell represents the number of machines that chose an affective-state group. The cells that are marked in grey represent the affective-state groups (*thinking* and *unsure*) that were inferred by the threshold method for this sentence.

TABLE V
AN EXAMPLE OF THE THRESHOLD METHOD FOR ONE SENTENCE LABELLED AS THE AFFECTIVE STATE *choosing* THAT BELONGS TO THE AFFECTIVE-STATE GROUP *thinking*. EACH CELL SHOWS THE RANKING OF ONE AFFECTIVE-STATE GROUP. DARK GREY MARKS AFFECTIVE-STATE GROUPS CHOSEN BY THE THRESHOLD METHOD (SELECTED BY 6-8 MACHINES).

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
choosing.wav	2	5	2	3	3	4	3	7	7

The inference results of the threshold method for the testing set appear in (Table VI). Using the threshold method, the accuracy of recognition of each affective-state group was at least 75% (random probability in this case is 14%). The overall accuracy was 83%. These inference results refer to subtle affective states.

TABLE VI
INFERENCE RESULTS USING THE THRESHOLD METHOD.

Inferred class / Actual class	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
joyful	89.3	3.6	3.6	21.4	35.7	25.0	17.9	21.4	7.1
absorbed	3.4	75.9	10.3	10.3	6.9	10.3	20.7	37.9	34.5
sure	10.7	17.9	89.3	10.7	7.1	50.0	17.9	7.1	3.6
stressed	17.4	4.3	8.7	78.3	26.1	17.4	34.8	30.4	13.0
excited	0.3	5.6	11.1	27.8	83.3	55.6	16.7	5.6	0.0
opposed	12.8	10.3	48.7	12.8	30.8	87.2	12.8	5.1	12.8
interested	10.7	17.9	7.1	21.4	7.1	14.3	75.0	32.1	21.4
unsure	6.9	27.6	0.0	34.5	13.8	17.2	24.1	82.8	31.0
thinking	0.0	50.0	10.7	17.9	3.6	7.1	14.3	42.9	85.7

The threshold method is more accurate in the sense that the label of the examined affective-state group is more likely to be included in the inference results, and it allows inference of co-occurring affective-state groups.

D. Comparison to other classification methods

The architecture of independent pair-wise machines was compared to other architectures, including a single machine for all the affective-state groups, i.e. a machine that chooses one of the nine affective-state groups (one-against-all classification). This machine was implemented with decision-tree classification, neural network, polynomial SVM and Gaussian SVMs. In the pair-wise architecture (one-against-one classification), an all-SVMs machine was also examined. Although the precision of these machines was in most cases relatively high (70%-90%), the true-positive values for some of the affective-state groups were low and the tenfold cross-validation results were close to random probability (for example, 13% in the neural network system and 11% in the all SVMs pair-wise classification).

Additional preliminary tests included various non-supervised classification methods for pairs of affective-state groups and for several affective-state groups at a time, for example PCA and EM that were mentioned before, and more. They were not successful, probably because the groups are not mutually exclusive and the ranges of many attributes are continuous.

Most of these methods allow the inference of only one affective-state group at a time while the pair-wise comparison method allows inference of more than one affective-state group for a single sample. Finding an optimised algorithm for each machine in a pair-wise system, improves the results in comparison to a single arbitrary algorithm (such as the all SVMs pair-wise system). The same applies for a single sub-set of metrics.

Because the training was done for each pair of affective-state groups, the machine training was relatively simple. As there was no definite definition of the optimal solution (100% recognition and 100% cross-validation are beyond the scope of realistic expectations), the outcome of *good enough* classification for each pair-wise machine was improved by the integration of multiple classification results from the different machines.

V. CO-OCCURRING AFFECTIVE STATES

The previous section demonstrated the ability of the combined machines to infer known entities, i.e. to recognise the classes that it was trained to recognise. However, the design allows the combined machine to simultaneously infer several affective-state groups and rank them (in effect it performs "semi-blind" multi-label classification), as demonstrated in Tables V and VI. In order to examine these capabilities, we extended the scope of the examined data and used different verification methods.

A. Inference within the training and testing sets

Annotation of subtle affective states, co-occurring and mixed affective-states is difficult. Devillers *et al.* [20] describe annotation of speech segments with one major affective state label and one optional secondary label from a group of 21 fine grained emotions that belong to 7 coarse emotion definitions, by two annotators. The textual content had a role in the annotation.

In the case of the Mind Reading database there were already over 700 fine-grained labels of affective states, and the text of each sentence aimed to be neutral [44]. It remained to encode the meaning and the expected behaviour of these affective states according to the nine inferred affective-state groups, or more precisely, to evaluate the automatically inferred or encoded sentences.

To evaluate the inferred combinations of affective-state groups we first looked at the inference results for the affective states that were part of the training and testing data. The ranking for each affective-state group was in the range 0-8. The highest score 8 means that an affective-state group was recognised in all the comparisons as the most probable candidate. Several classes or candidates could be automatically chosen with a relatively high number of comparisons other than 8. If the criterion was one standard deviation above the

mean, these ranks were 6 and 7. These affective-state groups were chosen in all but one or two of the comparisons (pair-wise machines). It means that they did not only appear and relate to the affective state but that they were also dominant. Several dominant affective-state groups could be recognised simultaneously. In the same manner it is possible to say that an affective-state group was *not* recognised as significant, or significantly was not recognised, if the ranking was in the range 0-2.

Table VII shows an example of the inference results for each sentence with the affective state *choosing* from the affective-state group *thinking*. It shows that in all the sentences labelled as *choosing*, the inferred affective-state group *thinking* was the most dominant. The affective-state group *unsure* was also dominant, it was chosen by six or more machines in four of the six sentences. Other affective-state groups, such as *stressed* and *opposed* were also recognised as dominant in some of the sentences. The affective-state group *sure* was recognised with a very low rate, or not recognised, consistently.

The next stage was to analyse affective states rather than single sentences. This stage was based on the assumption that a semantic concept (an affective state) can be better *characterised* by the inference results that are common to all or most of the six samples that represent it, rather than by a single sentence at a time. If the affective-state groups characterise the affective state or the behaviour related to it they should appear in most of the sentences that represent it.

TABLE VII
AN EXAMPLE OF INFERENCE OF CO-OCCURRING AFFECTIVE-STATE GROUPS FOR SENTENCES LABELLED AS THE AFFECTIVE STATE *choosing* THAT BELONGS TO THE AFFECTIVE-STATE GROUP *thinking*. EACH ROW SHOWS THE SPEECH SIGNAL (ON THE LEFT) AND THE NUMBER OF COMPARISONS IN WHICH EACH OF THE NINE AFFECTIVE-STATE GROUPS WAS CHOSEN. DARK GREY MARKS AFFECTIVE-STATE GROUPS CHOSEN BY 6-8 MACHINES. LIGHT GREY MARKS AFFECTIVE-STATE GROUPS CHOSEN BY 0-2 MACHINES (NOT RECOGNISED). THE FINAL DEFINITION OF THE AFFECTIVE STATE *choosing* IS AT THE 2 BOTTOM LINES, STATING THE NUMBER OF SENTENCES IN WHICH IT WAS RECOGNISED (OR NOT): • 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			6					4	6
Choosing			•					•	•

In order to check the statistical significance of the results, a Friedman test [58] that measures variance by ranks was applied to the ranked lists, i.e. the inference results. The hypothesis in the Friedman test is that all the columns are treated equally, i.e. all the affective-state groups are selected equally. If the Friedman test results are very small ($p< 0.05$), there is a strong evidence that the hypothesis is not correct [73], and there is a significant difference between the ranking results of the different columns. For example, the Friedman test result for the affective state *choosing* was ($p< 7 \cdot 10^{-5}$).

Friedman test can verify that all the affective-state groups behave in a significant manner. However, it does not specify the characteristics and meaning of this behaviour, i.e. what the characteristic ranking of each affective-state group is and how a combination of ranks characterises the analysed affective states.

Therefore, a double-threshold procedure was applied. *Dominant* affective-state groups were *recognised* by at least 6 machines in at least 4 (>66%) of the six sentences with the same label. The dominant affective-state groups are marked in the next tables by bullets •. Affective-state groups that were *recognised* in at least 6 comparisons in 3 of the same sentences are marked by empty circles ○ in the tables. They signify 50% of all the sentences and 50%-75% of the sentences in which a *dominant* affective-state group was recognised. These affective-state groups cannot be considered dominant but they may be influential. Recognised affective-state groups appear in dark grey. Similar procedure was applied to affective-state groups that were chosen by 0-2 machines in at least 4 sentences, and over-lapping 3 sentences respectively. These affective-state groups were *not* recognised for the examined affective state and their inference may add to its understanding by elimination. They appear in light grey.

The two rows at the bottom of Table VII summarise the inferred combination of affective-state groups that refers to the examined affective state *choosing*, i.e. *thinking* and *unsure*. The affective-state group *sure* was not inferred. The affective-state groups *stressed* and *opposed* are expected and accepted behavioural expressions of the affective state *choosing*, but they appeared only in a small number of sentences and therefore could not be considered significant or dominant for the affective state in general.

A representative combination that was automatically inferred for an affective state was compared to the lexical definition of the affective state labels in dictionaries [74] and thesaurus engines, or to the expected behavioural characteristics. The inferred combinations were often similar to the lexical definitions (more details in Section V-B). The results were checked by eight people and most of the results were also presented to audience. There was agreement regarding the justification of most of the results. In the given example, the inferred combination for *choosing* was agreed to be correct by nearly a hundred people. The lexical definition is *to decide what you want from a range of things or possibilities*, while the definition for *decide* is *to choose something, especially after thinking carefully about several possibilities* [74]. These definitions entail the uncertainty at the choosing stage and the lack of it upon making the decision, but not clearly. Affective state labels that have similar meaning often had a similar or an identical inferred combination.

Table VIII shows the inferred combinations for each affective state in the *thinking* group. The affective state labels that are marked in grey were not part of the training set. The samples of the other affective states were divided between training and testing. Friedman test results appear next to the affective-

state labels. The affective-state groups that were identified by 6-8 machine are marked in dark grey. The affective-state groups that were not identified with the examined affective states, i.e. chosen by 0-2 machines, appear in light grey.

As can be seen, some of the affective states appear as combinations of several affective-state groups. All the combinations correspond to the lexical definition of the affective states and to the expected behavioural patterns. This example demonstrates how the inference of co-occurring affective-state groups improves the recognition and characterisation of complex behaviour in comparison to inference of a single affective state. Even though many of the affective states were trained as a single affective-state group, the inference results distinguish between them.

TABLE VIII
INFERENCE RESULTS OF INDIVIDUAL AFFECTIVE STATES FROM THE *thinking* GROUP OF THE MIND READING TAXONOMY: AFFECTIVE STATES (AFFECTIVE STATES IN GREY LINES WERE NOT PART OF THE TRAINING SET); FRIEDMAN TEST RESULTS; ACCUMULATED INFERENCE RESULTS, DARK GREY SIGNIFIES RECOGNITION BY 6-8 MACHINES IN 4-6 SENTENCES, LIGHT GREY SIGNIFIES NO RECOGNITION (0-2 MACHINES).

Concept	**F.T.**	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
comprehending	$7{\cdot}10^{-3}$									•
deciding	$1{\cdot}10^{-3}$			•						•
regarding	$1{\cdot}10^{-1}$			•						•
thoughtful	n.a.			•		•				•
wool-gathering	$1{\cdot}10^{-4}$					•	○			•
calculating	$4{\cdot}10^{-4}$		○							•
dreamy	$9{\cdot}10^{-3}$		•			•				•
fantasising	$8{\cdot}10^{-4}$		•	•		○				•
brooding	$7{\cdot}10^{-5}$		•	○		•			○	•
considering	$1{\cdot}10^{-3}$	•	•	•		•	○		•	○
choosing	$7{\cdot}10^{-5}$			•					•	•
thinking	$3{\cdot}10^{-3}$			•					•	•
realising	$3{\cdot}10^{-3}$			•	•				•	•

B. Additional affective states

In order to explore the scope of the inference system and for additional validation, the system was applied to new affective states that were not used for the training of the system but belong to the same *emotion groups* [44], or meaning group, in the Mind Reading taxonomy and database. Examples of such affective states appear in grey in the Concepts column of Table VIII. As can be seen, the inference results in these affective states agree with the lexical meaning and with the expected behavioural expressions. For example, the accumulated inference results of the affective state *realising* included the affective-state group *stress* that can be associated with (unpleasant) surprise. The affective state *considering* was inferred as a combination of *absorbed* and *uncertain*, and possibly *thinking*. Its lexical definition is *to spend time thinking about a possibility or making a decision* [74]. It refers to a state which is more inward or absorbed than *choosing*. The combination was inferred automatically and agrees with the meaning of the concept. *Joy*, *excitement* and *certainty* were not identified with this affective state in any of the sentences. This additional information is not part of the definition or characterisation but it may indicate by elimination on properties of the affective state or its behavioural characteristics.

A Friedman test was first applied to the training and testing sets (633 sentences) and then to all the Mind Reading database (4400 sentences). In 306 affective states of the 749 affective states of the full Mind Reading database, much beyond the inferred affective-state groups, the rankings of the affective states were found to be significant for all the affective-state groups ($p<0.05$).

Most of the affective states (98%) that got significant results in the Friedman test were included in the group of affective states that passed the double-threshold procedure. In the other two percent of the affective states, the ranking of all the affective-state groups ranged mostly between 3-5, i.e. significantly close to random, meaning that the set of affective-state groups could not characterise them.

On the other hand, the Friedman test implies that *all* the affective-state groups behave in a significantly characteristic manner. The double-threshold procedure on the other hand, requires that *at least one* affective-state group will be dominant. For a large variety of affective states, it cannot be expected that all the nine affective-state groups will be significant and meaningful. Therefore, the double-threshold procedure characterises many more affective states than the Friedman test.

The double-threshold procedure characterised 570 (76%) of the affective states either as the inferred affective-state groups or by elimination. From these affective states, approximately 85% agreed with the lexical definition and the expected behaviour (by agreement of all eight examiners). Using the double-threshold criterion, at least four of the six sentences that define each affective state had the same recognised affective-state groups, i.e. total of 2280 sentences, from which at least 1784 sentences were characterised (not by elimination). Both the Friedman test and the double-threshold procedure are objective measures that verify the accuracy and consistency of the results. Both were applied automatically with no manual intervention. In many more single sentences one or more affective-sate groups were recognised by 6-8 machines, and the inference results agree with the lexical definition, but as a group they did not pass the double-threshold criterion.

Multiclass and multi-label classification was performed while the training was done with a single label at a time. Indeed, not every nuance is distinguished by the set of nine affective-state groups but it does characterise a wide range of affective states. The repeated results and the large number of different affective states imply that the inference method is meaningful. Such capabilities of an automated system have not been reported previously. One of the implications of these results is that the machine can be used for mapping and conceptualisation of affective states, i.e. to define the relations between affective states according to their vocal expressions.

C. Speaker variability

In order to verify that the system can be used for different speakers with no additional training, the recognition error rates

for each of the speakers in the Mind Reading database were examined. This test was used because each of the training sentences for a certain affective state was uttered by a different speaker, and the random selection of samples implied that nuances of affective states were trained on certain speakers and tested on others. There was no significant difference in accuracy between the different speakers over the whole Mind Reading database (it ranged between 85-90%). The system was also tested on recordings of six new speakers [59]. Each of the metrics in the samples was normalised for each new speaker, as described in Section III. No additional training of the inference system was required. The inference results were significantly correlated ($P<0.05$) to events and physiological cues [49].

D. Distinguishing affective states and comparison to human performance

The inference machine performed a task that is usually performed by people. Therefore its capability to distinguish between different affective states was compared to human performance on the CAM Battery Test (CBT) as reported by Golan *et al.* [44].

In the CBT, each question includes a recorded sentence and a choice of four labels of affective states. After listening to the recorded sentence the participants were asked to "choose the word that best describes how the person is feeling" from the four given labels (one true answer and three foils). The inference machine was applied to the same sentences or voices that were used for the CBT, and used the same foil affective states that were used in the battery questions. The foil affective states were represented by the accumulated inference results of the samples that represent the affective state in the Mind Reading database.

The CBT was tested on 21 participants with Asperger Syndrome (AS group) and on 17 matched controls (control group). The average number of sentences recognised by the control group was close to 43 sentences out of 50 and the participants in the AS group recognised on average less than 36 concepts by their vocal correlates [44]. In comparison, the inference machine successfully distinguished between the affective states and the foil affective states in 49 of the 50 sentences. In this case it outperformed humans. These findings imply that the machine could distinguish between complex affective states that were not necessarily part of the affective-state groups that it was trained to recognise.

VI. Summary and Discussion

We present a classification method for inference of co-occurring affective states from their non-verbal expressions in speech. The input to the classification system is a large set of metrics. The metrics are derived from the vocal features that are extracted from the speech signal.

The classification consists of pair-wise comparisons between affective-state groups (beyond the set of basic emotions, or the dimensions positive-negative, active-passive). Each pair-wise machine has its own set of metrics and classification algorithm. This stemmed from the observation that different vocal features distinguish different affective states and the training process verified it.

For each utterance, the pair-wise comparisons are consolidated into a single ranked list that reflects the number of comparisons in which each affective-state group is chosen. The ranked list represents inference of co-occurring affective states. The ranked list can be used in different ways for different applications. The system can be easily adapted to new affective states and to new speakers without affecting the existing machine.

Experiments on the Mind Reading database show that this method allows accurate detection of the affective-state groups from the speech signals. The paper presents examples of inference of affective states from the initial training and testing data set in comparison to their lexical definitions and to the expected behavioural patterns. The inference was successfully extended to new affective states that were not used for training. It was further used for characterising a large variety of affective states. The ability of the classification system to distinguish between complex affective states was compared to human performance in an independent test and was found to be superior.

The classification allows presentation of a very large number of expressions and nuances. However, the presented system is not complete. It does not represent the entire range of affective states and does not distinguish between all the existing definitions and all possible nuances. Nevertheless, this system shows that very few, carefully chosen, affective-state groups can increase the accuracy and the distinguishing capabilities of an automatic system to infer and characterise a very large range of affective states within and beyond the set of affective-state groups that it was trained to infer. The implication of independent training of each pair-wise machine is that additional affective-state groups will require a few more pair-wise machines while no re-training of the existing machines will be required. In this manner the system can be easily adapted to various applications.

The system generalises to new speakers without additional training due to the speaker-dependent normalisation process. The generalisation to new languages is supported by the combination of normalisation of speech metrics and of the representation method.

A training strategy is presented to deal with learning cases where a large set of predictor features are needed to disambiguate between the categories while also addressing the sparsity inherent in the problem. Many features and metrics contribute, overall, to the classification, yet only a few of these features may be actively employed by the speakers at any time to encode the characteristics of the expression they want to deliver. The paper presents and validates a strategy for modularising the learning problem by decomposing it into simpler learning sub-tasks which can carry out the learning effectively with a considerably smaller subset of features. These strategies can benefit applications that share a similar structure and challenges.

Inference of co-occurring classes can be beneficial in various fields; in particular, fields that relate to other aspects of human perception and cognition, such as colour retrieval

[75], [76], that share characteristics with the presented field of affective-state inference. This paper shows that a comprehensive solution that considers the representation method as a part of the classification is very powerful because it provides representation of different aspects of the complex information domain. Furthermore, it enhances the existing knowledge of the domain by presenting a new perspective. (In this case, presenting the relations between affective states as inferred from their vocal correlates). It also presents an objective tool (within the limitations of the training) for representing information that is often perceived as subjective. The architecture itself is simple to implement and provides both good classification performance and flexibility.

January 20, 2009

ACKNOWLEDGMENT

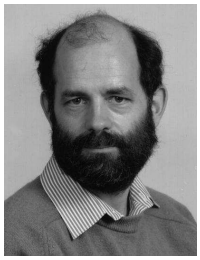
The authors thank Edna Schechtman, Yael Edan, Yehuda Werner and Boaz Lerner for their help in revising this paper. The authors thank AAUW Educational Foundation, Cambridge Overseas Trust, Girton College, The Computer Laboratory and Deutsche Telekom Laboratories at Ben-Gurion University for their partial support of this research.

REFERENCES

[1] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. XVLII, pp. 263–291, 1979.
[2] A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio, "Deciding advantageously before knowing the advantageous strategy," *Science*, vol. 275, pp. 1293–5, 1997.
[3] R. W. Picard, *Affective Computing*. Boston: MIT Press, 1997.
[4] C. Nass and S. Brave, *Wired for Speech: How voice activates and advances the human-computer relationship*. Boston: MIT Press, 2005.
[5] D. Premack and G. Woodruff, "Does the chimpanzee have a 'theory of mind'?" *Behaviour and Brain Sciences*, vol. 4, pp. 515–526, 1978.
[6] S. Baron-Cohen, A. Leslie, and U. Frith, "Does the autistic child have a theory of mind?" *Cognition*, vol. 21, pp. 37–46, 1985.
[7] B. Reeves and C. Nass, *The media equation*. Cambridge University Press, 1996.
[8] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.
[9] R. Cornelius, "Theoretical approach to emotion," in *ISCA Workshop on Speech and Emotion*, Belfast, 2000.
[10] A. Whiten, *Natural theories of mind*. Oxford: Basil Blackwell, 1991.
[11] S. Baron-Cohen, "The descent of mind: Psychological perspectives on hominid evolution," M. Corballis and S. Lea, Eds. Oxford University Press, 1999, ch. Evolution of a theory of mind?
[12] W. James, "What is an emotion?" *Mind*, vol. 19, pp. 188–205, 1884.
[13] K. R. Scherer, "Studying the emotion-antecedent appraisal process: An expert system approach," *Cognition and Emotion*, vol. 7, pp. 325–355, 1993.
[14] R. Zajonc, "Feeling and thinking: Preferences need no inferences," *American Psychologist*, vol. 35, pp. 151–175, 1980.
[15] M. V. den Noort, M. P. C. Bosch, and 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.
[16] 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.
[17] J. D. Haynes and G. Rees, "Decoding mental states from brain activity in humans," *Nature Reviews Neuroscience*, vol. 7, pp. 523–534, 2006.
[18] M. Slors, "Personal identity, memory, and circularity: An alternative for q-memory," *The Journal of Philosophy*, vol. 98, no. 4, pp. 186–214, 2001.
[19] K. Höök, "From brows to trust: Evaluating embodied conversational agents," Z. Ruttkay and C. Pelachaud, Eds. Kluwer, 2004, vol. 7, ch. User-centred design and evaluation of affective interfaces.
[20] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection," *Neural Networks*, vol. 18, pp. 407–422, 2005.
[21] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, L. Devillers, L. Vidrascu, N. Amir, L. Kessous, and 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.
[22] V. Petrushin, "Emotion in speech: Recognition and application to call centers," in *ANNIE*, 1999. [Online]. Available: citeseer.ist.psu.edu/petrushin99emotion.html
[23] P. Y. Oudeyer, "The production and recognition of emotions in speech: Features and algorithms," *International Journal of Human Computer Interaction*, vol. 59, no. 1-2, pp. 157–183, 2003.
[24] R. Fernandez and 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.
[25] L. Vidrascu and L. Devillers, "Five emotion classes detection in real-world call center data: the use of various types of paralinguistic features," in *Paraling2007*, 2007.
[26] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Automatic hierarchical classification of emotional speech," in *Multimedia Workshops, ISMW '07*, 2007, pp. 291–296.
[27] P. Ekman, "Handbook of cognition and emotion," M. Power and T. Dalgleish, Eds. Chihester, UK: Wiley, 1999, ch. Basic emotion.
[28] R. Fernandez and R. W. Picard, "Modeling drivers' speech under stress," *Speech Communication*, vol. 40, pp. 145–59, 2003.
[29] "Nemesysco Ltd.- Voice Analysis Technologies," http://www.nemesysco.com/, Israel, Sept 2006.
[30] C. A. Moore, J. F. Cohn, and G. S. Katz, "Quantitative description and differentiation of fundamental frequency contours," *Computer Speech and Language*, vol. 8, no. 4, pp. 385–404, 1994.
[31] M. Schröder, "Speech and emotion research: An overview of research frameworks and a dimensional approach to emotional speech synthesis," The Institute of Phonetics, Saarland University, Tech. Rep., 2004.
[32] K. R. Scherer, "Approaches to emotion," K. R. Scherer and P. Ekman, Eds. Hillsdale, 1984, ch. On the nature and function of emotion: a component process approach, pp. 293–317.
[33] C. M. Whissell, "Emotion: Theory, research, and experience," R. Plutchik and H. Kellerman, Eds. New York: Academic Press, 1989, ch. The dictionary of affect in language, pp. 113–131.
[34] J. Kim, "Robust speech recognition and understanding," M. Grimm and K. Kroschel, Eds. Vienna: I-Tech Education and Publishing, 2007, ch. Bimodal Emotion Recognition using Speech and Physiological Changes.
[35] M. Grimm and K. Kroschel, "Robust speech recognition and understanding," M. Grimm and K. Kroschel, Eds. Vienna: I-Tech Education and Publishing, 2007, ch. Emotion Estimation in Speech Using a 3D Emotion Space Concept.
[36] M. Y. M. Hoque and M. Louwerse, "Robust recognition of emotion from speech," in *6th International Conference on Intelligent Virtual Agents, Marina del Rey*, 2006.
[37] "Humaine deliverable d5f," HUMAINE Network of Excellence, EU's 6^{th} Framework Project, http://emotion-research.net/projects/humaine/deliverables, 2006.
[38] T. Sobol-Shikler and P. Robinson, "Visualizing dynamic features of expressions in speech," in *proceedings of ICSLP, Jeju, Korea*, 2004.
[39] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotions in speech," in *ICSLP 96*, 1996.
[40] F. Burkhardt and M. Schröder, "Emotion markup language: Requirements with priorities," W3C Incubator Group, http://www.w3.org/2005/Incubator/emotion/XGR-requirements/, May 2008.
[41] E. Rosch, C. B. Mevis, W. Gray and D. Johnston, "Basic objects in natural categories," *Cognitive Psychology*, vol. 8, pp. 382-439, 1976.
[42] R. I. Phelps and P. B. Musgrove, "A prototypical approach to machine learning," Technical Report TR/02/85, Brunei, 1985.
[43] S. Baron-Cohen, O. Golan, S. Wheelwright, and J. J. Hill, "Mindreading: The interactive guide to emotions," Jessica Kingsley Limited, http://www.jkp.com, London, 2004.
[44] O. Golan, S. Baron-Cohen, and 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*, vol. 23, pp. 7160–7168, 2006.

[45] R. el Kaliouby and 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.
[46] T. Sobol-Shikler and P. Robinson, "Recognizing expressions in speech for human computer interaction," in *Designing a More Inclusive World, S. Keates, J. Clarkson, P. Langdon and P. Robinson (Eds), Springer-Verlag*, 2004.
[47] E. Douglas-Cowie, N. Campbell, R. Cowie, and P. Roach, "Emotional speech: towards a new generation of databases," *Speech Communication*, vol. 40, pp. 33–60, 2003.
[48] S. Baron-Cohen, J. J. Hill, O. Golan, and S. Wheelwright, "Mindreading made easy." *Cambridge Medicine*, vol. 17, pp. 28–29, 2002.
[49] 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.
[50] Z. Xiao, E. Dellandrea, W. Dou, and L. Chen, "Two-stage classification of emotional speech," in *Digital Telecommunications, ICDT '06*, 2006.
[51] E. Allwein, R. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," in *17th International Conference on Machine Learning*. San Francisco, CA: Morgan Kaufmann, 2000.
[52] C. W. Hsu and C. J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE Transactions on Neural Networks*, vol. 13, no. 2, pp. 415–425, 2002.
[53] C. M. J. A. N. Marquis de Condorcet, "Essay on the application of analysis to the probability of majority decisions," 1786.
[54] "Condorcet method," Wikipedia, The Free Encyclopedia, http://en.wikipedia.org/wiki/, August 2008.
[55] J. Malkevitch, "The process of electing a president," AMS, American Mathematical Society, http://www.ams.org/featurecolumn/archive/elections.html, April 2008.
[56] I. H. Witten and E. Frank, "Data mining: Practical machine learning tools with Java implementations," in *Morgan Kaufmann, San Francisco*, 2000.
[57] T. Sobol-Shikler, R. el Kaliouby, and P. Robinson, "Design challenges in multi-modal inference systems for human-computer interaction," in *proceedings of the 2nd Cambridge Workshop on Universal Access and Assistive Tehnology (CWUAAT), Cambridge, UK*, 2004.
[58] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," in *The Annals of Mathematical Statistics*, vol. 11, 1940, pp. 67–73.
[59] T. Sobol-Shikler, "Analysis of affective expressions in speech," Technical Report, UCAM-CL-TR-740, Computer Laboratory, University of Cambridge, 2009. Patent pending.
[60] 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.
[61] E. Zwicker, G. Flottorp, and S. S. Stevens, "Critical bandwidth in loudness summation." *The Journal of the Acoustical Society of America*, vol. Volume 29, pp. 548–57, 1961.
[62] E. Zwicker, "Subdivision of the audible frequency range into critical bands (Frequenzgruppen)," *The Journal of the Acoustical Society of America*, vol. Volume 33, p. 248, 1961.
[63] Iamblichus, *On the Pythagorean life*. E. G. Clark Trans., Liverpool University Press, (c300 ad)/1989.
[64] P. Gorman, *Pythagoras, a life*. London: Routledge and K. Paul, 1979.
[65] Galileo, *Dialogues Concerning Two New Sciences*. New-York: Dover Publications Inc., 1954.
[66] D. A. Schartz, Q. C. Howe, and D. Purves, "The statistical structure of human speech sounds predicts musical universals," *The Journal of Neuroscience*, vol. 23, pp. 7160–7168, 2003.
[67] M. J. Tramo, P. A. Cariani, B. Delgutte, and L. D. Braida, "The cognitive neuroscience of music," I. Peretz and R. Zatorre, Eds. New-York: Oxford University Press, 2003, ch. Neurobiology of harmony perception.
[68] S. Visa and A. Ralescu, "Issues in mining imbalanced data sets - a review paper," in *Midwest Artificial Intelligence and Cognitive Science Conference*, 2005, pp. 67–73.
[69] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, 1982.
[70] J. Platt, "Advances in kernel methods - support vector learning," B. Schoelkopf, C. Burges, and A. Smola, Eds. Boston: MIT Press, 1998, ch. Machines using Sequential Minimal Optimization.
[71] J. R. Quinlan, "C4.5: Programs for machine learning." San Mateo, CA: Morgan Kaufmann, 1993.
[72] M. A. Hall, *Correlation-based feature sub-set selection for machine learning*. New Zealnad: Hamilton, 1998.
[73] M. Hollander and D. Wolfe, *Nonparametric Statistical Methods*. New York: J. Wiley, 1973.
[74] "Cambridge dictionaries online," Cambridge University Press, http://dictionary.cambridge.org/, 2008.
[75] M. Grundland and N. A. Dodgson, "Color search and replace," in *Computational Aesthetics 2005, EUROGRAPHICS, Girona, Spain*, 2005, pp. 101–109.
[76] M. Grundland, "Color, style and composition in image processing," Ph.D. dissertation, Computer Laboratory, University of Cambridge, 2007.

Tal Sobol-Shikler Received the BSc and MSc degrees in Electrical Engineering from Tel-Aviv University and the PhD degree in Computer Science and Technology from the University of Cambridge. She is currently with the Department of Industrial Engineering and Management at Ben-Gurion University of the Negev. Her research concerns the application of human cognition and communication cues to the enhancement of human-computer interfaces and of human-robot interactions. She is a member of IEEE.



Peter Robinson Peter Robinson is Professor of Computer Technology and Deputy Head of the Computer Laboratory at the University of Cambridge in England, where he leads the Rainbow Group working on computer graphics and interaction. His research concerns new technologies to enhance communication between computers and their users, and new applications to exploit these technologies. Recent work has included desk-size projected displays and inference of users' mental states from facial expressions, speech, posture and gestures. He is a Chartered Engineer and a Fellow of the British Computer Society.