Technical Report

UCAM-CL-TR-740
ISSN 1476-2986

Number 740

UNIVERSITY OF
CAMBRIDGE

Computer Laboratory

Analysis of affective expression in speech

Tal Sobol-Shikler

January 2009

15 JJ Thomson Avenue
Cambridge CB3 0FD
United Kingdom
phone +44 1223 763500

http://www.cl.cam.ac.uk/

This technical report is based on a dissertation submitted March 2007 by the author for the degree of Doctor of Philosophy to the University of Cambridge, Girton College.

Many of the features and techniques presented in this report are subject to patent applications.

Technical reports published by the University of Cambridge Computer Laboratory are freely available via the Internet:

http://www.cl.cam.ac.uk/techreports/

ISSN 1476-2986

Abstract

This dissertation presents analysis of expressions in speech. It describes a novel framework for dynamic recognition of acted and naturally evoked expressions and its application to expression mapping and to multi-modal analysis of human-computer interactions.

The focus of this research is on analysis of a wide range of emotions and mental states from non-verbal expressions in speech. In particular, on inference of complex mental states, beyond the set of basic emotions, including naturally evoked subtle expressions and mixtures of expressions.

This dissertation describes a bottom-up computational model for processing of speech signals. It combines the application of signal processing, machine learning and voting methods with novel approaches to the design, implementation and validation. It is based on a comprehensive framework that includes all the development stages of a system. The model represents paralinguistic speech events using temporal abstractions borrowed from various disciplines such as musicology, engineering and linguistics. The model consists of a flexible and expandable architecture. The validation of the model extends its scope to different expressions, languages, backgrounds, contexts and applications.

The work adapts an approach that an utterance is not an isolated entity but rather a part of an interaction and should be analysed in this context. The analysis in context includes relations to events and other behavioural cues. Expressions of mental states are related not only in time but also by their meaning and content. This work demonstrates the relations between the lexical definitions of mental states, taxonomies and theoretical conceptualisation of mental states and their vocal correlates. It examines taxonomies and theoretical conceptualisation of mental states in relation to their vocal characteristics. The results show that a very wide range of mental state concepts can be mapped, or described, using a high-level abstraction in the form of a small sub-set of concepts which are characterised by their vocal correlates.

This research is an important step towards comprehensive solutions that incorporate social intelligence cues for a wide variety of applications and for multi-disciplinary research.

Contents

Chapter 1

Introduction

1.1 Motivation

Human-computer interaction and human-mediated communication have become a major part of our lives, but still lack the basic means of recognising and responding to non-verbal cues of attitudes, emotions and mental states, that we take for granted in human communication and reasoning. They fail to appreciate the users' reactions and intentions, and therefore fail to respond and react to them. To utilise the full potential of these new technologies, user-aware interfaces that complement existing interaction methods are needed.

While language is crucial to human communication, in most interactions it is supplemented by various forms of expressive information, such as facial expressions, vocal nuances, gesture and posture. These features depend on the context of the interaction and are often accompanied by physiological reactions, such as changes in heart rate.

Reeves and Nass observed that people interact with computers in the same way that they interact with each other [1]. There is a requirement for human computer interface (HCI) applications that support human emotional needs [2–4] In addition, humans' reasoning is affected by emotions [5–7].

The performance of computers interfaces can be improved by recognising and responding to human emotions and mental states. It is especially important in speech interfaces. In these systems speech is used to convey commands and data, while natural behaviour also uses speech for dialogue acts, thinking out loud, expressions of frustration, misunderstanding, discomfort, and more. Most of these functions relate to nuances of expressions, and some of them are obvious only in speech. The term expressions is used to describe the exhibited non-verbal behavioural cues that are related to the mental states.

In order to develop systems that infer mental states from non-verbal behavioural and physiological cues, signal processing and analysis techniques have to be developed, while at the same time consolidating psychological and linguistic analyses of emotions and mental states [8]. This dissertation aims to address these issues.

1.2 Objectives and results

Research in the field of *affective computing* [9] relates specifically to emotion recognition, analysis and synthesis. Most research in this field is based on analysis and synthesis of basic and extreme emotions like joy, anger, sadness, fear, disgust and surprise [10–13]. A few other mental states such as stress, frustration and depression have also been investigated in this context.

From the pragmatic point of view, information technology has to deal with emotions and mental states as they occur in real settings. However, expressions in general, mental states, behavioural patterns and attitudes have not been thoroughly investigated. Nuances of expression and emotion blends have been mentioned but not explored, and the investigation of expression changes over time is in its very early stages.

For all these reasons, the focus of this dissertation is to investigate expressions in speech, which represent natural speech, expression mixtures and expression nuances for a variety of mental states and of expression changes in time.

The first part of the dissertation presents a comprehensive framework that incorporates all the design and implementation stages of automated inference systems. It describes an implementation of this framework from the definition and recording of a database, through the definition and extraction of vocal features and the derived statistical and temporal metrics, to the design of a flexible and expandable inference system and its evaluation.

The second part of the dissertation focuses on the analysis of the relations between expressions. It includes the application of the inference system to (multi-modal) sustained interaction analysis and to expression mapping and characterisation. Expression mapping refers to the presentation and conceptualisation of mental states and the relations between them and between their expressions.

One of the challenges of this research was to define vocal properties that can be used for synthesis as well as for inference and finding mathematical definitions of these properties. I developed new mathematical definitions and algorithms for the extraction of features such as the fundamental frequency. I added the harmonic properties that have not been used before in this context, based on findings from musicology, physics [14–16] and neuroscience [17] that indicate that people both hear and generate harmonic properties in speech. These properties proved to be good indicators of voice quality.

The features were used as input to an automatic inference system that recognises co-occurring emotions and complex mental states. The system maps input speech segments according to the expressions they convey and performed well in tests [18].

Mapping the relations between expressions was another objective of this research. Mapping is important for prediction of changes, for fine-tuning of inference and for browsing between large number of concepts for user interfaces and synthesis. The inference system was used for mapping the concepts of the MindReading database [19], arranged into 24 groups according to the MindReading taxonomy. It mapped 459 of these mental states into a smaller expressions space according to their vocal correlates. This mapping is different in approach from other mapping techniques [20–24]. The mapping results agree with the respective lexical definitions and meanings of both single concepts and concept groups, while other meaning groups have immerged as well. The mapping reveals interesting properties of expressions, of their acting and of the taxonomy.

One of the main assumptions in this research was that expressions co-exist and change asynchronously in time. My analysis shows that expressions change gradually throughout an interaction, and more specifically through a human-computer interaction. For this purpose I used the Hebrew multi-modal database, Doors, that was defined and recorded as part of this research, based on the Iowa Gambling Test [25]. The inference was compared to the computer game events and to behavioural cues such as mouse movements. This experiment also shows that for certain mental states, a system trained on one language can be used for inference in another language.

1.3 Dissertation structure

The research described in this dissertation draws inspiration from several disciplines. Chapter 2 presents different aspects of expressions, including theories on the social and cognitive roles of emotions and mental states. It discusses a novel presentation of the temporal characteristics of expressions, and a summary of previous work on the conceptualisation and mapping of emotions. It then discusses the subject of expressions in human-computer interaction, including both automatic inference and synthesis systems. It highlights the shortcomings of main stream research in dealing with mental states other than the basic emotions. The chapter also reviews the main approaches to analysis of paralinguistic cues. Additional background surveys for more specific topics are included throughout the dissertation.

Chapter 3 presents a general multi-stage framework for automatic paralinguistic systems, followed by an overview of its implementation in this dissertation, which is described in more detail in the following chapters.

Chapter 4 surveys the issue of defining and recording suitable databases for automatic analysis of expressions. It then describes the two corpora used throughout this dissertation, Doors and MindReading, and discusses their respective merits. It also describes the algorithm used for segmentation of raw sound-tracks.

Chapter 5 is the first of two chapters that discuss the implementation of an automated inference system. This chapter presents the definition and extraction of vocal features from the speech signal. It also presents the temporal and statistical characteristics of these features.

Chapter 6 describes the inference of complex mental states from single sentences. In addition to a single inferred expression it presents the possibility of expression mixtures, expressions that co-occur simultaneously. The system uses pair-wise classification machines. Each of these machines chooses between two expressions, using its own features and classification algorithm. A voting system combines their results. The system is flexible and expandable to accommodate additional features and expressions. This chapter is summarised by the presentation of various evaluation and validation results.

Chapter 7 is the first of two chapters that discuss further applications of the inference system. This chapter presents a novel approach to expression conceptualisation, using a small set of expressions to map a very large variety of mental state concepts, according to their vocal correlates. It also examines the vocal correlates of the MindReading taxonomy. In the first section of the chapter, the ability of the inference system to distinguish between complex mental states is compared to human performance as reported in an independent test. The second section presents the inference results for 459 mental states from the MindReading database, mapped according to concept groups that are defined by the MindReading taxonomy and to the recognisable vocal expressions. The third section presents the relations between the different concepts only according to the set of nine recognisable expressions. The inference and mapping reveal meaning of mental state concepts and the relations between different concepts, between concept groups based on lexical meaning and between concepts and their vocal (behavioural) expressions.

Chapter 8 demonstrates by examples how the scope of the system can be extended to inference in the interaction level, in which the single utterances are parts of sustained interactions. This inference is reinforced by multi-modal analysis of events and behavioural cues. The system's scope is extended to a different language, naturally evoked expressions over time and multi-modal information, using the Doors database.

Chapter 9 summarises the work presented here and its major contributions. It concludes the dissertation with directions for future research.

1.4 Publications

Some of the results in this dissertation and during this research have appeared in the following publications:

1. T. Sobol Shikler, P. Robinson, "Recognising 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

2. T. Sobol Shikler, R. El-Kaliouby, P. Robinson, "Design Challenges in multi-modal inference systems for human-computer interaction", *proceedings of the 2nd Cambridge Workshop on Universal Access and Assistive Tehnology (CWUAAT) 2004*, Cambridge, UK

3. T. Sobol Shikler, P. Robinson, "Visualising Dynamic Features of Expressions in Speech", *proceedings of ICSLP2004*, Jeju, Korea

4. T. Sobol Shikler, P. Robinson, "Affect Editing in Speech", *proceedings of the 1st International Conference on Affective Computing and intelligent Interaction (ACII)* 2005, Beijing, China

These publications include only preliminary observations and results. The rest of the results that appear in this dissertation are being processed for publication.

References

[1] Reeves B. and Nass C., *The media equation*, Cambridge University press, 1996.

[2] Picard W. Rosalind, "Affective computing challenges", *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[3] Hudlicka Eva, "To feel or not to feel: The role of affect in human-computer interaction", *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 1–32, 2003.

[4] Nass C. and Brave S., *Wired for Speech: How voice activates and advances the human-Computer relationship*, The MIT Press, 2005.

[5] Kahneman D. and Tversky A., "Prospect theory: An analysis of decision under risk", *Econometrica*, vol. XVLII, pp. 263–291, 1979.

[6] Bechara A., Damasio H., and Damasio A. R., "Emotion, decision making and the orbitofrontal cortex", *Cereb Cortex*, vol. 10, no. 3, pp. 295–307, 2000.

[7] Damasio A. R., *Descartes Error: Emotion Reason and the Human Brain*, Putnam Sons, New-York, 1994.

[8] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., and Taylor J. G., "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[9] Picard R. W., *Affective Computing*, Cambridge, Massachusetts: MIT Press, 1997.

[10] Ekman P., *Handbook of cognition and emotion*, chapter Basic emotion, Wiley, Chihester, UK, 1999.

[11] Fernandez R., PhD thesis, Media Arts and Sciences, Massachusetss Institute of Technology, 2004.

[12] Breazeal C. and Aryananda L., “Recognizing affective intent in robot directed speech”, *Autonomous Robots*, vol. 12, pp. 32–80, 2002.

[13] Dellaert F., Polzin Th., and Waibel A., “Recognizing emotions in speech”, in *ICSLP 96*, 1996.

[14] Iamblichus, *On the Pythagorean life*, Clark G. translator Liverpool, c300/1996.

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

[16] Schartz D. A., Howe Q. C., and Purves D., “The statistical structure of human speech sounds predicts musical universals”, *The Journal of Neuroscience*, vol. 23, pp. 7160–7168, 2003.

[17] Tramo M. J., Cariani P. A., Delgutte B., and Braida L. D., *The cognitive neuroscience of music*, chapter Neurobiology of harmony perception, Oxford University Press, New-York, 2003.

[18] Golan O., Baron-Cohen S., and Hill J., “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.

[19] Baron-Cohen S., Riviere A., Fukushima M., French D., Hadwin J., Cross P., Bryant C., and Sotillo M., “Reading the mind in the face: A crosscultural and developmental study”, *Visual Cognition*, vol. 3, pp. 39–59, 1996.

[20] Schrödr M., “Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis (ph.d thesis). vol. 7 of phonus, research report of the institute of phonetics”, Tech. Rep., Saarland University, 2004, http://www.dfki.de/∼schroe.

[21] Scherer K.R., *Approaches to emotion*, chapter On the nature and function of emotion: a component process approach, pp. 293–317, Hillsdale, NJ: Erlbaum, 1984.

[22] Whissell C. M., *Emotion: Theory, Research, and Experience*, chapter The dictionary of affect in language, pp. 113–131, New York: Academic Press, 1989.

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

[24] “Humaine deliverable d5f”.

[25] Bechara A., Damasio H., Tranel D., and Damasio A. R., “Deciding advantageously before knowing the advantageous strategy”, *Science*, vol. 275, pp. 1293–5, 1997.

Chapter 2

Background

This dissertation derives its inspiration from various disciplines, including psychology, sociology, physiology, philosophy, musicology, engineering and computer science. This chapter explores the main theories that explain the social and cognitive roles of emotions and mental states and their expression in human behaviour and communication. As part of the discussion on expressions, it also presents the temporal characteristics of expressions and the asynchronous nature of mental states mixtures. In addition, it briefly surveys other research on conceptualisation of emotions, which aims to map emotions and mental states using different metrics.

This chapter then surveys the subject of expressions in human-computer interaction. It highlights the importance of incorporating emotions and mental states in human-computer and human-machine interfaces and surveys previous research on automatic inference. In each of these areas it also presents the approaches taken in this research.

The final section of this chapter demonstrates how research into the automatic analysis of the expressions through the non-verbal aspects of speech is related to a wide range of research fields such as hearing, communication disorders, music, linguistics and speech recognition. These different approaches feature throughout the dissertation, especially in the definition of vocal features and their extraction methods.

2.1 Emotions, mental states and expressions

This review of emotions presents a few of the most popular theories about the existence and roles of emotions and mental states, and their observable expressions. It explains the difference between this research and other approaches used in the field of affective computing. This difference affects the choice of databases, the analysis methods and possible applications. The theories appear mostly in the chronological order of their original appearance.

Most research related to automated analysis of expressions is based on small sets of *basic emotions*, such as *joy*, *sadness*, *fear*, *anger*, *disgust* and *surprise* [1]. 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 different situations. This is also the evolutionary o *Darwinian theory*, based on Darwin's book of 1872 [2]. The implication of this theory is that if there is indeed a small number of basic or fundamental emotions, each corresponding to a particular evolved adaptive response pattern, then it should be possible to see the traces of evolution in various aspects of emotion, for example, in speech. The dominance of this theory in automatic inference of affect can be explained by the fact that they 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* ([3–5] and references within). Another explanation may be that stereotypical presentations of these expressions are perceived as easier to act and to recognise, and therefore useful for both quick acquisitions of datasets, and as a starting point for an emerging research field. This research perspective ignores the major part of human expressions. 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 *Jamesian* tradition, relying on James' article from 1884 [6], focuses on the experience of emotions by bodily reactions. Our bodies respond first and our experience of these changes constitutes what we call emotion. The implications are that first, major injuries affect the intensity of emotional experience. Secondly, there should be differentiation of emotions at the level of the autonomic nervous system. Finally, reliable changes in affect and autonomic activation appear to follow posed facial expressions of emotion and the bodily postures associated with affect. For example, smiles can improve the mood of the person who smiles, while the inability to smile often leads to depression [7]. Gauging reactions from physiological cues, for example skin conductivity (Galvanic Skin Response), GSR as indicators of stress and deception, and heart rate as indicator of arousal level and valence (positive/negative) can be regarded as related to this theory.

The dominant theory in the 1970s was the *cognitive* one, whose central assumption is that thought and emotions are inseparable. Accordingly, perception of emotions depends on *appraisal*, the process of judging whether events in the environment are good or bad for us. Scherer published several articles that establish the processes of appraisal [8], the connection between emotions and speech, and on the subject of mixed emotions as opposed to emotions as single separately occurring states [9–12]. This research takes into account the co-occurrence of different mental states and the ability to assign mixtures of emotions and mental states to expressions.

The cognitive approach to emotions started to change when focus was given to subconscious emotional reactions, such as preferences, that can exist without cognition [13]. Much of contemporary psychology has come to recognise that a great deal of human emotional functioning is rooted in unconscious processes. During the last two decades, many studies were conducted in this field. These studies, for example, showed that humans pick up the emotional content of unconscious facial expressions that are not intended to influence the perception of others. Other studies showed that humans evaluate objects (as for example "good" or "bad") at an unconscious level [14]. Bechara *et al.* show how certain brain injuries affect emotional and nervous system responses to events (measuring GSR), even before conscious awareness exists, and how they consequently affect decision-making and risk-taking [15]. They describe how response to a card game evolves over time. The Doors database, which is part of the research described in this dissertation, is based on this experiment. The cognitive role of affect and emotions and their influence on human cognition and decision making is demonstrated in the works of Kahneman and Tversky that show how emotions influence economic decision-making [16].

The last major perspective is the *social constructivist*, originating in the 1980s. Social constructivists believe that emotions are cultural products that owe their meaning and coherence to learned social rules. Social intelligence and social cognition encompass our abilities to grasp, remember, and interpret our and others' behaviours in terms of mental states (thoughts, intentions, desires, and beliefs) and to interact in the highly complex social world [17]. The *Theory of mind* [18, 19] or *MindReading* [20], that is the ability to attribute the full range of mental states to ourselves and to others, and to use such attributions to make sense of, to predict and to influence behaviour, can be regarded as part of this perspective. It comprises of two parts: perception of behavioural cues, and the ability to explain them. Höök [21] extends these definitions, and describes affect as human, rich, complex and ill-defined *experience*.

I adopt a comprehensive point of view, that considers expressions in general, both as part of social intelligence, social cognition, and MindReading [22]. According to this view, expressions are part of human communication and behaviour in society. Therefore, they usually appear in the context of an interaction. However, expressions are also related to physiological and environmental parameters. This multi-faceted approach is expressed in the inclusion of physiological cues in the Doors database, in addition to context awareness in the form of event recordings (described in Chapter 4). Another assumption in this work is that the *universality* or *generality* of expressions is not limited only to basic emotions but can be extended and applied also to complex mental states such as *thinking*.

The term 'expressions' refers in this work to the outer representation of emotions, mental states, moods, attitudes, physiological states, cultural display rules, dialogue acts, etc. It is the observable behaviour (conscious and unconscious) that we can perceive and would like to interpret. This perspective also takes into account the translation of cues from an inference system into synthesis systems that aim to imitate these expressions, and not their actual source.

Time-related characteristics of expressions

Natural human interactions are dynamic in nature. If the entire interaction is the observed time window, the different underlying causes for expressions have varying relative durations, and they should be considered in the analysis. Cowie *et al* presented a short description of time-based categories of expression [3], from autonomic changes and what they call expressions that last seconds, through attitudes, full-blown emotions that can last from minutes to hours to moods (hours to months) to long-term emotional disorders and traits that exist years if not a lifetime. However, they do not take into account the parallel nature of these sources and their interactions. I present here a schematic description of an interaction and the relative timing and durations of both expressions and their underlying causes (the word expression will be used for both) over the duration of the interaction as shown in Figure 2-1. Here the term 'expressions' is used again in its broadest definition, which includes both posture characteristics, that can be dominant for years, and micro facial expressions that can last milliseconds.

For example, the duration of the speaker's personality extends beyond the duration of the interaction, and affects the nature of responses and expressions generated throughout the interaction. General dispositions or mental states are also relatively stable throughout the duration of a single interaction. People who are depressed and sad, or in love and happy or even deep in thought, will usually keep this general attitude or mood, and show it, unless something changes it during the interaction. However, these changes do not frequently occur. These relatively long-term expressions may be important for diagnostic systems, such as in the case of clinical depression. The attitude towards the situation or towards the setup of the interaction is more dynamic in nature. For example, a new employee attending a review with his boss may feel stressed before and at the beginning of the interaction, while becoming more confident during the sustained interaction. A client using a call centre may become annoyed, frustrated and angry. In some applications, we would like to analyse these changes, and respond to them.

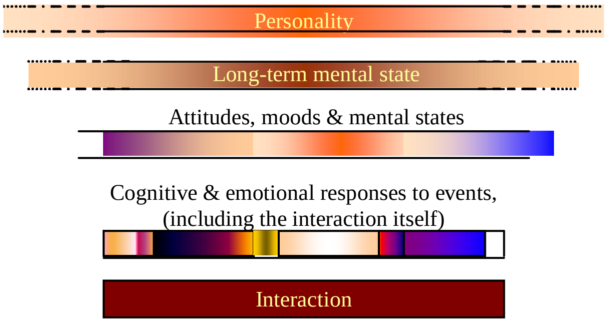
Personality
Long-term mental state
Attitudes, moods & mental states
Cognitive & emotional responses to events,
(including the interaction itself)
Interaction

Figure 2-1: Schematic description of the various underlying reasons for expressions, and their durations relative to the duration of an interaction. The time scale of the expressions is relative to the length of the analysed interaction (at the bottom).

There are local, short-term causes for expressions that a system should detect, but the impact on the output depends on the specific application. For example, we would like to detect short-term misunderstandings. A system may respond to such an event and change the course of the interaction (for example, in a teaching system). Another course of response is to keep the knowledge as a cue for a more continuous interpretation, without changing the inference of longer-term expressions (for example, dialogue systems). Therefore, the structure of the general inference system should allow the co-existence of these temporal variations alongside a longer-term analysis. Abrupt changes, such as surprise, should be able to change the course of the inference, but the duration of the change and its actual impact on the overall interaction can vary. Another implication of this approach is that a completely 'neutral' expression hardly ever exists. The manner of response is context dependent. A similar philosophy is expressed by Höök [21].

Conceptualising mental states and emotions

The relationship between expressions may have several uses for automatic recognition and synthesis of expressions. In particular, it can be useful for continuous tracing of expressions, assuming gradual changes over time. There are several approaches to conceptualising emotions and distinguishing them. The categorical approach postulates a limited number of universal basic emotion categories, from which all the rest are derived, either as a blend of basic emotions, or in conjunction with a cognitive process. An example of this approach can be seen in Plutchik's work as illustrated in 2-2 [23].

The dimensional approach is based on identifying emotions by their placement on a small number of dimensions such as valence (positive/negative), activity (passive/active), potency (strong/weak) and action tendencies (outward/inward) [3, 24, 25]. Whissel presented a list of 4000 words with associated ratings on the Activation and Evaluation dimensions. 3000 words are at a distance of more than the standard deviation from the neutral mean value. The word list has been widely cited and used to measure the emotional impact of different systems and texts [26]. Cowie *et al.* [3] expressed the notion that it is unrealistic to expect a machine to be able to reach such a level of discrimination.

The Feeltrace tool [27] is an extension of this approach. It enables a user to trace a small number of dimensions continuously over time, as can be seen in Figure 2-3 [28].

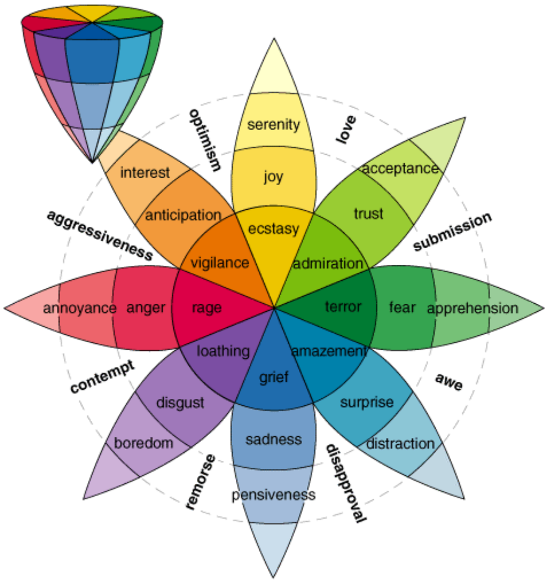

Figure 2-2: A three-dimensional circumplex model describes the relations among emotion concepts, which are analogous to the colours on a colour wheel. The cone's vertical dimension represents intensity, and the circle represents degrees of similarity among the emotions [23].

Douglas-Cowie *et al.* [29] proposed a list of 48 emotion categories for the HUMAINE project, these emotions are arranged into 10 groups, including *negative forceful*, *negative/positive thoughts*, *caring*, *positive lively*, *reactive*, *agitation*, *negative not in control*, *negative passive* and *positive quite*.

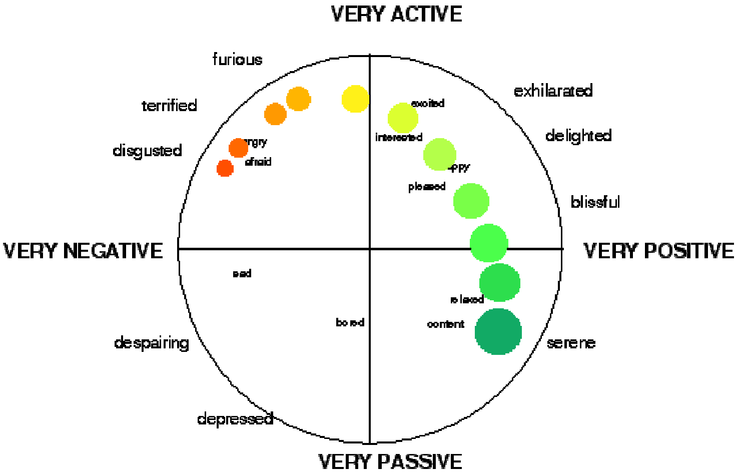

Figure 2-3: Feeltrace is a labelling tool for two emotion dimensions, developed by Cowie *et al* [28]. It allows for the tracking of a perceived emotional state continuously over time, on the two main emotion dimensions activation and evaluation.

Another approach is the prototype approach, which assumes that language and knowledge shape the way people categorise information. It is a compromise between the two other approaches. It has both contents of individual categories and the hierarchical structures among them. The MindReading database, which is one of the databases used in this work, is arranged according to this method [30]. Table 2-1 presents the main group categories of the MindReading. These methods do not address nor preclude the possibility of co-existence of different mental states.

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

Table 2-1: The 24 mental state groups that constitute the MindReading taxonomy of Baron-Cohen *et al.* [30]. Basic emotions are listed in the left column. The groups that are addressed in this dissertation are indicated with a *. Two groups were extracted from the interested group: interested and absorbed.

In this dissertation, an automatic mapping of expressions according to their vocal properties is presented. The mapping is into a space that is defined by the vocal nonverbal speech cues of an arbitrary set of emotions and complex mental states, including *joyful*, *sure*, *unsure*, *thinking*, *stressed*, *excited*, *opposed*, *interested* and *absorbed*. The last two groups are taken from the interested group in the MindReading taxonomy. This mapping reveals the vocal correlates of the MindReading taxonomy, and the relation of different expressions to the chosen set. In addition, co-existing expressions are considered

The role of emotions in human-computer interfaces

Another line of psychological and sociological research is concerned with the role of human emotions and human communication cues, such as facial expressions, and verbal and nonverbal speech cues in human-computer interfaces. Nass and Beeves, in the book *The media equation* [31], reveal that people interact with computers in the same way that they interact with each other. They expect politeness and proper responses to their attitudes and mental states. People assess differently and react differently to interfaces according to the 'personalities' they reveal. In the book *Wired for Speech* [32], Nass and Brave demonstrate how voice qualities and expressions in voice activate and advance the human-computer relationship. Picard stresses the emotional needs of users in human computer interfaces [33], while Hudlicka [34] explains why we need machines that 'feel', or address user affect. Höök's [21] perspective regards people's affective reactions as part of ongoing interactions embedded in a broader social context, and adapted to the current context. These works lay the ground for the research of automatic analysis and synthesis of human expressions, for integration in responsive and user-friendly machines. The inability to reason about mental states is considered one of the main inhibitors of social and emotional intelligence; it appears for example in people with Autism Spectrum

Disorders (ASD). It seems imperative to incorporate the ability to perceive behavioural cues and the ability to interpret them or to imitate human behaviour in automatic systems that aim to interact with people using human communication cues or to teach and assist people with disabilities or special communication needs.

Further motivation for integration of affect and the recognition of affect into computer systems is the influence of emotions on cognitive processes. Modelling this influence and creating agents that can predict or imitate human reasoning is an interesting possibility but is beyond the scope of this dissertation.

2.2 Automated expression-related systems

Automated systems relying on cues from human communication usually limit their attention to a small set of underlying theories. Work related to intelligent agents or socially adept agents usually refers to social psychology, and use natural language and gestures as cues of mental states [35, 36]. Other types of synthesised systems include humanoid robots and pets that use analogies from human behaviour, especially dialogue rules and basic affective speech [37–41]. The majority of research on automated recognition of implicit communication cues considers emotional content. This is true for physiological signals [42], speech cues [43–45], and facial expressions [3, 46–48]. In speech though, there is also a vast effort to identify stress, for example in pilots' and drivers' speech [49], for improved speech recognition [50], and in work related to depression diagnosis [51]. There are also commercial companies, for example Nemesysco [52] provides services such as identification of attempts at insurance fraud made over the phone, employee stress management and detection of deceit during job interviews.

Nevertheless, a wider view of expressions in general has not been fully researched in the context of automatic expression recognition and analysis. Current research efforts have only started to address the problem of changes in expressions over time [37, 53]. For example, Picard successfully recognised certain emotions from physiological reactions of an actor, recorded on different days [37]. In another example, el Kaliouby has developed a system that automatically recognises several mental states from short video streams of head gestures and facial cues, with the inferences changing in time [54].

The study of the wide range of natural expressions, including nuances of expression, mixtures of emotions and attitudes has not been thoroughly explored yet in the context of automatic recognition and synthesis. One of the main efforts in recent years in this direction was the JST/CREST ESP project, that involved recordings of people in real situations spanning 5 years [55]. This research focus shifted from what the speaker feels (emotions) to what the speaker is doing (relationships). They verified that speaking style depends upon who we are speaking to, and by how we feel about what we are saying. They also found that grunts (non-verbal utterances and sounds) constituted more than half the content of the recorded speech. Fernandez [43] has developed a system that recognises breaths and pauses. The emphasis has been on prosody, the non-verbal features of speech, and on the incorporation of time-varying parameters of speech at various time scales. However, these works are only the beginning. Understanding all of these features is vital for correct and fully automated analysis and synthesis of expressive speech, and the level of knowledge and tools is still relatively low.

There is a substantial body of work in affective speech synthesis, as can be seen in the review by Schröder [56]. Morphing of affect in speech, meaning regenerating a signal by interpolation of auditory features between two samples, was described by Kawahara and Matsui [57]. Their work explored transitions between two utterances with different expressions in the time-frequency domain. Further results on morphing speech for voice changes in singing were presented by Pfitzinger [58], who also reviewed other work and techniques related to morphing.

However, most studies have explored just a few extreme expressions, and not nuances or subtle expressions. The methods that use prosody characteristics consider a single definition for a whole sentence or utterance. Examples of the benefits of such manipulations for human-computer interaction, and voice characterisation are discussed by Nass and Brave [32]. They achieved relatively long-term effects such as perceived gender, and association of personality traits with synthesised voices. Only a few have integrated the linguistic prosody categorisations such as intonation contours [59, 60]. Examples of morphing, using two samples in order to generate intermediate expressions, consider very short utterances (one short word each), and include a few extreme acted expressions. None of these techniques has led to editing tools for general use.

The combination of input modalities, that is the integration of different communication and context cues such as gestures, posture, facial expressions, verbal and non-verbal speech, physiological cues, events, forums (audience identity and the relationship with the audience) and locations, is in its early stages because there are no simple tools for feature tracking and representation in each medium. The output modalities of such systems can also have different forms, including different types of feedback, voice synthesis and animation, all of which are subjects of research and development. It is impossible to ignore in this context the HUMAINE consortium [61] that aims to address many of these questions and more. Their toolbox includes entries related emotion description, emotion classification, signal analysis (face, gesture and bio-signals), databases, labelling tools and usability.

This dissertation presents an inference system that recognises a set of emotions and mental states beyond the set of basic emotions. It is applied to the recognition of a very large set of mental states, both acted and naturally evoked. It examines expression mixtures and dynamic analysis of expressions in time, in addition to combination with other modalities. From this point of view, this dissertation presents a step towards the integration of inference system and dialogue systems. The features chosen for analysis can be translated to synthesis. The analysis is based on the assumption (enhanced by observations) that there are temporal characteristics of features within utterances as well as between utterances. The dissertation presents a new approach to expression mapping using the automatic inference. A new tool for affect editing was also introduced during this research, but is beyond the scope of the dissertation.

2.3 Approaches to Expressive Speech Analysis

Expressive speech analysis is the analysis of the paralinguistic cues, that is the nonverbal vocal communication cues, such as tone of voice, intensity and the like. It can be regarded from several points of view, including signal processing, speech recognition and linguistics, hearing and neurology, music and psycho-acoustic. These different approaches influence the selection of speech features for expression analysis.

The first and most common approach is the production approach, which tries to model the speech signal according to its production system, which is the influence of the structure of the breathing mechanism, the vocal folds, mouth and noise on the speech signal properties. Another approach is from the perspective of perception, which analyses how the speech signal is perceived and processed by the human ear and brain. This latter approach is used in analysis of hearing mechanisms, both neurological and psychological, to treat hearing impairments, and in musical analysis. An example of the difference between the approaches is the difference between the term fundamental frequency, which is derived from the production approach, and the term pitch, which relates to the perception approach. The production approach relates to mathematical models of the vocal tract. Research involves X-ray photographs, ultrasound, and MRI images of the vocal tract during articulation [62, 63] and development of simplified models such as Linear Predictive Coding (LPC) [64]. According to this approach, expressive uses of speech depend on prosodic features as well as the production of phonemes. They are related to factors like subtle changes in the breathing muscles and the vocal fold, and to vocal-tract shaping factors, which relate to movements of the upper articulators: the velum (soft palate), tongue, teeth, jaw and lips. Figure 2-4 shows the location of the speech production mechanisms. The prosodic features are duration, intensity, fundamental frequency and spectral patterns. The fundamental frequency is the rate of the vocal fold vibrations, 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 [64–66].

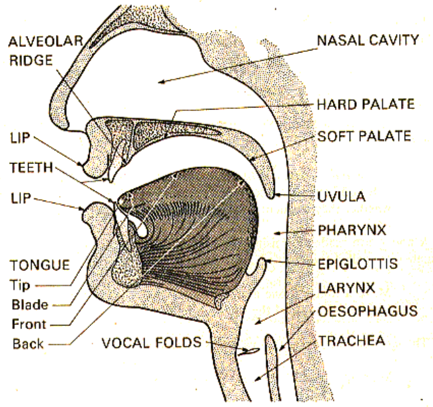

Figure 2-4: Vocal tract structure [66].

From the perspective of the listener, our brain processes faces and speech in different centres from those in which it processes other images and sounds, respectively [67, 68]. One aspect of the perception point of view is psycho-acoustic tests, in which people are asked to assess certain features of sounds, music, speech, and modified speech signals. Tests and features relevant to emotional speech analysis were described, for example, by Mozziconacci and by Murray *et al.* [69, 70], and in the context of human-computer interaction by Nass and Brave [32]. The most relevant features to recognition of emotions and expressions, according to these works, are properties of the pitch contour, the energy of the speech signals, and features related to the spectral content. Works that tried to estimate the combinations of features that correspond to different emotions yielded different and sometimes contradictory results ([3] and references within). However, pitch perception by people is not trivial, and is a matter of investigation by hearing and music researchers [71, 72].

Understanding the structure of the ear, and the processing in it may help to define the features of the speech signal that are used for analysis. For example, the spectral content of speech can be related to the cochlea's structure, in which frequencies are located on a logarithmic scale, with sensitivity frequency bands of logarithmically increasing width, or with periodicity that corresponds to a particular wavelength [73]. Human hearing can process sounds in the range of approximately 20 Hz - 20 kHz. The ear can also process sound pressure levels of 0 dB to 120 dB (equivalent to the range of air pressure of 20 μPascal to 20 Pascal), but the sensitivity changes with the frequency bands. Higher pressure is required in lower frequencies in order to be perceived. Works in the field of communication disorders showed the importance of information carried in relatively high frequencies (at least up to 8 kHz) of the speech signal [74–76]. Much of the perception is done in the brain, therefore we can use telephones that use a much narrower frequency range but the ability to perceive unfamiliar words decreases. The extent of the information loss in the case of expressions perception is much more difficult to assess.

The linguistic point of view includes relevant descriptions of intonational phrases or tone groups, for analysis of stress and accent in the pronunciation of words and sentences. It is also used for prosody analysis of various speech and dialogue acts, such as the distinction between questions (rising edge at the end of a sentence) and statements (falling edge), and the distinction among different types of questions in English [3], or logical arguments in Hebrew. These layers of information complicate still further the analysis of prosody in relation to emotions, attitudes to the situation and listeners, moods and personality. In English, prosodic units are variously known as information-units, tone-units, tone-groups, intonational phrases and word groups. The most important item in an utterance has different terms: accent, nucleus, stress, emphasis, which is in many cases a peak of the pitch contour. One factor that influences the perception of prosody is the differences in the pitch height of syllables within an utterance. Studies in that tradition associate different types of tones and tone shapes with emotions [3].

In recent years, one of the most widely used automatic models for pitch structures is the Tilt model, by Taylor [77]. Detection of intonation events involves determining where, in the speech signal, accent and boundary events are located. Using the ToBI model, which is a standard for modelling English prosody [78], the process involves not only determining whether the event is an accent or boundary, but what tones make up the event. The third task, placement, is the act of linking an event with a portion of linguistic text (e.g. syllable, word, phrase). The main differences though between the field of automated speech recognition and automated recognition of the non-verbal aspect of speech, is that in speech recognition there are layers of information with absolute labels of words and meanings. In the area of expressive speech recognition these higher levels do not exist, at least not in such an absolute and obvious form.

Another point of view for expressive speech analysis and prosody analysis is the interpretation of prosody as music, or more precisely with terminology and features that are used for the description of music. Musical terminology includes the term pitch, but it is argued that pitch in speech is different from the pitch of music. One argument is that the pitch constantly changes in speech. Another argument is that there are no mathematical relations between the pitches in a series that constitute a sentence. These arguments have not been examined in the context of expressive speech. Musical terminology also incorporates terms such as harmonics, which are multiples of the fundamental frequency. The number and amplitude of the harmonics, or the harmonic spectra, can define the sound of musical instruments in a way that can be described as *thin*, *bright* or *rich* [79]. As such terms are applied to expressions in human voice, it seems reasonable to expect similar spectral characteristics both in voice characteristics and in expressive speech sounds. A related definition is consonance and dissonance, or how pleasant and harmonious or unpleasant and inharmonious combinations of tones sound [80–84]. This point of view has been investigated recently in the context of emotional and expressive speech by Fernandez [85], who included dissonance, chords and rhythm calculations in his analysis. These features proved to be discriminative for a set of four basic emotions, including *afraid*, *angry*, *happy* and *sad*, and 'neutral' speech.

Other musical characteristics include tempo, rhythm, intensity, dynamics, and tonal structures or melodies that can be seen in this respect as pitch contour structures. A few of these features have been examined in emotional or expressive speech. In western music, melodies are divided into phrases, built of motifs. Different phrase structures are associated with different levels of strength and movement, which mostly depend of the manner of their completion [86]. Different combinations of characteristics create tension or relaxation, and aspects of the perception of these characteristics have even been found among very young babies [87].

The music perception point of view is also supported by findings from neurological research that concern the perception and analysis of music, emotions and speech in different centres of the brain, and certain disorders. This may explain why certain attributes of music have been found to be universal, and why music itself is something that exists universally. The structure of auditory centres in the brain is complicated and not yet fully understood. However, broadly speaking, there are centres that are tuned to periodicity. The right hemisphere is considered to be related to the analysis of music, while the left hemisphere is more related to the processing of language. The left hemisphere deals with complex spectral combinations that change relatively quickly, while the right hemisphere appears more attuned to melodies. Lesions in right cortical regions impair recognition of emotion in prosody, while lesions in other parts of the brain that are activated with expressive vocal stimuli, such as the amygdale, do not reduce recognition [88]. People can lose the ability to understand speech, while being able to understand prosody and voice, and vice-versa. This general approach also has the potential to explain findings from child language and prosody development.

All these points of view show how analysis of affect in speech is intricate and related to other aspects of human cognition, physiology, and the like. These different aspects should be considered for expressive speech analysis, because they may contribute both features and methods. This is required because there are still features of affective speech that have not been fully addressed, such as voice quality, and expressions that include speech acts, nuances of expressions and expressions of subtle emotions and mental states. Because of the high degree of complexity involved, this work aims to address parts of this broad research problem. I examined and adopted different points of view for the definition of features, and for setting feature properties for automated analysis of expressive speech, with respect to synthesis. These features and their origins are described in more detail in Chapter 5

2.4 Summary

This chapter presented the background for this dissertation, including the theories that are most commonly used in the area of affective and expressive systems, that is automatic systems that incorporate or recognise the expressions of emotions and mental states, and why such systems are required. It presented topics like the dynamic characteristics of expressions, or the behaviour of expressions over time, as well as methods and approaches for expressions characterisation and conceptualisation. This chapter surveyed current and past automatic affective systems, and presented the fields and approaches that can contribute to the development of inference systems from paralinguistic cues. These themes will be developed in the next chapter, which discusses the requirements and challenges involved in the development of such systems and presents a framework for the research presented in this dissertation.

References

[1] Ekman P., *Handbook of cognition and emotion*, chapter Basic emotion, Wiley, Chihester, UK, 1999.

[2] Darwin C., *The Expression of the Emotions in Man and Animals*, New-York, 1898.

[3] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., and Taylor J. G., "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[4] Cornelius R. and Cowie R., "Describing the emotional states that are expressed in speech", *Speech Communication*, vol. 40, pp. 5–32, 2003.

[5] Cornelius R., "Theoretical approach to emotion", in *ISCA Workshop on Speech and Emotion*, Belfast, 2000.

[6] James W., "What is an emotion?", *Mind*, vol. 19, pp. 188–205, 1884.

[7] VanSwearingen J. M., Cohn J. F., and Bajaj-Luthra A., "Specific impairment of smiling increases the severity of depressive symptoms in patients with facial neuromuscular disorders", *Aesthetic Plastic Surgery*, vol. 23, pp. 416 – 423, 1999.

[8] Scherer K. R., "Studying the emotion-antecedent appraisal process: An expert system approach", *Cognition and Emotion*, vol. 7, pp. 325–355, 1993.

[9] Scherer K. R., "Emotion effects on voice and speech: Paradigms and approaches to evaluation", in *ISCA Workshop on Speech and Emotion, Belfast, Northern Ireland*, 2000.

[10] Scherer K. R., "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.

[11] Johnstone T., Banse R., and Scherer K. R., "Acoustic profiles in prototypical vocal expressions of emotion", in *Proceedings of the XIIIth International Congress of Phonetic Sciences, ICPhS95, Stockholm, Sweden*, 1995, pp. 2–5.

[12] Johnstone T. and Scherer K. R., "The effects of emotions on voice quality", *Geneva Studie in Emotion and Communication*, vol. 13, pp. 3, 1999.

[13] Zajonc R.B., "Feeling and thinking: Preferences need no inferences", *American Psychologist*, vol. 35, pp. 151–175, 1980.

[14] Van den Noort M.and Bosch M. P. C. and Hugdahl K., "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.

[15] Bechara A., Damasio H., Tranel D., and Damasio A. R., "Deciding advantageously before knowing the advantageous strategy", *Science*, vol. 275, pp. 1293–5, 1997.

[16] Kahneman D. and Tversky A., "Prospect theory: An analysis of decision under risk", *Econometrica*, vol. XVLII, pp. 263–291, 1979.

[17] Brehm S. S. and Kassin S. M., *Social Pschology*, Houghton Mifflin Company, Genea, Illinois, 1996.

[18] Baron-Cohen S., *The descent of mind: psychological perspectives on hominid evolution*, chapter Evolution of a theory of mind?, Oxford University Press, 1999.

[19] Premack D. and Woodruff G., "Does the chimpanzee have a 'theory of mind'?", *Behaviour and Brain Sciences*, vol. 4, pp. 515–526, 1978.

[20] Whiten A., *Natural theories of mind*, Oxford: Basil Blackwell, 1991.

[21] Höök K., *Evaluating ECAs*, chapter User-centred design and evaluation of affective interfaces, Kluwer.

[22] Baron-Cohen S., Riviere A., Fukushima M., French D., Hadwin J., Cross P., Bryant C., and Sotillo M., "Reading the mind in the face: A crosscultural and developmental study", *Visual Cognition*, vol. 3, pp. 39–59, 1996.

[23] Plutchik R., "The nature of emotions", *American Scientist online*, vol. 89, no. 4, pp. 344, 2001, http://www.americanscientist.org/articles/01articles/Plutchik.html.

[24] Schrödr M., "Speech and emotion research: an overview of research frameworks and a dimensional approach to emotional speech synthesis (ph.d thesis). vol. 7 of phonus, research report of the institute of phonetics", Tech. Rep., Saarland University, 2004, http://www.dfki.de/∼schroe.

[25] Scherer K.R., *Approaches to emotion*, chapter On the nature and function of emotion: a component process approach, pp. 293–317, Hillsdale, NJ: Erlbaum, 1984.

[26] Whissell C. M., *Emotion: Theory, Research, and Experience*, chapter The dictionary of affect in language, pp. 113–131, New York: Academic Press, 1989.

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

[28] "http://www2.dfki.de/ schroed/feeltrace/".

[29] "Humaine deliverable d5f".

[30] Golan O., Baron-Cohen S., and Hill J., "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.

[31] Reeves B. and Nass C., *The media equation*, Cambridge University press, 1996.

[32] Nass C. and Brave S., *Wired for Speech: How voice activates and advances the human-Computer relationship*, The MIT Press, 2005.

[33] Picard W. Rosalind, "Affective computing challenges", *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 55–64, 2003.

[34] Hudlicka Eva, "To feel or not to feel: The role of affect in human-computer interaction", *International Journal of Human-Computer Studies*, vol. 59, no. 1-2, pp. 1–32, 2003.

[35] Dragoni A. F. and Puliti P., "Mental states recognition from speech acts through abduction", in *ECAI 94. 11th European Conference on Artificial Intelligence. Proceedings. Wiley*, Chichester, UK, 1994, pp. 183–187.

[36] Kikuchi H. and Shirai K., "Mechanism of generating meta-utterance with predicting changes of mental states in spoken dialogue", *Transactions of the Information Processing Society of Japan*, vol. 43, no. 7, pp. 2130–7, 2002.

[37] Picard R. W., Vyzas E., and Healey J., "Toward machine emotional intelligence: analysis of affective physiological state", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 10, pp. 1175–91, 2001.

[38] Cassell J., "Embodied conversational interface agents", *Communications of the ACM*, vol. 43, no. 4, pp. 70–8, 2000.

[39] Cassell J. and Vilhjalmsson H., “Fully embodied conversational avatars: making communicative behaviors autonomous”, *Autonomous Agents and Multi-Agent Systems*, vol. 2, no. 1, pp. 45–64, 1999.

[40] Isbister K., Nakanishi H., Ishida T., and Nass C., “Helper agent: designing an assistant for human-human interaction in a virtual meeting space”, in *CHI 2000 Conference Proceedings. Conference on Human Factors in Computing Systems. ACM*, New York, NY, USA., 2000, pp. 57–64.

[41] Tosa N. and Nakatsu R., “Life-like communication agents - emotion sensing character ”mic” and feeling session character ”muse””, in *IEEE Conference on Multimedia*, 1996, pp. 12–19.

[42] Healey J. and Picard R. W., “Digital processing of affective signals”, in *IEEE Conference on Multimedia*, ICASSP 1998, pp. 251–252.

[43] Fernandez R., PhD thesis, Media Arts and Sciences, Massachusetss Institute of Technology, 2004.

[44] Breazeal C. and Aryananda L., “Recognizing affective intent in robot directed speech”, *Autonomous Robots*, vol. 12, pp. 32–80, 2002.

[45] Dellaert F., Polzin Th., and Waibel A., “Recognizing emotions in speech”, in *ICSLP 96*, 1996.

[46] Yacoob Y. and Davis L. S., “Recognizing human facial expressions”, in *Image Understanding Workshop.*, San Francisco, CA, USA, 1994, pp. 827–35.

[47] Lisetti C.L. and Schiano D.J., “Automatic facial expression interpretation: Where human computer interaction, artificial intelligence and cognitive sciences intersect”, *Pragmatics and cognition*, vol. 8, no. 1, pp. 185–235, 2000.

[48] Yan Li F. Y., Ying-Qing X., Chang E., and Heung-Yeung S., “Speech driven cartoon animation with emotions”, in *ACM Multimedia 2001*, Ottawa, Canada, 2001.

[49] Fernandez R. and Picard R. W., “Modeling drivers’ speech under stress”, *Speech Communication*, vol. 40, pp. 145–59, 2003.

[50] Zhou G., Hansen J. H. L., and Kaiser J. F., “Classification of speech under stress based on features derived from the nonlinear teager energy operator”, in *IEEE ICASSP-98: Inter. Conf. on Acoustics, Speech, and Signal Processing*, Seattle, Wash., May 1998, vol. 1, pp. 549–552.

[51] Katz G. S., Cohn J. F., and Moore C. A., “A combination of vocal f0 dynamic and summary features discriminates between three pragmatic categories of infant-directed speech”, *Child Development*, vol. 67, pp. 205–217, 1996.

[52] “Nemesysco, http://www.nemesysco.com/, sept 2006.”.

[53] Klasmeyer G., “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.

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

[55] Douglas-Cowie E., Campbell N., Cowie R., and Roach P., “Emotional speech: towards a new generation of databases”, *Speech Communication*, vol. 40, pp. 33–60, 2003.

[56] Schröder M., “Emotional speech synthesis: A review”, in *Proceedings of Eurospeech 2001*, Aalborg, 2001, pp. 561–564.

[57] Kawahara H. and Matsui H., “Auditory morphing based on an elastic perceptual distance metric, in an interference-free time-frequency representation”, in *ICASSP’2003*, 2003, pp. 256–259.

[58] Piftzinger H. R., “Unsupervised speech morphing between utterances of any speakers”, in *Proceedings of the 10th Australian International Conference on Speech Science & Technology Macquarie University*, Sydney, December 2004, pp. 545–550.

[59] Burkhardt F. and Sendlmeier W. F., “Verification of acoustical correlates of emotional speech using formant-synthesis”, in *ISCA Workshop on Speech & Emotion*, Northern Ireland, 2000, pp. 151–156.

[60] Mozziconacci S. J. L. and D. J. Hermes, “Role of intonation patterns in conveying emotion in speech”, in *ICPhS 1999*, 1999, pp. 2001–2004.

[61] “http://emotion-research.net”.

[62] Fant G.C.M., *Acoustic theory of speech production*, Mouton and Co., 1960.

[63] Flanagan J.L., *Speech analysis synthesis and perception*, Springer-Verlag, 1972.

[64] Markel J.D. and Gray A.H. Jr., *Linear Prediction of Speech*, Springer, Berlin (west), 1976.

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

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

[67] Weiss D. J., Ghazanfar A. A., Miller C. T., and Hauser M. D., "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] Damasio A. R., Tranel D., and Damasio H., "Faces and the neural substrates of memory", *Annual Review of Neuroscience*, vol. 13, pp. 89–109, 1990.

[69] Mozziconacci S. J. L., "Modeling emotion and attitude in speech by means of perceptually based parameter values", *User Modeling and User-Adapted Interaction*, vol. 11, no. 4, pp. 297–326, 2001.

[70] Murray I. R. and Arnott J. L., "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–108, 1993.

[71] De Cheveigne A. and Kawahara H., "Multiple period estimation and pitch perception model", *Speech Communication*, vol. 27, no. 3-4, pp. 175–85, 1999.

[72] Scheirer E.D., *Music-listening systems*, Massachusetts Institute of Technology, Boston, Massachusettes, USA, 2000.

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

[74] Stelmachowicz P. G., Pittman A. L., Hoover B. M., and Lewis D. E., "Effect of stimulus bandwidth on the perception of s in normal- and hearing-impaired children and adults", *Journal of the Acoustical Society of America*, vol. 110, no. 4, pp. 2183–90, 2001.

[75] Yona D., "The effect of different frequency gain response on the perception of high frequency words", Master's thesis, Tel Aviv University, Tel Aviv, 1993.

[76] Apter-Yehezkely G., "Evaluation of speech discrimination in sensory-neural high-tone loss subjects", Master's thesis, Tel Aviv University, Tel Aviv, 1993.

[77] P. Taylor, "Analysis and synthesis of intonation using the tilt model", *Journal of the Acoustical Society of America*, vol. 107, pp. 1697–1714, 2000.

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

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

[80] Iamblichus, *On the Pythagorean life*, Clark G. translator Liverpool, c300/1996.

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

[82] Galileo, *Dialogues Concerning Two New Sciences*, Dover Publications Inc., New-York 1638, 1954.

[83] Schartz D. A., Howe Q. C., and Purves D., "The statistical structure of human speech sounds predicts musical universals", *The Journal of Neuroscience*, vol. 23, pp. 7160–7168, 2003.

[84] Tramo M. J., Cariani P. A., Delgutte B., and Braida L. D., *The cognitive neuroscience of music*, chapter Neurobiology of harmony perception, Oxford University Press, New-York, 2003.

[85] 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.

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

[87] Stephen McAdams and Emmanuel Bigad, Eds., *Thinking in sound*, Oxford University Press, New-York, 1993.

[88] Adolphs R. and Tranel D., "Intact recognition of emotional prosody following amygdale damage", *Neuropsychologia*, vol. 37, pp. 1285–1292, 1999.

Chapter 3

Automated vocal inference systems

The previous chapter discussed why automated paralinguistic systems could improve computer interfaces and the most common theories behind the current systems. It presented some of the complexity involved in their development due to the dynamic nature of expressions, and the asynchronous nature of mixture of expressions and expression nuances. In addition, it introduced the multiple related research fields that add to the complexity by supplying a large variety of features and methods whose relevance to the nature of such systems is not always clear or has not been fully tested yet.

One of the most fundamental questions in the area of recognising and synthesising expressions in the non-verbal aspects of speech is universality: will we ever be able to realise a single system suitable for every user, environment, scenario and expression? This chapter suggests a set of requirements from systems for automatic inference and synthesis of expressions from speech. It describes the challenges involved in their development and the merits of incorporating dynamic analysis of expressions in time in such systems.

This chapter presents a general multi-stage framework for automatic paralinguistic systems. It then places the framework for the research which is described here within the wider context of the general framework. Shortly summarising how each level of the general framework has been interpreted and implemented for the specific requirements of this work.

3.1 Requirements

The requirements from general automated paralinguistic systems, both inference and synthesis, can be summarised in two points: support natural behaviour, and be as general as possible. Both the complexity of the development process, as well as the number of possible applications and their financial potential require these features. This section describes these requirements.

User independent

A system should be easily adapted to recognise the expressions of most people, without an extensive training session. If there is a training session it should be done automatically. An affective synthesis system should be able to generate the desired expressions for every voice.

Range of expression

An ideal system would recognise and trace all the expressions that are required for the specific application. The recognisable expressions should include most of the spectrum of expressions.

Natural expressions

The goal is to recognise (and to synthesise) natural expressions and acted expressions. To allow recognition of *mixtures* of *asynchronous* subtle expressions and nuances of expressions of moods, attitudes, mental states, dialogue acts, etc.

Fully automated

The system should not require any manual intervention.

Real-time

It is desirable that an interface will respond to the user without a noticeable delay.

Text and context independent

The system should be able to recognise and/or to synthesise most expressions, no matter what the length of the utterance, or the uttered text.

Flexibility and generality

It is desirable that the development of one recognition or synthesis system will facilitate the development of other systems for different applications, expression ranges, speakers or languages (within limitations). Ideally, a system would be expandable and general. A system would be able to incorporate data that was trained using different databases, and be used to recognise expressions from new sources that it had not been trained on.

Multi-modal systems

Multi-modal analysis, which includes cues for context and various behavioural patterns and communication cues, is not a requirement, but it can enhance the accuracy of both inference and synthesis.

3.2 Challenges

The challenges involved in the development of systems that answer these requirements are derived from both the complex nature of human expressions and the ability or lack of ability to classify and label them precisely, and from the technological challenges involved. The technical part is highly dependent on the more abstract issues of definitions (which expressions or concepts does a label encompass) and labelling of expressions (give names to recorded expressions). For comparison, in the field of speech recognition, there are layers of information with absolute labels of words and meanings. However in the area of expressive speech recognition these higher levels do not exist, at least not in such an absolute and obvious form. Some of the main challenges which are inherent to the expressions domain are listed here. The more technical oriented challenges are briefly discussed in Section 3.3 that presents the general framework and in more detail in the next chapters.

Intricate information

Human expressions have various roles in human communication. There are unconscious expressions and there are conscious and intentional expressions. Human expressions reveal mental states, emotions and attitudes. They also reveal intentions, speech acts (greeting, apologising, describing, asking and the like) [1] and dialogue acts (such as statement, acknowledgement, question and answer) [2]. In addition, human expressions are affected by physiological states like discomfort, and environmental contexts, starting from background noise that affects the loudness or intensity of speech to social contexts that may affect the display rules. All this information is carried by the same basic cues that are also shared by the verbal speech.

No 'neutral' expression

Another implication is that there is *no 'neutral' expression.* Because of the assumption of prevailing moods and mental states, there is no real definition of neutral. Therefore, there could be no requirement to initiate a system using a certain 'neutral' expression according to which all the other expressions would be calibrated. The definition of 'neutral' as small variations of the same expressions may still exist, as well as *neutral* as opposed to *opinionated.*

Lexical definitions

Lexical definitions of emotions, mental states, speech acts and communication cues depend on language and cultural background. The difference in language and culture can mean that certain labels of expressions cannot be analysed or translated from one language to another, because the vocabulary has no parallel notions, so an expression may be interpreted in a slightly different manner [3].

Many subtle nuances of daily expressions have no name. Analysing and labelling mixtures of expressions and their underlying meanings are even more complicated. People recognise many of the underlying expressions, but often find them very difficult to name or label [4, 5]. Even a small subset of definitions, for example *anger*, which is considered one of the basic emotions [6], or the more elaborate set of *cold anger* and *warm anger*, cover a whole variety of nuances, among which we can distinguish by context, vocal cues and sometimes also lexically. For example, being angry because of a new tax, sounds different from frustration when something does not work properly or from being angry with a child, or with an adult, which are different in context and vocal cues. Furthermore, there are concepts that lexically belong to the *angry* group and can be distinguished by all three forms (lexical, vocal, context) such as *annoyed* and *furious*, but are often analysed together.

Lack of uniform or repetitive behaviour

We expect different reactions from different people and even from the same speaker when the context of the event that caused the reaction changes. Things as mundane as different weather conditions can change the underlying mood and therefore the subsequent reactions.

The probability of a person repeating precisely the same expression, with the same text and with the same underlying meaning or meanings, is very low. Different people respond in different manners to the same scenarios, and have different ways of expressing themselves according to their personality and background.

Context

Every utterance has a certain context, and analysis of human expressions is also related to context. The context may be the interaction, the audience or environmental parameters. These parameters influence the way the utterance is expressed, and the way it should be analysed. The interaction point of view influences the design in two major ways, the first is the time domain, and the second is the context domain

A complication may occur when the *target of the speech*, either *audience* or *medium* changes. For example, in the middle of a computer interaction that includes commands and dictated text the speaker may also think aloud. These intervals can take the form of local calculations, trying to remember, frustration, or trying to improve understanding by reading aloud. They are not necessarily related to the analysed interaction.

During a sustained interaction the speaker may speak to other people around him, in the immediate vicinity, in another room, or through various devices, such as the telephone, both as part of the interaction or in parallel to it. We may like to detect these target changes as well. Noise can be considered as a context dependent parameter.

Dynamic and asynchronous changes in time

The multi-layer scheme which is presented in Figure 2-1 describes the different time characteristics of human expressions in relation to the course of an interaction. *Mixtures of expressions* occur all the time. The occurrences and durations of these expressions may *change asynchronously* during an interaction and carry their own impact.

Most people do not speak continuously and endlessly, but rather in time segments whose duration and timing are set according to the personality and the nature of the interaction. An analysis of a continuous interaction should integrate the analysis of the current speech segment with the analysis of the previous transitions among speech segments or utterances.

On the other hand, certain expressions can be analysed in a 'stand-alone' manner. Furthermore, single syllable utterances, such as different manners of saying 'no', can be recognised without further knowledge of the context or longer-duration speech cues. In addition, some utterances can change the course of an interaction. For example, the expression of *surprise* can be inferred in many cases without special regard to the dynamics of the whole interaction.

Different *speech* and *dialogue* acts may follow or emphasise different expressive patterns, for example, in dialogues with salespersons questions may reveal a range from misunderstanding and puzzlement to developed understanding and interest, while answers or statements may reveal reactions to the other speakers' input, such as agreement or amazement. On the other hand both can reveal boredom, annoyance or interest.

Dynamic tracking of changes is significant for analysis of tendencies, such as escalations of situations, and emotions. An ideal system would track both long term states and gradual changes in addition to sudden changes. The manner of response to these changes depends on the specific application.

Ambiguity resolution

Another motivation for dynamic tracing is to *resolve ambiguities.* As a rule, people tend to be more sensitive to subtle changes than to levels. People can recognise a large variety of acted expressions which are mostly stereotypical and exaggerated expressions. People also recognise many natural expressions and when they cannot define the precise label, or pinpoint a single underlying state (mental, emotional, physical, etc.) they can usually assume one or several possible labels, simultaneously. For example, for subdued expressions the perceived meaning can be tired, bored, and/or sad, and it can also be interpreted as related to an introvert personality. Additional cues during sustained interactions can help to adjust and to refine the initial estimations.

Dynamic changes within an utterance

Another kind of dynamic behaviour can be found *within an utterance.* This level is related more to the technical side of the inference system. It includes various intonation and energy patterns. Short-term variations within an utterance can change our perception of the uttered expression in a way that statistics over the whole utterance do not reveal.

Implications for design

For all these reasons, the required solutions for both inference and continuous synthesis should enable both the estimation of single utterance expressions and an analysis at the interaction level. The proposed framework should be flexible enough to respond to local changes and short utterances, and also to long-term attitudes and moods and their transitions, without compromising short-term events and abrupt changes. The inference should comprise mixtures of expressions. The inference system should be flexible enough to encompass a large range of applications with little overhead of training for each specific set of speakers and situations.

A hierarchical model for dynamic analysis of expressions in non-verbal speech is a good candidate for the implementation of such a solution. Its different layers support the analysis of different time units within an interaction, and follow transitions among utterances. Furthermore, a hierarchical structure allows flexible development and adjustment of the different layers separately, and the use of sub-systems in different combinations for different applications. Furthermore, depending of course on the implementation, a general system that was developed for a certain dataset or several datasets could be adjusted to other datasets that include other expressions with similar dynamic characteristics, both at the inter-utterance level, and the interaction level, although the specific inference may be different among speakers, expressions and contexts.

In addition, the interaction level should include the analysis of the connectivity among expressions of related lexical labels, such as different types of anger. The mapping of parameters that influence lexical labels and the differences and transitions among them, such as different activity levels, and their correlation to the expressions, may add to the ability of the system to predict and analyse transitions. This multi-faceted analysis is part of a wider solution at the interaction level. Synthesis systems that incorporate such connectivity schemes would allow their users greater flexibility and control. Table 3-1 summarises the requirements, challenges, implications for design and the basic solution.

Requirements
• User independent • Large range of expressions: natural and acted, mixtures, asynchronous, extreme, subtle and nuance • Fully automated • Real-time • Text and context independent • Flexible and general: language, expression range, speakers, applications • Multi-modal
Challenges
• Intricate information: expression mixtures, no neutral expression, different expressions have different time spans • Dynamic and asynchronous changes: long-term tendencies, meaningful stand-alone utterances, different speech and dialogue acts may follow different dynamics, dynamic changes within an utterance • Lexical definitions: depend on language and cultural background, different concepts under one label, not every nuance is labelled • Lack of uniform or repetitive behaviour • Expressions and their interpretations are context dependent • Changes in context may occur during an interaction - target, medium, audience
Implications
• Utterance level analysis • Interaction level analysis • Track long and short term tendencies • Detect sudden and temporary changes • Identify significant states, changes and tendencies
Solution
• Hierarchical model that detect different dynamic characteristics • Different applications would have to change only specific layers of the model • Expandable to other expressions • Can use various datasets for expanding to different applications and expressions • Explore connectivity and relations among expressions and labels for better transition estimations

Table 3-1: Summary of the requirements, challenges and the derived implications for design for automatic inference systems of expressions from vocal cues.

3.3 General framework

The model presented here is designed for continuous analysis and inference of expressions in non-verbal speech and, with slight adaptations, to synthesis. The model consists of five layers of processing, depicted as a block diagram in Figure 3-1. For each level, there is a short description of the requirements in the design stage and the expected analysis process in the final system.

Each of the layers contributes to the whole solution, and to the compatibility of the system with a hierarchical model of the expressions. At the design stage each layer poses its own challenges, in addition to the challenges posed by their integration into different applications.

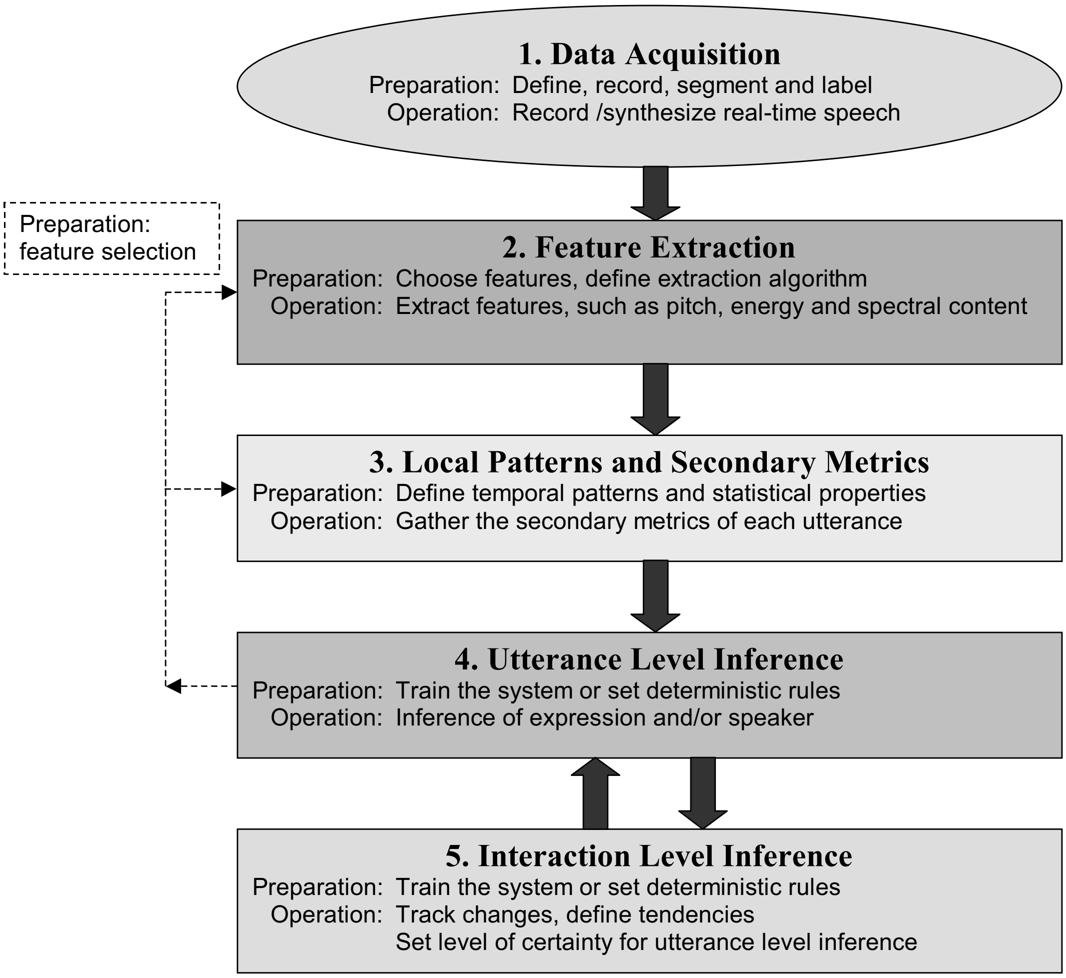
1. Data Acquisition
Preparation: Define, record, segment and label
Operation: Record /synthesize real-time speech
Preparation: feature selection
2. Feature Extraction
Preparation: Choose features, define extraction algorithm
Operation: Extract features, such as pitch, energy and spectral content
3. Local Patterns and Secondary Metrics
Preparation: Define temporal patterns and statistical properties
Operation: Gather the secondary metrics of each utterance
4. Utterance Level Inference
Preparation: Train the system or set deterministic rules
Operation: Inference of expression and/or speaker
5. Interaction Level Inference
Preparation: Train the system or set deterministic rules
Operation: Track changes, define tendencies
Set level of certainty for utterance level inference

Figure 3-1: A hierarchical framework for dynamic inference. "Preparation" refers to the requirements for building the inference machine. "Operation" refers to the inference processing of speech signals in the working system.

This framework is hierarchical in nature, from audio stream and segmentation through extraction of vocal features to inference of expressions. It also supports multiple levels of temporal inference, including vocal features with various temporal characteristics, utterance level inference and on to interaction level analysis. The asynchronous analysis cues from the different levels can be integrated in order to enhance the inference.

There are several reasons for using a hierarchical framework. First, analysis of the entire speech signal without any pre-processing would have been preferable, but is not practical. The sampled speech signals form very long arrays. Training a machine to recognise similarities and differences among such long arrays is not efficient. Furthermore, such inference could not reveal understandable and transferable information that may support further analysis of other datasets and might limit the application of the findings to synthesis.

The interpretation of each stage can be flexible and adjusted to the available data, and to the requirements of the application. This framework allows generalisation of solutions by integrating similar parts from systems that were built using different datasets and features. The framework is designed for inference systems, but can be easily translated to synthesis.

Stage 1: Data Acquisition

The first stage is the data acquisition layer. This layer sets the limitations for the rest of the processing. The framework addresses the design considerations for the acquisition of the expressive datasets, which relate to the dynamical nature of interactions. These datasets are different from datasets that include stand-alone utterances, or datasets of several extreme expressions or of a single mental state. The design of the dataset is crucial to the performance of the inference system. The acquisition of expressive data is not trivial and there are only a few reference datasets for comparison. Training and building a system that can be adapted to various scenarios, such as real-time inference and natural synthesis, has to use datasets that supply a wide range of observations for training. The design, recording and labelling of such datasets is a major design challenge. The goal is to create a system that can work on recorded speech in real-life and real-time situations, or generate natural-sounding synthesised speech.

Stage 2: Feature definition and extraction

The second stage is the extraction of basic physical metrics or features from the speech signal, such as the fundamental frequency (pitch, intonation), energy or intensity of the signal, and spectral content. This layer is common to most current inference systems. However, a good enough set of parameters that should be extracted in order to achieve a better coverage of expression nuances is yet to be defined. Furthermore, there are many definitions and algorithms for the extraction of some of the main features, especially for the extraction of the fundamental frequency and speech rate, and the number indicates the complexity of the problem, as most of the algorithms are not robust enough for all speakers and recording platforms. For example, features that describe voice qualities such as dull or warm still lack definition.

Stage 3: Local patterns and secondary metrics

The third layer processes local patterns and statistical characteristics of a whole utterance. These local patterns include parameters such as fundamental frequency (f_0) contours and energy patterns along an utterance. In addition, it includes temporal characteristics such as the durations of speech parts such as silence, voiced and unvoiced speech and speech rate. This level of analysis has not been thoroughly investigated yet and few systems currently use local patterns within utterances. The building stage of a system requires the definition of the statistical and temporal metrics. It may also require the definition of a set of patterns that are significant for the inference or synthesis of expressions, in addition to building a mechanism that can recognise or generate these patterns.

Stage 4: Utterance level inference

The fourth layer comprises inference at the utterance level, where local patterns, together with statistical metrics and inputs from the analysis of previous utterances, contribute to the analysis or to the synthesis of the current utterance. The inference output may include several observations in parallel, as we expect mixtures of expressions, and different layers of expressions, each of them with different temporal characteristics and with different probability or confidence levels, as described in the previous section. This approach is different from most of the currently published work in this area, where either one mental state is recognised, or one of a few extreme emotions, for each utterance by itself.

Stage 5: Interaction and expressions' mapping layer

The fifth layer is the interaction layer. This takes into account the connectivity among expressions, including gradual transitions between expressions and between utterances. This layer is the one that recognises the dynamic characteristics of the expressions during an interaction. In this layer, expression mapping, the analysis of the relations between expressions with different lexical definitions, can be integrated. The mapping can contribute to the prediction and verification of an inference. It can also define the gradual transitions among expressions for synthesis.

The implementation of this layer depends significantly on the requirements, goals and setups of the different applications because the scope of a general solution is very large whereas the previous layers are more general and can be easily used for various applications. In a multi-modal solution an integration of cues from different modalities can be performed at this stage or in the utterance layer [7].

This layer incorporates the analysis of the relation between expressions. Therefore, expression mapping is a part of this layer definition. It involves the application of the inference machine to the description or inference of a wide set of mental states and to analysis of the relations between their meanings and their related expressions.

3.4 The implementation framework

The previous section described a general framework for dynamic analysis of expressions. This section describes how this framework has been implemented in this research. It presents summaries of the implementation of each of the framework's layers. A schematic description of the areas in which this work is innovative appears in Figure 3-2. Detailed descriptions are given in the next chapters, in which the different topics are organised by their order in the framework.

Databases

Two datasets have been used in this research. The Doors database was defined and recorded as part of this work, and the MindReading database was used for training and for expression mapping. Table 3-2 summarises their properties. The MindReading database provides a very large number of labelled mental states, unlike most work in the field that explore only few expressions, and mostly 'basic emotions'. In this research, an extended version of the MindReading database which includes 756 concepts was used. The Doors database provides naturally evoked expressions and expression nuances, during a human-computer interaction situation, with multi-modal cues. Chapter 4 describes these datasets in more detail.

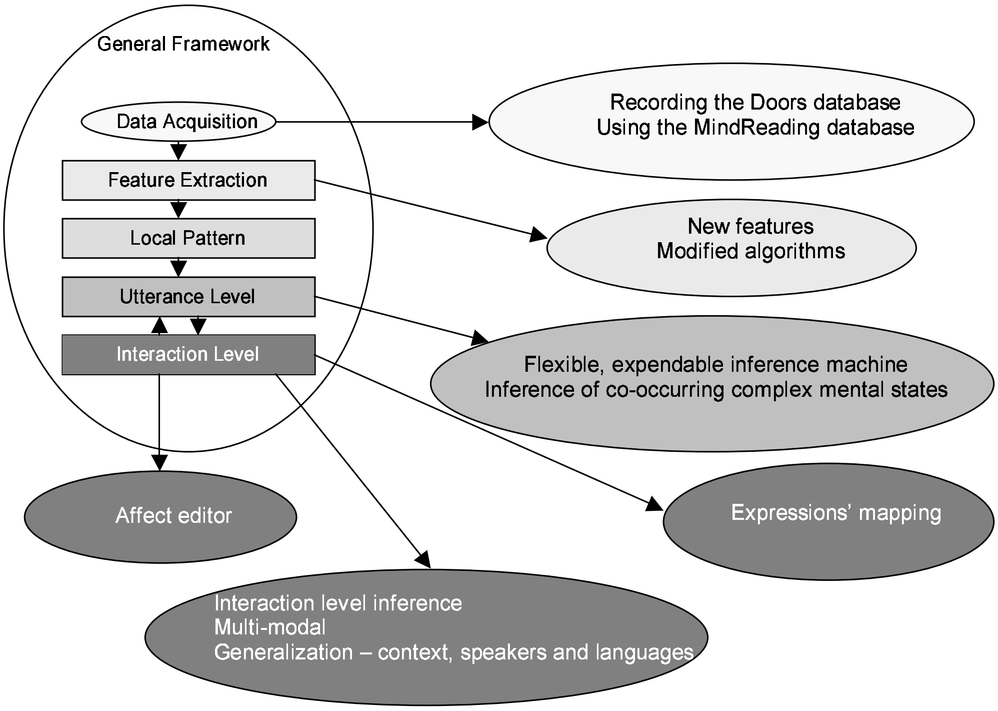
General Framework
Data Acquisition
Feature Extraction
Local Pattern
Utterance Level
Interaction Level
Recording the Doors database
Using the MindReading database
New features
Modified algorithms
Flexible, expendable inference machine
Inference of co-occurring complex mental states
Affect editor
Expressions' mapping
Interaction level inference
Multi-modal
Generalization – context, speakers and languages

Figure 3-2: This work and the general framework, the colour code is from light to dark according to processing stage. The general framework is described on the left. The contribution of this work which is relevant to each layer appears on the right in a similar colour and with an arrow to the equivalent layer in the general framework.

Features

The main features that are derived directly from the speech signal in this work are pitch (the fundamental frequency), energy and the spectral content or the energy of the signal in different frequency bands. Vocal features that have not been used before for expressive speech analysis were also used in this work, for example features that draw on similarities to musical properties such as harmonic properties, dissonance and consonance. All the mathematical definitions and extraction algorithms have been reviewed, and changes to existing algorithms have been introduced, for example in the pitch extraction algorithm.

Local and secondary metrics

The next stage is to derive secondary metrics that describe how the main vocal features, which are extracted directly from the speech signal, change over time and from one expression to another. Some of these secondary metrics are statistic measures of a feature value or duration throughout the utterance, such as mean, range, median, standard deviation, minimum and maximum values, or the number of occurrences.

	Doors	MindReading
Language	Hebrew	English
Text	2 Sentences, 100 repetitions each + unrestricted text + laughter, breathing, etc Neutral (task related) content	Different text for each sentence Mostly neutral content
Speakers	15 speakers age 25-55	12 speakers different age groups, including children
Natural/acted	Natural	Acted
Context	HCI interaction ~15 minutes for each person	Unrelated sentences
Labels	Unlabelled (partial manual labelling, and in correlation to other modalities)	Labelled: 6 sentences of each of over 756 mental state concepts, organised in 24 groups (the MindReading taxonomy)
Modalities	Multi-modal: voice, facial expressions, events physiological cues (BVP, GSR, ECG), mouse movements, reaction time	Voice only (includes also unrelated recordings of facial expressions and gestures)

Table 3-2: Properties of the two datasets used in this work, the MindReading and the Doors.

Another type of secondary metrics used here is related to behavioural patterns of the features within the utterance, and the statistical properties of these patterns or speech parts. A rule-based parsing algorithm was defined. It divides the utterance into speech parts that are linguistically significant, such as energy peaks with no pitch (consonants, breathing, etc), energy peaks with pitch (vowels or full syllables) and pauses or silence. The durations, the peak values, the distances and the value relations were considered, as well as the speech rate. Many of these metrics can be paralleled to musical properties such tempo and the relations between tones in a melody.

A further examination of local patterns in the form of an alphabet of fundamental frequency contours was undertaken by Laura Crockett for her final-year undergraduate dissertation under my supervision [8].

The vocal features, their extraction algorithms and the secondary metrics are described in detail in Chapter 5.

Utterance level inference

An inference system was built for emotions and complex mental states. It allows recognition of expression mixtures that change asynchronously. The machine was built to recognise an arbitrary set of nine expressions, including: *joyful*, *absorbed*, *thinking*, *interested*, *sure*, *unsure*, *opposed*, *excited* and *stressed*.

The inference system is built from pair-wise machines. Each machine distinguishes between two expressions from this set. The machines were built using either a tree decision algorithm (C4.5) or support vector machine (SVM). These algorithms were chosen because they yielded the best results compared to other classification algorithms and their implementation is simple. Each machine and each expression pair use a different set of features and metrics for the classification. There is no one set of features or metrics which is suitable for all the expressions. The training of each machine was based on finding a combination of classification algorithm and feature set that optimises the classification (inference) results.

A decision or voting mechanism then chooses the dominant expressions as revealed by the combined results of the pair-wise machines. Several expressions can be dominant at the same time, allowing for expression mixtures. The inference system was validated by 10% cross-validation and by 70%-30% split, over 546 sentences. Its accuracy is these tests was 81% (chance is 11%).

The system can be expanded to include more expressions by adding pair-wise machines. Additional pair-wise machines can be based on various classification methods and use different features and different datasets. The inference system is described in Chapter 6. Implementation of the system to various applications and further verification are described in the next section, in Chapter 7 and Chapter 8.

Interaction level inference, mapping and applications

Three different tests have been done for the last level of the framework, which includes the interaction level inference, expressions' mapping and applications.

The first test compares the ability of the machine to distinguish complex mental states to human performance.

The second is automatic expression mapping. The mapping is of concepts according to the vocal correlates of a small set of expressions (concepts). I describe here the mapping of 459 concepts of the MindReading taxonomy and database. The mapping was done according to the expression set which is recognisable by the inference machine. Two methods of mapping are presented the first combines the MindReading taxonomy and the taxonomy which is defined by the expressions that are recognisable by the inference system, and the second is only according to the recognisable expressions.

The third part includes tracking vocal expressions in time and comparing them to events and to other behavioural cues. The analysis of the Doors database provides a multi-modal perspective by comparing the recognised expressions to events and user decisions during the interaction, and by comparison of vocal expressions to other forms of behavioural expressions, such as delays and mouse movements.

This experiment demonstrates the flexibility and generality of the system. It shows an application of a system that was trained on a one dataset (MindReading) for the analysis of another dataset (Doors). These datasets are in different languages (Hebrew and English), one acted the other naturally evoked, one includes single utterances the other provides dynamic tracking in time, during an HCI interaction, of mixtures of asynchronous expressions and nuances of expressions.

The Affect Editor is an example of an synthesis related application. It is describe in a paper [9], but is beyond the scope of this dissertation.

3.5 Summary

This chapter explored the requirements and challenges involved in the development of automated paralinguistic systems that give global solutions. It concludes that a global solution is not necessarily plausible or feasible, and not particularly necessary. However, the inclusion of dynamic analysis can help to enhance the flexibility of partial solutions. Flexible and expandable solutions can enhance future broader systems.

A general framework for the development and structure of such systems was introduced, and the system and solutions which were developed in this research have been described along the guidelines of this framework and its possible applications. Detailed descriptions of all the stages of the research are given in the next chapters, including relevant references.

References

[1] Jhon Searle, *Speech Acts*, Cambridge University Press, 1969.

[2] Grosz B. and Sinder C., "Attenetion, intentions, and the structure of disclousre", *Computational Linguistics*, vol. 19, no. 3, pp. 175–204, 1986.

[3] Wiezrbicka A., *Emotion and Culture: Empirical Studies of Mutual Influence*, chapter Emotion, language and cultural scripts, American Psychological Association, Washington, 1994.

[4] Campbell N., "Perception of affect in speech - towards an automatic processing of paraligustic information in spoken conversation", in *ICSLP 2004*, 2004.

[5] Douglas-Cowie E., Campbell N., Cowie R., and Roach P., "Emotional speech: towards a new generation of databases", *Speech Communication*, vol. 40, pp. 33–60, 2003.

[6] Ekman P., *Handbook of cognition and emotion*, chapter Basic emotion, Wiley, Chihester, UK, 1999.

[7] Sobol Shikler T., El-Kaliouby R., and Robinson P., "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.

[8] Crockett L., "Alphabet for intonation in speech", Tech. Rep., Computer Science Tripos, Part II, T. Sobol-Shikler (supervisor), Computer Laboratory, University of Cambridge, 2005.

[9] Sobol Shikler T. and Robinson P., "Affect editing in speech", in *proceedings of the 1st International Conference on Affective Computing and intelligent Interaction (ACII), Beijing, China*, 2005.

Chapter 4

Data Acquisition

Ideally an inference machine should be able to infer an expression from any given utterance, or to analyse the expressions during the course of an interaction. A synthesiser should be able to generate a required expression or expressions. In order to build such machines, suitable training data is required.

A major challenge in this research area is the lack of conventional, public databases of naturally evoked, labelled expressions, both for single mode and for multi-modal analysis. This shortage requires each group or researcher to construct a new database and therefore findings cannot be easily translated from one project to another, and the performance of different systems cannot readily be compared. A more comprehensive review of this issue can be found in a paper by Cowie *et al.* [1]. In this paper, the writers identify four issues that should be considered in the development of emotional databases including scope, how many expressions and speakers; naturalness, acted or naturally evoked; context, such as text and situation and descriptors, the types of existing annotations for the database, such as expression labels and phonological descriptors. The development of databases to enhance the dynamic analysis of expressions should consider the same criteria.

Eliciting emotions

Many projects are based on staged expressions or on read paragraphs, using actors [2]. Another approach is to collect emotional episodes from films [3,4]. Several speech databases include nonsense speech, with the aim of eliminating the effect of text. Most of these databases focus on Ekman's basic emotions [5] , or on other small sets of extreme emotions.

The problem is that extreme emotions are rare in everyday life, whereas nuances are common. Everyday expressions may include a mixture of intentions, mental states and emotions. In addition, staged expressions are different from real expressions; an example is the difference between a facial expression of smile and the label of happiness.

Several approaches for eliciting natural or natural-like emotions have been developed. The method that obtains data which is most natural is to use recordings of people in real situations, for example during telephone conversations, or of pilots during flight. The CREST database [6] is a major effort of this type and includes recordings of people in their natural environment over a period of five years. A different method is to use photographs, film episodes or music that elicit certain emotions, and record people while they are watching or hearing them [7]. This method is used mainly for the recording of facial expressions. Another method is to record people who perform a given task, for example a frustrating computer game [8], solving mathematical problems while driving [9] (for stress investigation) or asking young mothers to perform a certain task with their babies [10]. This method provides the researchers with more control on the content and the setting, although most of the databases include only one modality, elicit a small variety of expressions, consider time-discrete events, and are proprietary.

Labelling

A common problem to all these methods is the association of names, labels, or descriptors with the recorded expressions. Often many sub expressions may be defined under one definition of an expression [11], for example different types of anger. Mixtures of expressions and emotions also pose a problem for labelling. Cowie and Cornelius mention most of the descriptors of expressions [12], and also the tool FEELTRACE which allows users to label dynamic transitions among expressions. More elaborate systems, multi-modal analysis and context awareness can help the designers of a database to define additional descriptors. Recently several systems that should facilitate manual labelling of different aspects of multi-modal systems, such as video and speech, have been developed, but automatic systems are not available yet.

Segmentation

Another issue that arises from continuous recording is segmentation. A system should be able to detect human speech, to distinguish between speech and other sounds, and to divide the speech into meaningful and manageable units.

The issue of segmentation is relevant also for the working stage of an inference system which works in real-time. Such a system may receive a stream of data and choose the relevant time segments for analysis. Since many interactions do not occur in a fully controlled environment, there may be a number of complicating factors, such as the presence of noise and sounds, multiple speakers and the like.

Two databases were used in this research, Doors and MindReading. The Doors database consists of natural expressions, evoked during a computer related task. It is built of utterances with identical text to exclude the influence of textual changes from the detection and analysis of expressive features. The MindReading database includes a large variety of acted and labelled expressions, arranged in 24 groups containing 756 concepts [13, 14]. These databases represent two different and complementary types of databases, and therefore represent some of the requirements from databases in general, the challenges in defining, recording and analysing databases, and the advantages and the disadvantages of these types of databases and their implementations. In addition they provide diverse testing material for the inference system, and opportunities for further analysis of the relations between expressions in time and concept. The following section provides a description of these databases, and the algorithm used for the segmentation of the voice segments from the Doors database.

4.1 The Doors database

The Doors database was constructed in collaboration with the Psychology and the Bio-Engineering Departments in Tel Aviv University [1] My aim was to record a multi-modal

[1] Rinat Bar-Lev and Matti Mintz from the Psychology Department used the Doors game to investigate the performance of children with behavioural problems using the computerised IGT, and wrote the

database of naturally evoked expressions in a controlled environment. The goal was to investigate affective non-verbal speech and facial expressions. Additional data was recorded, including physiological cues, mouse movements and interaction related events, in order to support the identification and labelling of expressions, and for multi-modal analysis. The participants group comprised 15 Hebrew speakers, both male and female; the range of ages was 24 to over 50. Figure 4-1 shows a schematic description of the recording system.

The database consists of recordings of people engaged in a human-computer task. The task was a computer game designed to evoke emotions and expressions, based on the Iowa Gambling Test (IGT) [15]. The game involved one hundred repeated events, in which the participant had to choose one of four doors. Each door had a different profit and loss expectation, unknown to the participant, and the participant's aim was to gain as many points as possible.

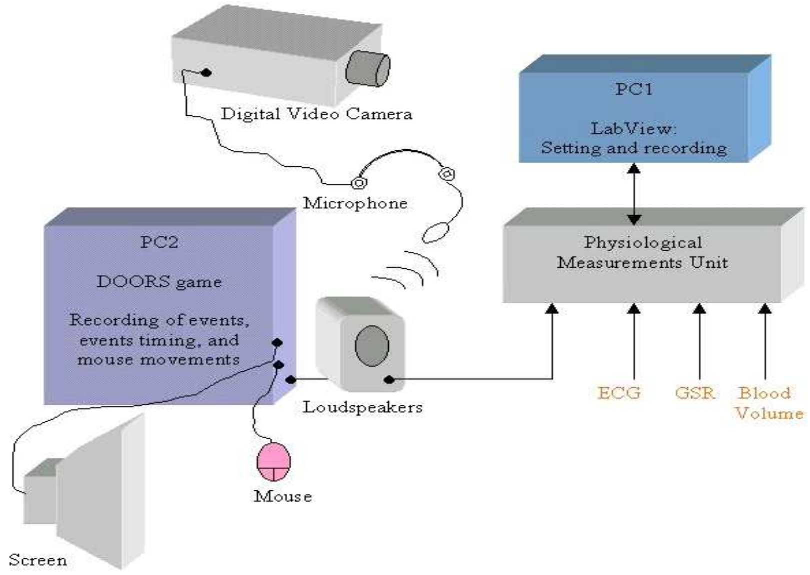

Figure 4-1: Database recording setup: video camera, microphone, a system for measurement of physiological cues, together with a PC running the computer game, which controls the loudspeakers, used for synchronisation.

Figure 4-2 shows four screen shots of the Doors game: Closed doors; an open door with high gain; an open door with low gain; The displayed gain value is set for each door, 50 points for two of the doors and 100 points for the other two. The last figure (bottom right) shows an open door revealing a loss appearing a few seconds after the opening of a door that at first showed a high gain. The loss value changed between trials and doors. The bar at the side shows a schematic estimate of the total gain so far (green) relative

softaware for it. Shay Davidi and Oded Barnea from the Bio-Engineering Department developed the tools for physiological measurements in order to compare BVP (Blood Volume at the Periphery) and ECG data as heart-rate indicators, and included the GSR for the benefit of the other participants. I defined the requirements of the video and audio recording, defined and arranged the settings. I also did the actual recordings with Shai Davidi with help from Vered Aharonson. I am the only participant who needed all the data and its synchronisation. The synchronisation method was defined by all the participants with the help of Noam Amir.

to the possible gain range (4 levels of pink), the white window at the top shows the total gain in points.

The speech consists of two repeated sentences, forming a corpus of 200 sentences for each participant, in addition to free speech sessions. The same sentence was repeated in order to allow extraction of features that are only related to expressions, eliminating differences due to textual content. The participant had to say one sentence when choosing a door 'open this door', phrased in Hebrew as 'open door this'. After seeing what was hidden behind the closed door, the participant had to say 'close door'. After 20 trials, we asked about the chosen strategy, and had a short interval of free speech, without text constraint, which was aimed at the human instructor, and not at the computer. The sampling rate of the recorded speech signals was 32,000Hz, 16 bit, mono.

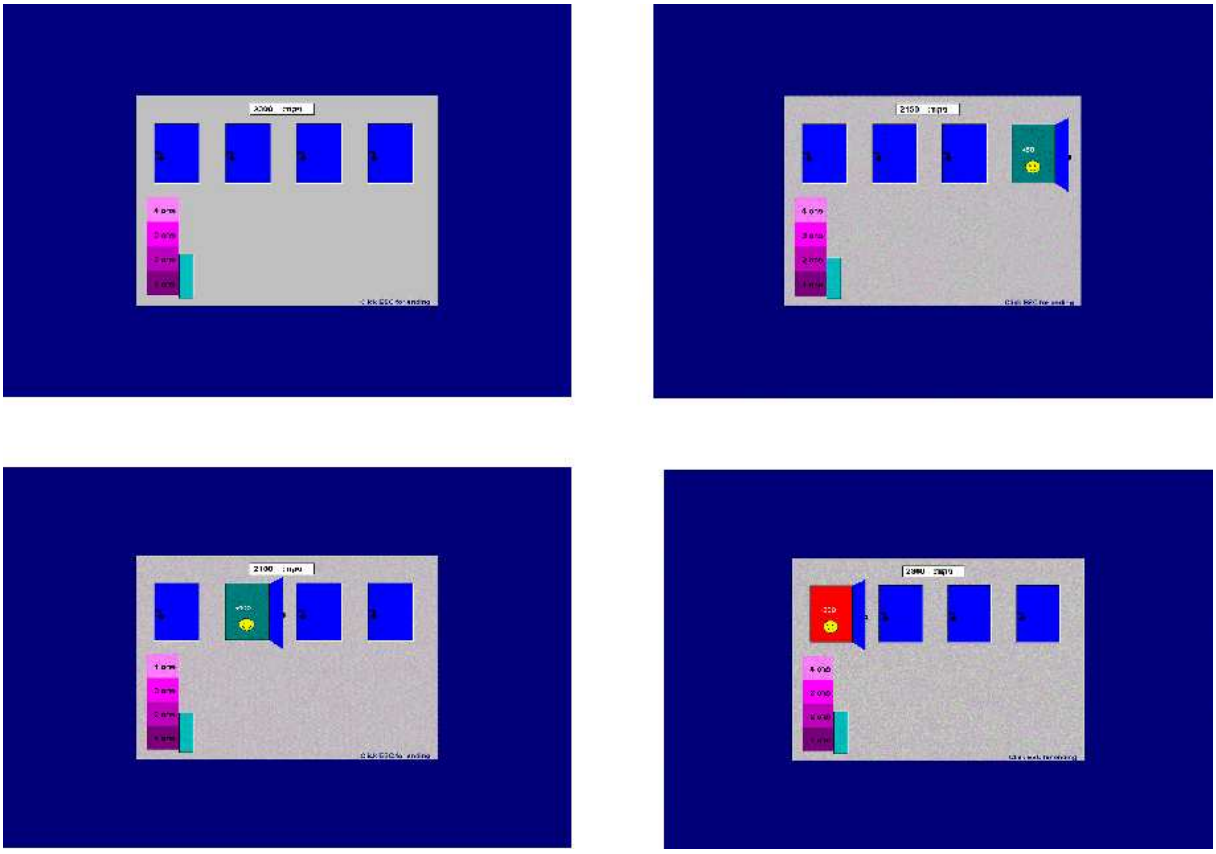

Figure 4-2: Screen shots of the Doors game (from left-upper corner, clockwise): Closed doors, an open door with high gain, open door with low gain. The gain value is set for each door, 2 doors 50 points and 2 doors 100 points. The picture on the lower right corner shows an open door showing loss following a display of high gain. The loss value changes between occurrences and doors. The bar at the side shows a schematic estimate of the gain (green) relative to the possible gain range (4 levels of pink), the white window at the top shows the total gain in points.

Figure 4-3 shows the recording system. The ECG (Electrocardiogram, equivalent to heart rate), GSR (Galvanic Skin Response, the measurement of skin conductivity, or humidity) and blood volume in the periphery were also measured and recorded by a special system built for this experiment. A delay of 3 seconds was set between the decision, i.e. mouse click, and the opening at the doors in order to allow the development of physiological reactions to expectations. The database captured the number and rate of mouse movements, calculated by the game program, and records of the participants' actions and the consequent results.

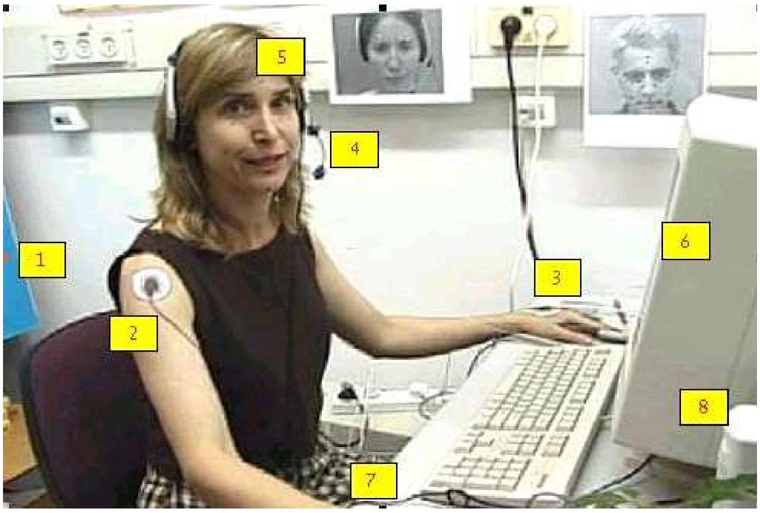

Figure 4-3: Database recording set-up from the participant's point of view: 1.blue background, 2. ECG electrodes on arm and waist, 3.GSR (Galvanic Skin Response) and blood volume electrodes, 4. Microphone, 5. Dots painted on face, 6. Computer screen, 7. Mouse, 8. Loudspeakers. The video camera was positioned above the computer screen.

Figure 4-4 shows an example of a frame from the video stream of facial expression and head gestures. The painted dots and blue background were designed to address the problem of tracking pose changes during analysis, as revealed in various related works [7, 16–18]. These measurements and recordings allow comparison among modalities; they also give a measure of arousal and can help explain the recorded expressions.

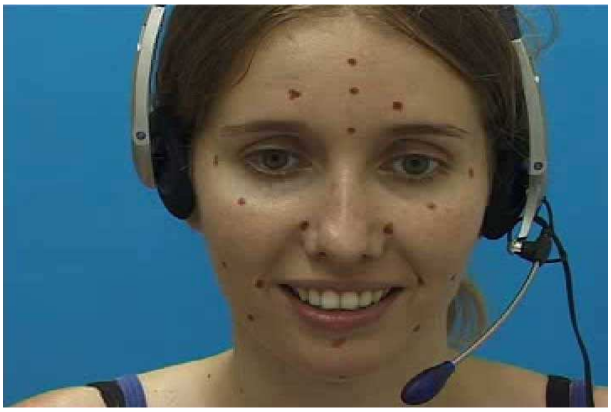

Figure 4-4: An example frame of the video recording of facial expressions and head gestures.

Segmentation

The Doors database includes speech recordings of 15 people. Each speech waveform includes 200 designated sentences, free dialogues, and 100 bell rings (which were used for synchronisation). An automated segmentation method was sought to facilitate the tedious task of extracting the designated sentences. Most segmentation methods relate to segmentation of speech signals into basic units of phonemes or into words [19]; here, however, the segmentation is into sentences, including one-syllable word sentences, such as 'no'.

The method chosen for segmentation of the speech and sound signals into sentences was based on the modified Entropy-based Endpoint Detection for noisy environments described by Shen *et al.* [20]. This method calculates the normalised energy in the frequency domain, and then calculates entropy, as minus the product of the normalised energy and its logarithm. In this way, frequencies with low energy get a higher weight. This corresponds to both speech production and speech perception, because higher frequencies in speech tend to have lower energy, and require lower energy in order to be perceived.

In order to improve the location of end-points I used a zero-crossing rate calculation [21] at the boundaries of the sentences identified by the entropy-based method. This corrected the edge recognition by up to 10msec in each direction

This method yielded very good results, recognising most speech segments (95%) for male participants, and bell rings, and eliminating most of the noise segments. However it requires different parameters for men and for women and therefore, most of the segmentation was done manually. Both sets required additional manual sorting. Principle Component Analysis (PCA) was applied to a small part of the database, in order to distinguish between the two sentences, and between speech segments and the bell ring segments, but it yielded good results only in extraction of the bell segments from the speech segments. The algorithm is described in detail in Algorithm 4-1.

Discussion

The Doors database, although it has many advantages, still has a few drawbacks:

Control

The test was not as controlled as it could have been. A speech recognition system could have been used to allow events to occur only when a certain sentence was uttered; to contain the participant to use the predetermined sentences. For simplicity, and in order to avoid extra stress on the participants, this was not implemented. As a result, some of the pre-designed sentences were altered by the participants throughout the experiment.

On the other hand, the timing of the text was not always synchronised with the actual reaction of the participants to the game's events that was expressed in other manners. It does not disguise long-term attitudes but sometimes miss spontaneous reactions. These spontaneous reactions are an advantage for analysis of natural behaviour.

Not all the participants experienced the same sequence of events. Different participants made different decisions about the doors, and therefore got different results, which elicited different expressions. In addition, different people react differently in similar situation, therefore, the expressions and labels change from one participant to the next.

In general, we expected higher motivation levels, and intended to tempt the participants to win using prizes. The actual recordings were done with volunteers, including graduate students and staff members, who could not easily be motivated by such prizes.

```
Segmentation Algorithm
Define:
•  The signal is divided into overlapping frames
•  FFT length 512, Hamming window of length 512.
For every frame of the signal, $X$:
•  $a = FFT(X \cdot Window)$
•  $Energy = abs(a)^2$
For non-empty frames calculate the normalised energy and the entropy :
•  $Energy_{norm} = \frac{Energy}{\sum_{frequency\ bins} Energy}$
•  $Entropy = -\sum Energy_{norm} \cdot \log_2 (Energy_{norm})$
•  $MinEntropy = \min(Entropy > 0)$
•  $Entropy_{th} = average(Entropy) + \mu \cdot MinEntropy \quad \mu = 0.1$
A speech segment is located in frames in which the $Entropy > Entropy_{th}$
For each segment:
•  Locate all short speech segment candidates and check if they can be unified with their neighbours.
   A segment shorter than 2 frames is not considered a speech segment.
   A short segment of silence in the middle of a speech segment becomes part of the speech segment.
•  Check that the length of the segment is longer than the minimum sentence length allowed;
from observation it was set to 0.1537 sec.
•  Calculate number of zero-crossing events at each frame, $Z_C$.
•  Define threshold of zero-crossing as 10% of the average $Z_C$ :    $Z_{Cth} = 0.1 \cdot average(Z_C)$
•  For each of the identified speech segment, check if the there are adjacent areas in which
   $Z_C > Z_{Cth}$. If there are, the borders of the segments move to the beginning and end as defined by
   the zero-crossing.
```

Algorithm 4-1. Segmentation algorithm.

Labelling

As in other naturally evoked databases, the database has no labels for expressions, i.e. it does not include the names of the related emotions, attitudes and mental states. One of the reasons for the recording of game events and multiple modalities was to help labelling.

Manual labelling with few labels was attempted for initial observations. These observations were conducted in order to verify the feasibility of nuances inference of complex mental states from non-verbal speech cues. At this stage, I tried to give one label to as many utterances as possible. In many cases, I could not define any specific label. The database was labelled manually and checked by four evaluators. Only utterances for which the evaluators agreed on a label were used. Some of the labels were justified by correlation with the game's events, such as participants' decisions, gains and losses, and special scenarios. In addition, I found high correlation among galvanic skin response (GSR) patterns, which are related to events that yielded the labelled expression of *uncertainty*.

In the second stage of the analysis, in which I used an inference machine that was trained and validated on another database, I used the results of the inference machine and the correlation of the vocal cues to the other modalities for analysis and validation. These results are reported in Chapter 8.

Processing

Another drawback, from the processing point of view, is the large amount of information: the raw data requires a lot of pre-processing before any analysis can be run on it. For example, the video recordings of facial expressions run continuously for 15 minutes, while many vision research groups still struggle with streams of a few seconds. The various data streams require segmentation, and synchronisation before any analysis can begin.

On the other hands it has many advantages:

Expressions in HCI

The database includes expressions from a specific type of HCI application.

Range of expressions

It focuses on nuances and on temporal changes, which have not been thoroughly investigated before. The recorded expressions, which differ among participants, include uncertainty, testing, determination, enthusiasm, and regained enthusiasm, a subdued expression, boredom, laughter that evolves from amazement, lack of understanding and disappointment, laughter that evolves from amusement, calculations, thinking aloud, and more.

Control

The only differences between the repeated sentences are related to their expressions. Therefore, the database is useful for the analysis of parameters that influence expressions in general, and for dynamic analysis of expressions in time. It represents natural expressions during an interaction with a computer. The two repeated sentences offer a measure of control for the experiment's generality. They also contribute to the dynamic analysis, as they provide different momentary expressions while following the same medium-term moods. We assume that nothing has changed the longer-term attitudes and mental states during the interaction because there was nothing in the interaction that could have influenced them.

Focus on medium and short-term expressions

In relation to the multi-layer model of expressions in time, the recording setup successfully generated changes in the medium and short-term attitudes and reactions that evolve from the interaction itself, as expected. Some of the expressions could be detected over several consecutive utterances or over a whole interaction, some lasted only one utterance, and some were revealed in unrelated intervals during an interaction. As expected many of the utterances reveal natural transitions within an expression.

Naturally evoked expressions

Although the setup is not entirely natural and does not necessarily represent the common HCI interaction, in the sense that people do not generally sit with electrodes being recorded, it reveals naturally evoked expressions and relations among them. The neutrality of the text and the lack of over motivation or extreme excitement seem to represent the majority of daily interactions.

Non-verbal sounds

The continuous recording of the whole interaction (vs. discrete utterances), revealed another aspect of dynamic speech and audio analysis, which is the wide spectrum and the significance of non-verbal sounds and breathing patterns. These reactions include sharp inhaling, exaggerated exhaling, laughter, free speech and non-verbal sounds. These cues are an integral part of the interaction and its affective impact. Their analysis could be part of a dynamic analysis of an interaction. In some cases we can also detect sounds that, without a sensitive microphone, we could not detect, such as subtle sounds while preparing the mouth for speaking. Such indicators of mouth dryness can be relevant to some applications.

Multi-modal

The cues from the different modalities complement each other. There are cues that appear only in non-verbal speech, and expressions that appear only in the facial expressions. Part of these differences relate to different durations of the expressions: we detect patterns of subtle facial expressions, like surprise, that last less than 200msec before returning to the original expression, while the generation of breathing patterns such as deep inhale require more effort and time and occur in more pronounced instances of surprise. In addition, facial features are active all the time while speech is segmented and appears in discrete instances, especially in the setup we used. On the other hand the control over larger facial expression acts is better and depends on display rules more than speech. Therefore the speech reveals subtle tendencies and changes that the facial expressions do not. In the database the additional modalities, even if they are not part of the actual inference system, add cues that help the designers to label the expressions in the speech utterances.

The properties of this database allow further research in the area of automatic multimodal inference of expressions and may be beneficial to research in other disciplines, such as linguistics.

In order to extend the range of the investigated expressions and of labelled expressions, an additional database was used, the MindReading database, which is described in the following section.

Additional recordings

Another short set of recordings for this work included recordings of 6 people saying 'lo' (which means 'no' in Hebrew), with 10 different unrestricted expressions each, in order to check very short expressive sentences. The observations from these recordings reveal that people change the length of the utterance and the intonation pattern in one word as they do in a longer utterance. It also emphasised the differences of voice spectrum among individuals, and that the change of expression should be estimated per person.

4.2 The MindReading database

The MindReading database was also used in this research.The MindReading DVD is an interactive computer-based guide to emotions. It was developed by a team of psychologists led by Professor Simon Baron-Cohen at the Autism Research Centre at the University of Cambridge, working with a London multimedia production company. The objective was to develop a resource that would help individuals diagnosed with ASD to recognise facial and voal expressions of emotions [13]. Many people diagnosed with Autism Spectrum Disorders (ASD) correctly recognise the basic emotions, but often fail to identify the more complex ones, especially when the signals are subtle and the boundaries between the emotional states are unclear [22]. Existing corpora of non-verbal expressions are of limited use to autism therapy, and to this dissertation, since they encompass only the basic emotions.

The MindReading database comprises of 4411 acted sentences, covering approximately 756 different concepts of emotions. For each emotion, there are 5-7 recorded sentences uttered by different actors. The database is based on taxonomy of mental states that are classified into 24 meaning or emotion groups. This makes it a valuable resource in developing an automated inference system for computer user interfaces and for expression mapping (the analysis and presentation of the relations between expressions and between mental state concepts).

In addition to the advantages there are two issues that should be considered in its analysis. The first evolves from the fact that the taxonomy is built by meaning groups. There are nuances of expressions that could fit into more than one of the 24 general groups. However, more precise categorisation among the groups is difficult and subject to different interpretations of the lexical meaning. In addition, a general conceptual group may include two extremes of a concept (positive/negative, active/passive and the like). For example, in the *unfriendly* group there are both *ignoring* and *argumentative.* In the *sad* group there are both *soulful* and *hysterical. Angry* includes *moody*, *touchy* and also *infuriated* and *wild. Excited* includes the range from *keen* and *refreshed* to *hysterical.* The *happy* group is more complicated, because it includes beside concepts related to happiness and amusement also terms that are related to contentment and well-being, such as *calm*, *safe* and concepts such as *relieved* and *exonerated*, which exist in cases of opposition or negative situations. Therefore, a connectivity analysis using these definitions is not straightforward, and trying to set expression groups in the MindReading database, in order to get larger groups for statistics posed some problems.

The second issue is the accuracy of the acted expressions and their relation to their assigned labels. Although several evaluators checked every utterance, and gave each the most probable label, nonetheless a different group of people could not attach the same labels to approximately 20% of a test set of 24 concepts [14]. The possible reasons for inaccuracy vary. The main reason is that it is difficult to act nuances, especially when they are also related to physiological reactions. For example, voice is much more difficult to control than facial expression. Evaluators are often given a set of definitions and have to choose the closest, which is not always the best. The task becomes more complicated when there are subtle expressions and thousands of utterances to evaluate.

Nevertheless, the MindReading database was successfully used in this research for both training and testing of the inference system, for expression mapping and for observations at the stage of vocal feature definition.

4.3 Summary

The Doors and MindReading databases are examples of databases that complement one another. I chose to use these databases because together they help to achieve a more generalised system. The Doors database supplies a detailed analysis of the dynamic characteristics in a natural interaction, and of the prosody parameters that define nuances of naturally evoked expressions. It is suitable for the analysis of short and medium- term expressions that change during an interaction. The database provides additional cues such as breathing patterns and free speech session. It is also built for multi-modal dynamic analysis which supports its labelling and helps to verify the inference in single modalities, in addition to defining the relations between different cues and their relative contributions to inference. This database has a large research potential also for linguistic research purposes. The MindReading database enlarges the scope of the analysed expressions by presenting a very large set of expression concepts and their linguistic definitions. Its labelling and group definition provide suitable training data. In addition, it allows analysis of the connectivity of expressions, both lexically and by using the prosody parameters. The two databases allow investigation of a large variety of research questions, but they also require a lot of additional work. In the scope of this work I could address only part of their potential.

References

[1] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., and Taylor J. G., "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[2] Petrushin V., *Intelligent Engineering Systems Through Artificial Neural Networks*, chapter Emotion in speech: Recognition and application to call centers, pp. 1085–1092, 1999.

[3] Polzin T. and Waibel A., "Emotion-sensitive human-computer interfaces", in *ISCA workshop (ITRW) on speech and emotion: a conceptual framework for research*, Belfast, North Ireland, 2000.

[4] Yan Li F. Y., Ying-Qing X., Chang E., and Heung-Yeung S., "Speech driven cartoon animation with emotions", in *ACM Multimedia 2001*, Ottawa, Canada, 2001.

[5] Ekman P., *Handbook of cognition and emotion*, chapter Basic emotion, Wiley, Chihester, UK, 1999.

[6] Douglas-Cowie E., Campbell N., Cowie R., and Roach P., "Emotional speech: towards a new generation of databases", *Speech Communication*, vol. 40, pp. 33–60, 2003.

[7] Cohn J.F. and Katz G.S., "Bimodal expression of emotion by face and voice", in *Workshop on Face / Gesture Recognition and Their Applications, The Sixth ACM International Multimedia Conference, UK*, 1998.

[8] Klein J., Moon Y., and Picard R. W., "This computer responds to user frustration: theory, design, and results", *Interacting with Computers*, vol. 14, no. 2, pp. 119–40, 2002.

[9] Fernandez R. and Picard R. W., "Modeling drivers' speech under stress", *Speech Communication*, vol. 40, pp. 145–59, 2003.

[10] Moore C. A., Cohn J. F., and Katz G. S., "Quantitative description and differentiation of fundamental frequency contours", *Computer Speech and Language*, vol. 8, no. 4, pp. 385–404, 1994.

[11] Wierzbicka A., "The semantics of human facial expressions", *Pragmatics and cognition*, vol. 8, no. 1, pp. 147–183, 2000.

[12] Cornelius R. and Cowie R., “Describing the emotional states that are expressed in speech”, *Speech Communication*, vol. 40, pp. 5–32, 2003.

[13] Baron-Cohen S., Riviere A., Fukushima M., French D., Hadwin J., Cross P., Bryant C., and Sotillo M., “Reading the mind in the face: A crosscultural and developmental study”, *Visual Cognition*, vol. 3, pp. 39–59, 1996.

[14] Golan O., Baron-Cohen S., and Hill J., “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.

[15] Bechara A., Damasio H., Tranel D., and Damasio A. R., “Deciding advantageously before knowing the advantageous strategy”, *Science*, vol. 275, pp. 1293–5, 1997.

[16] Yacoob Y. and Davis L. S., “Recognizing human facial expressions”, in *Image Understanding Workshop.*, San Francisco, CA, USA, 1994, pp. 827–35.

[17] Yacoob Y. and Davis L., “Computing spatio-temporal representations of human faces”, in *Proceedings 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA*, 1994.

[18] Tian Y., Kanade T., and Cohn J.F., “Recognizing action units for facial expression recognition”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[19] Gin-Der W. and Chin-Teng L., “A recurrent neural fuzzy network for word boundary detection in variable noise-level environments”, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, no. 1, pp. 97–115, 2001.

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

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

[22] Baron-Cohen S., Wheelwright S., Lawson J., Griffin R., and Hill J., *Handbook of Childhood Cognitive Development*, chapter The Exact Mind: Empathising and Systemising in Autism Spectrum Conditions, pp. 491–508, Oxford:Blackwell, 2002.

Chapter 5

Vocal features of emotions and mental states [1]

One of the goals of this research is to identify the features that characterise different expressions and nuances of expressions. The definition of features is an essential part of the analysis and has a significant effect on the inference accuracy. This chapter describes two stages of the general framework. It first describes the second stage of the framework which incorporates the definition and extraction of vocal features from the speech signal. It then describes the implementation of the third stage of the framework, the definition of secondary metrics, including statistical features and temporal characteristics which are extracted or derived from the vocal features.

This chapter includes a background section that explains why vocal features are required for the analysis, and reviews the most commonly used features. It then includes sections that explain the extraction of the main vocal features. It includes background that explains the selection of features and the selection of extraction algorithms. It surveys the algorithms that have been developed during the course of this work and modifications to existing algorithms. The extracted features include the fundamental frequency (f_0), energy, spectral content and harmonic properties. It is followed by a section that describes a parsing algorithm that divides the analysed utterance into several parts according to their vocal properties and the speech rates calculation which is derived from the parsing. The last section of this chapter defines the two sets of secondary features and metrics used for the inference in this work. The first set was used for the initial observations and includes mostly statistical metrics over the whole utterance. The second set was developed as a result of these observations and includes more time-related metrics. This set was used for the development of the inference machine which is described in Chapter 6.

I used MATLAB for the development of the algorithms and for their application to the speech signals.

5.1 Background

Analysis of the entire speech signal without any pre-processing would have been ideal, but is not practical. For example, a three second long utterance, with sampling rate of 32 kHz (which is relatively low) is presented as an array of 96,000 samples. Training a machine to recognise similarities and differences among such long vectors is not feasible. Furthermore, such inference could not reveal understandable and transferable information that might support further analysis of other datasets and that could be applied to synthesis. Therefore a smaller set of features that includes all the relevant information is

[1] The original features, metrics and algorithms described in this chapter are under patent pending

required.

Feature selection and extraction is not a trivial problem with known and accepted solutions. The specific contribution of each feature to expressions is still unclear. Psychological and psychoacoustic tests have examined the relevance of different features to the perception of emotions and mental states using features such as pitch[2] range, pitch average, speech rate, contour, duration, spectral content, voice quality, pitch changes, tone base, articulation and energy level [1–4]. However, different researchers have found different and sometime contradicting features to characterise the same global definition of expression. Furthermore, the definitions of expressions and emotions are not identical among different works, the datasets are different [1], and the feature descriptions are often qualitative in nature, especially those that address voice quality, for example *breathy*, *tense*, *grumbling* and the like.

The features most commonly used for automatic inference of emotions from speech are derived from the fundamental frequency, energy, spectral content, and speech rate. However, the determination of features for automatic prosody analysis is not well defined, and there is no uniform way or single algorithm to extract each feature. It is especially clear that for basic features, such as the fundamental frequency and speech rate, there are multiple definitions and several calculation methods, not all of them yielding the same results [5–10]. Additional features such as loudness, harmonies, jitter, shimmer, the fluctuations in f_0 frequency and in amplitude respectively, and rhythm have also been used. The accuracy of the calculation of these parameters is highly dependent on the recording quality, sampling rate and the time units and frame length for which they are calculated.

Any classification attempt relies heavily on these results, and therefore choosing the features to use is crucial for the rest of the research.

Few researchers have attempted to define features that describe voice quality in a mathematical or quantitative manner [11]. In the literature, the human voice is often described in terms such as *sharp*, *dull*, *warm*, *pleasant* and *unpleasant*. However, what creates these characteristics has not been discussed. There is no explanation as to how people can distinguish between real and acted expressions, including the distinction between real and artificial or constrained laughter.

The features for these expressions and their characteristics could be derived from any of the approaches to expressive speech analysis, including the production mechanism and signal processing, speech analysis, linguistic definitions, hearing mechanism, and speech perception. Additional features from a musical point of view, for example, tempo, harmonies, dissonances and consonances, have not been fully explored. Other features that can be derived from music perception include rhythm, dynamics, and tonal structures or melodies. Tonal related analysis includes chords, or the combination of several tones at each time unit, and the structure of the melody, or the contour, that is how the tones are arranged in a consecutive order in time. Different combinations of tones and tone-intervals are related to keys in music, which are related to the perception of consonance, dissonance and tension.

In addition to the problem of choosing features and defining robust extraction algorithms, most of the features have different values along the utterance. The next issue is to define a set of secondary features, metrics that will provide the means to analyse these

[2] Pitch in this context refers to the intonation, as in fundamental frequency (f_0).

features automatically. These include metrics that represent the behaviour of the features over a whole utterance, here referred to as *statistical features*.

The linguistic approach is that analysis of the temporal behaviour of the intonation and of other features should include reference to units such as syllables. In most datasets that are used for linguistic analysis the annotation of syllables is done manually. If the analysis is to be conducted without reference to speech recognition tools, an automatic parsing mechanism is required. The parsing here follows to a large extent the linguistic units, but it derives its motivation and definition from the vocal features including silence, energy peaks and energy peaks with pitch.

Different researchers use different sets of statistical features for automatic emotional speech analysis. For example, Dellaert *et al.* [12], use two sets of parameters. The first set included: mean, standard deviation, minimum, maximum and range of the pitch, slope and speaking rate. The second set comprised 17 features of smoothed and continuous pitch. Oudeyer [13] used another set of 200 features and later a sub-set of 15, including statistical features of pitch and of intensity of filtered signals. These two works were carried on different databases and different sets of expressions; there is no agreement about the calculation of the different features, nor about the role and the significance of each feature, which demonstrates the challenge involved.

5.2 Fundamental frequency

The central feature of prosody is the intonation. Intonation refers to patterns of the fundamental frequency, f_0, which is the acoustic correlate of the rate of vibrations of the vocal folds. Its perceptual correlate is pitch. People use f_0 modulation, i.e. intonation; in a controlled way to convey meaning.

As mentioned in the introduction to this chapter, there are many different extraction algorithms for the fundamental frequency. I examined three different methods for calculating the fundamental frequency, f_0, here referred to as pitch. All three methods are based on the autocorrelation of the signal. The autocorrelation is the cross-correlation of the signal with itself, or how well a signal matches a time-shift version of itself. Autocorrelation is useful for finding repeating patterns in a signal [14]. The first two methods are an autocorrelation method with inverse Linear Prediction Code (LPC) [7], and a cepstrum method [8]. Both methods of pitch estimation gave very similar results in most cases. The third method is based on Boersma's algorithm [10] which was used by him in the tool PRAAT [15], which in turn is used for emotions analysis in speech [11, 13], and by many linguists for research of prosody and prosody perception. Boersma pointed out that sampling and windowing cause problems in determining the maximum of the autocorrelation signal [10]. This method therefore includes division by the autocorrelation of the window. The next stage is to find the optimal sequence of pitch values for the whole sequence of frames by using the best time-shift candidates in the autocorrelation, i.e. those with the maximum values of the autocorrelation. This uses the Viterbi algorithm with different costs associated with transitions between adjacent voiced frames, in which there is pitch, and with transitions between voiced and unvoiced frames, in which there is speech but without vibrations of the vocal folds (these weights depend partially on the shift between frames). It also penalises transitions between octaves (frequencies twice as high or low).

The third method yielded the best results. However it still required some adaptations.

Speaker dependency is a major problem in automatic speech processing as the pitch ranges for different speakers can vary dramatically. It is often necessary to clarify the pitch manually after extraction. I have adapted the extraction algorithm to correct the extracted pitch curve automatically. The first attempt to adapt the pitch to different speakers includes the use of three different search boundaries, of 300 Hz for men, 600 Hz for women and 950 Hz for children, adjusted automatically by the mean pitch value of the speech signal.

```
Fundamental Frequency Extraction Algorithm
>>> Pre-processing:
        1.  Divide the speech signal Signal into overlapping frames
>>> Short term analysis:
        2.  Apply a Hamming window to the signal in the frame, so that the centre of the frame has a higher weight then the boundaries.
        3.  For each frame compute the normalised autocorrelation
        4.  Divide the signal autocorrelation by the auto correlation of the window
        5.  Find candidates for the pitch from the normalised autocorrelation signal - the first $N$ maxima values. Calculate parabolic interpolation with the autocorrelation points around it, in order to find more accurate maximum values of the auto correlation. Keep all candidates for harmonic properties calculation Algorithm 5-2
>>> Calculate in iteration an optimal sequence of $f_0$(pitch), for the whole utterance. Calculate for every frame, and every candidate in each frame, recursively, using the Viterbi algorithm. In each iteration, adjust the weights of the candidates according to:
        6.  Check if the candidates' frequencies are within the specific range, and their weights are positive. If not, they become unvoiced candidates, with frequency value 0.
        7.  Define the Strength as the relation between the average value of the signal in the frame and the maximal value of the entire speech signal. Calculate weights according to pre-defined threshold values and frame strengths for voiced and unvoiced candidates.
        8.  The cost for transition from voiced to unvoiced or from unvoiced to voiced
        9.  The cost of transition from voiced to voiced, and among octaves
        10. The continuity of the curve (adaptations to Boersma's algorithm): the adaption is achieved by adapting the strength of a probable candidate to the strength of the leading candidate.
              a. Avoid frequency jumps to higher or lower octaves
              b. Frequency changes greater than 10 Hz
              c. Eliminate very short sequences of either voiced or unvoiced signal.
              d. Adapt to speaker by changing the allowed pitch range.
>>> After $M$ iteration, the expectation is to have a continuous pitch curve.
```

Algorithm 5-1. An outline of the algorithm for the extraction of the fundamental frequency. The adaption to Boersma algorithm [10] are in the iteration stage (stage 10).

Although this improved the pitch calculations, the improvement was not general enough. The second change considers the continuity of the pitch curves. It consists of several observed rules. First, the maximum frequency value for the time-shift candidates in the autocorrelation is allowed to change if the current values are within a smaller or larger range. The lowest frequency default was set to 70 Hz, although automatic adaptation to 50 Hz was added, for extreme cases (Only very few sentences in the two datasets required a lower minimum value, mainly for men on specific events). The highest frequency was set to 600 Hz. Only very few sentences in the two datasets required a higher range, mainly children who were trying to be *irritating*.

Second, the weights of the candidates are changed if using other candidates with originally lower weights can improve the continuity of the curve. Several scenarios may cause such a change. First, frequency jumps between adjacent frames that exceed 10 Hz: In this case candidates that offer smaller jumps should be considered. Second, consider candidates exactly one octave higher or lower from the most probable candidate, with lower weights. In addition, in order to avoid unduly short segments, if voiced segments consist of no more than two consecutive frames the weights of these frames are reduced. Correction is also considered for voiced segments that are an octave higher or lower than their surrounding voiced segments. This algorithm eliminates the need of manual intervention in most cases, but it is time consuming. Algorithm 5-1 describes the outline of the algorithm.

Figure 5-1 shows two fundamental frequency curves, one as extracted by the original algorithm of PRAAT, and the other with the additional modifications.

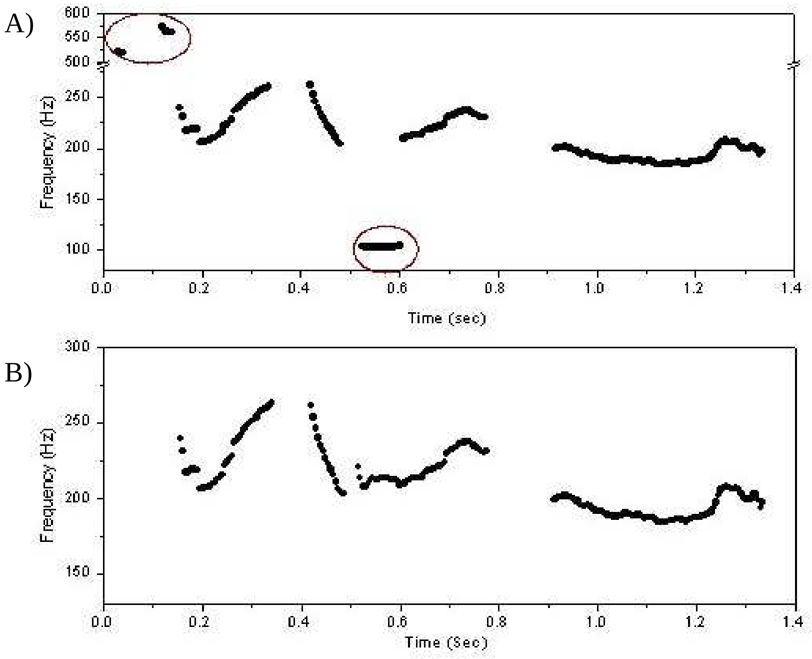

Figure 5-1: A) Pitch extraction using the PRAAT, (ringed regions indicate the outlier that require correction) B) after modification using algorithm 5-1.

5.3 Energy

The second feature that signifies expressions in speech is the energy, also referred to as intensity. The energy or intensity of the signal X for each sample i in time is:

$$Energy_i = X_i^2$$

The smoothed energy is calculated as the average of the energy over overlapping time frames, as in the fundamental frequency calculation. If $X_1 \ldots X_N$ defines the signal samples in a frame then the smoothed energy in each frame is:

$$SmoothedEnergy_{Frame} = \sum_{i=1:N} X_i^2$$

The first analysis stage considered these two representations. In the second stage only the smoothed energy curve was considered, and the signal was multiplied by a window so that in each frame a larger weight was given to the centre of the frame. This calculation method yields a relatively smooth curve that describes the more significant characteristics of the energy throughout the utterance:

$$SmoothedEnergy_{Frame} = \sum_{i=1:N} (X_i \cdot W_i)^2$$

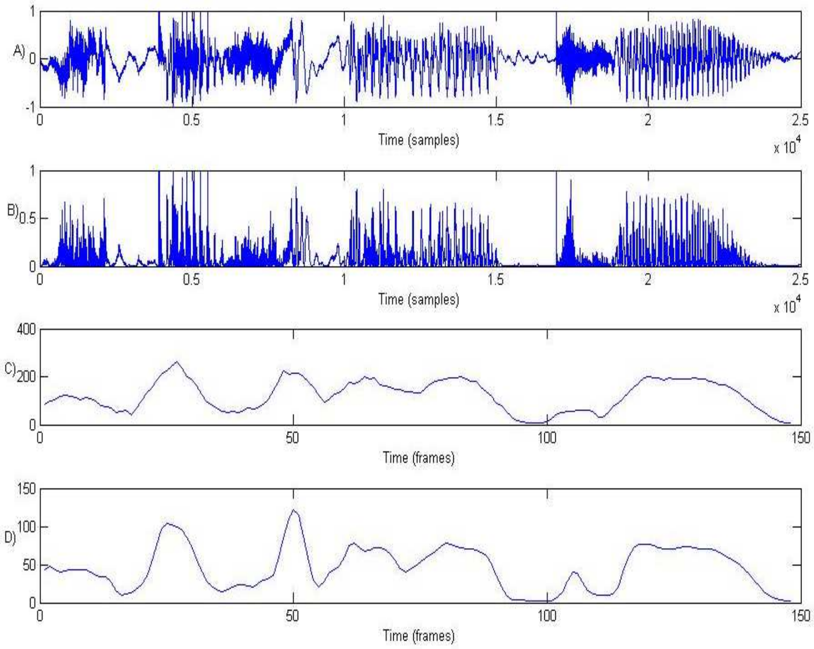

Figure 5-2: Different forms of energy calculations. A) speech signal, B) the energy of each sample, C) the average energy of each frame, D) the energy of frames with Hamming window.

Figure 5-2 shows a speech signal and the results of different energy calculations; the speech signal is shown in (A). Its energy (B), the smoothed energy (averaged) (C) and smoothed energy with a window (D). It can be seen that the smooth curves (C and D) give the general behaviour of the energy, or the contour of the energy, rather than rapid fluctuations that are more sensitive to noise, as in the energy calculation for each sample (B). The application of a window (D), emphasises the local changes in time, and follows more closely the original contour, as of the signal itself (A).

5.4 Spectral content

Features related to the spectral content of speech signals are not widely used in the context of expressions analysis. One method for the description of spectral content is to use formants, which are based on the speech production model [7]. I have refrained from using formants as both their definition and their calculation methods are problematic. They refer mainly to vowels and are defined mostly for low frequencies (below 4-4.5 kHz). The other method, which is the more commonly used, is to use filter-banks, which involves dividing the spectrum into frequency bands. There are two major descriptions of frequency bands that relate to human perception, and these were set according to psycho-acoustic tests - the Mel Scale [16] and the Bark Scale, which is based on empirical observations from loudness summation experiments [17,18]. Both correspond to the human perception of sounds and their loudness, which implies logarithmic growth of bandwidths, and a nearly linear response in the low frequencies. In this work, the Bark scale was chosen because it covers most of the frequency range of the recorded signals (effectively 100Hz-10 kHz). The Bark scale ranges from 1 to 24 and corresponds to the first 24 critical bands of hearing. The subsequent band edges are (in Hz) 0, 100, 200, 300, 400, 510, 630, 770, 920, 1080, 1270, 1480, 1720, 2000, 2320, 2700, 3150, 3700, 4400, 5300, 6400, 7700, 9500, 12000, 15500. The formula for converting a frequency f (Hz) into Bark is:

$$Bark = \arctan\left(\frac{0.76 \cdot f}{1000}\right) + 3.5 \cdot \arctan\left(\left(\frac{f}{7500}\right)^2\right)$$

In this work, at the first stage, 8 bands were used. The bands were defined roughly according to the frequency response of the human ear, with wider bands for higher frequencies up to 9 kHz. Figure 5-3 shows the energy in different bands of a speech signal using the eight bands. In the second stage the Bark scale up to 9 kHz was used.

5.5 Harmonic properties

One of the parameters of prosody is voice quality. We can often describe voice with terms such as *sharp*, *dull*, *warm*, *pleasant*, *unpleasant*, etc. However, the speech features that create these characteristics have not been identified. Concepts that are borrowed from music may describe some of these characteristics and provide explanations for phenomena observed in the autocorrelation of the speech signal.

Calculation of the fundamental frequency using the autocorrelation of the speech signal usually reveals several candidates for pitch, they are usually harmonies, multiplications of the fundamental frequency by natural numbers, as can be seen in Figure 5-4 and Figure 5-5 (a).

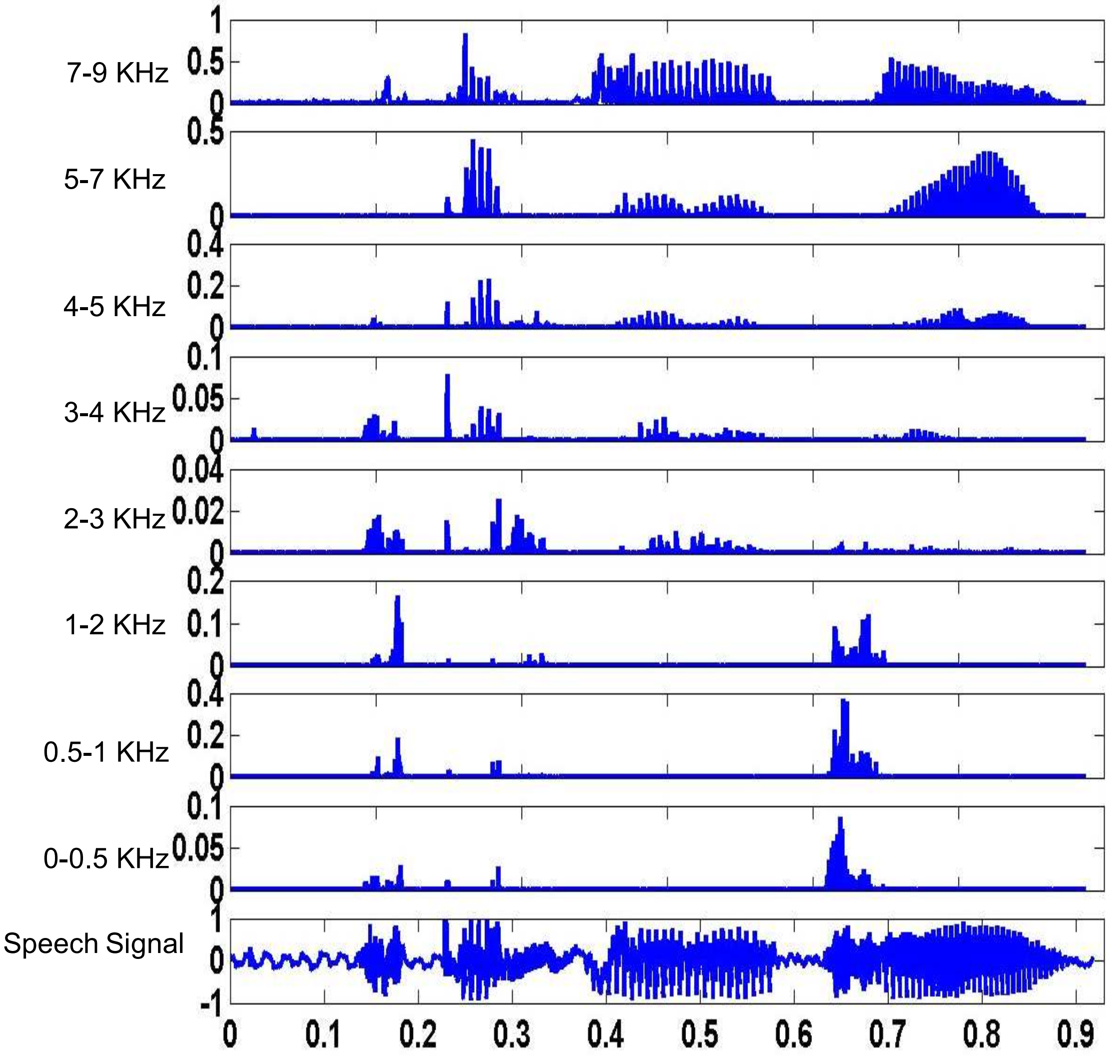

Figure 5-3: Energy in different frequency bands: 0-500 Hz (band 1), 500-1000 Hz (band 2), 1-2 kHz (band 3), 2-3 kHz (band 4), 3-4 kHz (band 5), 4-5 kHz (band 6), 5-7 kHz (band 7), 7-9 kHz (band 8), and the speech signal (at the bottom).

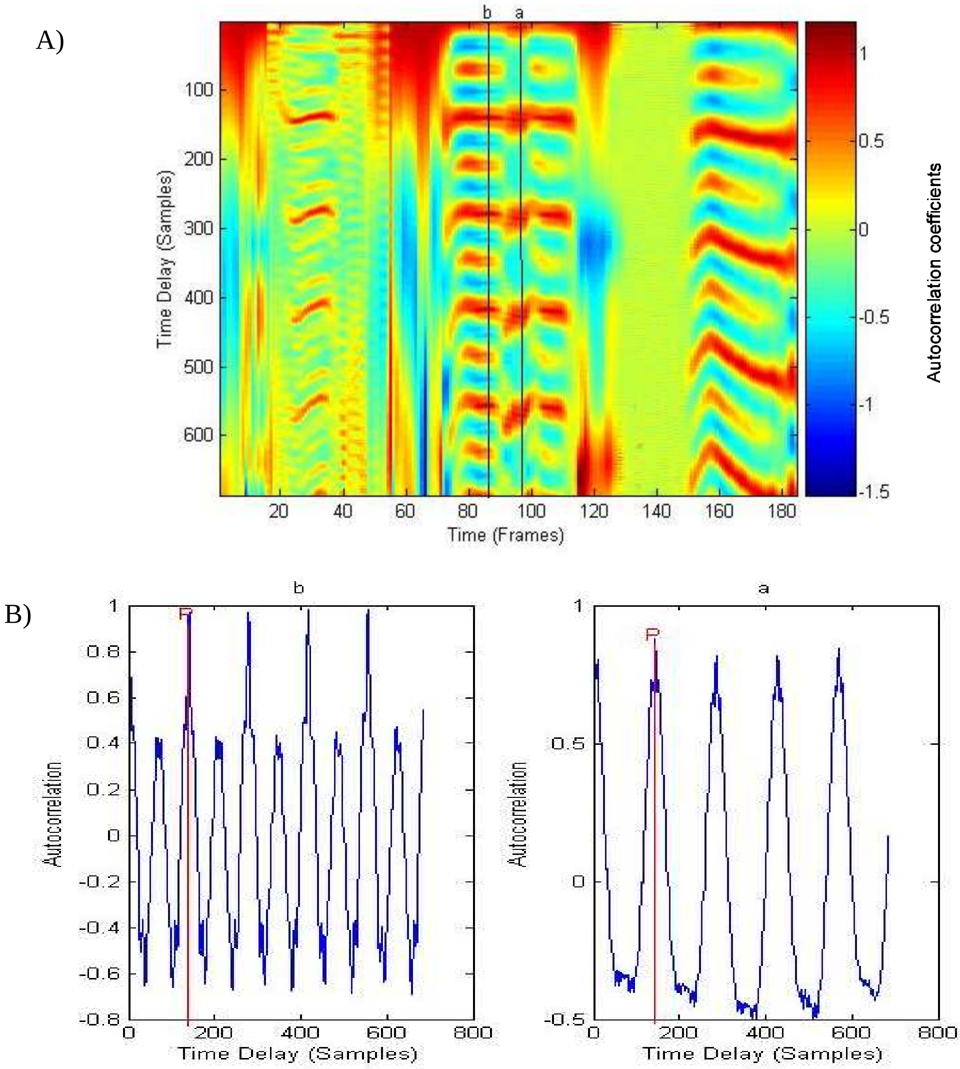

Figure 5-4: (A) Autocorrelation of a sentence uttered by a female speaker, calculated on overlapping time-frames and (B) the autocorrelation of specific time-frames from the speech signal in (A): a - pitch only, b - pitch and one significant harmonic interval which corresponds to ratio 3:2 to the f_0 frequency. The lines indicated by P denote the time delay of the pitch, i.e. $Pitch = SamplingFrequency/Time\,Delay(P)$.

In expressive speech, there are also other maximum values, which are considered for the calculation of the fundamental frequency, but are usually ignored if they do not contribute to it. Interestingly, in many cases they reveal a behaviour that could be associated with harmonic intervals, pure tones with relatively small ratio between them and the fundamental frequency, especially 3:2, as can be seen in Figure 5-4 (the line indicated by b). Other intervals, such as 4:3 and more complicated patterns also appear, as can be seen for example in Figure 5-5. These candidates do not exist in all speech signals, and can appear only in parts of an utterance. It seems as if these relations might be associated with the musical notations of consonance.

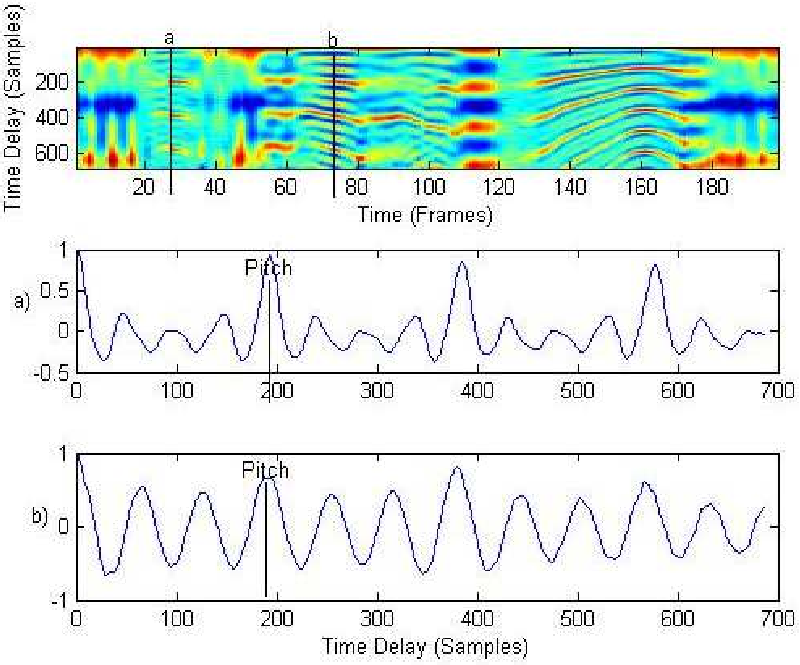

Figure 5-5: Harmonic intervals in the autocorrelation of expressive speech. The top figure is the autocorrelation of the whole speech signal, a) and b) are the autocorrelation at the time frames marked in the top figure. The time delay of the points marked as 'Pitch' are the points for which the pitch or fundamental frequency is calculated: $FundamentalFrequency = SamplingFrequency/TimeDelay(Pitch)$.

In other cases, the fundamental frequency is not very 'clean', and the autocorrelation reveals candidates with frequencies which are very close to the fundamental frequency. In music, such tones are associated with roughness or dissonance. There are other ratios that are considered unpleasant.

The main high-value peaks of the autocorrelation correspond to frequencies that are both lower and higher than the fundamental frequency, with natural ratios, such as 1:2, 1:3 and their multiples. In this work, these ratios are referred to as sub-harmonies, for the lower frequencies, and harmonies for the higher frequencies, intervals that are not natural numbers, such as 3:2 and 4:3 are referred to as harmonic intervals. Sub-harmonies may suggest how many precise repetitions of f_0 exist in the frame, which can also suggest how pure its tone is. (The measurement method limits the maximum value of detected sub-harmonies for low values of the fundamental frequency). I suggest that this phenomenon appears in the speech signals and may be related to the harmonic properties, although the terminology which is used in musicology may be different.

One of the first applications of physical science to the study of music perception was Pythagoras' discovery that simultaneous vibrations of two string segments sound harmonious when their lengths form small integer ratios (e.g. 1:2, 2:3, 3:4) [19, 20]. These ratios create consonance, blends that sound pleasant. Galileo postulated that tonal dissonance, or unpleasant, arises from temporal irregularities in eardrum vibrations that give rise to "ever-discordant impulses" [21]. Schwartz *et al.* [22] have shown that statistical analysis of the spectrum of human speech sounds show that the same ratios of the fundamental frequency are obvious in different languages. Tramo *et al.* [23] explain the neurobiology of harmony perception. They show that information about the roughness and pitch of musical intervals is present in the temporal discharge patterns of the Type I auditory nerve fibres, which transmit information about sound from the inner ear to the brain. These findings indicate that people are built to both perceive and generate these harmonic relations.

The ideal harmonic intervals, their correlate in the 12 tones system of western music and their definitions as dissonances or consonances are listed in Table 5-1. The table also shows the differences between the values of these two sets of definition. These differences are smaller than 1% [24]. Terhardt [25] suggested that the different scales are approximations.

Number of Semitones	Interval Name	Consonant?	Intonation Ratios	Equal Temperament	Difference
0	unison	Yes	1/1=1.000	$2^{0/12} = 1.000$	0.0%
1	semitone	No	16/15=1.067	$2^{1/12} = 1.059$	0.7%
2	whole tone (major)	No	9/8=1.125	$2^{2/12} = 1.122$	0.2%
3	minor third	Yes	6/5=1.200	$2^{3/12} = 1.189$	0.9%
4	major third	Yes	5/4=1.250	$2^{4/12} = 1.260$	0.8%
5	perfect fourth	Yes	4/3=1.333	$2^{5/12} = 1.335$	0.1%
6	tritone	No	7/5=1.400	$2^{6/12} = 1.414$	1.0%
7	perfect fifth	Yes	3/2=1.500	$2^{7/12} = 1.498$	0.1%
8	minor sixth	Yes	8/5=1.600	$2^{8/12} = 1.587$	0.8%
9	major sixth	Yes	5/3=1.667	$2^{9/12} = 1.682$	0.9%
10	minor seventh	No	9/5=1.800	$2^{10/12} = 1.782$	1.0%
11	major seventh	No	15/8=1.875	$2^{11/12} = 1.888$	0.7%
12	octave	Yes	2/1=2.000	$2^{12/12} = 2.000$	0.0%

Table 5-1: Harmonic intervals, also referred to as just intonation, and their dissonance or consonance property, compared with equal temperament, which is the scale in western music. The intervals in both systems are not exactly the same, but they are very close.

When two tones interact with each other and the interval or ratio between their frequencies create a repetitive pattern of amplitudes, their autocorrelation will reveal the repetitiveness of this pattern. Figure 5-6 and Figure 5-7 [23] show different combinations of two pure tones and their harmonies in the time and frequency domains. Minor second (16:15) and triton (7:5=1.4, 45:32=1.40625 or 1.414, the definition depends on the system in use) are dissonances while perfect fifth (3:2) and forth (4:3) are consonances. Minor second is an example of two tones of frequencies that are very close to each other, and can be associated with roughness (Figure 5-6(I), Figure 5-7 (E)), perfect fourth and fifth create nicely distinguishable repetitive patterns (Figure 5-6 (J), (L), Figure 5-7 (F), (H)), which are associated with consonances. Tritone, which is considered a dissonant, does not create such a repetitive pattern (Figure 5-6 (K)), while creating roughness (signals of too close frequencies) with the third and fourth harmonies (multiplications) of the pitch (Figure 5-7 (G)).

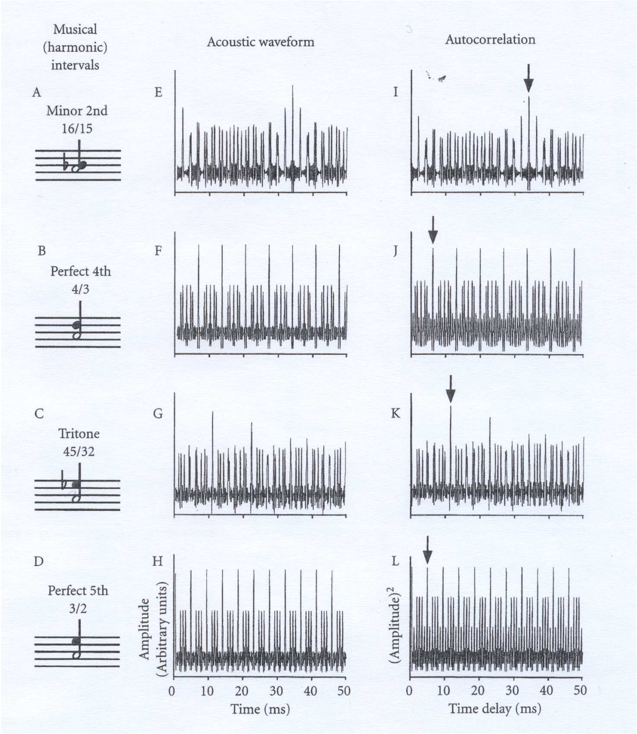

Figure 5-6: Combinations of two harmonic tones with different ratios (A-D) the tones with musical notations, (E-H) the tones in the time domain (I-L) the autocorrelation of the tone [23].

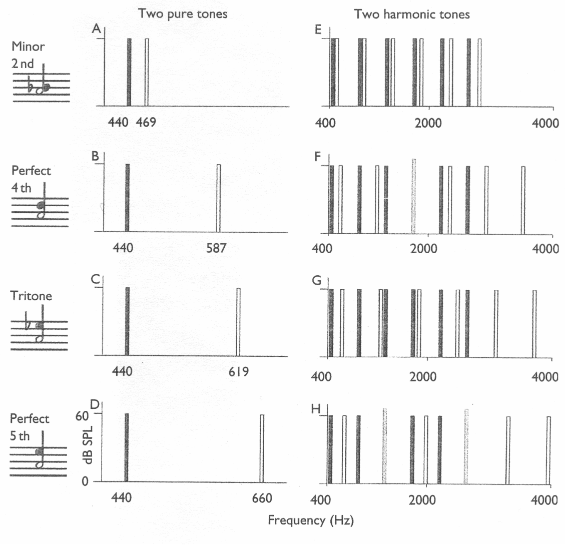

Figure 5-7: Combinations of two harmonic tones with different ratios (A-D) the tones in the frequency domain, (E-H) the tones and their harmonies in the frequency domain [23].

Fernandez [11], used a model which was developed by Plompt and Levelt [26] for the description of dissonance or roughness, that claimed that consonance is the absence of dissonance. Dissonance as a function of the ratios between two pure tones can be seen in Figure 5-7. The curve of the dissonance perception has a minimum at unison, rises fast to maximum and decays again. It rises faster as the lower frequency in the ratio is higher.

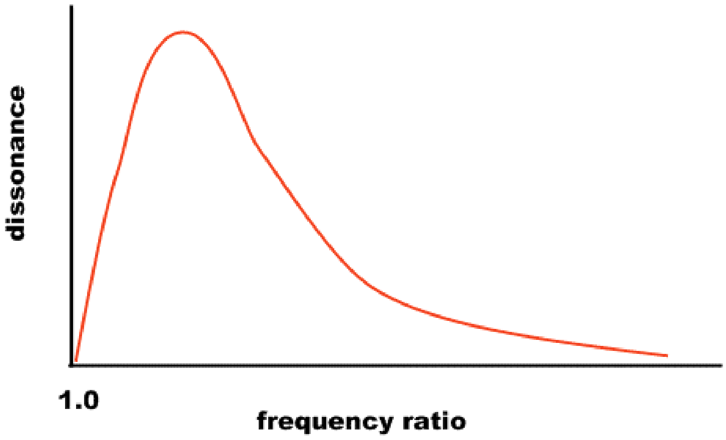

Figure 5-8: Dissonance as a function of the ration between two tones [26]

However, there seem to be well-known and robust results regarding the perceived sense of intervals when two pure tones of different frequencies interact with each other [19–23, 25].

Two tones are perceived as pleasant when the ear can separate them clearly and when they are in unison, for all harmonies. Relatively small intervals (relative to the fundamental frequency), are not well-distinguished and perceived as 'roughness'. The autocorrelation of expressive speech signals reveals the same behaviour, therefore I included the ratios as appeared in the autocorrelation to the extracted features, and added measures that tested their relations to the documented harmonic intervals.

The harmonies and the sub-harmonies were extracted from the autocorrelation maximum values. The calculation of the autocorrelation follows the sections of the fundamental frequency extraction algorithm (Algorithm 5-1), that describes the calculation of candidates. The rest of the calculation, which is described in Algorithm 5-2 is performed after the calculation of the fundamental frequency is completed:

Extracting ratios

For the candidates calculated in Algorithm 5-1, do:

If $Candidate > f_0$ then it is considered as harmony, with ratio:

$harmonies = \frac{candidate}{f_0}$

Else, if $Candidate < f_0$ then it is considered as sub-harmony, with ratio:

$sub\_harmonies = \frac{f_0}{candidate}$

For each frame all the Candidates and their weights, $CandidateWeights$, are kept.

Algorithm 5-2. Extracting ratios: the definitions of 'harmonies' and 'sub-harmonies' as used in this work.

The next stage is to check if the candidates are close to the known ratios of dissonances and consonances (Table 5-1), having established the fact that these ratios are significant. I examined for each autocorrelation candidate the nearest harmonic interval and the distance from this ideal value. For each ideal value I then calculated the normalised number of occurrences in the utterance, i.e. divided by the number of voiced frames in the utterance.

The ideal values for sub-harmonies are the natural numbers. Unfortunately, the number of sub-harmonies for low values of the fundamental frequencies is limited, but since the results are normalised for each speakers this effect is neutralised.

These features may be able to explain how people can distinguish between real and acted expressions, including the distinction between real and artificial laughter, including behaviour that is subject to cultural display rules or stress. The distance of the calculated values from the ideal ratios may reveal the difference between natural and artificial expressions. The artificial sense may be derived from inaccurate transitions while speakers try to imitate the characteristics of their natural response. However, this aspect requires further investigation.

The results of my experiments reveal that the harmonic related features are among the most significant features for distinguishing between different types of expressions, as shown in the features automatically allocated for the inference machines, as described in Chapter 6.

5.6 Parsing

Time variations within utterances serve various communication roles. Linguists and especially those who investigate pragmatic linguistics use sub-units of the utterance for observations. Speech signals (the digital representation of the captured/recorded speech) can be divided roughly into several categories. The first is *speech* and *silence*, in which there are no speech or voice. The difference between them can be roughly defined by the energy level of the speech signal. The second category is *voiced*, where the fundamental frequency is not zero, i.e. there are vibrations of the vocal folds during speech, usually during the utterance of vowels, and *unvoiced*, where the fundamental frequency is zero, which happens mainly during silence and during the utterance of consonants such as /s/,/t/ and /p/, i.e. there are no vibrations of the vocal folds during articulation. The linguistic unit that is associated with these descriptions is the syllable, in which the main feature is the voiced part, which can be surrounded on one or both sides by unvoiced parts. The pitch, or fundamental frequency, defines the stressed syllable in a word, and the significant words in a sentence, in addition to the expressive non-textual content. This behaviour changes among languages and accents.

In the context of non-verbal expressiveness, the distinction among these units allows the system to define characteristics of the different speech parts, and their time-related behaviour. It also facilitates following temporal changes among utterances, especially in the case of identical text. The features that are of interest are somewhat different from those in the purely linguistic analysis, such features may include, for example the amount of energy in the stressed part compared to the energy in the other parts, or the length of the unvoiced parts.

I have decided to try two approaches to parsing. In the first I tried to extract these units using image processing techniques from spectrograms of the speech signals and from smoothed spectrograms. Spectrograms present the magnitude of the Short Time Fourier Transform (STFT) of the signal, calculated on (overlapping) short time-frames. For the parsing I used two dimensional (2D) edge detection techniques including Canny [27] and zero crossing. However, most of the utterances are too noisy, and the speech itself has too many fluctuations and gradual changes so that the spectrograms are not smooth enough and do not give good enough results.

Parsing Rules

1. Define silence threshold as 5% of the maximum energy.
2. Locate peaks (location and value) of energy maximum value in the smoothed energy curve (calculated with window), that are at least 40 msec apart.
3. Delete very small energy peaks that are smaller than the silence threshold.
4. Beginning of sentence is the first occurrence of either the beginning of the first voiced part (pitch), or the point prior to an energy peak, in which the energy climbs above the silence threshold.
5. End of sentence is the last occurrence of either pitch or of the energy getting below the silence threshold.
6. Remove insignificant minimum values of energy between two adjacent maximum values (very short-duration valleys without a significant change in the energy. In a 'saddle' remove the local minimum and the smaller peak.)
7. Find pauses- look between two maximum peaks and find if the minimum is less than 10 percent of the maximum energy. If it is true then bracket it by the 10 percent limit. Do not do it if the pause length is less than 30 msec or if there is a pitch in that frame.

Algorithm 5-3. Parsing an utterance into different speech parts.

The second approach was to develop a rule based parsing.From analysis of the extracted features of many utterances from the two datasets in the time domain, a few rules for parsing were defined. These rules follow roughly the textual units. Several parameters have been considered for their definition, including the smoothed energy (with window), pitch contour, number of zero-crossings, and other edge detection techniques.

Algorithm 5-3 describes the rules that define the beginning and end of a sentence, finds silence areas and significant energy maximum values and locations. The calculation of secondary time-related metrics is then done on voiced part, where there are both pitch and energy, places in which there is energy (significant energy peaks) with no pitch, and on durations of silence or pauses.

5.7 Speech rate

Speech rate, which seems to be an intuitive parameter of paralinguistic speech (fast, slow, etc.), does not have a uniform mathematical definition. There are many different definitions of speech rate in the speech analysis literature. In general, most researchers set their own definitions. For example, the average length of the voiced part of speech [28],the sum of voiced frames in which the energy is above threshold divided by the number of words in the utterance, etc. The definition in this work is:

$$Speech\_rate = \frac{Voiced}{Sentence\_length}$$

In the first set of features that was used for initial observations, the sentence length was defined only from the beginning until the end of the voiced parts of the sentence. In the second set, the definition was according to the parsing algorithm. It seems to be more robust and to represent the speech rate more accurately.

5.8 Statistical and time-related metrics

The vocal features extracted from the speech signal reduce the amount of data because they are defined on (overlapping) frames, thus creating an array for each of the calculated features. However, these arrays are still very long and cannot be easily represented or interpreted. Two types of secondary metrics have been extracted from each of the vocal features. They can be divided roughly into statistical metrics which are calculated for the whole utterance, such as maximum, mean, standard deviation, median and range, and to time-related metrics, which are calculated according to different duration properties of the vocal features and according to the parsing, and their statistical properties. Figure 5-9 demonstrates the relations between the sampled signal, the vocal features which are extracted from overlapping time frames of the speech signal and the secondary metrics which are calculated from these features throughout the utterance.

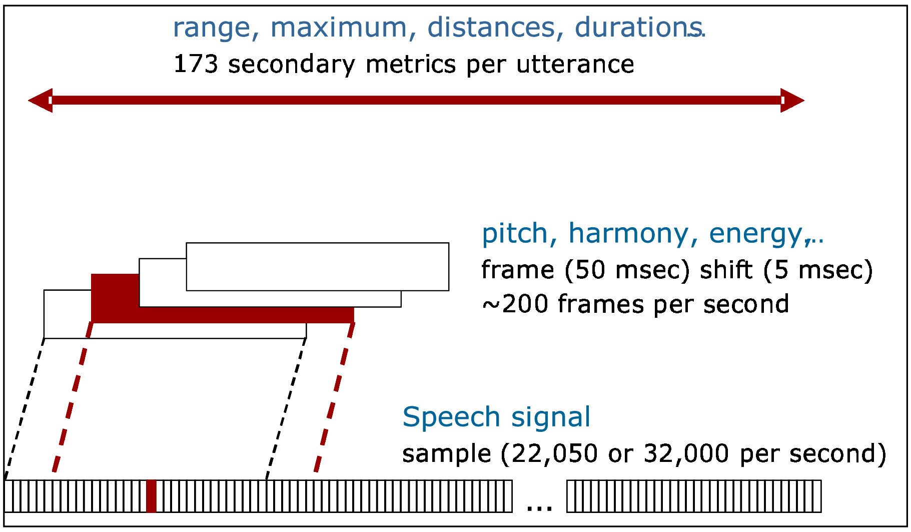
range, maximum, distances, durations..
173 secondary metrics per utterance
pitch, harmony, energy,..
frame (50 msec) shift (5 msec)
~200 frames per second
Speech signal
sample (22,050 or 32,000 per second)
...

Figure 5-9: A schematic description of the temporal relations between the original sampled speech signal, the vocal features extracted per frame, and the secondary metrics which are extracted from these features.

Feature definition

In order to locate the significant properties of the vocal features several types of observations have been made on both datasets. The Doors database was very helpful because it allowed observation of similar sentences, by the same speakers with different expressions, so that only the changes that evolve from the expression change are observed. For this purpose two types of comparisons were made - between different expressions and between consecutive utterances with subtle and gradual expression changes. The changes along the interaction are discussed in more detail in Chapter 8. The main types of observations were made using visualisation of the time-frequency domain of the speech signals and of the pitch contours. A few examples are presented in this section.

Time-frequency variations

A way of visualising the differences among expressions in speech is to use spectrograms, which are presentations of the Short-Time Fourier Transform over overlapping windows of the signal. Although the fundamental frequency, which reveals the intonation, is not very clear in this representation, it reveals other parameters of time and frequency variability which cannot be as clearly observed using other processing methods. An example is shown in Figure 5-10.

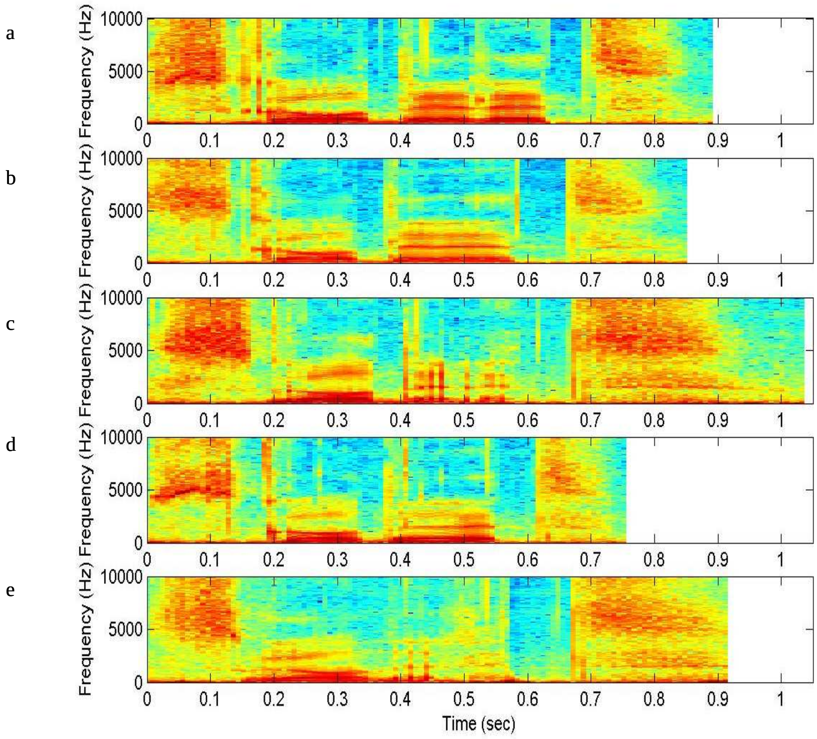

Figure 5-10: Spectrograms of the sentence *sgor de-le-t* (close door), uttered by the same person, with naturally evoked expressions of uncertainty (a,b), testing (c), cheered (d), and down (e). The colours represent energy levels. Red for high energy, decreasing through orange, yellow, green to turquoise and blue.

The five spectrograms are of sentences of *sgor de-le-t* , meaning 'close door' in Hebrew, uttered by the same speaker, with the expression labelled as *uncertainty.* The first two are of the same labelled expression of *uncertainty*, the third relates to an expression of *testing*, the fourth is *cheered* and the last one relates to *down*, or *withdrawn.* It can be seen that although the first two utterances are not identical, they are very similar to each other, especially when compared to the rest of he samples.

The significant parameters are the sentence length, the length of each uttered part, and the distances among these parts. The intensity or the amount of energy in each uttered part, and the distribution of the energy along the frequencies' spectrum, in various points in time. Whereas *cheered* is signified by a short sentence, short units and very distinct energy bands (red). *Testing* reveals a broken third part and a lengthy and blurred ending part, all the parts of this expression are intensified, especially the last part. *Down*, as may be expected, is pronounced in a longer sentence, with longer uttered parts and longer pauses, the uttered parts contain less energy, and the third uttered part nearly disappears.

These observations clearly show the relevance of duration and intensity of the whole signal and of certain frequency bands to characterisation of expressions and subtle expressions.

Pitch contour (Intonation)

An example for the complexity of pitch contours can be seen in Figure 5-11. It shows four pitch contours extracted from utterances of the same text, uttered by the same male speaker, but with different expressions.

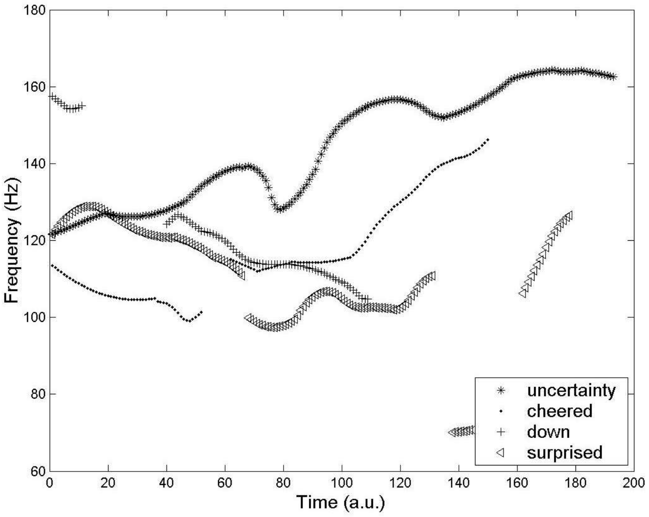

Figure 5-11: Pitch contours of four utterances, of the same text, by the same male speaker, with four different expressions: *uncertain, cheered, down* (subdued), and *surprised.*

It can be seen that although global characteristics like pitch range and utterance duration are of similar magnitude (the actual utterances start before the beginning of the voiced parts, which are shown in this graph), the local patterns change drastically between different curves. Even if the slopes of the whole contours have the same tendency, there are local characteristics that can be used to distinguish between them. For example, *surprised*, *cheered* and *uncertain* have the same rising slope towards the end of the utterances. The intonation of the *surprise* expression starts with a sharp rising slope, which changes to a descending slope, and then the curve bands up again, with several minor fluctuations along the curve, and in a few separated sub-curves. The *uncertainty* curve looks like rising steps, and the voiced part is continuous throughout the utterance, while the intonation curve of the *cheered* expression descends and then rises, with relatively few fluctuations along the curve.

These observations reveal that the significant characteristics can be gathered from durations and the relation between pitch frequencies in different parts of the utterance. It can also be seen that it is not enough to describe a contour as rising, because contours rise in different manners to convey different expressions. In a way it is similar to music. For example, people detect when a sentence is ended at a 'wrong' note or tone. There are certain relations between consecutive tones that are more pleasant than others and used in different contexts. Pitch contours can change by an octave, or double the frequency on occasions. However, I have not managed to find a precise manner to describe these relations mathematically as done in western music, and therefore decided to use the extreme values of pitch at the locations of extreme values of the signal's energy, the relations between the values, durations and the distances (in time) between consecutive extreme values.

Feature sets

I have examined mainly two sets of features and definitions. The first set, listed in Table 5-2 was used for initial observations, and it was improved and extended to the final version with which the inference system was defined, as listed in Table 5-3.

The final set includes the following secondary metrics of pitch: voiced length, voiced length - the duration of instances in which the pitch is not zero, and unvoiced length, in which there is no pitch. Statistical properties of its frequency were considered in addition to up and down slopes of the pitch, i.e. the first derivative or the differences in pitch value between adjacent time frames. Finally, analysis of local extremum (maximum) peaks was added, including the frequency at the peaks, the differences in frequency between adjacent peaks (maximum-maximum and maximum-minimum), the distances between them in time and speech rate.

Similar examination was done for the energy (smoothed energy with window), including the value, the local maximum values, and the distances in time and value between adjacent local extreme values. Another aspect of the energy was to evaluate the shape of the energy peak, or how the energy changes in time. The calculation was to find the relations of the energy peaks to rectangles which are defined by the peak maximum value and its duration or length. This metric gives a rough estimate for the nature of changes in time and the amount of energy invested.

Temporal characteristics were estimated also in terms of 'tempo', or more precisely in this case, with different aspects of speech rate. Assuming, based on observations and music related literature that the tempo is set according to a basic duration unit whose products are repeated throughout an utterance, and this rate changes between expressions and different speech parts of the utterance. The assumption is that different patterns and combinations of these relative durations play a role in the expression.

The initial stage was to gather the general statistics and check if it is enough for inference, which proved to be the case. Further analysis should be done for accurate synthesis. The 'tempo' related metrics used here include the shortest part with pitch, that is the shortest segment around an energy peak that includes also pitch, the relative durations of silence to the shortest part, the relative duration of energy and no pitch and the relative durations of voiced parts.

Feature #	**Feature name**	**mean**	**std**	**range**	**med**	**max**	**up**	**1^{st}positive derivative**	**1^{st} negative derivative**	**Relative part**
	Pitch features									
1	Speech rate									
2-3	Voiced length	√	√							
4-5	Unvoiced length	√	√							
6-13	Pitch	√	√	√	√	√	√	√	√	
14-17	Pitch maxima	√	√	√	√					
18-21	Pitch minima	√	√	√	√					
22-25	Pitch extrema distances (time)	√	√	√	√					
	Energy features									
26-29	Energy	√	√	√	√					
30-32	Smoothed energy						√	√	√	
33-36	Energy maxima	√	√	√	√					
37-40	Energy maxima distances (time)	√	√	√	√					
	Energy in bands									
41-45	0-500Hz	√	√	√	√					√
46-50	500-1000Hz	√	√	√	√					√
51-55	1000-2000Hz	√	√	√	√					√
56-60	2000-3000Hz	√	√	√	√					√
61-65	3000-4000Hz	√	√	√	√					√
66-70	4000-5000Hz	√	√	√	√					√
71-75	5000-7000Hz	√	√	√	√					√
76-80	7000-9000Hz	√	√	√	√					√

Table 5-2: Extracted speech features, divided to pitch related features energy in time and energy in frequency bands. The ticked boxes signify which of the following was calculated for each extracted feature: mean, standard deviation, range, median, maximal value, relative length of increasing tendency, mean of 1st derivative positive values (up slope), mean of 1^{st} derivative negative values (down slope), and relative part of the total energy.

The harmonic related features include a measure of 'harmonicity', which is the sum of harmonic intervals in the utterance, the number of frames in which each of the harmonic intervals appeared (as in Table 5-1), the number of appearances of the intervals that are associated with consonance and those that are associated with dissonance and the sub-harmonies.

Feature #	Name	Description	N°	mean	std	median	range	max	min
	Pitch								
1	Speech rate	$\frac{Voiced\_length}{Sentence\_length}$							
2-3	voiced length	$(pitch\_ends_n - pitch\_starts_n) \cdot shift$		✓	✓				
4-5	unvoiced length	$(pitch\_starts_n - pitch\_ends_{n-1}) \cdot shift$ if there is an unvoiced part before the start of pith it is added		✓	✓				
6-10	Pitch value	Value of $pitch$ when $pitch > 0$	✓		✓	✓	✓	✓	
11-12	up slopes	$(pitch_n - pitch_{n-1}) > 0$	✓			✓			
13-14	down slopes	$(pitch_n - pitch_{n-1}) < 0$	✓			✓			
15-17	max pitch	Maximum pitch values			✓	✓	✓		
18-20	min pitch	Minimum pitch values (non zero)			✓	✓	✓		
21-23	max jumps	Difference between adjacent maximum pitch values			✓	✓	✓		
24-26	extreme jumps	Difference between adjacent extreme pitch values (maximum and minimum)			✓	✓	✓		
27-30	max dist	Distances (time) between pitch peaks			✓	✓	✓		✓
31-34	extreme dist	Distances (time) between pitch extremes			✓	✓	✓		✓
	Energy								
35-38	Energy value	Smoothed energy+ window			✓	✓	✓	✓	
39-41	max energy	Value of energy at maximum peaks			✓	✓	✓		
42-44	energy max jumps	Differences of energy value between adjacent maximum peaks			✓	✓	✓		
45-47	energy max dist	Distances (time) between adjacent energy maximum peaks			✓	✓	✓		
48-50	energy extr jumps	Differences of energy value between adjacent extreme peaks			✓	✓	✓		
51-53	energy extr dist	Distances (time) between adjacent energy extreme peaks			✓	✓	✓		
	'Tempo'								
54	shortest part with pitch	$min(parts\,that\,have\,pitch)$							
55-58	*'tempo'* of silence	$\left(\frac{Silence\_parts\_lengths}{shortest\_part}\right)$			✓	✓	✓		✓
59-62	*'tempo'* of energy and no pitch	$\left(\frac{energy\_no\_pitch\_parts\_lengths}{shortest\_part}\right)$			✓	✓	✓		✓
63-66	*'tempo'* of pitch	$\left(\frac{pitch\_parts\_lengths}{shortest\_part}\right)$			✓	✓	✓		✓
67-70	resemblence of energy peaks to squares	$\left(\frac{energy\_peak\_area}{peak\_max \cdot peak\_duration}\right)$			✓	✓	✓		✓

Feature #	Name	Description	N°	mean	std	median	range	max	min
	Harmonic properties								
71	harmonicity	$\left(\frac{number\_of\_harmonic\_parts}{length(pitch)}\right)$ number of frames with pitch and harmonic interval	√						
72-83	harmonic intervals	Number of frames with each of the harmonic intervals	√						
84	consonance	Number of frames with intervals that are associated with consonance	√						
85	dissonance	Number of frames with intervals that are associated with dissonance	√						
86-89	sub-harmonies	Number of sub-harmonies per frame			√	√	√	√	
	Filter-bank								
90-93	central frequency	101 Hz			√	√	√	√	
94-97	central frequency	204 Hz			√	√	√	√	
98-101	central frequency	309 Hz			√	√	√	√	
102-105	central frequency	417 Hz			√	√	√	√	
106-109	central frequency	531 Hz			√	√	√	√	
110-113	central frequency	651 Hz			√	√	√	√	
114-117	central frequency	781 Hz			√	√	√	√	
118-121	central frequency	922 Hz			√	√	√	√	
142-145	central frequency	1079 Hz			√	√	√	√	
142-145	central frequency	1255 Hz			√	√	√	√	
142-145	central frequency	1456 Hz			√	√	√	√	
142-145	central frequency	1691 Hz			√	√	√	√	
142-145	central frequency	1968 Hz			√	√	√	√	
142-145	central frequency	2302 Hz			√	√	√	√	
146-149	central frequency	2711 Hz			√	√	√	√	
150-153	central frequency	3212 Hz			√	√	√	√	
154-157	central frequency	3822 Hz			√	√	√	√	
158-161	central frequency	4554 Hz			√	√	√	√	
162-165	central frequency	5412 Hz			√	√	√	√	
166-169	central frequency	6414 Hz			√	√	√	√	
170-173	central frequency	7617 Hz			√	√	√	√	

Table 5-3: The set of secondary metrics, extracted from the vocal features, which were used for the inference machine.

The last group includes the filter bank and statistic properties of the energy in each frequency band. The centres of the bands are at 101 ,204, 309, 417,531, 651, 781, 922, 1079, 1255, 1456, 1691, 1968, 2302, 2711, 3212, 3822, 4554, 5412, 6414 and 7617 Hz. Although the sampling rate in both databases allowed for frequency range that reaches beyond 10 kHz, the recording equipment not necessarily does, therefore no further bands were required.

5.9 Summary

This chapter presents the vocal features and secondary metrics which were defined and used in this research for the inference of expressions from non-verbal speech. The motivation for defining features was taken from various disciplines such as psycho-acoustic research, linguistics and musicology. Some of the features presented here are commonly used in speech technologies, some have new or modified extraction algorithms and others are new and have not been used before in this respect. The secondary metrics are calculated from the extracted vocal features and represent temporal and statistical properties of these features. These features and metrics may be beneficial to other speech technologies, especially to affective synthesis. The next chapter presents their use for the inference of expressions from non-verbal speech.

References

[1] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., and Taylor J. G., "Emotion recognition in human-computer interaction", *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[2] Nass C. and Brave S., *Wired for Speech: How voice activates and advances the human-Computer relationship*, The MIT Press, 2005.

[3] Murray I. R. and Arnott J. L., "Toward the simulation of emotion in synthetic speech: a review of the literature on human vocal emotion", *Journal of the Acoustical Society of America*, vol. 93, no. 2, pp. 1097–108, 1993.

[4] Petrushin V., *Intelligent Engineering Systems Through Artificial Neural Networks*, chapter Emotion in speech: Recognition and application to call centers, pp. 1085–1092, 1999.

[5] Terhardt E., Stoll G., and M. Seewann, "Algorithm for extraction of pitch and pitch salience from complex tonal signals", *J. Acoust. Sec. Amer.*, vol. 71, pp. 679, 1982.

[6] Moore C. A., Cohn J. F., and Katz G. S., "Quantitative description and differentiation of fundamental frequency contours", *Computer Speech and Language*, vol. 8, no. 4, pp. 385–404, 1994.

[7] Markel J.D. and Gray A.H. Jr., *Linear Prediction of Speech*, Springer, Berlin (west), 1976.

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

[9] Zhao W. W. and Ogunfunmi T., "Formant and pitch detection using time-frequency distribution", *International Journal of Speech Technology*, vol. 3, no. 1, pp. 35–49, 1999.

[10] Boersma P., "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.

[11] Fernandez R., PhD thesis, Media Arts and Sciences, Massachusetss Institute of Technology, 2004.

[12] Dellaert F., Polzin Th., and Waibel A., "Recognizing emotions in speech", in *ICSLP 96*, 1996.

[13] Oudeyer P. Y., "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.

[14] "http://en.wikipedia.org/wiki/autocorrelation".

[15] Boersma P. and Weenink D., "Praat: doing phonetics by computer, www.pratt.org".

[16] Stevens S. S., Volkmann J., and Newman E. B., “A scale for the measurement of the psychological magnitude pitch”, *The Journal of the Acoustical Society of America*, vol. Volume 8, no. 3, pp. 185–190, 1937.

[17] Zwicker E., Flottorp G., and Stevens S. S., “Critical bandwidth in loudness summation.”, *The Journal of the Acoustical Society of America*, vol. Volume 29, pp. 548–57, 1961.

[18] Zwicker E., “Subdivision of the audible frequency range into critical bands (frequenzgruppen)”, *The Journal of the Acoustical Society of America*, vol. Volume 33, pp. 248, 1961.

[19] Iamblichus, *On the Pythagorean life*, Clark G. translator Liverpool, c300/1996.

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

[21] Galileo, *Dialogues Concerning Two New Sciences*, Dover Publications Inc., New-York 1638, 1954.

[22] Schartz D. A., Howe Q. C., and Purves D., “The statistical structure of human speech sounds predicts musical universals”, *The Journal of Neuroscience*, vol. 23, pp. 7160–7168, 2003.

[23] Tramo M. J., Cariani P. A., Delgutte B., and Braida L. D., *The cognitive neuroscience of music*, chapter Neurobiology of harmony perception, Oxford University Press, New-York, 2003.

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

[25] Terhardt E., “Pitch, consonance, and harmony”, *The Journal of the Acoustical Society of America*, vol. 55, no. 5, pp. 1061–1069, 1974.

[26] Plomp R. and Levelt W. J. M., “Tonal consonance and critical bandwidth”, *The Journal of the Acoustical Society of America*, vol. 38, pp. 548–560, 1965.

[27] Canny J., “A computational approach to edge detection”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[28] Feng Y. L., Xu Y. Q., Chang E., and Shum H. Y., “Speech driven cartoon animation with emotions”, in *ACM Multimedia, Otawa, Canada*, 2001.

Chapter 6

Utterance level inference system

This chapter presents the inference system that was designed and tested in this research. This inference machine was designed to answer most of the requirements of a global inference machine as stated in Chapter 3 and it deals successfully with the challenges involved. The developed machine is based on an underlying taxonomy of a few expressions that relate to groups of mental states, whose combinations represent a large variety of emotions and complex mental states and their related behavioural (vocal) cues. It recognises mixtures of expressions and can recognise asynchronous changes and subtle expressions. This approach to inference systems is novel because it uses different sets of features for different expressions, and does not assume one 'optimal' set of features that distinguishes between all expressions. This approach allows for a flexible and extendable machine, allowing integration of new machines that use different sets of features, tested on different datasets and more, without affecting the existing machine. It can easily be adapted to new expressions and to new speakers.

This chapter describes the inference machine and the algorithms upon which it is based. It also describes some of the most common approaches and preliminary experiments and observations that were carried out before the inference system was constructed in order to explore methods and requirements. The inference machine and the initial experiments were built using the data from the Doors and the MindReading databases that are described in Chapter 4, using the feature sets described in Chapter 5.

6.1 Background

The fourth stage of the framework for the inference of expressions, which is described in Chapter 3, comprises the inference of an expression or expressions from the current utterance in an interaction or in a stand-alone manner, which is the common practice. It gets as inputs the statistical and temporal metrics of the vocal features that are calculated over the whole utterance. In more complicated systems it can include feedback from the interaction level analysis in order to enhance the accuracy of the recognition of each utterance.

The most common approach to inference systems is to generate one machine or several models of a single processing method, using one set of prosody features in order to recognise one expression such as stress or a small set of basic emotions. The first analysis issue is feature selection and dimension reduction, i.e. trying to define which properties of the speech signal are the most relevant for expression recognition. The main method used for feature selection is to use classification results. For example, Dellaert [1] used three classification methods and three selection methods. No unsupervised or semi-supervised technique was examined in that research. Oudeyer [2], who tried to locate the best algorithms for the task in addition to the most significant features, examined a large range of data-mining and machine-learning techniques. Most works are based on small sets of basic expressions, and assume that one set of features is suitable for the whole range of expressions.

6.2 Preliminary observations and results

This section describes some preliminary experiments, their results and how they contribute to the definition of the inference machine.

Blind clustering (unsupervised learning)

As a first stage, I tried a straightforward analysis, i.e. to use blind clustering and dimension reduction methods on the entire MindReading database and on whole sessions, of a 100 sentences each, from the Doors database. I have investigated several methods of blind clustering and dimension reduction. The first method is Principal Component Analysis (PCA), in which the principal components or vectors are linear combinations of the attributes (features) and are ranked according to their variance. Other methods included dimension reduction using fractals, based on the correlation dimension [3], which retains the original features, Expectation Maximization (EM), and more. However, these methods did not yield significant results because in both databases the speech utterances are mostly on a continuous scale, there are few extreme changes and most of the changes are gradual.

Figure 6-1 demonstrates this situation. It shows 25 utterances or sentences with the same text, uttered by the same male speaker, but with different expressions. Each utterance is a mark on the graph, the different expressions are marked with different colours and shapes. The expressions are *uncertain*, *cheered* and *enthusiastic*. The features for classification are the median of minimum pitch values, and the relative time of rising pitch. The features and rules for classification were derived by the C4.5 algorithm, using the J48 package of the Weka data-mining tool [4] (the algorithm is explained in section 6.3). It can be seen that some of the samples that were labelled manually as *uncertain*, are closer to the segments labelled as *cheered* or *enthusiastic*, more than to the other samples of *uncertainty*. These relative locations make blind clustering inefficient in this case.

The conclusions were that for the MindReading database larger groups of related concepts should be defined and that other analysis methods are required (supervised learning).

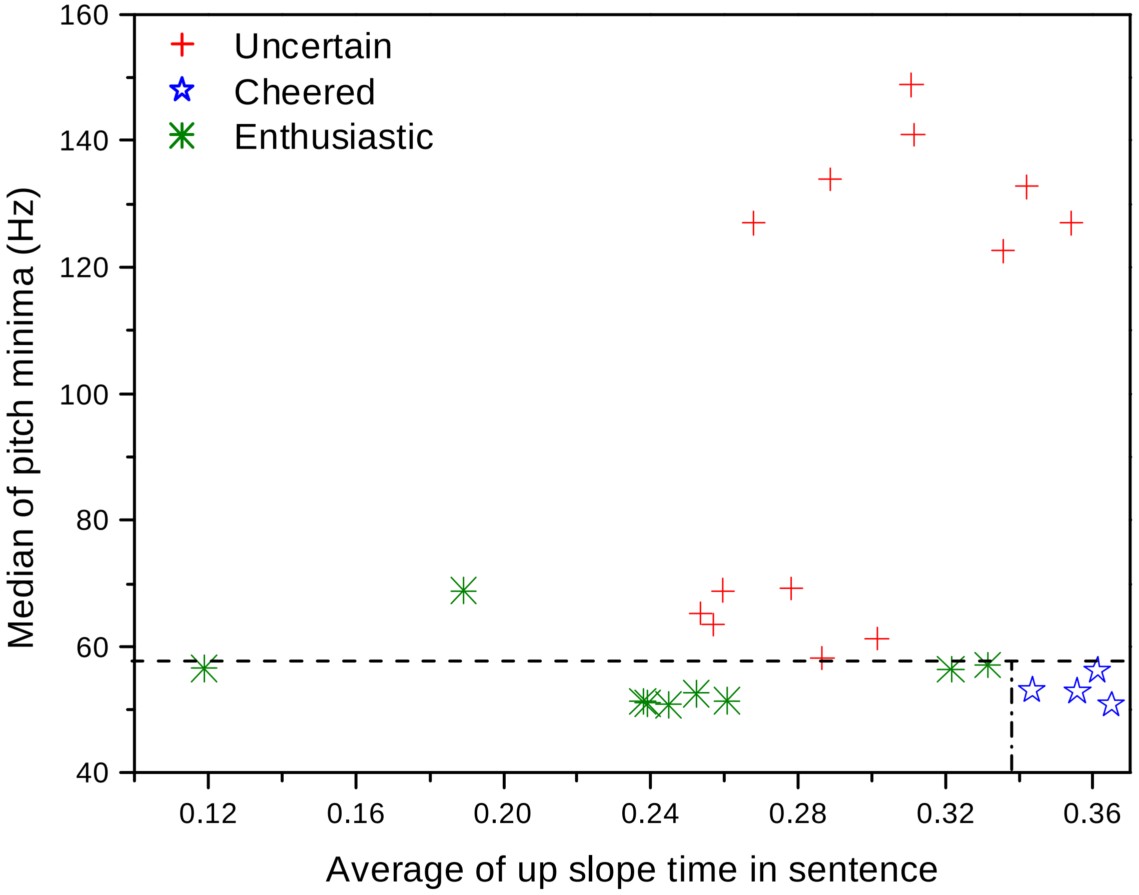

Figure 6-1: The expressions of *uncertainty*, *cheered*, and *enthusiastic*, found by the J48 classification package of Weka [4]. The transitions between expressions are marked by dashed partitioning lines. The features that distinguish between the groups are the median of local minimum pitch values, and the relative time of pitch rising. The samples were taken from an utterance of a male speaker in the Doors database.

Subtle expressions in HCI

Several stages of experiments and subsequent observations were performed in order to better understand and define the requirements from an inference system for complex mental states, and to verify its feasibility. The first stage was to verify that subtle expressions and nuances could be inferred from the non-verbal aspects of speech.

I have examined the sentences uttered by participants of the Doors database, which were manually labelled, and demonstrated that the situation of human computer interaction evokes a variety of expressions. These expressions are related not only to the narrow definition of emotions but also to mental states and attitudes. The labels included labels such as *enthusiastic* (or *vital*), *cheered* (or *encouraged*), *down* (meaning *remote* or *subdued* due to *boredom*, *tiredness* and maybe *disappointment*), *uncertain*, *choosing*, *calculating*, *testing*, *thinking* and more. These labels define subtle expressions that have not been examined before in this context. The labels were found to be highly related to the recorded events of the Doors game.

The experiments showed that these expressions are distinguishable using the non-verbal cues of speech such as statistical features of pitch, energy and spectral content, that were calculated over the whole utterance [5]. Although the recognition was good, the need for other features that further reflect voice quality and time-related changes was noticed. The initial experiment demonstrated the requirements and potential abilities of a more general machine based on labelled samples (utterances) from more speakers.

Features

Time-frequency visualisation of speech signals and sequences of pitch contours of sentences with the same text but with different expressions confirmed the initial assumption that different expressions are characterised by different paralinguistic features. As can be seen in Figures 5-10 and Figure 5-11 in the previous chapter.

In many cases a certain expression can share certain features properties with another (second) expression while the same features distinguish between this expression and a (third) different expression. Therefore, different sets of features are required in order to distinguish between different expressions.

The implementation of the inference machine was built according to these preliminary conclusions with regard to the requirements and challenges, as described in Chapter 3. The machine is described in the next sections.

6.3 The inference machine

This section describes the architecture of the inference machine, the expressions that were chosen to demonstrate its abilities, the algorithms and the paralinguistic features used in its application.

Expressions

The inference machine was built using groups of concepts from the MindReading database. The groups were chosen by their relevance to the expressions recognised in the Doors database. They represent basic emotions, attitudes and complex mental states. The chosen set of expressions include: *joyful*, *absorbed*, *sure*, *stressed*, *excited*, *opposed*, *interested*, *unsure* and *thinking*. Table 6-1 lists the concepts that were used for training each of these groups. It also states the groups to which these expressions belong according to the MindReading taxonomy.

Most of the expression groups correspond to one of the taxonomy concept groups. *Absorbed* (*concentrating*) and *interested* are both taken from the *interested* group. Typically though, in the literature of cognition and emotion, they are referred to separate mental states [6–8]. I have selected two classes from the same group to test the system's ability to distinguish between finer shades of mental states. Also, *absorbed* denotes that the user's attention is directed towards himself, while *interested* refers to attention directed outward and to questions. Kaliouby made similar distinction for inference of mental states from facial and head gestures [9].

The *unsure* group includes concepts from the *surprised* and *thinking* groups in addition to concepts from the *unsure* group. These concepts have lexical definitions that include *unsure* as synonym. The *stressed* group includes mostly concepts from the *bothered* group and concepts from the *afraid* group whose lexical definitions are synonym to *bothered* or *stressed*, such as worried.

Expression	Concepts	Group
joyful	tickled, carefree, amusing, overjoyed, festive, merry, delighted, enjoying, felicity, glad, happy, joking, joyful, jubilant, rejoicing, triumphant	happy
absorbed	absorbed, engaged, committed, concentrating, focused, thorough, involved	interested
sure	adamant, assertive, confident, sure, convinced, decided, knowing, determined, resolved	sure
stressed	bothered, hurried, hampered, overwrought, overrun, pressured, rushed, stressed, flustered, impatient, tense, pestered, restless, turmoil, worried	bothered, afraid
excited	alert, dynamic, lively, exhilarated, excited, inspired, invigorated, adventurous	excited
opposed	argumentative, confrontational, contradictory, contrary, disagreeing, disapproving, disinclined	unfriendly
interested	asking, curious, fascinated, probing, questioning, quizzical, scrutinizing, interested	interested
unsure	confused, clueless, faltering, hesitant, indecisive, unsure, undecided, insecure, ambivalent, puzzled, baffled, considering, debating	unsure, surprised, thinking
thinking	fantasising, thinking, thoughtful, brooding, choosing, deciding, wool-gathering, calculating, comprehending, realising	think

Table 6-1: The 9 mental state groups that constitute the expression groups recognised by the inference machine (left), the corresponding concepts, taken from the MindReading database that were used for training the machine (middle), and the concept groups as defined by the MindReading taxonomy, from which they were taken (right).

6.4 Architecture

The inference machine was designed in two-level architecture, as described in Figure 6-2.

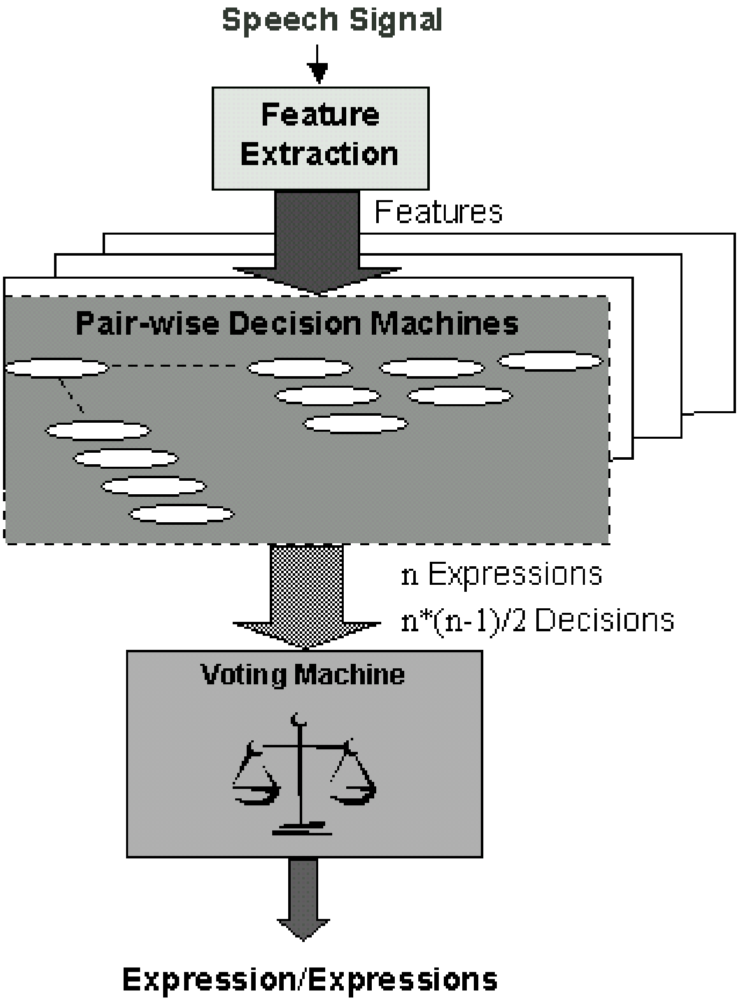

Figure 6-2: Schematic description of the inference machine. The machine extracts features from the input speech signal. Different groups of these features are used as attributes for the different pair-wise decision machines that decide between every pair of expressions which is the most likely. The decisions of all the pair-wise machines are fed into a voting machine that decides which are the most probable expressions which can be related to the processed speech signal.

Pair-wise decision machines

The first level consists of a set of pair-wise decision machines. Each machine distinguishes between a pair of expressions. If the number of expressions to be recognisable by the machine is n, the number of pair-wise machines is:

$$(n-1)+(n-2)+\ldots+1=\frac{n\cdot(n-1)}{2}$$

The results of the pair-wise machines enter into a voting machine which decides which of the expressions were elected by the majority of the machines. Two voting methods were examined in this work.

A set of 173 prosody features was used for the implementation of the machines, as described in Table 5-3. The features were normalised per speaker. For each speaker, each of the features was normalised with mean 0 and standard deviation 1. It means that for new speakers the system requires only little training. The training data included sentences uttered by 10 speakers (males, females, adults and children) and different text in each utterance.

In order to build the pair-wise machines two supervised classification mechanisms were used: C4.5 decision-tree, as implemented in the J48 package of the Weka data-mining tool [4], and linear Support Vector Machines (SVM), using the Weka tool [4]. The algorithms, which are described below, were chosen for the simplicity of the decision-machines implementation. For each pair of expressions the two classification schemes were examined, and the one with the best performance, measured by precision and ten-fold cross-validation (with high True-Positive for both expressions) was chosen. The two different types of algorithms were chosen to demonstrate the flexibility and extendibility of the system and its ability to integrate systems that are built with different data, features and algorithms. Other algorithms have been examined, including SVM with Gaussian and Polynomial kernels, they yielded similar classification results but their implementation is more complicated and therefore they were not pursued.

C4.5

The C4.5 and C5.0 algorithms, which were developed by Quinlan [10], construct a decision tree using a divide and conquer strategy. It starts with a root in which all the instances set, that is the set of all the labelled speech utterances, are in one group. From the root the instances are split into two sub-groups. Each of these sub-groups is a node that can be further split into sub-groups recursively.

At each node one attribute (in this case one paralinguistic feature) defines the test by which the instances at the node are split by. The two sub-groups are the sets of instances with attribute value above and below a certain local threshold. The attribute with the highest information gain is selected to be the splitting test at the node. The information gain is relative to the splitting of the instances at the node. It has maximal value if the sub-groups are of the same size.

The classification error of a node is calculated as the sum of the errors of the child nodes. If the result is greater than the error of classifying all the instances at the node as belonging to the most frequent class, then the node is set to be a leaf, and all sub-trees are removed. Ideally the process terminates when all the instances at the leaf-nodes belong to the same class (or when the data cannot be further split). Pruning is required in order to avoid over-fitting.

An example of a decision-tree pair-wise machine can be seen in Figure 6-3. It represents the distribution of the class labels *joyful* and *absorbed* using the vocal features as attributes. The top node represents all the data. The training set consists of 37 samples (sentences) for the *joyful* class and 41 samples for the *absorbed* class. The classification tree algorithm concludes that the best way to separate between the two classes is by using the variable (attribute, or feature) ”Energy range below 100Hz”, and its threshold value for decision is -0.19 (using the normalised features). This variable can have two values ‘yes’ or ‘no’. The training set is separated into two branches. In the ‘yes’ branch there are 33 *joyful* samples and 10 *absorbed* samples, while in the ‘no’ branch there are 4 *joyful* samples and 31 *absorbed* samples. Each of these branches is then separated, in the same manner but with different variables, until all the samples are separated according to classes, as best as possible. This last stage is the leaves level of the tree. There are 6 leaves. Only 3 of the 78 samples were wrongly classified in the training set.

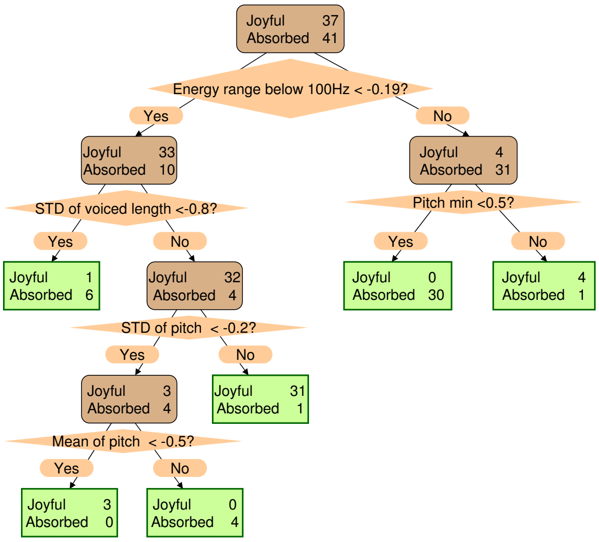

Figure 6-3: The decision-tree pair-wise machine that distinguishes between the vocal expressions associated with the mental states *joyful* and *absorbed*. The top node is the root that represents all the data. The intermediate branches are in brown, and the leaves, the end of the classification process, are in green (like an upside-down tree).

Support vector machines

Support Vector Machines were invented by Vapnik in 1979 [11]. In its simplest, linear form, an SVM is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin. In the linear case, the margin is defined by the distance of the hyperplane to the nearest of the positive and negative examples. The formula for the output of a linear SVM is: $u = w \cdot x - b$
w is the normal vector to the hyperplane and x is the input vector. The separating hyperplane is the plane $u = 0$. The nearest points lie on the planes $u = \pm 1$. The margin m is thus: $m = \frac{2}{|w|}$ so we want to minimise w. A schematic description of SVM hyperplanes can be seen in Figure 6-4.

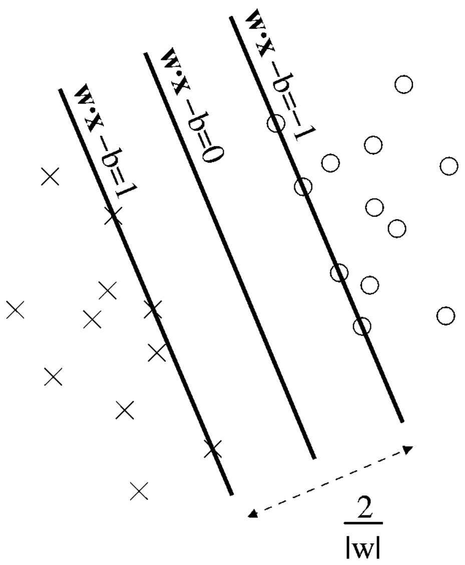

Figure 6-4: Support vector machine optimal hyperplanes and training samples [12].

After the SVM has been trained, it can be used to classify unseen 'test' data. This is achieved using the following decision rule;

$$C = \begin{cases} 1, & if\ w \cdot x + b \geq 0 \\ -1, & if\ w \cdot x + b \leq 0 \end{cases}$$

If the multiplication of the attribute values (features) that were extracted and calculated from the utterance, with the vector coefficients that were derived from the training, is equal or greater than zero, the instance (utterance) is classified into one of the classes ($C = 1$) and if it is smaller it belongs to the other class ($C = -1$).

Feature selection and dimension reduction

The goal was to create relatively compact and efficient machines, rather than use all the 173 features. It was also desirable to find and define the features that distinguish expressions, as much as possible, without compromising the machine efficiency. Therefore, the chosen machines were not necessarily the optimal machines for the data (if such machines exist), but rather the compact machines that yielded good enough results. Each machine was built of a relatively small set of features, as can be seen in Table 6-2.

Another hypothesis was that different features distinguish different expressions, or different expressions are characterised by different features, and one set of features cannot characterise all the expressions (unless it is very large).

This approach was found to be correct. Different features were required for different pairs of expressions. In order to check the validity of this observation, I checked how borrowing features that distinguish between one pair of expressions distinguish between another set of expressions. The result was that often a set of features that yielded near optimal results for one pair of expressions yielded no more than chance for another pair of expressions. The two algorithms also required different sets of features for the same expression pairs. For example, a set of features that was used to distinguish between *sure* and *absorbed* yielded precision of 89% using the C4.5 algorithm, 81% using SVM and only 73% for the expression pair *sure* and *excited*, using the C4.5 algorithm (the feature numbers are: 3,19,26,48,86,90 and 96, their descriptions appear in detail in Table 5-3). In a few cases, two different sets of features yielded nearly similar results for the same type of algorithm. The feature selection method depends on the classification algorithm. Feature selection is an inherent feature of the C4.5 algorithm, in which properties of features are the parameters for decision making. An additional stage of feature selection was required for SVMs. Various methods were used to choose a relatively small set of features, including using several arbitrary attribute selection algorithms of Weka, integrating sets of features from tree classifiers and other methods, choosing the most significant features of SVM machine with the full set of features, choosing features by meaning groups, such as pitch related features, tempo, harmonicity and spectral content, and other methods and combinations.

This new approach allows a combined inference engine to be based on different algorithms and different sets of features for different expressions and enables integration of systems that were developed in different manners for expendable and flexible machines that can be used for a large variety of applications.

Pair-wise decision machines - summary

Table 6-2 summarises the pair-wise machines' properties. For each expression pair it denotes the classification precision (percent), tenfold cross-validation and ROC area that were calculated by Weka [4] using the training data. The table describes the type of machine that was used for the classification: decision-tree (C4.5) or SVM, and the number of features or attributes that were used. The average number of features is 10. There are 6 SVM-based machines and 30 tree-based machines.

Features significance

The next stage was to check the significance of the various features for the classification. One of the methods to measure significance is to count the number of machines in which each feature appears. The features that appeared in the largest number of machines are harmonicity measure (feature no. 71 from the feature list in Table 5-3) that appeared in 15 of the 36 pair-wise machines, which indicates voice quality, pitch median (6), and standard deviation of the energy in the first filter-band (91) that appeared 11 times each. Minimum value of energy durations where there is no pitch (59), and the range of the energy in the first and second filter bands (92, 96) appeared in 8 machines each. The features that did not appear in any of the machines were 3 of the 12 harmonic intervals (78, 80 and the 83), length of pitch down slopes and the range of the sub-harmonies, in addition to properties of several of the spectral-bands.

	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
joyful		P: 95 C: 82 R:0.949 SVM 19	P: 98 C: 83 R:0.994 Tree 8	P: 98 C: 61 R:0.998 Tree 18	P: 96 C: 60 R:0.982 Tree 8	P: 99 C: 71 R:0.997 Tree 7	P: 90 C: 77 R:0.898 SVM 40	P: 99 C: 75 R:0.997 Tree 8	P: 98 C: 72 R:0.993 Tree 13
absorbed			P: 90 C: 84 R: 0.9 SVM 12	P: 98 C: 87 R:0.983 Tree 5	P: 96 C: 81 R:0.978 Tree 6	P: 98 C: 78 R:0.989 Tree 6	P: 96 C: 82 R:0.98 Tree 5	P: 99 C: 64 R:0.996 Tree 6	P: 99 C: 73 R:0.999 Tree 10
sure				P: 90 C: 84 R: 0.9 SVM 12	P: 96 C: 79 R:0.981 Tree 7	P: 96 C: 72 R:0.99 Tree 8	P: 100 C: 78 R: 1 Tree 8	P: 100 C: 78 R: 1 Tree 8	P: 97 C:75 R:0.988 Tree 7
stressed					P: 99 C: 73 R:0.999 Tree 9	P: 96 C: 84 R:0.979 Tree 7	P: 99 C: 66 R:0.995 Tree 7	P: 99 C: 68 R:0.999 Tree 15	P: 98 C: 72 R:0.993 Tree 15
excited						P: 98 C: 74 R:0.988 Tree 9	P: 98 C: 71 R:0.99 Tree 9	P: 97 C: 64 R:0.994 Tree 8	P: 98 C: 79 R:0.997 Tree 8
opposed							P: 96 C:75 R:0.981 Tree 8	P: 82 C:79 R:0.817 SVM 6	P: 98 C: 81 R:0.987 Tree 6
interested								P: 98 C: 72 R:0.99 Tree 8	P: 99 C: 83 R:0.997 Tree 8
unsure									P: 89 C: 89 R:0.888 SVM 22
thinking									

Table 6-2: Details of the 36 pair-wise machines, including: P- precision (percent), C-cross-validation of 10% (percent), R-ROC area, machine type: Tree (C4.5) or SVM, and the number of features it includes.

The combined inference machine

Each expression can be recognised by 0-8 machines of the 36 pair-wise machines, therefore several mental-states can be recognised at the same time. The next stage is to combine the output of these machines to a set of inferred expressions that represent the emotions or mental states that are conveyed by the utterance. Two methods were considered for the combined machine. One uses the *Condorcet voting method* [13] and the second uses a threshold on the number of machines that recognised each expression.

Condercet voting method

A Condorcet method is a single winner election method in which voters rank candidates in order of preference. It is a voting system that always elects the candidate whom voters prefer to each of the other candidate. This candidate can be found by conducting a series of pair-wise comparisons.

In certain circumstances an election has no Condorcet winner. This occurs as a result of a kind of tie known as a 'majority rule cycle', described by Condorcet's paradox. The manner in which a winner is then chosen varies from one Condorcet method to another. Condorcet completion method is a method used to find a winner when there is no Condorcet winner. In voting systems, the Smith set [14, 15] is the smallest non-empty set of candidates in a particular election such that each member beats every other candidate outside the set in a pair-wise election. Ideally, this set consists of only one candidate, the Condorcet winner. Conversely, an occurrence of Condorcet's paradox implies that the set has more than one member. Voting systems that always elect a candidate from the Smith set pass the Smith criterion and are said to be "Smith-efficient". Smith-efficient methods necessarily meet the Condorcet criterion

Copeland's method [16] is a Condorcet method in which the winner is determined by finding the candidate with the most pair-wise victories. When there is no Condorcet winner (when there are multiple members of the Smith set), this method often leads to ties. For example, if there is a three-candidate majority rule cycle, each candidate will have exactly one loss, and there will be an unresolved tie between the three. Critics argue that it also puts too much emphasis on the quantity of pair-wise victories and defeats rather than their magnitudes.

According to the Condorcet method, only an expression that is chosen by all the largest number of machines should be inferred. However, if the maximum number of machines that recognise an expression is smaller than 8 (all the comparisons vs. other expressions), all the expressions that are recognised by the maximum number of machines are elected.

Threshold method

However, in this case it is not necessary to solve conflicts to get only one winning candidate, because several of the candidates can co-exist. The only exception in this particular system is a possible conflict between *sure* and *unsure*, which indicates that the level of confidence is not a significant parameter of the specific mental state or speech segment.

The second method infers expressions that are chosen by a number of machines which is above a threshold. The threshold is defined as one standard deviation above the mean number of machines, which means that at least six machines recognised an expression. Several expressions can be elected by 6-8 machines. Therefore a combination of expressions can be inferred.

6.5 Inference evaluation

Table 6-3 shows the confusion matrix of the machine that combines all the pair-wise inference machines, for 9 expressions of mental-states, using the Condorcet voting method. The probability of randomly choosing an expression is 11%. As can be seen all the expressions were recognised with a much higher rate than that. Table 6-4 shows the inference results of the threshold method. 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.

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

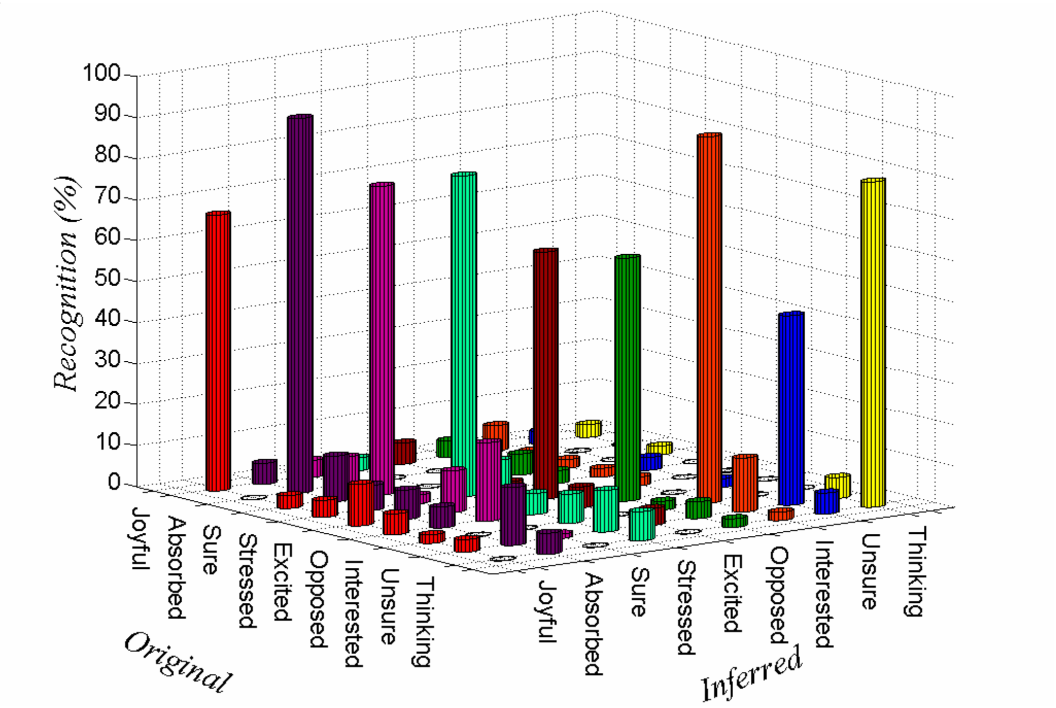

Table 6-3: Confusion matrix of the inference machine using the Condorcet method. The recognition is defined as the expression that was chosen by the maximum number of sub-machines. The Columns present the recognised mental-state. The rows represent the class of the analysed data. The results are presented as percent of recognition out of the introduced data class.

The pair-wise machines were trained using similar size groups of entities that were chosen randomly for each pair of expressions from a larger set of entities from the same classes, using 380 entities (sentences). The recognition accuracy of the combined system on the same set of entities was 81%. The testing was done using the full sets, using 546 speech segments, which in effect is 70%-30% split (between testing set (546) and the training sets (380)). In this set 79% of the expressions were recognised correctly.

This architecture was compared with another machine that built a decision tree based on the C4.5 algorithm to distinguish between all the expressions. The tenfold cross-validation of the single machine was close to the random probability, which is much worse than the cross-validation of the combined machine.

Furthermore, the approach of the combined system is preferable because it is more flexible. It can be extended without changing the existing structure and it can infer mixtures of expressions.

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

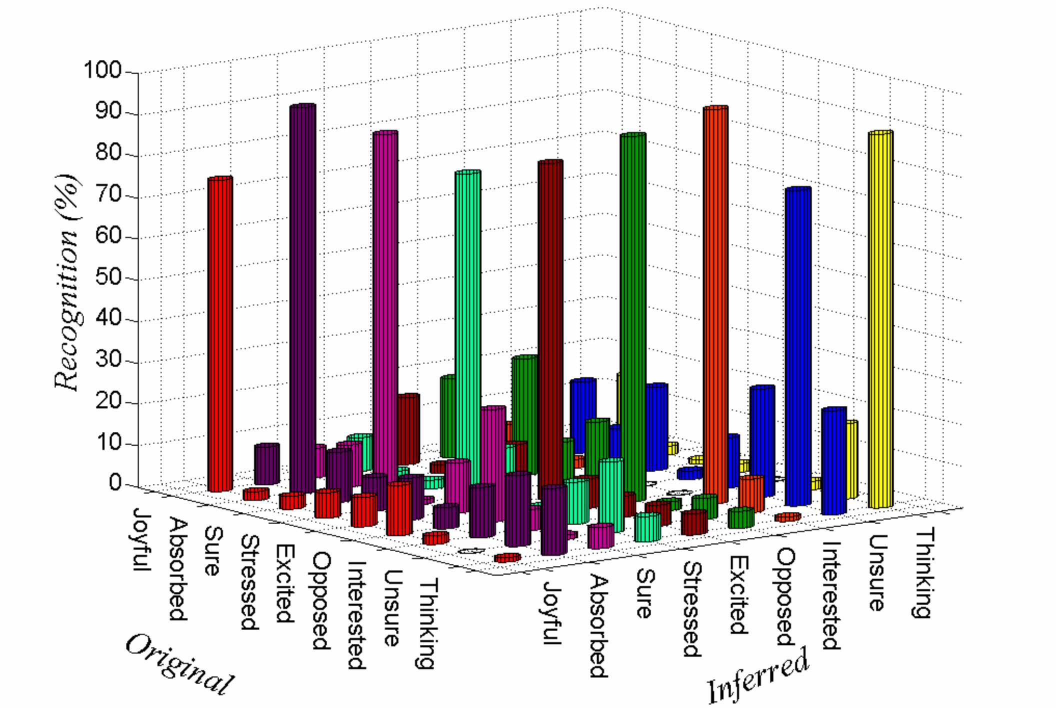

Table 6-4: Confusion matrix of the inference machine using the threshold method. The Columns present the recognised mental-state. The rows represent the class of the analysed data. The results are presented as percent of recognition out of the introduced data class. Tested on 546 speech segments, of which 432 recognised correctly.

6.6 Summary

This chapter presents the design and validation of an inference machine of expressions from non-verbal speech. This machine is the product of all the design and implementation layers that are described in the previous chapters. Its structure is flexible and allows extension to new speakers, behavioural cues, expressions and datasets. Its statistical validation results are high, around 80%.

The system is based on a novel approach which includes the recognition of multiple expressions beyond the set of basic emotions and the recognition of expression mixtures. The approach to implementation is based on the observation that different features characterise different expressions and the relations between them. The implementation itself consists of combination of various classification techniques in one system which contributes to its adaptability and expendability.

The inference system was further tested on speakers who were not part of the training set, on different expressions, text and even on another language. It was tested on acted text and on naturally evoked expressions, on stand-alone utterances and on utterances that are part of an on-going interaction. Its findings were compared to human performance, using the CAM Battery test as reference [17] and to other behavioural and contextual cues, as discussed in the next chapters.

References

[1] Dellaert F., Polzin Th., and Waibel A., “Recognizing emotions in speech”, in *ICSLP 96*, 1996.

[2] Oudeyer P. Y., “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.

[3] Caetano T. Jr., Agma T., Leejay W., and Christos F., “Fast feature selection using the fractal dimension”, in *XV Brazilian Symposium on Databases (SBBD), Paraiba, Brazil*, 2000.

[4] Witten I. H. and Frank E., *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[5] Sobol Shikler T. and Robinson P., “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.

[6] Ekman P., *Human Ethology*, chapter Brows: Emotional and Conversational Signals, pp. 169–200, London: Cambridge University Press, 1999.

[7] Ellsworth P. C., “Confusion, concentration and other emotions of interest”, *Emotion*, vol. 3, no. 1, pp. 81–85, 2003.

[8] Rozin P. and Cohen A. B., “High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of naturally occurring facial expressions of americans”, *Emotion*, vol. 3, no. 1, pp. 68–75, 2003.

[9] el Kaliouby R. A., “Mind-reading machines: automated inference of complex mental states”, Tech. Rep., Computer Laboratory, University of Cambridge, 2006.

[10] Quinlan J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

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

[12] AnAj A.A., “http://en.wikipedia.org/wiki/image:svm_margins.png”, 2006.

[13] Caritat M. J. A. N. marquis de Condorcet, “Essay on the application of analysis to the probability of majority decisions”, 1786.

[14] Smith J. H., “Aggregation of preferences with variable electorate”, *Econometrica*, vol. 41, pp. 1027–1041, 1973.

[15] Ward B., "Majority rule and allocation", *The Journal of Conflict Resolution*, vol. 5, no. 4, pp. 379–389, 1961.

[16] Copeland A. H., "A 'reasonable' social welfare function", Seminar on Mathematics in Social Sciences, University of Michigan, 1951.

[17] Golan O., Baron-Cohen S., and Hill J., "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.

Chapter 7

Expression taxonomy and mapping

This chapter presents the application of the inference machine to expression mapping. Expression mapping refers here to the description, presentation and analysis of the relations between expressions. This term is slightly different in meaning from the terms 'conceptualisation' and 'taxonomy' because it refers to the 'location' of expressions in relation to other expressions using various criteria and to both lexical meanings and their behavioural characteristics.

Mapping the relations between expressions may have several uses for automatic recognition and synthesis of expressions. In particular, it can be useful for continuous tracing and prediction of expressions, assuming gradual changes over time. Such mapping could facilitate speech editing and enhance expression related interfaces in which the user can navigate between desired expression labels. The mapping can integrate several types of description as described in Chapter 2.

Throughout the chapter the term *concept* is used for the label and lexical meaning of the mental state. The term *expression* refers to their behavioural characteristics, in this case, the vocal correlates. A *recognisable expression set* is a set of concept groups whose expressions are recognised by the inference machine according to their vocal correlates. Here the recognisable expressions are: *joyful*, *absorbed*, *sure*, *excited*, *opposed*, *stressed*, *unsure*, *interested* and *thinking*.

The first section of this chapter describes a test that lays the ground for applying the inference machine to the whole MindReading database and its application to concept and expression mapping. The inference machine performs a task which is usually performed by people. It is therefore interesting to compare its performance to human performance as a validation measure. Generalisation to other expression groups seems to be the next validation target. It is desirable to check if the inference machine can distinguish between expressions and concepts beyond the recognisable expression set and to compare its performance to human performance on similar task. Therefore, the inference machine was tested on the CAM Battery Test [1]. The machine's performance was found comparable or even to outperform human abilities, as they are described by Golan *et al* [1].

This chapter then describes expression mapping of the full MindReading taxonomy by applying the inference system to the voice part of the MindReading database. I show two mapping or presentation methods. The first examines the characteristics of the *concept groups*, the meaning groups defined by the MindReading taxonomy.

From this presentation additional connections between concepts from different groups become apparent. Therefore, the second mapping method presents the concepts in the space which is defined by the recognisable expression set.

The chapter begins with a short description of the MindReading taxonomy that is used throughout the chapter. It then presents the method which was used to infer the combination of recognisable expressions that represents each mental state concept. Section 7.1 presents the CAM Battery test, the method that was used to evaluate the machine's capability to distinguish between expressions of complex mental states and its comparison to human performance. Section 7.2 lists the inference results of 459 mental states, using the recognisable expressions. The mental states are mapped or arranged according to the concept groups of the MindReading taxonomy. The expressive characteristics of each concept group are also defined and analysed. In addition, similarities are found between different groups, for example groups that are related to cognitive processes vs. groups that are associated with basic emotions. However, new groups emerge through similarities between groups, therefore Section 7.3 presents mapping of the same mental states according to the recognisable expressions only. This mapping reveals other similarities and meanings, across and beyond the meaning groups, as defined by the MindReading taxonomy. Some of these characteristics resemble other conceptualisation methods. At the end of each section there is a short summary of the results, Section 7.4 highlights and summarises them.

The MindReading taxonomy

As described in Chapter 4, the MindReading taxonomy of emotions [2] comprises of over 740 emotion and mental state concepts, including all the emotion terms in the English language, as well as mental states with an emotional dimension (e.g. doubting). Mental states that could be a purely bodily state (e.g. hungry) and states with no emotional dimension (e.g. reasoning) are not included. These concepts are grouped into 24 concept groups (such as the *happy* group, the *thinking* group, the *sneaky* group, etc). The list of 24 emotion groups is shown in Table 7-1. Developmental testing resulted in the emotion concepts being further subdivided into six different levels, on the basis of word frequency in the English language and verbal comprehension. The six levels represent an age range from pre-school through to adulthood. Each concept is demonstrated by 6 instances, or sentences. In the version used in this research there are 4412 speech segments. The full list of emotions, according to emotion groups and developmental levels can be found in papers by Baron-Cohen *et al.* [3, 4].

Afraid	Excited	Liked	Surprised
Angry	Fond	Romantic	Thinking
Bored	Happy	Sad	Touched
Bothered	Hurt	Sneaky	Unfriendly
Disbelieving	Interested	Sorry	Unsure
Disgusted	Kind	Sure	Wanting

Table 7-1: The 24 groups included in the emotion taxonomy (adapted from Baron-Cohen *et al* [4]).

Method

The mapping was done by running the inference machine on the MindReading database. For each sentence, the number of machines that chose or recognised each of the recognisable expressions was recorded. Only expressions that were chosen by at least six machines were considered for further analysis. That means more than one standard deviation above the mean of the machines number. In this case, no significance was associated with the number of machines that recognised an expression as long as it was above the standard deviation threshold.

Each concept in the database is represented by six sentences. A concept was considered *recognised* if at least one of the recognisable expressions was recognised in at least four of these sentences (>66%). For the recognised concepts, other expressions that appeared in three of the same sentences (50%) were also considered for characterising the concept.

An example can be seen in Table 7-2. It shows the inference and mapping results for the concept *heated* from the *angry* group. The first five rows present the inference results for each of the sentences that represent the concept in the database. The expression *excited* was chosen by six or more machines in four of the six sentences. The expression *opposed* was chosen by more than six machines in three of these four sentences. Therefore, these are the inferred or recognised expressions for this concept. This method was used in all the experiments presented in this chapter.

Concept	**group**	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
0401601P4Theated.wav	angry	4	5	2	3	5	3	2	5	7
0401601P5Theated.wav	angry	3	3	3	5	7	7	4	3	1
0401601P6Theated.wav	angry	5	2	2	5	6	8	2	2	4
0401601R5Theated.wav	angry	6	3	0	7	7	5	3	4	1
0401601Z5Theated.wav	angry	5	3	6	2	6	7	3	2	2
Heated	angry					4	3			
Heated	angry					●	○			

Table 7-2: An example of the mapping method. For each of the speech signals for the concept heated the number of machines for each recognisable expression is stated. Green 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. The turquoise marking is the convention used throughout this chapter, ● recognition in 4-6 sentences, ○ recognition in 3 sentences.

7.1 Comparison to human performance

This section describes the performance of the inference machine in the voiced part of the CAM Battery Test, as described by Golan *et al.* [1]. This experiment had two goals: to measure the ability of the inference machine to distinguish between complex mental states, beyond the limited set of expressions that it was trained to recognise, and to compare it to human performance.

The CAM Battery Test (CBT) tests recognition of 20 complex emotions and mental states. It is based on utterances from the MindReading database, which was used for training the inference machine. The test is aimed at adults with Asperger Syndrome, who can recognise simple emotions and pass basic theory of mind tasks, but have difficulties recognising more complex emotions and mental states. The CBT was tested on males and females with Asperger Syndrome (AS group) and on matched controls (control group).

The inference machine's ability to distinguish between expressions of concepts is compared to that of the participants in the control group.

The Cambridge MindReading (CAM) Face-Voice Battery

For better understanding of the comparison results, this section reviews the CAM battery experiment as reported by Golan *et al.* [1]. From hereon I refer only to the voice part of the battery.

The CAM battery evaluates a selection of 20 concepts, taken from the MindReading taxonomy, representing 18 of the 24 concept groups. The battery includes mental states with different valence (positive, negative or neutral) and emotions of varying intensity (subtle and intense). Selection of the concepts followed three principles: concepts should be (a) selected from all 24 concept groups, (b) mainly subtle, and (c) important for everyday social functioning. The larger concept groups (*unfriendly* and *sad*) were represented by two concepts each.

The test questions consisted of a choice between four concept labels for each of the recorded sentences. The questions were created by a computer programme which randomly selected the sentences and three foil words, ensuring that foils were not from the same emotion group as the target answer. In order to make sure that the chosen concepts are taken from the adult emotional repertoire, they were all selected from the higher levels of the taxonomy, which include concepts that are usually only understood by people over the age of 15. After listening to a sentence the participants were asked to "choose the word that best describes how the person is feeling".

It is important to note that although the MindReading database was labelled by agreement of several people, additional validation of each sentence of the chosen sentences was conducted before carrying out the group analysis. The data from the 21 adults in the control group was first analysed as follows: An item was considered valid if at least 11 out of 21 (>50%) of the participants selected the target word and no more than six (<33%) selected any one of the foils. Using these criteria, five of the concepts were excluded from the battery. Eight other concepts had one invalid item each, and these items were removed. In order to keep the same number of items for all the concepts, one item was randomly removed from each of the remaining 12 concepts of the CAM battery, so that the final battery comprised 20 concepts.

These results imply that the accuracy level of the labelling of the MindReading database in not consistent. Therefore, strict conditions were taken in the expression mapping which is described in the second part of this chapter.

Performance comparison

The participants in the CAM battery test used for comparison were the control group which comprised 17 adults recruited from a local employment agency, including 12 males and 5 females. Table 7-3 presents their results in the voice part of the battery. The average number of sentences recognised by the control group was close to 43 sentences out of 50 [1]. The AS group comprised 21 people as well. Their performance was consistently lower than the performance of the control group. On average the participants in this group recognised less than 36 concepts by their vocal correlates.

		AS	Control	Total
Females	Mean	33.17	44.80	38.45
	SD	9.62	2.59	9.27
Males	Mean	36.73	41.92	39.04
	SD	4.22	3.96	4.81
Total	Mean	35.71	42.76	38.87
	SD	6.19	3.78	6.29

Table 7-3: Average number of concepts that were recognised out of 50 sentences in the vocal test by males and females in the AS and control Groups [1].

The inference machine was applied to the same sentences or voices that were used for the CAM battery, and used the same foil concepts that were used in the battery questions.

Its defined task was to distinguish between the voice of the concept and the voices of the foil concepts, i.e. the concepts and the foil concepts should have yielded different combinations of the recognisable expressions. The expressions recognised for the concept sentence by at least six machines were compared to the expressions that were recognised for the foil concepts, using the six sentences of the foil concept following the method which was described in the previous section. An example of the vocal correlates of one of the battery questions, as depicted by the inference machine, is presented in Table 7-4.

The example shows one of the two questions for the concept *lured*. The foil concepts are: *comprehending*, *so-so* and *rejecting*. The concept *lured* is recognised as a combination of *thinking* and *unsure*, *comprehending* as *thinking*, *so-so* as *thinking* and *absorbed*, and *rejecting* as *sure* and *opposed*. The different combinations allow us to distinguish between the different examined concepts.

The inference machine successfully distinguished between the concepts and the foil concepts in 49 of the 50 sentences. In this case it outperforms humans.

Furthermore, the meaning of the recognised combinations is reasonable. It either conforms to the concepts' lexical meaning or at least does not contradict it.

These findings imply that the machine can distinguish between complex mental states that are not necessarily part of the recognisable expression set on which it was trained. These mental states are characterised by combinations of the recognisable expressions. Therefore, the machine can be used for finding relations between expressions, and be applied to a larger set of expressions, as described in the next sections.

	Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
Concept Sentence	lured	2	4	3	2	5	2	5	7	6
Concept (6 sentences)									●	●
Foil Concepts	comprehending									●
	so-so		●							●
	rejecting			●			●			

Sentence file name	The question			
0903302A6Tlured.wav	1. lured	2. comprehending	3. so-so	4. rejecting

Table 7-4: The question from the CAM battery (bottom) and the vocal correlates of the examined expressions as depicted by the inference machine (top): first row from the top - number of pair-wise machines that chose each expression for the concept sentence. Second row- recognised expressions for the whole concept (*lured*). Bottom rows - recognised expressions for the foil concepts (*comprehending*, *so-so* and *rejecting*). ● - expression chosen by more than 6 machines in at least 4 sentences, as in Table 7-2.

Expression mapping

The following sections describe the application of the inference machine to expression mapping. The goal was to examine the relation between expressions, their meaning and the relations between them.

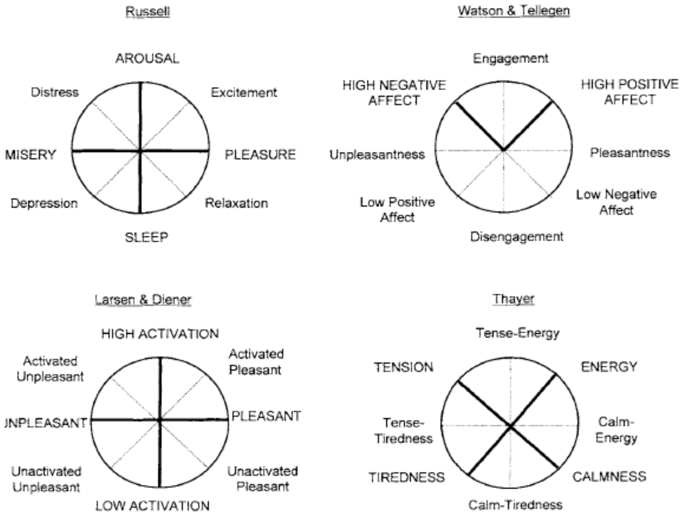

Figure 7-1: Four descriptive models of core affect. From Yik, Russell, and Feldman Barrett [5]

One of the reasons for the relevance of expression mapping to automatic systems was stated by Cowie *et al.*: *'The obvious goal for an automatic recogniser is to assign category labels that identify emotional states. However labels as such are very poor descriptions. In addition, humans use a daunting number of labels to describe emotions.'* [6]. Recognising the relations between concepts and categories may facilitate the recognition and the navigation between concepts.

This research area has been approached by linguists and psychologists for decades, for different reasons and from different points of view. Russell and Barrett wrote *"Unfortunately, there is no consensus on what that structure (of emotions) should be. Indeed, there is every appearance of disagreement: Some researchers use categories, some dimensions; some use bipolar concepts, some unipolar ones; and some presuppose simple structure, some a circumplex, and some a hierarchy."* [5]. Figure 7-1 presents some of the circumplex methods. Additional methods are reviewed in Chapter 2.

In the next sections I examine two descriptive frameworks for the mapping of concepts and their related expressions. The first is the hierarchical presentation of the MindReading taxonomy and its 24 concept groups and the second is the nine-dimensional space which is defined by the recognisable expressions set.

Mapping according to a small sub-set and vocal correlates

The mapping suggested here uses small set of concepts and their expressions, including *joyful*, *absorbed*, *sure*, *excited*, *opposed*, *stressed*, *unsure*, *interested* and *thinking*, which consist the recognisable expressions set. The most interesting and important feature of this mapping is that it is done through the vocal correlates of these concepts.

These concepts are not commonly used for mapping, although they include several variations of the terms which are mentioned in the literature. For example, *joyful* (*pleasure*, *high positive affect*, *joy*), *excited* (*arousal*) and *stressed* (*tension*). The recognisable expression set also includes terms related to both complex mental states such as *thinking*, *absorbed* and *interested* and to confidence level like *sure* and *unsure*, that do not appear in other descriptor sets. The term *opposed*, although indicating *negative affect*, also refers to more cognitive related terms such as *disagreement*.

The recognisable expression set allows using day-to-day descriptors that discern at least part of the complex phenomena of emotions and mental states, albeit in a fuzzy manner rather than along axes. It also allows getting more information about the emotional content or the nature of an expression beyond its location in a certain concept group in a hierarchical taxonomy.

The expression inference machine was applied to the full MindReading database, using the method described at the beginning of this chapter. It was done in order to explore the expression mapping capabilities of the inference system, to check the vocal correlates of the MindReading taxonomy, and to find other possible ways of grouping expressions and finding the relations between them.

Investigating the relations between expressions strongly depends on the manner of presentation. One method is summary for all the groups, including all the instances (sentences/utterances) belonging to them, according to the MindReading database. This method was used for preliminary observations. It gives a broad idea of the characteristics of each group according to the inference system. However, it is sensitive to outliers. Therefore, the more strict criteria for defining recognised concepts, as described above, were devised. 459 concepts were recognised in this manner. As demonstrated in the next sections, most of the inference results (>85%) correspond to the lexical definitions of the concepts, or at least do not contradict them. These numbers indicate that the inference results are significant and mostly reliable.

The two mapping methods which are presented in this chapter refer only to the recognised concepts. The first method examines properties of the MindReading taxonomy, as defined by the recognisable expression set. In the second method, the recognised concepts are organised only according to combinations of the recognisable expressions. These methods examine different aspects of the relations between concepts, their groups and their meaning (linguistic and behavioural). They examine the mapping capabilities of the inference machine and its limitations. They may also suggest properties and limitations of acting various mental states.

7.2 Vocal mapping of the MindReading taxonomy

This section presents the first mapping method which examines properties of the MindReading taxonomy. The presentation is arranged according to the taxonomy concept groups. In each group, the concepts are arranged according to combinations of the recognisable expressions. This mapping method allows the examination of three parameters according to the recognisable expression set:

1. The inference of each concept and its relations to the concept group
2. The relations between concepts within concept groups
3. The characteristics of the concept groups and the relations between them.

The taxonomy is based on meaning based concept groups such as *romantic*, *unfriendly*, etc. These groups often include opposite extremes of the more commonly used mapping dimensions, such as activation (passive/ active) and valence (positive/ negative). Therefore, this presentation can also reveal meanings and their related behavioural cues beyond the scope of these dimensions.

Method and Results

Each of the next tables shows the inference results for a concept group, as defined by the MindReading taxonomy and database. The presentation format is demonstrated by Group Table 1 which presents the mapping results of the *sad* group.

In the table each row shows the combination of expressions that were recognised for the concepts which are stated on the left. For example, in the *sad* group, the inferred expression for the concepts *despairing* and *distraught* is *stressed*. As in the previous tables, • signifies recognition in at least four sentences (>67%), while ◦ signifies expressions that were recognised in 3 of these sentences (50%). Expressions recognised in three sentences are perceived as less significant and less accurate.

The colours are an added interpretation of the results. In the nine expression columns on the right, turquoise is used for expressions that could be expected for the concept or for its acting. Yellow is used for unusual or less expected expressions and expression combinations. In most group tables there are also grey columns. These indicate expressions that were not recognised as related to most of the concepts in the examined concept group. The colours in the left column indicate concepts that may belong to other groups, according to their inferred combination. Orange indicates, in most cases, that other similar (unusual or unexpected) combinations appear also in other groups. Yellow indicates that they do not necessarily belong to the examined group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
lonely, pining, reconciled, weak								●	
overwrought				●				●	
pitied		●						●	○
exhausted		●						○	○
lovelorn		○					○		●
agonising		●							○
grieving		○							●
discouraged, soulful		●							
disillusioned, subdued			●						
grave			●						●
empty, tired, self-pitying, suffering, lovesick									●
distraught, turmoil				●					
troubled		●		●					
tormented				●	●			●	
despairing	●			●					
hysterical	●			●		●			
suicidal	●					●			

Group Table 7-1. The sad group results. On the left - the mapped concepts. The nine columns on the right are the recognisable expressions. ● -expressions recognised by at least 4 sentences, ○ - expressions recognised in three of these sentences. Turquoise cells indicate probable expression, yellow cells indicate less probable expression for the concept. Grey columns mark expressions that did not appear in a significant manner for any of the concept in the group. The concepts marked in yellow display unusual combinations of expressions. Orange coloured concepts may belong to new cross-taxonomy groups.

Admittedly, these interpretations are rather subjective. Unfortunately, there are no predefined definitions that correspond to the lexical terms used for this mapping.

As can be seen in the Group Table 7-1 and in the following tables, most of the combinations of the recognisable expressions agree with the lexical concept that they represent. The inferred expressions do not necessarily give the full lexical definition, but it cannot be expected from the recognisable expression set. In other cases, inferred expressions suggest properties of the concepts that are not directly derived from their lexical meaning. They can evolve from situations in which they appear or even from the fact that they are acted and the manner of acting (single actor, speaks to oneself). In many cases, the inference highlights concept properties that are not always obvious from the lexical definitions and therefore it is interesting to review these results.

Most of the inference results are reasonable and acceptable under these conditions. Therefore I focus on the less expected mapping results. For each of the Group Tables I highlight a few of the concepts' and group's characteristics and discuss the findings at the end. The *sad* group is discussed here in more detail in order to demonstrate a few of the inference and mapping characteristics that are highlighted and referred to later on.

In the *sad* group, the most recognised expressions are *absorbed*, *stressed*, *unsure* and *thinking*. The other expressions occur only in few concepts. Furthermore, expressions such as *joyful* and *excited* are not expected to be inferred for the *sad* concept group. The combination of *joyful* and *opposed* is also not expected.

The expression *interested* was only partially recognised in one concept, *lovelorn*. Although the concept *sad* is mostly associated with self-absorbance, being interested (in somebody) or asking questions (about the situation) can be an integral part of the concept *lovelorn*.

The expression *excited* was recognised only in the concept *tormented*. It appears in combination with *stressed* and *unsure*. Linguistically, this concept does not necessarily belong to the *sad* group. It may belong to another group such as *bothered* in which this combination appears in several concepts, or to a new group altogether. Many of these concepts are related to extreme expressions.

The less expected combinations of *joyful* with *stressed* and *opposed* appear in the concepts *hysterical*, *suicidal* and *despairing*. These concepts are also related to very extreme expressions and can be associated with lack of control. These combinations occur also in other concept groups.

The next tables present the inference results of all recognised expressions and groups in alphabetical order. A few of the possibly less expected expression combinations are highlighted for further discussion at the end. The large majority of concepts though, reveal reasonable and interesting combinations that highlight various aspects of these concepts.

Group Table 7-2 shows the *afraid group*. The dominant expressions in this group are *absorbed*, *stressed* and *unsure*. The expressions *joyful*, *sure*, *excited* and *opposed* do not appear. The exception is the concept *frantic* with a combination of *joyful*, *stressed* and *excited*, as discussed in the *sad* group. The expression *thinking* is also not an obvious part of this group. It is interesting to note that *afraid* is associated with *basic emotions* rather than with complex mental states such as *thinking*. The concept *interested* may indicate *questioning*, especially when appearing with *unsure* and *stressed*. It is more difficult to find an explanation for the inference of the concept *afraid*.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
afraid							●		
dreading		●							
cowed		●						●	
daunted		●						○	
disturbed		●						●	●
shaken, watchful								●	
worried				●			●	●	
nervous				●			●	●	
stressed				●				●	
cowardly, jumpy, panicked, terrified, pressured				●					
frantic	●			●	○				

Group Table 7-2. Afraid

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
displeased		○					●		
angry, indignant, miffed, wild						●			
bitter			●			○			
grumpy, disagreeable			○			●			
embittered, moody			●			●			
heated					●	●			
needled					○	●			
explosive	○					●			
dudgeon	●			○					
raging	●			○					
complaining, exasperated, frustrated				●					

Group Table 7-3. Angry

The *angry* group is presented in Group Table 7-3. *Anger* is usually considered one of the *basic emotions*. *Opposed* is the dominant expression, as can be expected. It appears in combination with *sure*, *excited*, and *stressed*. Apart from the *displeased* concept, the expressions *absorbed*, *interested*, *unsure* and *thinking* do not appear in this group, There is a group of concepts that represent extreme expressions, including *dudgeon* (in a very angry or irritated mood), *explosive* and *raging*. This group includes combinations of *joyful* with *opposed* or *stressed*, as can be seen also in the *sad* and *afraid* groups. The inference of *displeased* is a combination of *absorbed* and *interested*, which is different from the other concepts in this group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
blank			●						
distant, negligent, unimpressed, unthinking			●						
vacant							●		
passive							●		○
jaded, noncommittal, vague									●
listless								○	●
unenthusiastic								●	
inattentive				●				○	
complacent			●		●	●			

Group Table 7-4. Bored

The next group is *bored* which can be seen in Group Table 7-4. This group is not well represented by the recognisable expression set. Although the least recognised expressions here are *joyful*, *stressed*, *excited* and *opposed*, which describe a depth of response. As expected, a few explanations are still required.

The concepts of *vacant* and *passive* appear to elicit mostly the *interested* expression. These concepts belong to the same domain, but as opposites.

The concept *complacent* appears in this group in the formal publications of the MindReading database, in an earlier version it was located in the *sure* group. The recognised combination for it (*sure*, *excited*, *opposed*) conforms to the lexical definition ('self-satisfied usually in an unreflective way and without being aware of possible danger' or alternatively, 'eager to please'), which indicate that it should be indeed located in the *sure* group.

Inattentive ('careless, not paying attention or taking proper care') can be associated with *stressed* and *unsure* as the reasons for the attitude rather than with *boredom*.

The association of *blank* with *sure* is acceptable and depends on the interpretation of the actor.

The *bothered* group appears in Group Table 7-5. This group is mostly recognised as *stressed*. Most of it was also used for training the *stressed* expression. The expressions that do not appear in this group are *absorbed*, *sure*, *opposed* and *thinking*. *Unsure* may appear in *overrun*, and *joyful* in *hurried* ('over-hasty, doing something too quickly because of a real or perceived lack of time'), which will then add hurried to the group of extreme expressions that are difficult to act, but the lack of evidence, only 3 (50%) of the sentences is not enough for conclusive results.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
flustered, impatient, pestered, restless, ruffled				●					
bothered				●	●				
overrun				●	○			○	
hampered				●			●		
rushed				●				○	
hurried	○			●					

Group Table 7-5. Bothered

Group Table 7-6 shows the results for the *disbelief* group. Only properties of 2 concepts out of 8 were recognised (25%).

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
guarded		●							
doubtful		●							●

Group Table 7-6. Disbelieving

Of the *disgusted* group only one concept from a total of 4 was recognised (25%), as can be seen in Group Table 7-7, although the *opposed* expression appear in many of the utterances of this group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
averse			●			●			

Group Table 7-7. Disgusted

Group Table 7-8 shows the results for the *excited* group. *Excited* and *stressed* are the dominant expressions here. The mixture of them is understandable, especially in the negative concepts such as *hysterical* and *uncontrolled*. *Sure* and *interested* are not characteristics of this group. An interesting characteristic is the recognition of *opposed* in the concepts that describe *motivation* and *enthusiasm*. *Intense* and *possessed* may belong better to the *absorbed* group, as it is defined in the taxonomy.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
keen, motivated, reckless						●			
intense		●						●	
possessed		●						●	
emotional				●					
hysterical				●				○	
titillated	●			●					
uncontrolled	●			●	○				
enthusiastic				○	●	●			
upbeat				●	●				
spirited, alert, dynamic, excited, invigorated, vibrant					●				
lively	○				●				
inspired		○			●				
exhilarated					●	○	○		

Group Table 7-8. Excited

The next group is *fond* which is presented in Group Table 7-9. The dominant expressions are *absorbed*, *unsure* and *thinking*. The expressions that are related to high arousal levels such as *joyful*, *excited*, *stressed* and *opposed* are not characteristics. It is interesting to note that although *joyful* and *excited* are the only obviously positive recognisable expressions, they are not apparent in this group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
trusting							●	●	
adoring								●	●
loving									●
fond		●							○
worshipping, affectionate, devoted, doting		●							
regard		●	○			○			

Group Table 7-9. Fond

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
nonchalant, proud, boosted						●			
flexible				●		●			
amenable			●			○			
easy-going			●			●			
comfortable, gratified, mischievous, protected			●						
privileged			○		●				
tolerant			●			●			
tickled	●		○		○				
happy	●				○				
jubilant	○			○	○				
rejoicing	●				●	○			
festive	●					○			
amusing, delighted, enjoying, felicity, glad, joking, joyful, triumphant, outgoing, unabashed, playful	●								
overjoyed	●						●		
merry	●						○		
blameless, familiar							●		
calm		●					○		
peaceful		○					●		
charismatic		●							
relieved							●		●
languid								●	
pleasure								●	●
accepting								●	

Group Table 7-10. Happy

The *happy* group is next. This group consists of a large variety of concepts. Some of the concepts are indeed related to happiness and joy. Others indicate different levels of well-being. For example, *safe* which is integral to many other expressions but apparent more in the context of opposite situations (danger), or *calm* which signifies low arousal level rather than positive disposition which *happy* or even *content* imply. In Group Table 7-10 the concepts marked in orange could probably belong to at least one different group of concepts.

Group Table 7-11 shows the *hurt* group. The dominant expressions are *absorbed*, *unsure*, *thinking* and *stressed*. The inferred expressions for the concept *terrorised* suggest that it could either belong to the *afraid* group or to the emerging group of extreme expressions. The concept *teased* seems to be interpreted by the actors as a positive rather than a negative expression and should have been in the *happy* group. *Misunderstanding* and *judged* seem to be acted from the point of view of power and contradicting opinions, as some concepts of the *unfriendly* group, rather than as hurt feelings. Other than these concepts, as expected, the expressions *joyful*, *sure*, *excited* and *opposed* are not among the characteristics of this group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
betrayed		●							
belittled, demoralised									●
neglected		○							●
downtrodden, compelled		●							●
isolated		●						●	●
deflated								●	●
scolded, resented, disbelieved, hurt, patronised, scrutinised								●	
excluded		○					●	●	
criticised							●		
corrected, mistreated, questioned, scorned, tortured, trapped				●					
umbrage				●	●				
terrorised	●			●	●	●			
misunderstood						●			
judged			●						
teased	●								

Group Table 7-11. Hurt

Group Table 7-12 shows the *interested* group. As expected the main expressions are *interested* and *absorbed* with a few that include *thinking* and *unsure*. As discussed in Chapter 6, the expressions *interested* and *absorbed* were trained on data from the *interested* concept group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
nosy, spellbound, tantalised								●	
obsessed					●				
tempted, fascinated, curious, interested, listening, probing, quizzical							●		
asking				●			●		
intrigued				●					
concentrating, conscientious, dazzled, thorough, involved, obedient		●							
focused, hypnotised		●							●
engaged		●		○				○	
awed		○						●	
absorbed		●						●	○
sensitive		●						○	○
lured								●	●

Group Table 7-12. Interested

Group Table 7-13 shows the *kind* group. This group is supposed to convey a positive attitude towards somebody else in a behavioural manner towards them-(active), rather then in the feeling towards them (passive concept group *fond*). *Interested* and *sure* are the most pronounced and obvious expressions in this group although *absorbed* and *stressed* are also quite apparent. There are a few problems with the definition of some of the concepts here, especially *acknowledge* and *fatherly*. Their meaning, acting and inference are sensitive to interpretation and context.

The next table, Group Table 7-14, presents the *liked* group. The reaction to being liked in rather complex and includes a whole variety of expressions, from *sure* to *unsure*, *excited* and *opposed*. *Joyful*, *thinking* and *stressed* are not among the characteristics in this group. Here again the context may play a major role in the interpretation. Arguably, open cases of adoration and flattery can create reactions of uncertainty (*unsure*), *modesty* and *opposing* arguments.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
praising, kind, caring				●					
charitable	○			●					
encouraging				○			●		
polite				●			○		
excusing, concerned, sensitive, warm							●		
receptive		○					●		
patient, pitying, protective		●							
soft				○				●	
persuading		○						●	
comforting		●						●	
indulgent		●							●
forgiving									●
tender			●					●	●
calming			○						●
responsive			●						
forthcoming			●		○				
complimenting		○	●						
acknowledging			●			●			
fatherly			●			○			
gallant						●			

Group Table 7-13. Kind

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
adored								●	
comforted, preferred		●							
favoured									
accepted, conciliated, reassured, trusted							●		
acclaimed			●			○			
indulged			○			●			
flattered					○	●			

Group Table 7-14. Liked

The *romantic* group is presented in Group Table 7-15. The dominant expressions in this group are *absorbed*, *thinking* and *unsure*. The concept *attractive* should have been in the *sure* group, as it was defined in earlier versions of the taxonomy.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
aroused, lustful, seductive									●
entrancing		○							●
fancying		●							●
loving, lecherous, seduced		●							
flattering, attracted		●						●	
sexy, romantic		●						●	●
intimate		●						○	○
bewitched		○					○	●	●
enticed								●	
attractive			●						

Group Table 7-15. Romantic

The *sneaky* group is presented in Group Table 7-16. The dominant expressions are *thinking*, *unsure* and *absorbed*. The recognition of *insincere* as *joyful* may be misleading. The concept *fawning* can belong to this group or to the *kind* or even to the *unsure* groups because it has two lexical interpretations: 'seek favour by flattery' and 'try to please'.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
insincere	●								
sneaky, lying, tempting									●
luring, manipulative		●							●
secretive		●						●	●
mysterious								●	●
furtive								○	●
fawning								●	

Group Table 7-16. Sneaky

The *sorry* group is presented in Group Table 7-17. This group has concepts shared with the groups *hurt* and *unsure*, and even *stressed*. The membership of the specific terms in the *sorry* group is arguable. Most of them are expressed in an introspective manner, including *thinking*, *unsure* and *absorbed*.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
mortified				●					
humiliated, disgraced								●	
regretful								●	●
ashamed, ignominious		●						●	●

Group Table 7-17. Sorry

The *surprised* group in presented in Group Table 7-18. Only one of the concepts in this group could be described as positive and was not recognised. The combination of *excited* and *stressed* is expected. The appearance of *joyful* in the concepts *shocked*, *appalled* and *scandalised* associates them with the new group of extreme expressions. The concept *dazed* has several meanings (confused, stunned, shocked, surprised, bemused, bewildered, etc), not all of them are related to *surprise* and the interpretation of the actors could have been different.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
startled					●				
shocked	●							○	
appalled	●			○	●				
scandalised	○			●					
horrified				●					
dazed							●	○	●

Group Table 7-18. Surprised

Group Table 7-19 presents the mapping of the concept group *sure*. The recognised expressions are mostly *sure*, *excited*, *opposed* and to a lesser extent *absorbed*. The concept *vain* can in corporate *joyful* if it was acted as gloating. The expression *committed* was used to train the expression of *absorbed*. The more forceful concepts such as *dogmatic*, *dictatorial* and *forceful* include elements of *excited*. It is interesting to note that according to the taxonomy many of the concepts in this group relate to confidence in case of opposition.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
pompous, sure, convinced, decisive, enduring, prepared, resisting, staunch			●						
persuaded, cocky, compelling, inexorable, egocentric, firm, immovable, relentless						●			
wilful						●	○		
opinionated			○			●			
confident, bullish, insistent, knowing, narcissistic, resolved, uncompromising			●			○			
adamant, arrogant, assertive, conceited, decided, determined, egotistical, purposeful, righteous			●			●			
unworried		●	●			●			
controlled		●	●		●				
boastful, challenging			○		●	●			
dogmatic			●		○	●			
competitive					●	○			
dictatorial					○	●			
forceful					●	●			
self-confident					●				
sincere								●	
committed		●						○	
convincing		●							
forward				●					
vain	●								

Group Table 7-19. Sure

The *thinking* group appears in Group Table 7-20. As expected it includes the expression *thinking*, with elements of *unsure* in concepts such as *considering*, *choosing*, *brooding* and *debating*. It also includes elements of *absorbed* in concepts such as *dreamy*, *fantasising* and also in *considering* and *debating*. *Stress* appears in *realising* that may have an element of *surprise* ('become aware of something'). The concept *preoccupied* might have been more suitable in the *bothered group*.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
comprehending, thoughtful, deciding, regarding, wool-gathering								●	●
calculating		○							●
dreamy, fantasising		●							●
considering		●						●	○
brooding		●						○	●
choosing, thinking								●	●
realising				●				●	●
preoccupied				●				●	
debating		○						●	

Group Table 7-20. Think

Group Table 7-21 presents the *touched* group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
effusive				●					
corny			○			●			
touched								●	●
nostalgic								○	●
wistful		●						●	●
sentimental		●							●

Group Table 7-21. Touched

Most of the concepts are recognised as combinations of *thinking*, *unsure* and *absorbed*, as expected. The concept *effusive* reflects excessive and extravagant expression of feelings, recognised here as *stressed*, while the other terms refer to more subtle concepts, which is also reflected in the recognised expressions. *Corny* seems to belong to the wrong group, and acting in a corny manner has a completely different (negative) meaning.

Group Table 7-22 shows the 'all encompassing' *unfriendly* group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
unapproachable							●		
frustrating		●					●		
condescending		●							●
coercing		●						●	●
mean, picky		●							
sadistic		●	●						
slighting			●		●				
antagonistic, intimidating, persecuting, snotty, unreceptive			●						
cold			●				○		
remote			●			○	○		
bombarding, disagreeing, malicious, rejecting, sarcastic, thwarting, unkind			●			●			
irritating, argumentative, discouraging, despising, uncooperative			○			●			
acidic, condemning, gruff, uncaring, stern			●			○			
punishing	○		●			●			
resentful	○		○			●			
boorish			●		●	●			
inhospitable		●	●		○	○			
vindictive			●		●	○			
defiant			○		○	●			
gleeful			○		●	●			
disruptive					●	○			
disobedient, odium, unruly					●	●			
forbidding	●				●	○			
contradictory				○	○	●			
insulting, brazen, confrontational, contrary, critical, derisive, devastating, difficult, disapproving, disconcerting, illiberal, rude, interfering, obstructive, vulgar, unwilling, violent, overpowering?						●			
blaming, disinclined	○					●			
hostile	●					●			

Group Table 7-22. Unfriendly

The *unfriendly* group includes 118 concepts, of which 64 were recognised (54%). The main recognised expressions here are *opposed* and *sure*, some are coupled with *excited* and other with *absorbed*. Interestingly, the concepts that mean no interest such as *remote*, *cold* and *unapproachable* were recognised with elements of *interested*. Several of the stronger concepts, including *punishing*, *resentful*, *forbidding*, *blaming* and *hostile* include element of the expression *joyful*.

The *unsure* group, as presented in Group Table 7-23 is just as complicated. The main recognised expressions are *unsure*, *absorbed*, *thinking*, *interested* and *stressed*. There are a few expressions that do not appear to belong in this group. For example, *deferential* can be in the *kind* group ('showing polite respect'). *Silly*, can refer to acting in a silly manner, as a joke, therefore its recognition as *joyful* is acceptable. *Self-deprecating* can mean a person who is *sure* of being guilty. *Neurotic* is inferred as a combination of *joyful* and *stressed* that is observed in other groups for extreme expressions.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
confused, clueless, insecure, humble, indecisive, inferior, innocent, unsure, baffled, disorientated								●	
inadequate				○				●	
puzzled							●	●	
faltering							○	●	○
undecided		○						●	
hesitant		○						●	○
hopeless		○							●
shy		●							●
naive		●							
deferential		●					●		
So-so									●
mystified							○		●
helpless				●	○				
bewildered	○			●	○				
sensitive				●					
denying neurotic	●			●					
silly	●								
self-deprecating			●						

Group Table 7-23. Unsure

The last group is the *wanting* group which is presented in Group Table 7-24. This group includes mainly the expressions *unsure*, *thinking*, *absorbed* and *opposed*. *Interested* and *excited* are not part of this group which includes concepts from *begging* to *demanding* and therefore from *unsure* to *sure*. One of the lexical definition of the concept *faddy*, is 'having strongly held but brief, enthusiasms', therefore it does not necessarily belong in this group.

Concept	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
needy, wishful									●
longing								○	●
appealing, greedy, begging								●	
requesting		○						●	
suppliant		●						●	
particular		●							
acquisitive						●			
demanding, rapacious			○			●			
faddy	●			●					

Group Table 7-24. Wanting

Discussion

This section describes the automatic mapping of mental state concepts and expressions. The concepts are arranged according to the Mindreading taxonomy whose properties are also examined. The inference was done by the inference machine, which is discussed in Chapter 6, and its recognisable expression set.

The expression mapping can be examined according to the following criteria:

- The inference of each concept.
- The inclusion of a concept in a concept group as defined by the taxonomy.
- The relations between concepts within a concept group.
- The characteristics of each concept group.
- The relations between concept groups.

Several conclusions can be drawn from the tables above as well as a few repetitive issues that require additional discussion. The main conclusions are:

1. Most of the inferred combinations for most of the concepts seem to correspond to their lexical meanings, or reasonable in respect to the situations in which they could be expected (approximately 85%).

2. The grouping of concepts into concept groups reveals common characteristics within each group. That means that there are vocal characteristics that are related to meaning. Or in other words, that meanings can be distinguished by vocal characteristics and by the recognisable expression set.

3. There are characteristics that are common to different concept groups. For example, groups that are related to *basic emotions* (such as *afraid* and *angry*), for which expressions such as *thinking*, *interested* and *absorbed* are rarely inferred, versus groups that are more related to cognition and self-absorbance and reflection in which such complex mental states are common.

However, not all the inference results are clear or straightforward. The main conclusions that arise from the examination of the tables and require additional discussion are the following:

1. The recognisable expression set is too coarse and does not include enough expressions for more precise labelling.

2. Misclassification. According to the inference results, certain concepts appeared to be in the wrong concept group, although their inference seems reasonable and correct. The reasons for the apparent or suspected misclassifications can be:

   - A concept has several lexical meanings. The taxonomy referred to one, while the database referred to another. For example, *complacent* (*bored* and *lazy* vs. *content* and *sure*).
   - There is no provision in the taxonomy for certain expressions. It intentionally does not include expressions that relate only to physiological reactions, therefore concepts that incorporate an obvious physiological reaction appear in a lexical group that relate only to one aspect of their interpretation. Their classification is more a judgemental interpretation of the concept meaning rather than its content. An example is the concept *calm* which indicates low arousal level rather than valence (positive/negative) which was placed in the *happy* group.
   - The same meaning has different behavioural expressions in different contexts. For example, *displeased* (*angry* vs. *absorbed*).
   - The concept definition does not belong to any of the groups, and the nearest group is a matter of interpretation. For example, *tormented* (*sad* vs. *bothered*).

3. Certain groups, such as *romantic*, *bored*, *disbelieving* and *fond*, are characterised by dominant combinations of *absorbed* and *thinking*. The inference in these cases does not necessarily convey the meaning of these groups but rather their expressions, especially in the context of a single actor who reflects (or fantasies) in a certain manner. The set of recognisable expressions is relatively small, and is not enough to distinguish between all the subtle expressions, or to define all the meaning subtleties.

4. Indirect interpretations of concepts appear, especially in relation to confidence level. For example, the concept *lonely* and its inferred expression *unsure*. It may also indicate lack of more precise descriptors.

5. Possible effects of acting that cannot be isolated. For example, the inferences of the expression *unsure* for a certain concept can evolve from the concepts' meanings, from its interpretation by the actors or from uncertainty in acting it (the actors express their own mental states beyond the expressions they try to convey).

6. An artefact of interpretation and acting expression is that many positive meanings are more apparent in the case of opposition. For example, most of the *sure* group and concepts which are related to motivation (like *enthusiastic* and *motivated*) reveal expressions that can be interpreted as defiance. In the same manner, concepts such as *trusting* include the inference of *unsure*. Concepts that are more easily acted by

asking questions invoke the inference of the expressions *interested* and *unsure*. Not as easily explained is the appearance of *absorbed* in active concepts that usually indicate interaction with other people, such as *persuading*, *comforting*, *protective* and *indulgent*.

7. Labelling and acting subtle expressions, especially those that relate to well-being is difficult. It appears in the inference of the concepts which are related to contentment in the *happy* group. An unofficial labelling verification attempt with 4 people revealed that indeed many of these expressions are acted in a manner that implies negative rather than positive expressions. The evaluators were asked if the label is correct, if the voice is positive, negative or neutral and whether the voice sounds acted or natural. In both the *happy* and in the *unfriendly* groups there were many mismatches.

8. In the *Happy* group I would have expected the confidence level (*sure*) to be high, but it is not at all a parameter in the inference. It could be an effect of the machine design, or an indication that this aspect is not significant in these expressions.

9. In certain cases, the inferred expression was from the same domain (for example interest level), but on the opposite side of the scale. For example, the concepts of *vacant* and *passive* elicit mostly the *interested* expression. It can be a case in which the dominant expression relates to the domain itself.

10. New groups appear. The more obvious groups are those with unexpected expression combinations, such as *joyful* and *stressed*, *joyful* and *opposed* and *excited* and *stressed*. These combinations are inferred for concepts from different concept groups. These concepts convey extreme emotions and expressions. There are several possible reasons for the inferred combinations:

    - The machine selects the closest expressions. *Joyful*, and to a lesser degree, *opposed*, are the most associated with extreme emotions and expressions. *Joyful* includes laughter and other forms of extreme expressions that signify less control of the speaker. *Excited* and *stressed* are also related to basic emotions as they are related to arousal level and connection to physiological reactions.
    - The combinations indeed characterise the concepts and their expressions. Another aspect that could be considered is that people can experience bipolar emotions such as *happy* and *sad* [7].
    - The combination is an artefact of the acting or the interpretation of the expression by the actors. It is difficult to act (to control) expressions that relate to lack of control, very extreme expressions and expressions that have associated physiological reactions. Many people do not even experience or reveal extreme expressions, such as: *hysterical*, *suicidal*, *neurotic* and *raged*. (note: It is interesting to note that in many plays, movies and soap operas, expressions such as laughter are not common. One of the reasons may be that it is difficult to reliably imitate expressions such as real laughter. The same may be correct for the acting of other expressions that indicate a certain lack of control).

It is important to note in this respect that there is no correlation and no inherent dependency between recognisable expressions.

The reason behind these combinations is not as significant as the fact that they are repeated in the inference of multiple concepts with common meaning-related characteristics.

Summary

As can be seen in the tables and discussion above, most of the recognisable expressions are related to the lexical definitions of the defined concepts. In addition there are definite and reasonable characteristics to the MindReading taxonomy groups that relate to their meaning.

It is a rather surprising result if considering the very small number of sentences per concept. Of the 749 concepts, 459 are mapped (61%), which is a very high number of concepts, and most of them (approximately 85%) are recognised correctly. 20% of the samples were excluded from the CAM Battery Test, so these results are very satisfactory. Table 7-5 shows the percentage of recognised concepts from each group from the MindReading taxonomy.

The fact that so many concepts are recognised increases the validity of the results and justifies further examination.

Many of the concepts that have been mapped with elements from expressions that are not precisely part of their linguistic meaning, but the inference can be justified. In certain cases acceptable and expectable behavioural characteristics were inferred rather than expressions that relate to the linguistic content and meaning.

In some cases the recognition pinpoints some of the challenges which are incorporated in defining taxonomy, in recording and labelling a database that includes such a large variety of concepts and expressions, and in the inference machine and the scope of the nine-dimensional expression set.

The results show that the mapping using this machine and technique is good, mostly reliable and gives a lot of information about the nature of expressive speech and the relationship between lexical meaning and expression manner. It also shows that concepts within concept groups in the MindReading taxonomy have many common characteristics. These characteristics, include both the expression they convey and the vocal characteristics of this expression and meaning.

They also reveal that the groups are not mutually exclusive and that the relations between concepts are more complicated and can be on several planes.

Group	# concepts	# recognised concepts	% recognised	Concepts not belonging to group
Afraid	24	16	67	
Angry	26	18	69	
Bored	19	14	74	2
Bothered	14	10	71	
Disbelieve	9	2	22	
Disgusted	4	1	25	
Excited	27	20	74	
Fond	14	9	64	
Happy	61	39	64	20 (?)
Hurt	58	27	47	3
Interested	32	26	81	
Kind	55	27	50	
liked	27	11	41	
Romantic	18	16	89	1
Sad	44	27	61	
Sneaky	16	10	62	1
Sorry	14	6	43	1
Sure	71	48	68	
Surprised	9	6	67	1
Thinking	19	15	79	1
Touched	8	6	75	1
Unfriendly	118	64	54	
Unsure	43	28	65	3
Wanting	19	13	68	
Summary	749	459	61	

Table 7-5: The number of concepts and the percentage of concepts for which vocal characteristics of the recognisable expressions have been mapped.

This mapping method reinforces the connection between meaning and expression, because it better demonstrates the differences between concept groups and concepts within groups of the MindReading taxonomy. However, from this presentation additional connections between concepts from different groups become apparent.

Therefore, the next section presents another mapping method that defines and presents the relations between concepts according to the recognisable expression set.

7.3 Expression based mapping

This section presents the second mapping method. This method consists of concept mapping in the nine-dimensional space which is defined by the recognisable expression set. Each of the recognisable expressions is presented as a core with its different combinations (it could be translated to a graphical representation of a root from which the combinations spring). The recognised concepts, their locations in the space and the relations between them are defined by the inferred combinations of recognisable expressions. It is not the usual definition of space because distances are not defined in it, although they could be added.

This mapping method allows the examination of the following parameters:

1. The relations between concepts with the same (or nearly the same) inferred combinations of expressions.
2. The characteristics of concepts whose inference includes a (core) expression.
3. The relations between expressions.

This mapping method reveals relations between expressions of concepts from different concept groups, as indicated in the previous section. These connections may be associated with meanings that have not been considered in the taxonomy. The inference of each concept and its relations to the inferred expressions could be also examined here, but this examination was performed in the previous section and therefore is not repeated.

Method

Each of the next tables includes the mapping with one of the recognisable expressions as a core and its combinations.

The column on the left includes the concepts whose inference includes the core expression. The second column indicates the concept group according to the MindReading taxonomy, as reference. The nine columns on the right consist of the recognisable expressions. Marked cells in one of these columns indicate that an expression was inferred for the concepts on the left. As in the previous tables, • signifies recognition in at least four sentences (>67%), while ◦ signifies expressions that were recognised in 3 of these sentences (50%). Expressions recognised in three sentences are perceived as less significant and less accurate. Only concepts in which the core expression was recognised in more than 4 sentences appear in its related table.

The colours aim to highlight some of the properties of the inference and mapping and therefore they are rather subjective. In the nine expression columns on the right, turquoise indicates expressions that could be expected for the concept or for its acting. Yellow indicates unusual or less expected expressions and expression combinations (as in the tables in the previous section). The grey columns mark expressions that were not recognised in combination with the core expression. In the two columns on the right, red text indicates one of the following occurrences: an unexpected inference result, misclassification in the taxonomy (as discussed in the previous section) or that a concept's meaning is significantly different from the concepts around it.

Results

The first recognisable expression presented in Expression Table 7-1 is *joyful*. This expression was recognised for both positive and negative concepts. Most of the positive concepts appear as *joyful* by itself or in combination with *excited* and *interested* with a possible recognition of *sure* in one concept. Five positive concepts include also recognition of either *stressed* or *opposed* like most of the negative concepts. The negative concepts mostly present extreme expressions of emotions. The expressions *thinking* and *absorbed* do not co-occur with *joyful*. *Sure* and *unsure* are not obvious in these combinations.

The expression *joyful* appears only in less than half of the groups in the MindReading taxonomy.

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
teased	hurt									
insincere	sneaky									
amusing, delighted, enjoying, felicity, glad, joking, triumphant, outgoing, unabashed, playful, joyful	happy	●								
vain	sure									
silly	unsure									
happy	happy	●				○				
tickled	happy	●		○		○				
jubilant	happy	●			○	○				
appalled	surprised	●			○	●				
frantic	afraid	●			●	○				
uncontrolled	excited	●								
terrorised	hurt	●			●	●	●			
hysterical	sad	●			●		●			
despairing	sad	●			●					
titillated	excited									
denying, neurotic	unsure									
faddy	wanting	●			○					
raging	angry									
forbidding	unfriendly	●				●	○			
rejoicing	happy	●				●	○			
festive	happy	●					○			
hostile	unfriendly	●					●			
suicidal	sad									
dudgeon	angry									
overjoyed	happy	●						●		
merry	happy	●						○		
shocked	surprised	●							○	

Expression Table 7-1. The recognisable expression *Joyful* and its combinations. Grey columns - expressions that do not appear in combination with the expression *joyful*. Yellow - expressions that cannot be easily explained. Red - concepts and groups that appear in combinations that do not agree with their lexical meaning, or different from the other concepts with the same combination.

The expression *absorbed* and its combinations appear in Expression Table 7-2. *Joyful*, *stressed* and *excited* do not usually co-occur with it. This expression appears mostly by itself and with *thinking*, *sure* and *unsure*. *Interested* also hardly appears, maybe because they present opposites in one dimension, the centre of *interested* is outside the speaker

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
troubled	sad		●		●					
engaged	interested		●		○				○	
committed	sure		●						○	
daunted	afraid									
cowed	afraid		●						●	
comforting	kind									
possessed	excited									
flattering, attracted	romantic									
suppliant	wanting									
sensitive	interested		●						○	○
intimate	romantic									
exhausted	sad		●						○	●
brooding	thinking									
pitied	sad		●						●	○
absorbed	interested									
considering	thinking									
coercing	unfriendly		●						●	●
disturbed	afraid									
isolated	hurt									
sexy, romantic	romantic									
secretive	sneaky									
ashamed, ignominious	sorry									
wistful	touched									
agonising	sad		●							○
fond	fond									
condescending	unfriendly		●							●
indulgent	kind									
intense	excited									
focused, hypnotised	interested									
shy	unsure									
downtrodden, compelled	hurt									
fancying	romantic									
dreamy, fantasising	thinking									
luring, manipulative	sneaky									
doubtful	disbelieving									
sentimental	touched									

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
sadistic	unfriendly		●	●						
regard	fond		●	○			○			
unworried	sure		●	●			●			
inhospitable	unfriendly		●	●		○	○			
controlled	sure		●	●		●				
calm	happy		●					○		
frustrating	unfriendly		●					●		
deferential	unsure									
charismatic	happy									
mean, picky	unfriendly									
convincing	sure									
dreading	afraid									
discouraged, soulful	sad									
patient, pitying, protective	kind									
concentrating, conscientious, dazzled, thorough, involved, obedient	interested		●							
naive	unsure									
betrayed	hurt									
worshipping, affectionate, devoted, doting	fond									
loving, lecherous, seduced	romantic									
distant, negligent, unimpressed, unthinking	bored									
particular	wanting									
guarded	disbelieving									
comforted, preferred	liked									

Expression Table 7-2. The expression *absorbed* and its combinations.

while *absorbed* is mostly centred on the speaker. However, there are a few concepts whose focus is outside such as *charismatic*, *convincing* and *protective* that include *absorbed* in the recognised combinations. They may be interpreted as aspects of concentration, and therefore indeed belong here. In this group there are many concepts for which the expression *absorbed* seems to relate to the manner of acting rather than to the concept definition.

The expression *sure* appears in Expression Table 7-3. *Sure* appears mostly by itself and in combination with *opposed*, *absorbed* and *excited*. The combination of *sure* and *excited* appears mostly in concepts which relate to negative attitude. The dominant groups from the MindReading taxonomy are *unfriendly*, *happy*, *angry*, *kind*, *sure*, *liked*, *disgusted* and *sad*. The concept *attractive* from the *romantic* group could have been also in the *sure* and *liked* groups. Certain expressions from different concepts groups in the MindReading taxonomy could be grouped into new meaning groups, such as concepts from the concept groups *kind* and *happy* or *liked* and *happy*, as well as concepts from the concept groups *sure* and *unfriendly*.

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
cold	unfriendly			●				○		
remote	unfriendly			●			○	○		
acidic, condemning, gruff, uncaring, stern	unfriendly									
amenable	happy									
bitter	angry									
confident, bullish, insistent, knowing, narcissistic, resolved, uncompromising	sure			●			○			
fatherly	kind									
acclaimed	liked									
bombarding, disagreeing, malicious, rejecting, sarcastic, thwarting, unkind	unfriendly									
easy-going	happy									
embittered, moody	angry									
adamant, arrogant, assertive, conceited, decided, determined, egotistical, purposeful, righteous	sure			●			●			
acknowledging	kind									
judged	hurt									
averse	disgusted									
punishing	unfriendly	○		●			●			
unworried	sure		●	●			●			
inhospitable	unfriendly		●	●		○	○			
dogmatic	sure			●		○	●			
vindictive	unfriendly			●		●	○			
boorish	unfriendly			●		●	●			
complacent	bored									
forthcoming	kind			●		○				
slighting	unfriendly			●		●				
tolerant	happy									
controlled	sure			●		●	●			
sadistic	unfriendly		●	●						
complimenting	kind		○	●						

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
antagonistic, intimidating, persecuting, snotty, unreceptive	unfriendly									
self-deprecating	unsure									
attractive	romantic									
disillusioned, subdued	sad									
accepted, conciliated, reassured, trusted	liked			●						
blank	bored									
comfortable, gratified, mischievous, protected	happy									
pompous, sure, convinced, decisive, enduring, prepared, resisting, staunch	sure									
responsive	kind									
grave	sad			●						●
tender	kind			●					●	●

Expression Table 7-3. The expression *sure* and its combinations.

The recognisable expression *stressed* and its combinations are presented in Expression Table 7-4. It does not appear in combination with the expression *sure*. There is a group of concepts that includes combinations with *joyful* and *excited* which are negative in nature. There are also more sedate combinations with *unsure* and *interested*. Again this expression appears mostly in very specific meaning groups, mostly negative, from to the MindReading taxonomy, including: *bothered*, *sad*, *excited*, *afraid*, *unsure*, *hurt*, *sad*, *angry* and *surprised*.

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
overrun	bothered				●	○			○	
tormented	sad				●	●			●	
upbeat	excited									
umbrage	hurt				●	●				
bothered	bothered									
helpless	unsure				●	○		○		

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
polite	kind				●			○		
rushed	bothered									
asking	interested				●			●		
hampered	bothered									
worried	afraid				●			●	○	
nervous	afraid				●			●	●	
hysterical	excited				●				○	
inattentive	bored									
overwrought	sad				●				●	
stressed	afraid									
preoccupied	thinking									
realising	thinking				●				●	●
troubled	sad		●		●					
flexible	happy				●		●			
hysterical	sad	●			●		●			
terrorised	hurt	●			●	●	●			
bewildered	unsure	○			●	○				
frantic	afraid	●			●	○				
uncontrolled	excited									
charitable	kind	○			●					
scandalised	surprised									
hurried	bothered									
despairing	sad	●			●					
titillated	excited									
denying, neurotic	unsure									
faddy	wanting									
distraught, turmoil	sad				●					
complaining, exasperated, frustrated	angry									
forward	sure									
cowardly, jumpy, panicked, terrified, pressured	afraid									
praising, kind, caring	kind									
emotional	excited									
intrigued	interested									
sensitive	unsure									
corrected, mistreated, questioned, scorned, tortured, trapped	hurt									
flustered, frantic, impatient, pestered, restless, ruffled	bothered									
mortified	sorry									
effusive	touched									
horrified	surprised									

Expression Table 7-4. The expression *stressed* and its combinations.

The expression *excited*, as can be seen in Expression Table 7-5, seems to signify a higher arousal level in comparison to *stressed*. It also has more positive aspects. It appears mostly in combination with *joyful*, *stressed*, *sure* and *opposed*. There are a few cases in which the excitement is related to *absorbed*, such as *inspired*, or to *interested*, such as *exhilarated*. Extreme stress appears in combinations of both *excited* and *stressed*, occasionally accompanied by *joyful*, which also indicates high arousal levels. The more complex mental states which are related to cognition and thinking, such as *thinking* and *interest* do not appear in most cases. Nor do *absorbed* and *unsure*.

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
inspired	excited		○			●				
controlled	sure		●	●		●				
slighting	unfriendly			●		●				
tolerant	happy									
privileged	happy			○		●				
gleeful	unfriendly			○		●	●			
boastful, challenging	sure									
vindictive	unfriendly			●		●	○			
boorish	unfriendly			●		●	●			
complacent	bored					●				
exhilarated	excited					●	○	○		
disruptive	unfriendly					●	○			
competitive	sure					●				
disobedient, odium, unruly	unfriendly									
heated	angry					●	●			
forceful	sure									
forbidding	unfriendly	●				●	○			
rejoicing	happy	●				●	●			
appalled	surprised	●			○	●				
terrorised	hurt	●			●	●	●			
enthusiastic	excited				○	●	●			
tormented	sad				●	●			●	
upbeat	excited									
umbrage	hurt				●	●				
bothered	bothered									
obsessed	interested									
startled	surprised									
self-confident	sure					●				
spirited, alert, dynamic, excited, invigorated, vibrant	excited									
lively	excited	○				●				

Expression Table 7-5. The expression *excited* and its combinations.

The expression *opposed* is the only expression that projects a decidedly negative attitude.

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
unworried	sure		●	●			●			
irritating, argumentative, discouraging, despising, uncooperative	unfriendly									
grumpy, disagreeable	angry									
opinionated	sure			○			●			
demanding, rapacious	wanting									
corny	touched									
indulged	liked									
bombarding, disagreeing, malicious, rejecting, sarcastic, thwarting, unkind	unfriendly									
easy-going	happy									
embittered, moody	angry									
adamant, arrogant, assertive, conceited, decided, determined, egotistical, purposeful, righteous	sure			●			●			
acknowledging	kind									
judged	hurt									
averse	disgusted									
defiant	unfriendly			○		○	●			
dogmatic	sure			●		○	●			
gleeful	unfriendly			○		●	●			
boastful, challenging	sure									
boorish	unfriendly			●		●	●			
complacent	bored									
needled	angry									
dictatorial	sure					○	●			
flattered	liked									
disobedient, odium, unruly	unfriendly									
heated	angry					●	●			
forceful	sure									
contradictory	unfriendly				○	○	●			
enthusiastic	excited				○	●	●			
flexible	happy				●		●			
hysterical	sad	●			●		●			
terrorised	hurt	●		●	●	●	●			
rejoicing	happy	●				●	●			
hostile	unfriendly									
suicidal	sad	●					●			
festive	happy									
dudgeon	angry									

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
blaming, disinclined	unfriendly	○					●			
explosive	angry									
resentful	unfriendly	○		○			●			
punishing	unfriendly	○		●			●			
wilful	sure						●	○		
menacing	unfriendly						●			●
maudlin	sad									
nonchalant, proud, boosted	happy									
insulting, brazen, confrontational, contrary, critical, derisive, devastating, difficult, disapproving, disconcerting, illiberal, rude, interfering, obstructive, vulgar, unwilling, violent, overpowering	unfriendly									
angry, indignant, miffed, wild	angry									
persuaded, cocky, compelling, inexorable, egocentric, firm, immovable, relentless	sure						●			
gallant	kind									
keen , motivated, reckless	excited									
misunderstood	hurt									
acquisitive	wanting									

Expression Table 7-6. The expression *opposed* and its combinations.

It appears in combinations mostly with *sure*, but also with *stressed*, *excited* and *joyful*. The combinations with *sure* are mostly concepts that express unfriendliness, or difference of opinion. Its combinations with *excited* express anger or challenging. The combinations with *joyful* indicate lack of control in most cases. The expression *opposed* and its combinations encompass concepts from 13 groups from the MindReading taxonomy.

The expression *interested* is related to more complex mental states, and is trained to infer different types of questions. Therefore *sure* does not appear in combinations with *interested*. *Interested* appears in combinations with *unsure*, *absorbed* and *stressed*. The last combination is common mostly to *stress* related concepts, and the inclusion of *interested* may relate more the content in the form of questions. There are hardly any combinations of *interested* with high excitement levels such as in *joyful*, *opposed* and *excited*.

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
helpless	unsure				●	○		●		
encouraging	kind				○			●		
asking	interested				●					
hampered	bothered									
worried	afraid				●			●	○	
nervous	afraid				●			●	●	
puzzled	unsure							●	●	
trusting	fond									
dazed	surprised							●	●	○
excluded	hurt		○					●	●	
peaceful	happy									
displeased	angry		○					●		
receptive	kind									
frustrating	unfriendly		●					●		
deferential	unsure									
relieved	happy							●		●
passive	bored							●		○
overjoyed	happy	●						●		
unapproachable	unfriendly									
blameless, familiar	happy									
afraid	afraid									
excusing, concerned, sensitive, warm	kind									
tempted, fascinated, curious, interested, listening, probing, quizzical	interested							●		
criticised	hurt									
vacant	bored									
questioning	disbelieving									
favoured	liked									

Expression Table 7-7. The expression *interested* and its combinations.

The next expression is *unsure*. Its concepts and combinations appear in Expression Table 7-8. The dominant expressions that appear in combination with *unsure* are *thinking*, *interested*, *absorbed* and *stressed*. The expressions *joyful*, *opposed*, *sure* and *excited* are apparently related to higher levels of confidence and self-esteem and there are very few combinations of *unsure* that include them.

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
excluded	hurt		○					●	●	
dazed	surprised							●	●	○
faltering	unsure							○	●	●
bewitched	romantic		○					○	●	●
hesitant	unsure		○						●	○
pitied	sad		●						●	○
absorbed	interested									
considering	thinking									
coercing	unfriendly		●						●	●
disturbed	afraid									
isolated	hurt									
sexy, romantic	romantic									
secretive	sneaky									
ashamed, ignominious	sorry									
wistful	touched									
persuading	kind		○						●	
awed	interested									
undecided	unsure									
debating	thinking									
requesting	wanting									
cowed	afraid		●						●	
comforting	kind									
possessed	excited									
flattering, attracted	romantic									
suppliant	wanting									
tender	kind					●			●	●
pleasure	happy								●	●
lured	interested									
deflated	hurt									
adoring	fond									
choosing, thinking	thinking									
mysterious	sneaky									
regretful	sorry									
touched	touched									
realising	thinking				●				●	●
tormented	sad				●	●			●	
soft	kind				○				●	
inadequate	unsure									
overwrought	sad				●				●	
stressed	afraid									
preoccupied	thinking									

Concepts	**Group**	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
nervous	afraid				●				●	●
puzzled	unsure								●	●
trusting	fond									
murderous	unfriendly								●	
lonely, pining, reconciled, weak	sad									
accepting	happy									
sincere	sure									
shaken, watchful	afraid									
nosy, spellbound, tantalised	interested									
confused, clueless, insecure, humble, indecisive, inferior, innocent, unsure, baffled, disorientated	unsure									
adored	liked									
scolded, resented, disbelieved, hurt, patronised, scrutinised	hurt									
humiliated, disgraced	sorry									
enticed	romantic									
unenthusiastic	bored									
fawning	sneaky									
appealing, greedy, begging	wanting									

Expression Table 7-8. The expression *unsure* and its combinations.

The last recognisable expression is *thinking*. It combinations appear in Expression Table 7-9. It encompasses a large variety of concepts and meanings. It appears mostly in combination with *absorbed*, as expected, and in combinations with both *sure* and *unsure*. There are also a few combinations with *interested*. The expressions that relate to high arousal and activation levels, including *joyful*, *opposed*, *stressed* and *excited* do not appear in these combinations, although they include both negative and positive concepts from 18 of the 24 groups in the MindReading taxonomy.

Concepts	**Group**	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
grieving	sad		○							●
hopeless	unsure									
neglected	hurt									
entrancing	romantic									
calculating	thinking									

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
condescending	unfriendly									
indulgent	kind									
intense	excited									
focused, hypnotised	interested									
shy	unsure									
downtrodden, compelled	hurt		●							●
fancying	romantic									
dreamy, fantasising	thinking									
luring, manipulative	sneaky									
doubtful	disbelieving									
sentimental	touched									
lovelorn	sad		○					○		●
bewitched	romantic		○					○	●	●
exhausted	sad		●						○	●
brooding	thinking									
coercing	unfriendly									
disturbed	afraid									
isolated	hurt									
sexy, romantic	romantic		●						●	●
secretive	sneaky									
ashamed, ignominious	sorry									
wistful	touched									
realising	thinking				●				●	●
listless	bored									
furtive	sneaky								○	●
longing	wanting									
nostalgic	touched									
pleasure	happy									
lured	interested									
deflated	hurt									
adoring	fond								●	●
choosing, thinking	thinking									
mysterious	sneaky									
regretful	sorry									
touched	touched									
tender	kind			●					●	●
grave	sad			●						●
calming	kind			○						●
menacing	unfriendly						●			●
relieved	happy							●		●
mystified	unsure							○		●
discomforted	afraid									

Concepts	Group	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
forgiving	kind									
empty, tired, self-pitying, suffering, lovesick	sad									
so-so	unsure									
languid	happy									
belittled, demoralised	hurt									
loving	fond									•
aroused, lustful, seductive	romantic									
comprehending, thoughtful, deciding, regarding, wool-gathering	thinking									
jaded, noncommittal, vague	bored									
sneaky, lying, tempting	sneaky									
needy, wishful	wanting									

Expression Table 7-9. The expression *thinking* and its combinations.

Discussion and summary

This section describes a new approach to automatic mapping of mental state concepts and expressions. This method maps the concepts into the nine-dimensional space which is defined by the recognisable expression set. It aims to explore how a very large set of concepts can be mapped using a relatively small sub-set of concepts, which are defined by their vocal correlates.

The expression mapping can be examined according to the following criteria:

- The relations between concepts with the same (or nearly the same) inferred combinations of expressions.
- The characteristics of concepts whose inference includes a (core) expression.
- The relations between expressions.

The conclusions are: There are common characteristics to concepts with the same or similar combinations of inferred expressions. In many cases, the common behavioural characteristics indicate a common meaning.

In other cases, the combinations reveal similar expressive characteristics of concepts with different meanings. The common characteristics could indicate for example arousal level, self-absorbance, and more.

Certain combinations tend to indicate meaning related properties such as valence (positive / negative).

Certain expressions represent or appear in relation to specific meaning groups from the MindReading taxonomy. Furthermore, certain expressions are more associated with concepts and concept groups that are commonly defined as basic emotions, while there are other characterising expressions for concepts and groups that are associated with cognitive processes and complex mental states.

Not every combination of expressions is observed. Table 7-6 summarises the more common combinations:

There are recognisable expressions that never co-exist is this database and others that are very dominant. The fact that a combination does not appear in the database does not imply that such a combination does not exist. A few unexpected combinations describe extreme or special concepts. Some of them relate to inherent meanings of the concepts.

As mentioned before, the definition of the recognisable expression set is relatively coarse, more precise grouping could be advantageous for further distinctions between meanings.

This method represents a novel approach to expression mapping by using a small set of concepts for describing the relations within a large set of concepts. The appealing property of this method is the link of the concepts to behavioural characteristics. This method presents the role of each of the recognisable expressions to inference and to mapping. It reveals new relations and groups within, across and beyond the definitions of the MindReading taxonomy. It emphasises relations between concepts. It demonstrates some of the shortcomings of acted emotions. It also enables to see that additional information or inferred expressions can be useful in order to better distinguish between expressions and expression groups, as the set of nine expressions is by no means complete.

	Joyful	**Absorbed**	**Sure**	**Stressed**	**Excited**	**Opposed**	**Interested**	**Unsure**	**Think**
Joyful				√	√	√	√		
Absorbed			√			√	√	√	√
Sure					√	√			√
Stressed					√	√	√	√	
Excited						√			
Opposed									
Interested								√	
Unsure									√
Think									

Table 7-6: Summary of the existing combinations between expression pairs.

7.4 Summary

This chapter introduces three sets of results. Each of them has its own significance. The first part shows the ability of the inference machine to distinguish between complex mental states, beyond the set of expressions that it was trained to recognise. In this test the machine's ability to recognise different combinations of expressions by which different concepts are distinguishable outperformed human results.

This ability of the machine to characterise concepts by combinations of recognisable expressions was then used to examine the vocal correlates of the MindReading taxonomy.

The inference was tested on 749 concepts, of which 459 concepts of emotions and mental states had consistent recognisable combinations. At least 85% of these combinations agree with the lexical meaning of the recognised concept, or can be explained by the limitations posed by their acting. The fact that so many concepts were correctly characterised demonstrates the validity of these results. These results show that this machine and technique supply information about the nature of expressive speech, the relation between lexical meanings and the way they are expressed by people.

The concepts and their inferred expressions were mapped according to the concept groups which are defined by the MindReading taxonomy. This mapping shows that indeed many of the concepts in the concept groups in the taxonomy have common characteristics that are often different from the characteristics of other meaning or concept groups. They include common meanings, common conveyed expressions and the vocal characteristics of these expressions. However, they also show that these groups are not mutually-exclusive and the relations between concepts are more complicated and can be presented by different systems and methods. Therefore, an additional mapping technique is presented.

The second mapping technique defines the relations between concepts according to combinations of the recognisable expressions. It reveals new relations and differences between concepts. Many concepts that are located in different concept groups in the MindReading taxonomy are close to each other according to this mapping. It also shows that certain combinations of expressions are more common than others and that some expressions never co-exist.

A further conclusion is that the acting can be very influential. For example, the acting of many expressions that relate to human relations was delegated into monologues and introspective moods, while it is not necessarily the case in daily situations. Certain concepts were more difficult to act, especially those that describe lack of control, and yielded several irregular combinations. It is also clear that these nine expressions are not enough to fully represent and distinguish among a large variety of concepts, although they present correctly, or map, an amazingly large number of concepts and their relations.

This mapping presents a novel approach to the conceptualisation and presentation of multiple emotion and mental state concepts and the relations between them. The ability of the machine to define properties of expressions using a limited set of expressions, in a manner that corresponds to the lexical definitions of the new expressions defines a new angle for mapping expressions. The recognisable expression set is a small arbitrary set of expressions of complex mental states which are not necessarily *basic emotions*. Their definition according to their vocal characteristics provides the behavioural characteristics of lexical concept and meaning-based groups.

This mapping has interesting implications and the approach is appealing for several reasons. The mapping allows navigation among a very large set of expressions. The approach resembles learning of new expressions by extending the knowledge using the existing knowledge as a basis. The mapping allows the examination of various descriptive schemes and theories, because it compares the ideas to the behavioural characteristics of the expressions. It offers a new way for examining conceptualisation theories using concrete evidence from related human behaviour. The automatic inference machine is used here as an automatic mapping machine. In this capacity it presents an objective tool for analysing *meaning.* The relations of lexical oriented grouping of expressions with paralinguistic characteristics of the same groups or with expressions from other groups can be used for a large variety of applications and disciplines, such as educational and therapeutic systems.

References

[1] Golan O., Baron-Cohen S., and Hill J., “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.

[2] Baron-Cohen S., Hill J. J., O. Golan, and S. Wheelwright, “Mindreading made easy.”, *Cambridge Medicine*, vol. 17, pp. 28–29, 2002.

[3] S. Baron-Cohen, Golan O., Wheelwright S., and Hill J. J., “Mindreading: The interactive guide to emotions”, London: Jessica Kingsley Limited (www.jkp.com), 2004.

[4] Baron-Cohen S., Golan O., Hill J. J., and S. Wheelwright, “Developmental changes in the emotion lexicon: a research note”, submitted.

[5] Russel J. A. and Barrett L.F., “Core affect, prototypical emotional episodes, and other things called emotion: Dissecting the elephant”, *Journal of Personality and Social Psychology*, vol. 76, no. 5, pp. 805–819, 1999.

[6] Cowie R., Douglas-Cowie E., Tsapatsoulis N., Votsis G., Kollias S., Fellenz W., and Taylor J. G., “Emotion recognition in human-computer interaction”, *IEEE Signal Processing Magazine*, vol. 18, pp. 32–80, 2001.

[7] J. T. Larsen, A. P. McGraw, and Cacioppo J. T., “Can people feel happy and sad at the same time?”, *Journal of Personality and Social Psychology*, vol. 83, no. 684-696, pp. 684–696, 2001.

Chapter 8

Interaction level multi-modal inference

This chapter describes the inference and analysis of expressions in the context of a sustained interaction. Interaction level analysis is part of the fifth layer of the framework which is described in in Chapter 3. It focuses on the dynamic analysis of a sustained interaction and how transitions are made between sentences and expressions. The interactions are human-computer interactions during a computer game, with intervals of human-human interactions and intervals of thinking aloud. The analysis enables the understanding of behavioural patterns during the interaction and has potential for predicting future behaviour and consequent events based on previous observations. The implementation of this level in a system and the manner in which the cues are interpreted and used, depend on the required application. Interaction analysis can be regarded as a step towards the integration of the technology of affect inference with dialogue systems.

An aspect for examination is the universality of the system or its ecological validity. The questions are whether the inference system can be used for a large variety of expressions, acted and natural, and its suitability for different languages and contexts.

The inference machine that was trained on an English database of acted expressions by British actors is applied to a Hebrew database that consists of naturally evoked expressions of non-actors Israelis in an HCI situation. It is therefore required to establish the fact that the inference results for this database are valid or at least reasonable. The labelling of naturally evoked expressions and mixtures of subtle expressions is not trivial, as discussed in Chapter 4. Therefore, the verification is done by comparison of the inferred expressions to events, to other physiological and behavioural cues and to the verbal content.

The first section of this chapter presents preliminary observations of various vocal cues during interactions and draws requirements for the implementation of sustained interaction analysis. The second section presents analysis of recorded interactions in the Doors database. It presents both statistical correlation results and analysis of co-occurrences of the inferred expressions with the various recorded cues throughout the interactions. It also demonstrates patterns of other behavioural cues throughout an interaction as an example of how the analysis can be extended to multi-modal inference systems.

8.1 Expressions in time

This section describes some of the observations, performed on the Doors database, that demonstrate temporal characteristics of expression. These experiments were carried out in order to verify the initial assumption regarding the importance of the interaction to inference of single utterances, and the relation between utterances in a sequence, during an interaction. It includes observations of intonation curves, spectrograms and initial classification result. The classification is based on manual labelling of a single possible dominant label per utterance. Because of the subtle nature of the expressions, and probably due to their complex combinations, only part of the utterances throughout the interaction could be labelled in such a definite manner. In order to allow labels consistency, six labels were assigned by the first listener and the rest of the listeners had to decide which of these six labels agreed with the examined utterances. As mentioned in Chapter 4, this is the common practice for such labelling. These labels are slightly different from the labels used by the inference machine.

Continuity and thresholds

One of the main observations is related to the changes of expressions over time. In most cases there is indeed a gradual change between consecutive utterances. However, the experiments showed that sometimes, especially during transitions, samples from different groups, labelled as different concepts, are closer to each other then to other samples from the same group. One reason is the gradual transitions between expressions and within nuances of expressions over time, as discussed in Chapter 2. The second is that different expressions are noticed if a certain threshold has been passed. These observations have strengthened the initial assumption that continuous analysis can reveal tendencies in the interaction that certain applications should be able to change in an early stage, such as boredom or frustration. They also reinforced the assumption that investigating mixtures of expressions is desirable.

Figure 8-1 shows an example of these two scenarios. The figure shows four consecutive pitch contours of the sentence *sgor de-le-t*, meaning 'close door' in Hebrew, uttered by a male speaker. Three of them were labelled as *uncertain*, while the fourth was labelled as *determined* and has a different contour. It can be seen that the *uncertain* curves change gradually toward the next expression of *determined*. The change is gradual, and can be seen in different characteristics, such as the length of the utterance, the main slope of the intonation, the range of f_0, and the slope, duration and shapes of the different parts. There is a threshold beyond which these gradual changes turn into a new inferred expression.

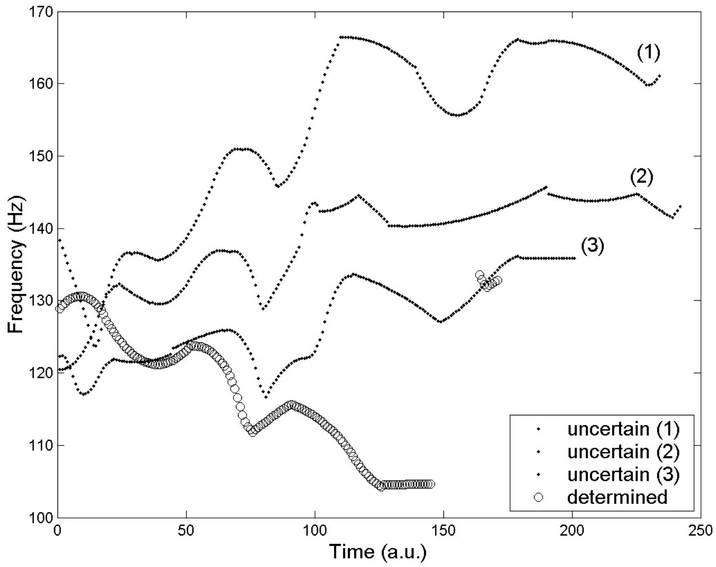

Figure 8-1: Pitch contours of consecutive utterances. Three with expression of *uncertain* and the fourth of *determined*.

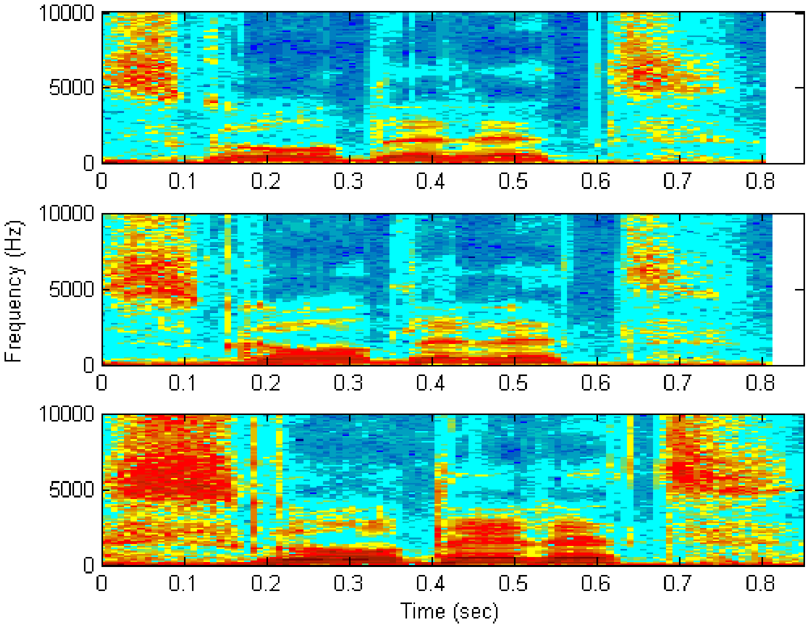

Figure 8-2: Spectrograms of three consecutive utterances with expression of *uncertainty*.

Figure 8-2 shows spectrograms of three consecutive utterances with the label *uncertain*. They reveal additional information, which is related to local patterns and their changes during the interaction, such as the gradual change in the duration of voiced and unvoiced parts, which increases for example, for the first part of the sentence (0 - 0.2 sec), between utterances. Other parameters that change among the utterances include the distances between consecutive speech parts, the total energy, and its distribution in the frequency domain.

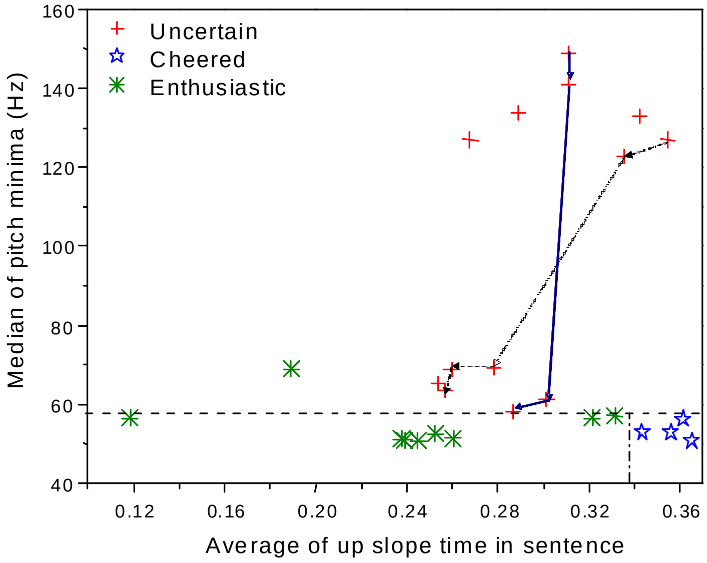

Figure 8-3: Classification using C4.5 algorithm of several sets of utterances of *sgor-de-let* by the same male speaker, with the labels *uncertain*, *cheered* and *enthusiastic*. Each point is a sentence (utterance). The arrows indicate consecutive utterances in time.

Figure 8-3 shows classification of several sets of utterances of *sgor-de-let* by the same male speaker, with the labels *uncertain*, *cheered* and *enthusiastic*. Each point is a sentence (utterance). The algorithm used for classification is C4.5 from the J48 package of Weka [1]. The arrows indicate gradual transitions between consecutive utterances in time. In some cases the transition is towards a different class or expression. In the case of *uncertain*, it is clear that the classification is changed by crossing a threshold rather then by using definite class group locations. Some of the utterances are closer to the next expression than to the main body of the utterances with the same expression, and would have classified as *enthusiastic* if blind clustering was used in this case.

These observations show how subtle changes in expression occur gradually over time, until a certain point is reached beyond which a different expression is perceived. It also demonstrates the importance of following an interaction and not only considering stand-alone utterances. The changes are not always gradual though. In extreme cases we can find consecutive utterances which are completely different from each other.

8.2 Analysis of the Doors Database

This section describes inference results of expressions during human-computer interactions, and interim human-human interactions. The experiment was designed so that only the participants' choices, the resulting game events and participants' expressions change during the interactions. The hypothesis was that the inferred expressions would follow or precede game events and participants' choices. This section presents statistical results that show significant differences in expression recognition that co-occur with certain game events. It also presents several interaction analyses in time, showing gradual changes, tendencies and the effect of single-occurrence events.

The goal here is not to build or demonstrate a new machine for interaction analysis, but rather to demonstrate how the inference machine could be applied to such analysis. It also demonstrates what types of information could be extracted and inferred from analysis of speech and multi-modal cues during an interaction. Another conclusion derived from the results is that the machine could be used for analysis of other languages, as validated by the correlation between the various cues.

The experiment

This section summarises the features of the Doors database that are relevant to the understanding of the experiment, including details of the computer game which was used to evoke expressions and characteristics of the participants.

The Doors game is based on the Iowa Gambling Test (IGT) by Bachara *et al.* [2, 3]. Computerised and manual versions of this game are commonly used in psychology research [4].

As in the IGT, the participants' aim was to win as many points as possible. The game requires a series of 100 door selections, one door at a time. After choosing each door, the subjects received some points. The amount was only announced after the door had opened. On certain occasions, after opening a door, the subjects were given points and then a penalty (points taken). The amount varied with the door according to a schedule unknown to the subjects. Choosing doors one and two yielded 100 points; choosing doors three and four yielded 50 points. However, the penalty amounts were higher in the high paying doors. Doors 1 and 2 were "disadvantageous" ('bad'), because they cost the most in the long run, while Doors 3 and 4 were "advantageous" ('good'), because they resulted in an overall gain in the long run. Therefore, an advantageous strategy would be to choose only these doors. After 20 trials and at the end of the game, the participants were asked what they knew about the game. With a few exceptions, no other interruptions were made.

This chapter does not aim to reproduce the original IGT results. The participant group was relatively small, and the setting was not completely identical to the original test, therefore no general behavioural characteristics of the participants are defined here. On the other hand, the test was designed so that for each participant the only things that changed during the interaction were the participant's choices and therefore the game events and the participant's reactions and expressions.

The Doors database offers natural expressions with controlled text. The repetition of the same text, by the same speaker, many times (100 repetitions for each sentence), with relations to context, allows examination of the dynamic changes of only the expressions during the interaction, including subtle changes and subtle expressions. Typical duration of an interaction is approximately 15 minutes.

Because the game supplies a large number of repetitive trials, the variety of situations and expressions per participant is relatively large. The experiment can therefore be based on comparisons between different sections of an interaction of each participant. The correlation between the recognised expressions and the game events in each scenario shows very interesting characteristics and the overall inference over time corresponds to the game events. Therefore, the fact that each participant behaved in a different manner enlarges the scope of the experiment. Each interaction can be regarded as an experiment by itself and the inference system proves to support different interaction manners.

Participants

As each interaction is an experiment, it is important to understand the differences between different interactions. These differences evolve from the characteristics of the different participants, their prior knowledge of the game, and the course of the interaction as revealed by their choices and the resulting game events.

This section presents the analysis results of 6 participants: 3 male and 3 female. All the participants described here were from the Engineering faculty. All the male participants were graduate students in electrical engineering of the age group 20-30. They had no prior knowledge of the game, unlike the report by Maia and McClelland [5], in which the participants were psychology students who had prior knowledge of the game. Furthermore, the venue was the Engineering faculty, and the experimenters were colleagues from the Engineering faculty. These facts are important because they may have influenced the participants' expectations of the game's requirements.

Two of the male participants (Male 1, Male 2) did not manage to devise any strategy to win the game, although their behaviour suggested otherwise in certain parts of the game. Male 3 on the other hand used a 3 doors rotation very successfully until the 99^{th} trial, when he lost many points.

The female participants were more diverse. The first participant (Female 1) was a graduate student in Electrical Engineering who managed to understand the rules towards the end of the game. The second participant (Female 2) was part of the experimenters' group and had prior knowledge of the game. The third female participant (Female 3) was an academic staff member who apparently had prior knowledge of the IGT. This assumption is inferred from her ability to describe the rules after 20 trials without experimenting all the rules at this stage. This participant was asked to use only one sentence ('open this door'). The other participants used two sentences, when they remembered to do so.

These different game patterns, although not conforming with the usual IGT results, supply interesting evidence regarding the ability to automatically analyse different interactions. The different instructions give additional knowledge for interpretation of the game events and participants' reactions.

Analysis Method

The examination of the inferred expressions here is slightly different from the method described in the previous chapters because the sum of the pair-wise machines is used for inference. This method is used because the utterances throughout the interaction are comparable in the sense that participants' short-term expressions and the game events are the only things that change. Therefore the assumption is that the machines reflect only the change in expression and every change should be considered.

T-tests were used for the significance measures. T-tests require normal distributions; therefore the distribution of each result group was examined. The interaction analysis was done by considering both specific events, and averaging of the current inference with the results of the last 5 utterances for better visualisation of tendencies.

General comparisons

This section presents an example of a general cross-participants comparison. It presents the participants' reactions to the common events of total gain above and below zero (positive or negative).

Table 8-1 shows the expressions depicted when the total gain was negative vs. positive, i.e. the total number of points gain until this event. As can be seen, different people reacted in different ways. It is not surprising because they also experienced different scenarios (in addition to different personalities).

Participant	Event	Temporary total gain	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
Male 1	Open door	negative	4.1	5.0	3.5	3.0	3.4	3.3	4.1	4.3	5.3
		positive	4.4	4.4	2.8	2.7	4.0	2.8	6.1	5.0	3.8
	Close door	negative	3.4	3.3	1.5	5.8	4.0	3.9	4.0	5.4	4.8
		positive	4.2	1.9	5.3	3.4	4.3	4.9	4.2	4.1	3.7
Male 2	Open door	negative	4.6	3.8	4.6	2.8	3.2	3.6	4.0	4.4	5.0
		positive	4.1	4.8	3.4	3.1	2.6	3.6	4.9	4.6	4.8
	Close door	negative	4.3	2.8	3.5	4.5	2.5	4.2	3.9	5.6	4.7
		positive	3.2	2.8	5.7	4.0	3.7	4.5	3.6	5.3	4.8
Male 3	Open door	negative	4.4	4.3	3.3	2.9	3.1	2.8	4.5	5.2	5.5
		positive	5.0	4.2	3.9	2.4	3.4	3.4	4.0	4.8	4.9
	Close door	negative	3.1	3.2	3.2	4.4	4.4	4.1	3.6	4.9	5.2
		positive	3.3	3.0	2.7	4.5	4.8	4.5	3.7	4.6	4.9
Female 1	Open door	negative	3.9	4.8	2.8	3.3	3.0	3.8	3.8	5.4	5.2
		positive	4.3	4.0	1.9	4.2	2.5	3.9	4.0	5.5	5.6
	Close door	negative	3.8	3.5	5.2	2.6	3.7	5.5	4.3	3.4	3.9
		positive	3.5	3.6	4.0	3.5	4.1	5.2	3.6	4.0	4.5
Female 2	Open door	negative	3.9	4.9	2.8	3.3	3.0	3.7	3.8	5.3	5.3
		positive	4.2	3.9	1.9	4.2	2.4	3.9	3.9	5.5	5.5
	Close door	negative	2.8	4.8	4.0	4.0	2.5	2.3	4.3	5.0	6.5
		positive	4.0	2.9	2.9	5.0	4.2	3.0	3.9	5.1	5.1
Female 3	Open door	negative	4.0	4.1	3.5	3.5	3.2	3.9	3.8	5.0	5.1
		positive	4.2	4.3	3.3	3.1	3.1	3.8	4.3	5.0	5.0

Table 8-1: Participants' reactions to positive and negative total gain (above and below 0).The significant expression differences are marked by both colour and '*', the numbers are the average (mean) number of machines that recognised each expression for the events on the left.

It can also be seen that most values are relatively low (below the mean which is 4). In the table, the events are described on the left, the recognisable expressions on the right. The tables include the mean values of machines that recognised an expression in each event group and the significant differences between expressions of different events (marked by both colour and '*').

Male 1 is the most expressive participant. When making a decision, he shows more interest and excitement when his gain is positive and thinks more when it is negative. After seeing the current result he shows more joy and confidence while being more *absorbed*, *stressed*, *unsure* and *thinking* when the gain is negative. The inferred mean values for the expression *opposed*, which are higher for positive gain, could have several explanations, for example - the positive result did not agree with the participant's expectation (disbelief, misunderstanding and the like). It could be related to the expressions of *enthusiasm* and high motivation, as in the MindReading. The label *enthusiastic* was used in the manual labelling of the database. On the other hand, it could manifest a difference between two speech cultures.

Most participants were more sure or confident or determined when they opened a door while losing. The expressions of Female 1 and 2 are similar when choosing a door. They project low overall confidence, but there is still a significant difference in confidence (sure) between the cases of total profit and loss. They are more determined or *sure*, and more *excited* and *absorbed* when loosing, while more *stressed* when they have positive gain. These results could also suggest that the gender dependency of reactions could be explored with such tools.

The participant Male 3 seems more *joyful* upon opening the next door when the gain is positive. Female 3 reveals no significant differences between these two cases.

There are expressions that do not change significantly, but on average where recognised more, such as *unsure*, *thinking* and *stressed*.

Individual participants

This section describes in more detail the main behavioural characteristics of 3 participants. It reveals significant behavioural changes that correspond to their individual experiences. The different presentation and analysis methods which are presented in this section provide different observation angles of the data and the events during the interaction and different types of information.

Female 2

Female 2 was part of the experimenters group and had prior knowledge of the game. She pretended to gamble to the 20^{th} trial, tried to win until the 60^{th} trial (her choice) and was then asked to gamble on the 'bad' doors. She shows completely different expressions between the game parts in which she gambled and between the part in which she could choose the safe and advantageous doors. Table 8-2 shows that she becomes more and more *absorbed* throughout the game, especially when choosing a door, as the levels of stress and confidence decrease. On the other hand, closing the door reveals other tendencies. There are similarities between the free gambling at the beginning and compulsory wrong choices at the end. The *joyful*, the *excitement* and the profit are higher with good choices. *Interest* and *thinking* are higher at the beginning, with the free gambling, and so is the opposition or displeasure.

	Measure	Event	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
Open door	Mean	1-20 free	5.4	4.1	2.5	3.7	3.2	5.1	3.8	4.2	4.2
		21-60 'good'	5.4	5.2	1.6	3.5	3.3	4.1	4.1	4.4	4.5
		61-100 'bad'	5.3	6.0	1.7	2.6	3.7	4.0	4.2	4.3	4.2
	Significant difference	1-20 vs. 21-60		*	*			*			
		1-20 vs. 61-100		*	*	*		*			
		21-60 vs. 61-100		*		*					
Close door	Mean	1-20 free	3.3	2.9	3.9	4.7	3.3	2.3	4.8	4.9	5.7
		21-60 'good'	4.3	2.9	2.4	5.1	4.6	3.2	3.5	5.1	4.9
		61-100 'bad'	3.3	3.5	3.6	4.8	3.2	2.6	4.2	5.3	5.5
	Significant difference	1-20 vs. 21-60	*		*		*	*	*		*
		1-20 vs. 61-100									
		21-60 vs. 61-100	*		*		*				
All speech signals	Mean	1-20 free	4.6	3.6	3.0	4.0	3.2	3.9	4.2	4.6	4.9
		21-60 'good'	4.8	3.9	2.1	4.2	4.0	3.7	3.9	4.7	4.7
		61-100 'bad'	4.6	5.4	2.2	3.1	3.5	3.7	4.2	4.5	4.6
	Significant difference	1-20 vs. 21-60			*		*				
		1-20 vs. 61-100		*	*	*					
		21-60 vs. 61-100		*		*	*				

Table 8-2: Female 2 had prior knowledge. Gambled freely in trials 1-20, chose only 'good' doors in trials 21-60 and was asked to choose only 'bad' doors in trials 61-100. The significant expression differences are marked by both colour and '*', the numbers are the average (mean) number of machines that recognised each expression for the events (left).

Figure 8-4 shows smoothed curves of the expression inference over time in comparison to the total profit and chosen doors. It includes all the sentences uttered by the participant during the interaction, including free test intervals (marked by gaps in the total profit curve). This perspective offers slightly different information. For example, it is obvious that the stress level has decreased drastically after the decision to gamble on the 'bad doors' was been taken. This observation is supported by galvanic skin response (GSR) measurements, in which the base line has also drastically changed at this point in time. The participant shows a combination of *joyful* and *opposed* and soon after she becomes more *absorbed* at this stage, and later the *stress* level returns to its previous values while her dominant expression turns to *thinking*. The most dominant expression changes throughout the interaction were between *unsure*, *joyful/happy*, *thinking* and *absorbed*. The expression *sure* is significantly low throughout the interaction. It is slightly higher at the beginning and during the second interview.

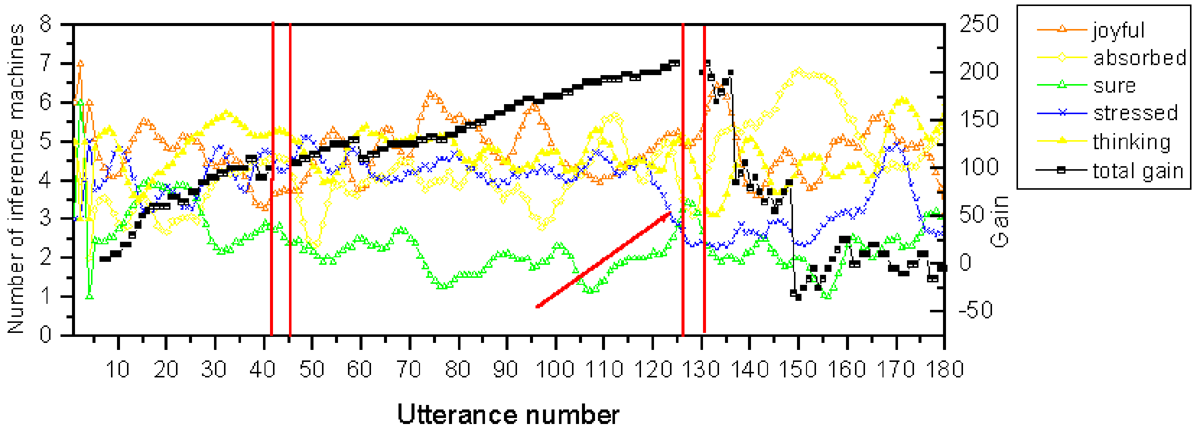

Figure 8-4: Female 2, expressions in time, total profit (total) and chosen doors (door index). The red arrow indicates the decrease in stress level apparent in the speech. The red lines indicate the borders of the two interviews and strategy changes.

More detailed observation, without averaging, shows yet another type of information, like the relation between the inferred expression and the verbal content. For example, during the second interval, in one sentence the most dominant expression is *interested*, and indeed the participant asks a question. In the other sentences during this interval, the dominant expression is *thinking*, when the participant tries to explain things and also tries to evaluate the suggestion to change tactics. These short intervals are not apparent in the long-term analysis.

These observations show that the inference machine offers a lot of information about the participant during an interaction. However, there are different ways to present and analyse the data in time. Each method provides different information. The information can be compared to other behavioural (text, choices) and physiological cues of the participant and to context related information.

Female 3

From the results it seems that Female 3 was aware of the game rules. She fully explained the rules after the first 20 trials with only partial results to support them. She was asked to continue to play, while complaining about boredom until the 71^{st} trial when she stopped playing. During the game she was asked to say only the sentence upon choosing the door, *ptah de-let zo*, meaning literally ’open door this’ (content: open this door), in Hebrew. However, there are significant differences between her expressions during the first 20 trials, before the interview and after the interview, in which that knowledge was proved correct and her inferred confidence level increased, as can be seen in Table 8-3. There is no significant difference between the inferred expressions before and after starting to gamble intentionally on the ’bad’ doors towards the end, after 34 trials (the text of the free speech utterances refers to boredom as the reason). The participant was more joyful before the confirmation, less opposed and less confident. Table 8-3 shows statistical results of both the designated phrase (open door) alone and with the free-text sentences. There is a difference between the game directed speech and the free speech in all the recognisable expressions. Another observation for this participant is the different types of laughter that could be distinguished by the inference machine. It was inferred in utterances with and without verbal content. The two main types were *joyful* and a combination of *joyful* and *stressed*.

Speech content	Events	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
Open door	1-20 before	4.6	4.4	3.1	3.2	2.9	3.3	4.3	5.0	5.0
	21-71 after	3.8	4.2	3.5	3.3	3.2	4.1	3.9	5.0	5.0
All sppech signals	1-20 before	4.2	3.8	2.6	4.0	3.6	3.6	4.4	4.9	5.0
	21-71 after	4.0	3.8	3.3	3.4	3.3	4.2	4.1	5.2	4.8

Table 8-3: The participant suspected prior knowledge which was confirmed after the first 20 trials.

The examination of this participant's interaction, combined with the different cues, allowed deducting information about her prior knowledge which was not available before. It demonstrates how multi-modal information can be used to improve understanding. It also shows that non-verbal utterances can be analysed automatically.

Male 1

As can be seen in Table 8-1 Male 1 was the most expressive participant. He spoke a lot to himself and laughed during a couple of intervals. Table 8-4 presents the reactions of Male 1 to the gain of the last opened door (positive or negative). There are different *joyful* reactions, one upon opening doors after a loss and after seeing current profit. The second is expected. The first can be explained by laughter which started after observing and digesting losses, as can be understood from the free text utterances. This laughter lasted into the next open-door event. In addition, the expression of *unsure* increases after temporary losses as well as for long term losses.

Participant	Event	Temporary gain	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
Male 1	Open door	negative	6.0	5.5	3.0	2.5	2.0	4.0	4.5	4.0	4.5
		positive	4.1	4.9	3.4	2.9	3.5	3.2	4.5	4.5	5.0
	Close door	negative	2.8	3.5	1.8	5.4	4.2	3.9	4.1	5.8	4.5
		positive	3.7	2.8	2.6	5.1	4.0	4.2	4.0	5.0	4.5

Table 8-4: Male 1 reactions to positive and negative last door's gain (above and below 0).

Participant	Event	Chosen door ('good' or 'bad')	joyful	absorbed	sure	stressed	excited	opposed	interested	unsure	thinking
Male 1	Open door	bad	4.2	4.8	3.4	3.0	3.5	3.1	4.1	4.7	5.1
		good	4.0	5.2	3.3	2.7	3.2	3.4	5.0	4.2	4.9
	Close door	bad	3.9	2.6	2.4	5.2	4.3	4.4	4.1	5.0	4.2
		good	3.1	3.5	2.8	5.0	3.5	3.9	3.9	5.3	5.0

Table 8-5: Male 1, Participant's reactions to his own choice of door (good or bad doors).

Table 8-5 shows the influence of the participant's own decision on his expressions, or maybe the other way around, the mental state that may have influenced the decision. The decision of which door to open happens in parallel to a significant change in the vocal expression of interest that significantly decreases for 'bad' doors. The reaction to the decision result is slightly different from the reaction to the gain. 'Good' doors encouraged thinking and reduced the excitement and laughter.

These results can be also explained by the time factor, the total gain in this interaction was below zero for most of the game, from the 17^{th} event until the end, which means that the deterioration in interest and excitement levels can also be related to the length of the game. The expression *joyful* increases after temporary losses. In this case the participant started laughing after several temporary losses, which can be seen in the video recordings of the interaction. These assumptions can be verified by examination of behaviour over time as presented in Figure 8-5. This figure shows the participant's expressions upon realising profit or loss (the sentence 'close door'). For better observability the recognisable expressions are roughly arranged into groups: positive - *joyful* and *sure* (orange), negative - *unsure* and *stressed* (grey), and neutral expressions that are related to thinking and cognition - *thinking*, *absorbed* and *interested* (green). *Opposed* or *disagree* is somewhat between the think and the negative groups. Again it can be seen that the positive beginning changes fast to alternating intervals of *thinking* and *uncertainty*. The confidence level (*sure*) that was high at the beginning went down later on. There are several intervals in which the participant seems to be *joyful*, some of them are intermingled with *stress*. He becomes more *absorbed* during the game although the interest level fluctuates. In the middle of the interaction and towards the end, the *joyful* and intervals of *thinking*, *stressed* and *unsure* become more dominant. Towards the end he started to gamble on the right door, he *thinks* more and regains *interest* for a short while. In parallel, although the participant said at the end of the game that he did not understand the advantageous strategy, after the 83^{rd} event he chose 14 'advantageous' doors (78%) vs. 4 'disadvantageous' doors (22%). This is one of the game's stages which was recognised in the original IGT.

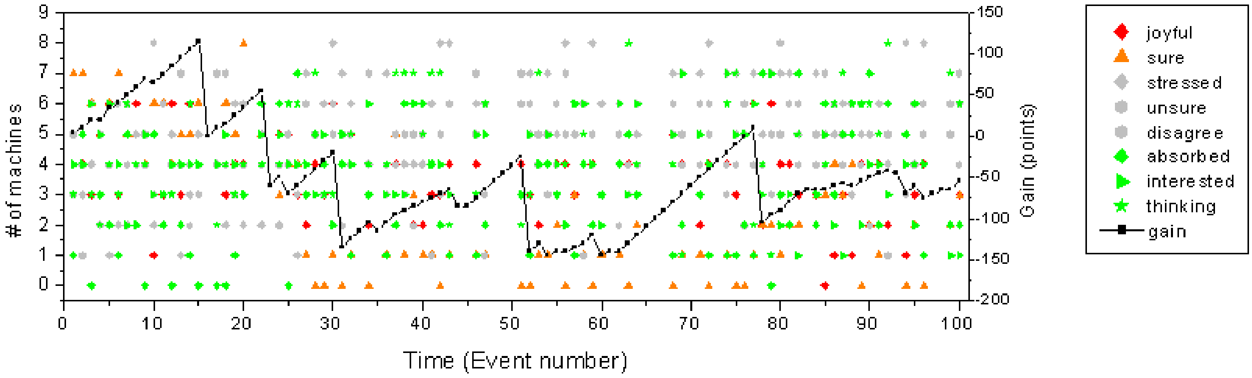

Figure 8-5: Expressions vs. Gain, upon realising gain (profit or loss). Green -*thinking*, *absorbed*, *interested*, Grey - *Stressed*, *opposed*, *unsure*, orange - *joyful*, *sure*.

In addition to these presentations of the expressions over time and in parallel to game events and decisions, other multi-modal information can be derived. In this case I demonstrate the relations between the game events and the behavioural cues which consist of the delay before the next decision and mouse movement rate.

Figure 8-6 shows the delay in time units and the amount of mouse movement per time unit vs. the total gain graph. The delay and mouse movements between event 20 and 21 are irrelevant, in this gap the participant was interviewed. The delay is on average longer before choosing 'bad doors', and when the total gain is below zero. The delay and number of mouse movements increases after temporary loss. These observations proved to be statistically significant.

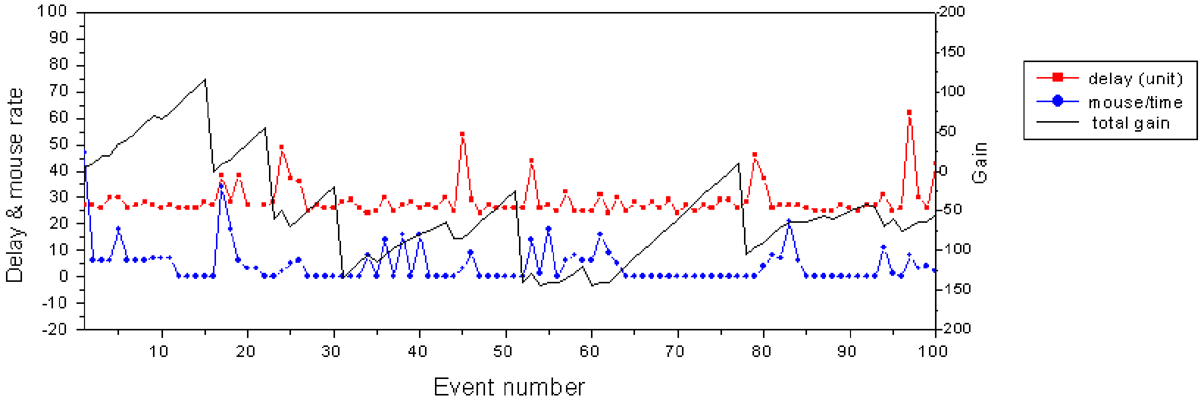

Figure 8-6: Mouse movements and delay vs. Gain. After most major losses the mouse movements and delays increase. The delay and mouse movements between event 20 and 21 were intentional, in this gap the participant was interviewed.

8.3 Summary and discussion

This chapter shows examples of gradual changes among utterances during an interaction, and the prosody parameters that create these changes. The proposed solution for an inference system should support both stand-alone expressions and utterances, and dynamic changes and transitions. In general, it should allow the system to follow the multi-layer structure of expressions in time during the course of an interaction.

This chapter presents statistical analysis of interactions of six participants from the Doors database. This analysis shows significant differences in the inferred expressions that correspond to game events and to participants' choices. The direction of the changes can be expected or at least reasonably explained in most cases.

Two conclusions can be drawn from these results:

It can be assumed that the inference is correct and can be used for interaction analysis.

It shows that the inference machine is general in the sense that it can infer both acted expressions and naturally evoked subtle expressions in two different languages. This implies that complex mental states and their acted and naturally evoked expressions are not unique to a specific language or culture. They can be recognisable across cultures and languages. Therefore machines can be trained and used on sources from different languages. However, the machine was trained on expression groups, rather than on nuances whose definition and perception can change between languages and cultures. The existence of these groups in most languages contributes to the reasonable results. It is not necessarily true for every language, background or situation.

In addition to statistical analysis of events that occurred many times throughout the interaction and the relations of expressions to specific physiological and behavioural cues, the interaction analysis includes step-by-step observations of the expressions and analysis of tendencies by averaging or smoothing of the inference results.

The inference machine recognised tendencies and subtle nuances of co-occurring expressions rather than one label of expression. It recognised changes also in the least obvious and least recognised expressions, such as slight confidence level changes.

Several expressions and changes are related to significant physiological reactions (GSR) and to the uttered text. These findings enhance the validity of the analysis. Such changes can occur few times during an interaction and therefore statistical analysis is not suitable for their interpretation.

Averaging or smoothing the inference results reveals relatively long-term tendencies. These tendencies can often be associated with other events and cues.

The different analysis methods contribute to better understanding of the participants' mental states, behaviour and text. From this information further information can be deduced for example regarding the participants' knowledge and even gender.

This chapter presents a few examples that demonstrate the advantages of using multimodal information, including the relation between behavioural cues such as mouse movements and delay in reaction to mental states, physiological cues and vocal correlates, and between vocal correlates, events and decisions. The conclusion is that inference can be improved by drawing on multiple sources.

References

[1] Witten I. H. and Frank E., *Data Mining: Practical machine learning tools with Java implementations*, Morgan Kaufmann, San Francisco, 2000.

[2] Bechara A., Damasio H., Tranel D., and Damasio A. R., "Deciding advantageously before knowing the advantageous strategy", *Science*, vol. 275, pp. 1293–5, 1997.

[3] Bechara A., Damasio A. R., Damasio H., and Anderson S. W., “Insensitivity to future consequences following damage to human prefrontal cortex”, *Cognition*, vol. 50, pp. 7–15, 1994.

[4] Bowman C. H., Evans C. E. Y., and Turnbul O. H., “Artificial time constraints on the iowa gambling task: The effects on behavioural performance and subjective experience”, *Brain and Cognition*, vol. 57, pp. 21–25, 2005.

[5] Maia T. V. and McClelland J. L., “A reexamination of the evidence for the somatic marker hypothesis: What participants really know in the iowa gambling task”, *PNAS*, vol. 101, no. 45, pp. 16075–16080, 2004.

Chapter 9

Summary and conclusions

This chapter presents a summary of the dissertation, its principal contributions and directions for future work.

9.1 Summary

With the increasing integration of computers and computer interfaces in our lives, arises the need of computers to be able to recognise and respond to human communication and behavioural cues of emotions and mental states. Furthermore, the expressions in speech are as meaningful as the verbal content and their inference and analysis may enhance the performance of speech and natural language technologies.

This dissertation addresses the problem of automated inference of complex mental states from expressions in speech. In particular, the expressions that relate to the affective and cognitive states of the mind that are not part of the basic emotions set. This is a challenging endeavour because of the uncertainty inherent in the inference of hidden mental states from behavioural cues, because the automated analysis of expressions in speech is an open signal processing problem, and because there is a lack of knowledge and agreement about the vocal cues composition of complex mental states.

This dissertation presents a framework for the design and implementation of automated inference (and synthesis) of mental states from vocal expressions. This framework is based on the dynamic nature of expressions during an interaction. The dissertation examines and defines new vocal features for the characterisation of expressions. It then describes an inference system that takes into consideration that different vocal features characterise different expressions and the existence of expression mixtures. The implementation of this system is by combination of pair-wise machines. These machines are based on two different classification algorithms. This approach to implementation contributes to the flexibility and expandability of the inference system. It allows it to integrate additional machines that incorporate additional expressions from other training databases and other recording setups, new speakers and different classification methods. The inference machine performs well in tests and can characterise and distinguish between expressions beyond the recognisable expression set, the set of expressions that it was trained to recognise.

The approach to inference in this dissertation is from a wide scope that includes the analysis of the relations between expressions. The dissertation demonstrates the application of the automatic inference to expression mapping and to multi-modal interaction analysis. For this analysis two complementing databases are used, MindReading and Doors. The Doors database was defined and recorded as part this research. These databases represent a large variety of acted and naturally evoked expressions and nuances of expressions, in a stand-alone manner and as part of an HCI interaction, in two languages, Hebrew and English. The ability of the inference machine to give reasonable results on such diverse data is evidence to its generality and robustness

The fact that complex mental states were automatically inferred in a language that is different from the language that was used for training suggests that the vocal correlates of complex mental states are not unique to one language. This result may advance the field of expression analysis, can facilitate the implementation of inference and synthesis systems and widen their scope. It may also enhance their economical validity.

The analysis of expressions in a context, as part of an interaction, enhances the understanding of the mental state of the speaker. The multi-modal analysis further enhances this understanding. The dissertation presents the analysis of human-computer interactions from the Doors database. This analysis demonstrates the dynamic natures of expressions and the relations between the inferred expressions, events, behavioural and physiological cues. The interaction analysis is a step towards the integration of the inference of expressions from speech with human-computer interfaces, with other speech technologies and with dialogue systems.

Expression mapping refers to the analysis and presentation of the relations between lexical concepts, the relations between behavioural expressions and the relations between concepts and expressions, or between meaning and its behavioural expression. The inference machine was used to examine the vocal correlates of the MindReading taxonomy, using the MindReading database. It inferred expression combinations that agree with the lexical definitions of the analysed concepts and with the meanings of the concept groups that are defined by the taxonomy.

Another mapping method was examined. This method is based on a new approach to expression mapping that uses only the recognisable expression set, a small sub-set of concepts and their related vocal characteristics, for mapping the relations between expressions and concepts. This method reveals additional connections between expressions and concepts. The two mapping techniques reveal various properties of concepts, their expressions and their interpretation by actors. They reveal the connection between meaning and behavioural expressions. The ability of an automated inference system to discern meaning is of extreme importance.

This research is innovative in its approach to the problem and to the implementation of the solution. It therefore presents an important step towards the integration of social and behavioural cues in a large variety of automated systems and user interfaces. It may contribute to speech technologies and to research in other fields.

The next sections summarise the principle contributions of this research and present directions for future work.

9.2 Contributions

The contributions of the dissertation are in its approach to analysis of expressions in speech, in the manner of implementation and in the resulting conclusions. The principal contributions are:

- The introduction of a framework for automatic expression recognition and analysis.

- The definition and recording of a database for multi-modal analysis of naturally evoked expressions.

- The definition of new vocal features, new temporal metrics and extraction algorithms for the characterisation of expressions.

- An implementation of an automatic inference system which is based on the approach that different sets of vocal features distinguish between different expressions.

- A computational model that is flexible and expandable to new speakers, expressions, languages, environments and features.

- Automatic inference of expressions from non-verbal speech that encompasses a wide variety of expressions of emotions and complex mental states, expression mixture and subtle nuances of expression. These expressions are beyond the set of basic emotions.

- Automatic inference and multi-modal analysis of expressions during a human-computer interaction. This expressions change dynamically in an asynchronous manner throughout the interaction.

- Two new methods of expression mapping using a small sub-set of mental state concepts (not basic emotions) and their vocal correlates.

- An analysis of expression characteristics and of the relations between vocal expression and meaning.

- An investigation of the generality of complex mental states across languages and cultures, in this case English and Hebrew.

- An algorithm and framework for inference of co-occurring classes

9.3 Future work

Future work can be carried in various directions. These directions relate both to enhancement of the system itself and to its applications. The main directions are:

Extension of the system

- Extend the scope of the system to fully automatic analysis of interactions, and to dialogue systems.
- Enable real-time processing. This phase is in process and requires mostly new implementation of an extraction algorithm for the pitch.
- Extend the scope of the system to new speakers in real-time. The suggested algorithm mimics human behaviour. It includes learning from known speakers with similar vocal properties at the beginning, and fine tuning to the new speaker through interactions.
- Improve the mapping capabilities of the system. It can be done by adding recognisable expressions, using finer expression definitions, and by introducing the definition of distances.
- Explore the suitability of the system to other languages and situations.

Applications

- Support automatic inference from more modalities and context cues to improve the recognition power of the system. In Sobol-Shikler *et al.* [1] we draw attention to the various issues inherent in building a multi-modal system for the recognition of a range of user mental states. Integrating different modalities in the computational model of mind-reading poses many research challenges with respect to building sensors and classifiers of the individual modalities, and developing a coherent model that integrates these modalities efficiently.
- Extend the scope of the findings to other speech technologies, especially to expression synthesis. An example is the Affect Editor, a tool for editing expressions of recorded and synthesised speech [2].
- Apply the inference system and the mapping to various applications. Specific application areas include assistive technologies, learning, security, e-commerce, the entertainment and the automobile industries. Other applications in mainstream (horizontal) computing domains include computer-mediated communication, ubiquitous computing and wearable devices.
- Apply the new features to other speech technologies.

Other applications can be to research in other disciplines such as linguistics, especially in the field of pragmatics and psychology.

References

[1] Sobol Shikler T., El-Kaliouby R., and Robinson P., “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.

[2] Sobol Shikler T. and Robinson P., “Affect editing in speech”, in *proceedings of the 1st International Conference on Affective Computing and intelligent Interaction (ACII), Beijing, China*, 2005.

Glossary

Notation	Description
Affective computing	A branch of artificial intelligence that deals with the design of devices which can process emotions, including emotion recognition, analysis and synthesis.
Asperger syndrome	A condition on the autistic spectrum. It is typically characterised by issues with social and communication skills. Also referred to as high-functioning autism
Basic emotions	Emotions that have distinct and universal expressions, such as happiness, sadness, anger, surprise, fear and disgust.
Behavioural cues	Parameters of human behaviour that contribute to the expression of mental states. Include speech, vocal cues, facial expressions, gestures, posture, decisions and more.
Blind clustering	Unsupervised learning. Dividing unknown and unlabelled data samples into clusters.
BVP	Blood volume at the periphery. Measurement of blood volume at the finger tip using non-invasive optical techniques.
C4.5	A decision tree based classification algorithm
CAM Battery	A battery of tasks, testing recognition of 20 complex emotions and mental states from faces and voices.
Classification	Supervised learning. A statistical procedure or system in which individual items are placed into classes or groups based on quantitative information (metrics) inherent in the items. It is based on a training set of previously labeled items, i.e. in the training stage (preparation) of the system each item has a class label.
Cognitive mental state	Mental states that refer to knowledge and to cognitive processes such as thinking, interest and concentration
Complex Mental States	Mental states that are not part of the basic emotion set

Notation	**Description**
Concept groups	Categories of mental states. It refers to 24 groups or categories of mental states, as defined by the MindReading taxonomy.
Concepts, Mental States concepts	The lexical definition and meaning of mental states. Generally, a concept is typically associated with a corresponding representation in language that denotes all of the objects in a given category or class of entities, and the relationships between them. Concepts are abstract in that they omit the differences between the entities, treating them as if they were identical.
Conceptualisation	Attempts to explain how knowledge is represented, or how objects are recognised, differentiated and understood and the relations between them.
Dialogue acts	Discourse structures, such as statement, acknowledgement, question and answer, turn taking and more.
Doors	A multi-modal database that was defined and recorded as part of this research. It includes recordings of 15 volunteers during a computer game with intervals of interviews. It consists of video recordings of facial expressions, audio recordings with and without controlled text in Hebrew. The controlled text includes 2 sentences repeated 100 times by each participant. In addition it includes recordings of GSR, BVP, ECG, mouse movements, reaction delays and game events. It provides naturally evoked expressions and is unlabelled.
ECG	An electrocardiogram -a graphic representation of the electrical activity measurement of the heart over time.
Energy	The intensity of the speech signal as it changes over time. It is expressed as X2, when X is the amplitude of the speech signal sample (which is related to the pressure of the speech signal).
Expression mapping	The presentation and conceptualisation of mental states, the relation between them and the relations between them and their expressions.
Expression mixture	Expressions that co-occur simultaneously

Notation	**Description**
Expressions	The behavioural characteristics, the outer representation or display of emotions, mental states, moods, attitudes, physiological states, cultural display rules, dialogue acts and the like. This work refers mostly to vocal expressions, as revealed by vocal features such as intonation.
Fundamental frequency, f_0	The rate of the vocal fold vibrations. It 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. Its modulation by the speaker creates the intonation.
Generalisation	A system's performance when trained on one corpus of data and tested on a different one.
GSR	Galvanic skin response, measurements of the skin conductivity which is changed by changes in skin humidity, as a result of stress, arousal and the like. It is used for example in lie detectors.
Harmonic intervals	Pure tones with relatively small ration between them and the fundamental frequency, such as 3:2 and 4:3.
Harmonic properties	The relations between the fundamental frequency and other frequencies and the effects these combinations create in the sound.
HCI	Human-computer interaction
Human-mediated communication	Human-computer-human communication, human-human communication through or with the help of computers.
IGT	Iowa gambling test. It is a card game which is extensively used for psychology research, it intends to elicit emotions and monitor decision making. In this game, the participants have to choose one card of four stacks of cards repeatedly 100 times. Each stack has different gain expectancy unknown to the participant. The goal of the participant is to maximise the gain.
Inference	The act or process of deriving a conclusion based solely on what one already knows. Here it refers to the automatic inference of mental states from their expressions in speech.

Notation	**Description**
Interaction	Sustained interaction. The interaction can be with other people, with a computer or machine, or even with oneself.
Labelling	Associating names, labels or descriptions with the recorded expressions
Mental states	States of mind that people experience, exhibit, express and attribute to each other. It is a general term that refers to emotions, cognitive states, intentions, beliefs, desires, moods, focus of attention and the like. The top-most level of the computational model.
MindReading Database (DVD)	An interactive guide to emotions that includes a comprehensive voice and video collection of mental state enactments.
MindReading taxonomy	Hierarchical conceptualisation method, which aims to represent mental states and emotions according to their meaning into 24 concept or meaning groups.
Multi-modal analysis	Analysis that involves comparison to various different cues. Multi-modal analysis of mental states and their expressions consists of comparison to additional behavioural cues, physiological cues, verbal cues (text) and context.
Naturally evoked expressions	Expressions of emotions and mental states that are not acted, but rather evoked due to the speaker's mental state.
Non-verbal expressions	Expressions that are exhibited by through behavioural cues other than text and language
Pair-wise machines	Classification method of multiple classes, that uses comparisons between every two classes independently.
Paralinguistic	The non-verbal elements of communication used to modify meaning and convey emotion , through pitch, volume, intonation and the like.
Parsing	Dividing a sentence to units such as voiced, unvoiced and silence
Physiological cues	Measurable bodily changes that indicate mental states, such as GSR, BVP, ECG.

Notation	**Description**
Pitch	The perceptual correlate of the fundamental frequency, or how people hear and perceive the sound created by the vibrations of the vocal fold. The term is often used instead of the term fundamental frequency.
Prosody, prosodic features	Features that influence the expressive uses of speech, derived from the production mechanism of speech and voice. They are related to changes in the breathing muscles and the vocal fold, and to shaping factors of the vocal tract, i.e. movements of the velum (soft palate), tongue, teeth, jaw and lips. Prosodic features include: fundamental frequency, intensity, spectral content, their durations and magnitude.
Psycho-acoustic tests	Psychological tests in which people are asked to assess certain features of sounds, music, speech and modified speech signals.
Recognisable expressions	The underlying taxonomy. A set of (nine) expressions that the inference machine was trained to recognise. Each of these expressions consists of a group of mental state nuances and their related vocal expressions. Their combinations describe a large variety of complex mental state expressions.
Sampling rate	The recording rate of the speech signal. The unit is Hertz - the number of samples that are recorded in a second. It defines the time intervals between samples.
Secondary metrics	Metrics that are derived from the vocal features. The second stage of processing. Include statistical metrics and temporal metrics. The statistical metrics include characteristics of the vocal features over the whole utterance, such as mean, standard deviation, range, median, minimum and maximum values. The temporal metrics include time related characteristics such as the durations of different speech parts within the utterance, including voiced, unvoiced and silence.
Segmentation	Detecting speech and vocal utterances in the raw recorded sound track and distinguishing them from other sounds.
Silence	Parts of the sentence in which there is no energy (no speech signal).

Notation	Description
Spectral content	The spread of energy across different frequency bands. The signal consists of a mixture of frequencies. Spectral content describes how much of the signal is located around different frequencies or tones.
Speech acts	Actions that are performed by the speech itself and directly or indirectly related to the content or text, such as labeling, repeating, answering, requesting (action or answer) , greeting, protesting, practicing, apologizing, describing and the like.
Speech segments	Sentences, utterances
Speech signal	The digital representation of the recorded speech or voice. It comprises a series of samples of the amplitude of the signal that are recorded at identical time intervals.
Statistical metrics	Statistical parameters, such as mean and standard deviation that describe the behaviour of the vocal feature over the whole utterance
SVM	Support vector machines. Classification algorithms that define hyperplane on which the distance between the samples of two classes is maximised.
Synthesis	Artificial generation of speech.
Taxonomy	Classification into categories and the principles underlying the classification. It also represents relations scheme such as hierarchical relations.
Temporal metrics	Metrics that are derived from time related properties of the vocal features.
Unvoiced	A sound in which the vocal fold does not vibrate. The fundamental frequency is zero, but there is sound and therefore energy.
Utterance	Something said or emitted as a vocal sound. In this work it refers mostly to a sentence.
Vocal cues, vocal features	The basic characteristics of expressions in the voice or speech. They are extracted directly from the speech signal by means of signal processing, i.e. mathematical processing of the digital representation of the recorded speech signal.
Voiced	A sound in which the vocal fold vibrates, the fundamental frequency is not zero.

Notation	Description
Voting algorithm	A method of decision making, selecting one or multiple winners from among candidates using a certain criteria of preference.