A prototype for a Conversational Companion for reminiscing about images

Yorick Wilks, Roberta Catizone, Simon Worgan, Alexiei Dingli, Roger Moore, Weiwei Cheng

Title: A prototype for a Conversational Companion for reminiscing about images

Authors: Yorick Wilks, Roberta Catizone, Simon Worgan, Alexiei Dingli, Roger Moore, Weiwei Cheng

PII: S0885-2308(10)00033-1
DOI: doi:10.1016/j.csl.2010.04.002
Reference: YCSLA 449

To appear in:

Received date: 8-9-2009
Revised date: 31-3-2010
Accepted date: 1-4-2010

Please cite this article as: Wilks, Y., Worgan, S., A prototype for a Conversational Companion for reminiscing about images, *Computer Speech & Language* (2008), doi:10.1016/j.csl.2010.04.002

A prototype for a Conversational Companion for reminiscing about images

Yorick Wilks, Roberta Catizone, Simon Worgan, Alexiei Dingli, Roger Moore and Weiwei Cheng

{r.catizone; s.worgan; y.wilks; r.k.moore; w.cheng}@dcs.shef.ac.uk,
alexiei.dingli@um.edu.mt

Abstract

This paper describes an initial prototype of the Companions project (www.companions-project.org): the Senior Companion (SC), designed to be a platform to display novel approaches to:

1) the use of Information Extraction (IE) techniques to extract the content of incoming dialogue utterances after an ASR phase;
2) the conversion of the input to RDF form to allow the generation of new facts from existing ones, under the control of a Dialogue Manager (DM), that also has access to stored knowledge and knowledge accessed in real time from the web, all in RDF form;
3) a DM expressed as a stack and network virtual machine that models mixed initiative in dialogue control;
4) a tuned dialogue act detector based on corpus evidence;

The prototype platform was evaluated, and we describe this; it is also designed to support more extensive forms of emotion detection carried by both speech and lexical content, as well as extended forms of machine learning. We describe preliminary studies and results for these, in particular a novel approach to enabling reinforcement learning for open dialogue systems through the detection of emotion in the speech signal and its deployment as a form of a learned DM, at a higher level than the DM virtual machine and able to direct the SC's responses to a more emotionally appropriate part of its repertoire.

Keywords: Dialogue, Human-Computer Interaction, Dialogue Management, ASR and Emotion

1 Introduction

COMPANIONS is an EU project that aims to change the way we think about the relationships of people to computers and the Internet by developing a virtual conversational 'Companion'. This will be an agent or 'presence' that stays with the user for long periods of time, developing a relationship and 'knowing' its owners preferences and wishes. The Companion communicates with the user primarily through speech, but also using other technologies such as touch screens and sensors.

This paper describes the functionality and system modules of the Senior Companion (SC), one of two initial prototypes built in the first two years of the project. The SC

provides a multimodal interface for eliciting, retrieving and inferring over personal information from elderly users by means of conversation about their photographs. The Companion, through conversation, elicits their life memories and reminiscences, often prompted by discussion of their photographs; the aim is that the Companion should come to know a great deal about its user, their tastes, likes, dislikes, emotional reactions etc, through long periods of conversation. It is assumed that most life information will soon be stored on the internet (as in the Memories for Life project: http://www.memoriesforlife.org/) and we are linking the SC directly to photo inventories in Facebook (see below). The overall aim of the SC is to produce a coherent life narrative for its user from these materials, although its short-term goals, reported here are to assist, amuse, entertain and gain the trust of the user. The Senior Companion uses a hybrid approach to dialogue management as well as intelligent adaptation of the user's emotional state which plays an important part in gaining the user's trust.

The technical content of the project is to use a number of types of machine learning (ML) to achieve these ends in original ways, initially using a methodology developed in earlier research: first, by means of an Information Extraction (IE) approach to deriving content from user input utterances (Catizone et al., 2002); secondly, using a training method for attaching Dialogue Acts to these utterances (Webb, et al., 2008) and lastly, using a specific type of dialogue manager (DM) that uses Dialogue Action Forms (DAF) to determine the context of any utterance, and a stack of these DAFs as the virtual machine that models the ongoing dialogue by means of a shared user and Companion initiative and generates appropriate responses (Catizone et al., 2003). The SC is not a robot and could be embodied in a screen, a handbag or a mobile phone while retaining the same "personality": it is more a very high level internet agent, dedicated to a single user over the long term. In the first part of this paper, we shall:

- describe the current SC prototype's functionality;
- set out its architecture and modules, focusing on the Natural Language Understanding module and the Dialogue Manager.
- set out our short term plans to enhance Dialogue Management performance with direct Internet access and initial ML experiments.

In the second part of the paper we shall describe experimental work linking the DM to emotional considerations.

2 The Senior Companion System

The Senior Companion (SC) prototype (Wilks 2007, 2008; Wilks et al., 2008) was designed to make a rapid advance in the first two years of a project so as to be the basis for a second round of prototypes embodying more advanced ML. This strategy was deliberately chosen to avoid a well-known problem with experimental AI systems: that a whole project is spent in design so that a prototype never emerges until the very end, which is then never fully evaluated and, most importantly, nothing is ever built upon the experience obtained in its construction. The central function of the

SC is engaging the user in discussion about their photographs: where and when they were taken, details about the people in them and their relationship to the user and each other. The SC extracts and stores facts obtained from the user's input and is able to pick up discussion with the user where the system left off in later user sessions. In addition to allowing reminiscing, the SC also permits the user to do basic photo management including selecting particular images or groups of images by pointing and organising the photos by means of a dialogue.

Once a photo is loaded, it is processed with face recognition software to identify any faces in it. The recognition software, OpenCV[1], provides positional information by identifying the face coordinates and this information is exploited in the Dialogue Manager by making explicit reference to the position of people in the photograph (the person on the left, right, center, etc.) as well as recognizing when there are groups of people. The system discusses properties of the photo as well as properties and relationships of the people in the photos.

The SC also contains a news reading feature which adds an interesting accompaniment to the photo domain and demonstrates the ability of the system to handle more than one kind of application at a time, and news has, of course, an unconstrained vocabulary . It is taken via RSS feeds from the BBC news website and includes news from three popular categories : politics, sports and business, and the user can choose between them, stop and start the feed by speaking etc. The system can also tell jokes on request, from a potentially endless internet source.

The following is the middle part of a sample dialogue generated by the system when discussing a group photo from a vacation in Zanzibar, and after the system had already stored from an earlier conversation that the user had a daughter Zoe. One of the people in the photo has been tagged as 'doctor' using Facebook :

SC: What is your relationship to the doctor?
User: There is no relationship.
SC: Please tell me how you know the doctor
U: Uh, he's a friend.
SC: How old was the doctor when this photo was taken?
U: I'm not sure but I think about 30.
SC: How do you know Octavia?
U: She's my daughter.
SC: Aha! I see. Octavia's sister is Zoe.
SC: How old was your daughter Octavia when this photo was taken?
U: Octavia was 19 years old.
SC: What was the occasion of this photo?
U: This was another family holiday.
SC: Where was this photo taken?
U: This photo was taken in Tanzania.
SC: have you ever been to Zanzibar?
U: Yes, I have.

[1] http://opencv.willowgarage.com/wiki/

SC: Great place, I hope you saw the beaches. When was this photo taken?
R: last year.

It is a key feature of the SC that it can make simple inferences from family relationships it already knows (e.g. that daughters of the same parent are siblings) and that it can access real-time information about places to show that it has some knowledge of what is being talked about, in this case the beaches on Zanzibar, and how this is done is discussed below. This real time access to unconstrained place information on the internet is an attempt to break out of classic AI systems that only know the budget of facts they have been primed with.
This basic system provides the components for future development of the SC, as well as its main use as a device to generate more conversation data for machine learning research in the future. Key features of the SC are listed below followed by a description of the system architecture and modules:

- A visually appealing multi-modal interface (Figure 1) with a character avatar to mediate the system's functionality to the user.
- Interacting with the user with multiple modalities – speech and touch.
- Includes face detection software for identifying the position of faces in the photos.
- Accepts pre-annotated (XML) photo inventories as a means for creating richer dialogues more quickly.
- Engages in conversation with the user about topics within the photo domain: when and where the photo was taken, discussion of the people in the photo including their relationships to the user.
- Reads news from three categories: politics, business and sports.
- Tells jokes taken from an internet-based joke website.
- Retains all user input for reference in repeat user sessions, in addition to the knowledge base that has been updated by the Dialogue Manager on the basis of what was said.
- Contains a fully integrated Knowledge Base for maintaining user information which contains:
    - Ontological information which is exploited by the Dialogue Manager and provides domain-specific relations between fundamental concepts.
    - A mechanism for storing information in a triple store (Subject-Predicate-Object)-the RDF Semantic Web format---- for handling unexpected user input that falls outside of the photo domain, e.g. arbitrary locations in which photos might have been taken.
    - A reasoning module for reasoning over the Knowledge Base and world knowledge obtained in RDF format from the internet; the SC is thus a primitive Semantic Web device (see Wilks, 2008)
- Contains basic photo management capability allowing the user in conversation to select photos as well as display a set of photos with a particular feature.

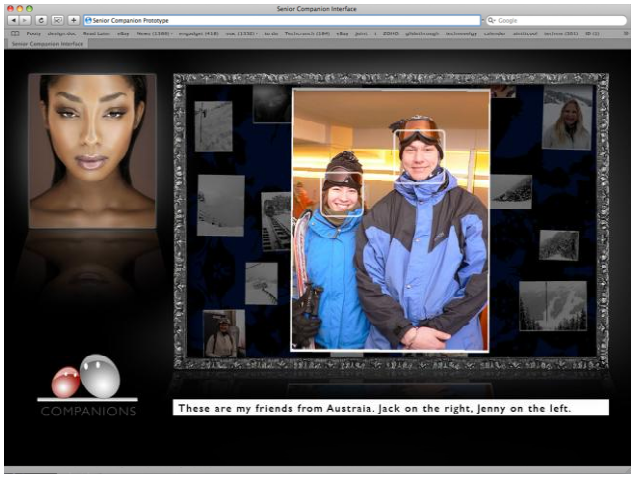


Figure 1: The Senior Companion Interface

2.1 *System Architecture*

In this section we will review the components of the SC architecture. As can be seen from Figure 2, the architecture contains three abstract level components – Connectors, Input Handlers and Application Services –together with the Dialogue Manager and the Natural Language Understander (NLU).

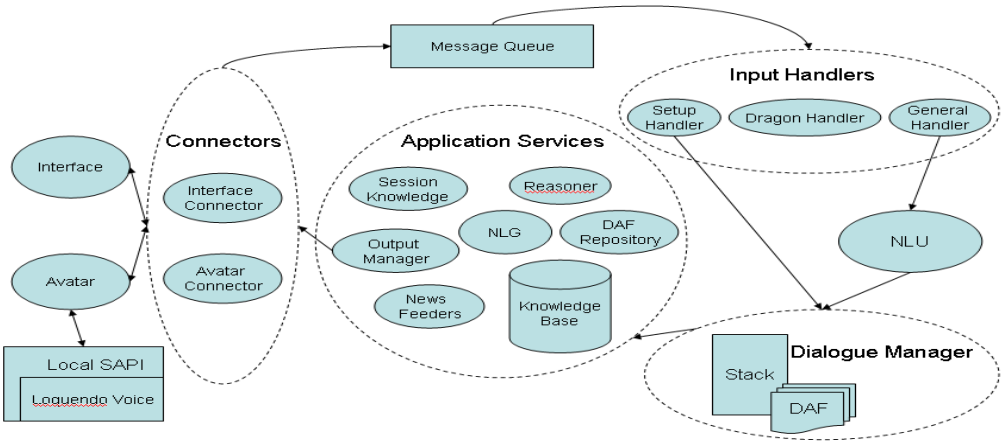


Figure 2: Senior Companion system architecture

Connectors form a communication bridge between the core system and external applications. The external application refers to any modules or systems which provide a specific set of functionalities that might be changed in the future. There is one connector for each external application. It hides the underlying complex communication protocol details and provides a general interface for the main system to use. This abstraction decouples the connection of external and internal modules, makes changing and adding new external modules easier. At this moment, there are two connectors in the system – Napier Interface Connector and CrazyTalk Avatar Connector. Both of them are using network sockets to send/receive messages.

Input Handlers are a set of modules for processing messages according to message

types. Each handler deals with a category of messages where categories are coarse-grained and could include one or more message types. The handlers separate the code handling inputs into different places and make the code easier to locate and change. Three handlers have been implemented in Senior Companions system – Setup Handler, Dragon Events Handler and General Handler. Setup Handler is responsible for loading the photo annotations if any, performing face detection if no annotation file is associated and checking the Knowledge Base if the photo being processed has been discussed in earlier sessions. Dragon Event Handler deals with dragon speech recognition commands sent from the interface while the General Handler processes user utterances and photo change events of the interface.

Application Services are a group of internal modules which provide interfaces for the Dialogue Action Forms (DAF) to use. It has an easy-to-use high level interface for general DAF designers to code associated tests and actions as well as a low level interface for advanced DAFs. It provides the communication link between DAFs and the internal system and enables DAFs to access system functionalities. Following is a brief summary of modules grouped into Application Services.

News Feeders are a set of RSS Feeders for fetching news from the internet. Three different news feeders have been implemented for fetching news from BBC website Sports, Politics and Business channels. There is also a Jokes Feeder to fetch Jokes from internet in a similar way. During the conversation, the user can request news about particular topics and the SC simply reads the news downloaded through the feeds.

DAF Repository is a list of DAFs loaded from files generated by the DAF Editor. A fresh copy of a DAF can be obtained by passing the DAF name to this module.

NLG is responsible for randomly selecting a system utterance from a template. An optional variable can be passed when calling methods on this module. The variable will be used to replace special symbols in the text template if applicable. For example, a template utterance "How do you know $?" will be returned as "How do you know John?" if passing variable "John" when calling generation method of this module.

Session Knowledge is the place where global information for a particular running session is stored. For example, the name of the user who is running the session, the list of photos being discussed in this session, the list of user utterance and etc.

Knowledge Base is the data store of persistent knowledge. It is implemented as an RDF triplestore using a Jena implementation. The triplestore API is a layer built upon a traditional relational database. The application can save/retrieve information as RDF triples rather than table records. The structure of knowledge represented in RDF triples is discussed later.

Reasoner is used to perform inference on existing knowledge in the Knowledge Base (example in next section).

Output Manager deals with sending messages to external applications. It has been implemented in a publisher/subscriber fashion. There are three different channels in the system – the text channel, the interface command channel and the avatar

command channel. Those channels could be subscribed by any connectors and handled respectively.

The original work behind the SC was based on a closed world where the user converses with the system. The SC initiates further conversation and in so doing elicits the discovery of tacit knowledge from the user. Whilst conducting the initial tests, the limitations of this approach immediately became evident. As soon as the user interacted with the SC, the conversation quickly went through unexpected paths which required more knowledge than was stored within the knowledge-base. For example, when an elderly person is speaking with the SC about an old photograph taken during WWII, the person would easily recall events of the period. Our Hybrid-World approach tackles this issue. Initially, it makes use of a closed-world where all the information is stored in the Knowledge Base, but the system's second version was able to access open internet knowledge, initially in unstructured (text) format, which we convert to, or locate in, RDF format. Before discussing this we shall discuss the SC's basic process for extracting content from dialogue input.

2.2 *Dialogue input understanding and inference*

Every utterance is passed through the Natural Language Understanding (NLU) module for processing. This module uses a set of well-established natural language processing tools such as the GATE (Cunningham, et al., 1997) system. The basic processes carried out by GATE are: tokenizing, sentence splitting, POS tagging, parsing and Named Entity Recognition. These components have been further enhanced for the SC system by adding new and improved gazetteers. These include new locations and family relationships. The Named Entity (NE) recognizer is a key part of the NLU module and recognizes the significant entities required to process dialogue in the photo domain: PERSON NAMES, LOCATION NAMES, FAMILY RELATIONS and DATES. Apart from the gazetteers mentioned earlier and the hundreds of extraction rules already present in GATE, about 20 new extraction rules using the JAPE rule language were also developed for the SC module. These included rules which identify complex dates, family relationships, negations and other information related to the SC domain. The following is an example of a simple rule used to identify relationship in utterances such as "Mary is my sister":

```
Macro: RELATIONSHIP_IDENTIFIER
(
({Token.category=="PRP$"}|{Token.category=="PRP"}|{Lookup.majorType=="person_first"}):person2
({Token.string=="is"})
({Token.string=="my"}):person1
({Lookup.minorType=="Relationship"}):relationship
)
```

Using this rule with the example mentioned earlier, the rule interprets person1 as referring to the speaker so, if the name of the user speaking is John (which was known from previous conversations), it is utilized. Person 2 is then the name of the person

mentioned, i.e. Mary. This name is recognised by using the gazetteers we have in the system (which contain about 40,000 first names). The relationship is once again identified using the almost 800 unique relationships added to the gazetteer. With this information, the NLU module identifies Information Extraction patterns in the dialogue that represent significant content with respect to a user's life and photos.
The information obtained (such as Mary=sister-of John) is passed to the Dialogue Manager (DM) and then stored in the knowledge base (KB). The DM filters what to include and exclude from the KB. Given, in this example, that Mary is the sister of John, the NLU knows that sister is a relationship between two people and is a key relationship. However, the NLU also discovers syntactical information such as the fact the both Mary and John are nouns. Even though this information is important, it is too low level to be of any use by the SC with respect to the user, i.e. the user is not interested in the part-of-speech of a word. Thus, this information is discarded by the DM and not stored in the KB. The NLU module also identifies a Dialogue Act Tag for each user utterance based on the DAMSL set of DA tags and prior work done jointly with the University of Albany (Webb et al., 2008).

The KB is a long-term store of information which makes it possible for the SC to retrieve information stored between different sessions. The information can be accessed anytime it is needed by simply invoking the relevant calls. The structure of the data in the database is an RDF triple, and the KB is more commonly referred to as a triple store. In mathematical terms, a triple store is nothing more than a large database of interconnected graphs. Each triple is made up of a subject, a predicate and an object. So if we had to take the previous example, Mary sister-of John; Mary would be the subject, sister-of would be the predicate and John would be the object. If we had to imagine this graphically, Mary and John would be two distinct points in a 3D space and the sister-of relationship would be the line (or relationship) that joins these two points in space. There are various advantages to using this structure; first, the relationship between different objects is explicitly defined using the predicates in the triples. The second advantage is that it is very easy to perform inferences on such data. So if in our KB, we add a new triple which states that Tom is the son of Mary, we can easily infer (by using the previous facts) that John is the uncle of Tom. The inference engine is an important part of the system because it allows us to discover new facts beyond what is elicited from the conversation with the user.

Uncle Inference Rule:
(?a sisterOf ?b),
(?x sonOf ?a),
(?b gender male) -> (?b uncleOf ?x)

Triples:
(Mary sisterOf John)
(Tom sonOf Mary)

Triples produced automatically by ANNIE (the semantic tagger):
(John gender male)

Inference:
(Mary sisterOf John)
(Tom sonOf Mary)

(John gender male)
->
(John uncleOf Tom)

This kind of inference is already used by the SC and we have about 50 inference rules aimed at producing new data on the relationships domain. This combination of triple store, inference engine and inference rules makes a system which is weak but powerful enough to mimic human reasoning in this domain and thus give the SC minimal appearance of intelligence. For our prototype we are using the JENA Semantic Web Framework for the inference engine together with a MySQL database as the knowledgebase. However, this knowledge of family relationships is not enough to cover all the possible topics which can crop up during a conversation . In such circumstances, the DM switches to an open-world model and instructs the NLU to seek further information online.

2.3 *The Hybrid-world approach*

When the DM requests further information on a particular topic, the NLU first checks with the KB whether the topic is about something known. At this stage, we have to keep in mind that any topic requested by the DM should already be in the KB since it was preprocessed by the NLU when it was mentioned in the utterance. So, if the user informs the system that the photograph was taken in Paris, (in response to a system question asking where the photo was taken), the utterance is first processed by the NLU which discovers that "Paris" is a location using its semantic tagger ANNIE (A Nearly New Information Extraction engine). The semantic tagger makes use of gazetteers and IE rules in order to accomplish this task. It also goes through the KB and retrieves any triples related to "Paris". Inference is then performed on this data and the new information generated by this process is stored back into the KB.

Once the type of the information is identified, the NLU can use various predefined strategies: In the case of locations, one of these strategies would be to seek for information in Wiki-Travel or Virtual Tourists. The system already knows how to query these sites and interpret their output by using predefined wrappers. A wrapper is essentially a file, which describes where a particular piece of information is located. This is then used to extract that information from the webpage. So a query is sent online to these sites and the information retrieved is stored in the triple-store and this information is then used by the DM to generate a reply. In the previous example, the system managed to extract the best sightseeing spots in Paris. The NLU would then store in the KB triples such as [Paris, sight-seeing, Eiffel Tower] and the DM with the help of the NLG would ask the user "I've heard that the X is a very famous spot. Did you see it when you were there?" Obviously in this case, X will be replaced by the "Eiffel Tower".

On the other hand, if the topic requested by the DM is unknown or the semantic tagger is not capable of understanding the semantic category, the system uses a normal search engine. A query is sent to the search engines and the top pages are retrieved. These pages are then processed using ANNIE and the different attributes are analyzed. The standard attributes returned by ANNIE include information about Dialogue Acts, Polarity (i.e. whether a sentence has positive, negative or neutral

connotations), Named Entities, Semantic Categories (such as dates and currency), etc. The system then filters the information collected by using generic patterns and generates a reply from the resultant information. So if the user is talking about cats, the system searches for cats online. It processes the pages and its current strategy is to identify all the statements by using Dialogue Acts. So in our example, the system would retrieve the following statements:

Cats may be the most popular pet in the world
Cats recover quickly from falls
Some people don't like Persian Cats

These statements are then checked for polarity and only the most polarity-distinct statements are kept (i.e. if the statements are prevailingly negative then the system will give a negative answer, and so on). The polarity checking is performed by using a list of words with negative or positive connotations and counting which words prevail in the sentence. A sentence with a prevailing number of positive words is considered a positive sentence. The opposite occurs for negative words. In this example, the first two statements are prevailingly positive because of words such as "popular" and "recover" so the answer returned will be a positive one. The NLU would then select one of these two statements at random, send it to the DM and using the NLG, it would reply "You know that I've heard that X" where X is replaced with "cats may be the most popular pet in the world". ANNIE's polarity methods have been shown to be an adequate implementation of the general word-based polarity methods pioneered by Wiebe and her colleagues (see e.g. Akkaya et al., 2009)

In synthesis, this hybrid world approach allows us to focus on the closed world of images that exists between the user and the system but, when necessary, the system is allowed to venture cautiously in the open world, thus enriching the user experience. This is, as we noted, an important step towards breaking down the traditional closed-world assumptions of practical AI systems. Initial experimental results show that on average the system adopts the open world approach 20% of the time. The open world approach adds facts to the database only when the topic under discussion is known to the system. So, in the previous example where the system asks whether the user visited the Eiffel Tower, a positive or negative reply will be stored in the database and used during later conversations. However, when the topic is unknown such as in the case of the cats, the response of the SC is quite generic thus the conversation is not stored. During interactions with the system, it was noticed that the use of an open world approach (even when the subject was unknown by the system) produced a rather more realistic conversation than a system without the open world model. Users reported that they were amazed by the system possessed so much knowledge about the topic being discussed. We believe that the hybrid world model is potentially useful as a way of improving the interaction between the user and such systems.

2.4 *Emotion in the Senior Companion*

Traditionally, spoken dialog systems have been able to learn and reason over uncertainty through a partially observable Markov decision process (POMDP) that uses reinforcement learning (RL) to modify its strategy according to some reward function (e.g. Young, 2002). In restricted domains with a clearly defined goal (e.g. placing a pizza order in Young's case) RL proceeds by offering a reward when the

task is successfully completed. In the open-ended scenarios like that proposed by the COMPANIONS project how do we capture this reward function? We propose that the Companion should be tasked with maintaining the user's positive emotional state, ensuring continued user satisfaction. Given the goal of maintaining this state emotion becomes a central motivating factor requiring an internal representation of both itself and its user.

We considered two proposed representations of emotion, the discrete theory of basic emotion (Ekman, 1999) and the continuous theory, which maps a range of emotions onto a two dimensional space (Wundt, 1913, Cowie, Douglas-Cowie, Savvidou, & McMahon, 2000). We have opted for the continuous theory, which, instead of arbitrarily defining a number of discrete emotional categories, is open to a data driven approach. By creating a self-organizing map (SOM) (Kohonen, 1982), capturing the variability of emotional speech, we can train the system on a wide range of emotional utterances forming distinct attractors within the defined emotional space. Having trained the system these attractors will then be tied to a subset of dialogue acts, reflecting the agent's current emotional state. RL can then proceed by allowing the Companion to learn how its own movement in this space affects the user. However, this is only possible if the Companion can place a representation of the state of the user into this space, accordingly we need to ascertain the user's current emotional disposition.

2.5 *Emotion and speech in the Senior Companion*

The human speech signal contains a wide range of paralinguistic information and the interpretation and exploitation of this information presents a fascinating challenge. This section seeks to demonstrate that emotion is central to the construction of an open-ended dialogue system. Emotion forms a useful, practical, metric that enables an agent to both maintain a user's positive emotional state and allow it to judge and refine its current dialogue strategy.

One way to determine the emotional state of the user is through an interpretation of the speech signal. Accordingly, this section details a method of detecting emotion within speech and then exploiting this information to drive an open-ended dialogue system. After presenting the outline of an advanced open-ended dialogue system we test an initial model to judge the validity of this approach. This simplified model perceives a number of real speech utterances with varying emotional content (ranging from stressed to happy) and learns to manipulate the emotional state of an artificial user through reinforcement learning. The agent acquires the users emotional response to its replies and is able, through a balancing of exploration and exploitation, to maintain the users positive disposition. This initial work is encouraging and clearly shows that emotion can motivate open-ended dialogue. However, a number of substantial challenges remain, including the perception of natural, as opposed to acted, speech and the development of an unsupervised learning approach.

This section details how the detection of a users stress levels can be exploited to form a persistent reinforcement-learning goal for an open-ended dialogue system. These stress levels will be determined from the paralinguistic features of the users utterances and a simplified model will be constructed to demonstrate the feasibility of this

approach. Building on this implemented proof of concept model we will outline how this work can be expanded to provide a rich emotional representation and motivation for the COMPANIONS project in future work. This is necessary as the companions project seeks to construct an agent capable of open-ended conversation with a specific user.

As we noted in the last section, spoken dialog systems have been able to learn and reason over uncertainty through a partially observable Markov decision process (POMDP), as shown in Figure 3, using RL to modify its strategy according to some reward function (Young, 2002). POMDP's attempt to address "the problem of choosing optimal actions in partially observable stochastic domains" (Kaelbling et al.,1998, p. 100) but in practice they frequently prove to be computationally intractable. Accordingly, many researchers settle for a Markov decision process augmented by "a compression of the current belief state"(Roy et al.,2000, p. 94), to ensure that actions are taken in an uncertain environment in reasonable time. This compressed belief state, Bs, drives the reinforcement learning process and its success is entirely dependent upon the derived reward function. In this section the system beliefs, Bs, have been modified to represent a belief about the user's emotional state. These are formed through an acquired observation, O, of the emotional content of the utterance.

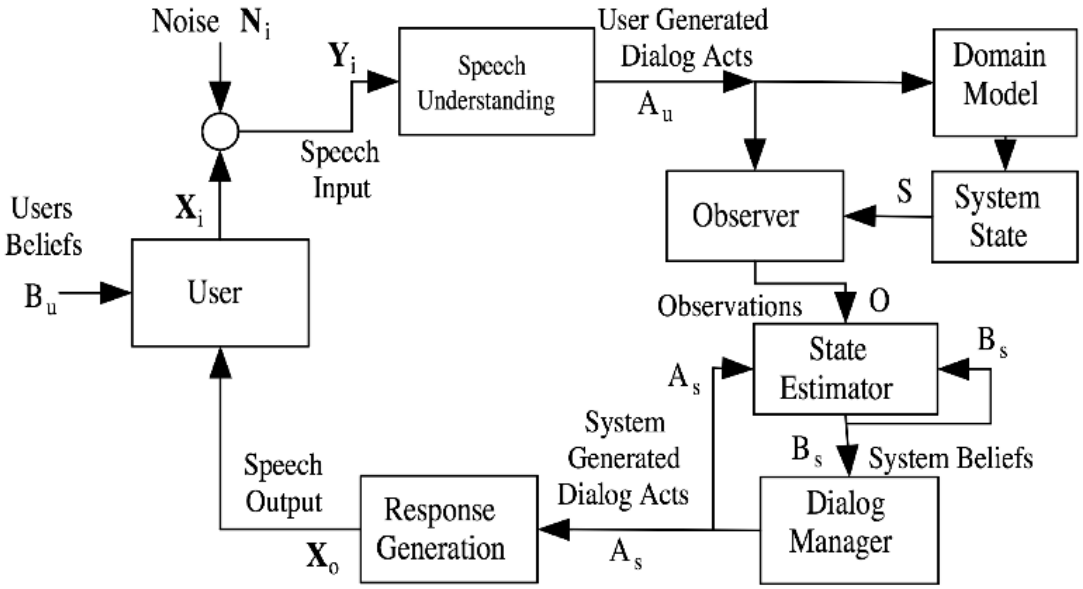


Figure 3. The overall POMDP framework

The following discussion is composed of two parts: first, is a proposed dialogue model modifying the ASR module that should enable an agent to operate intelligently within a continuous emotional space and an implemented simplification demonstrating the feasibility and possibility of this proposal. The model will be constructed to test the underlying assumptions of the more complex proposal, demonstrating an initial first step that acknowledges the technical challenges (Greasley et al., 2000,;Cowie et al., 2001; Hopkins et al., 2005; Barreto et al., 2007) remaining within this field. We believe that this implementation demonstrates how various, future, technical improvements can be exploited by a complete dialogue

model. In the implementation of the SC in section 1 above, the ASR was all performed by a tuned Nuance system to a single speaker, which is a plausible constraint since Companions are for individual owners. In this section we describe models where the ASR interacts actively with the DM to give more plausible emotional response. This model has been tested separately, see below, but not yet implemented in the main SC prototype described above.

We will first discuss the challenges of maintaining open-ended dialogue and how this relates to the COMPANIONS project. As part of this discussion we will propose a model that has the potential to navigate the full range of emotional states and adjust its dialogue strategy accordingly. As a first step towards this complex model we will outline an initial prototype in section 2.2, giving details of a system that can learn and manipulate the simple emotional states of an artificial user. The results of this system will then be discussed in section 2.3 and the shortcomings and future work highlighted in section 2.4

.

As we noted earlier, in Young's work (ibid.) reinforcement learning (RL) proceeded as in Figure 3 by forming a belief, Bs, about the current progress towards the successful end of the conversation and assigned an appropriate reward. We have modified this approach, as the belief state is now an understanding of the current emotional state of the user and the reward function is modified accordingly. At present, the Companions system proceeds through a series of tests and actions arranged into dialogue action forms (DAFS), as the dialogue proceeds DAFS are pushed onto or popped from a conversational stack, allowing the dialogue with the user to precede. This strategy will remain in what follows; instead of using a POMDP system to learn the specific dialogue strategy we will use RL to acquire an intelligent movement through a partitioned emotional 'DAFS space'. This effectively decreases the resolution of the POMDP system, allowing it to learn an overall strategy leaving the details to the existing DAFS. So for example, the sadness of a given situation could be reflected by an empathetic tone of voice and conciliatory paralinguistic features but the dialogue itself remains the providence of the DAFS.

Before deploying a complete representation of this continuous emotional space we will present a 'proof of concept' model that responds to the stress levels of the user and acquires a mapping between its actions and the resulting emotional state of the user. This decision was taken as, despite some success in the classification of large emotional corpora (Schuller et al., 2003; Park and Sim, 2003; Lin and Wei, 2005; Casale et al., 2007), an accurate and rapid understanding of the users' current emotional state will require further research. Building initially on the work of others (Paeschke. 2004; Cowie et al., 2000) we will reduce the proposed emotional space to a single dimension representing the current stress level. By detecting key features of stressful speech (e.g. high mean F0, wide pitch range, high energy and fast tempo) it will then be the task of the Companion to move this representation of the user along the continuum towards an unstressed state. Initially, an artificial user, one that possesses a complex mapping between the Companions' actions and its own emotional state, produces this representation. The task of the Companion is to learn this mapping, through an online reinforcement learning procedure, and to exploit this knowledge to move the user along the simple speech stress continuum into a 'contented' state. The challenges of perceiving the user's emotional state will be simulated by having the artificial user select an utterance from a previously annotated

speech stress database (Scherer et al., 2008). This selection will accurately reflect the user's current emotional condition and by using real speech we have captured the uncertainty inherent in this method. Testing and training then proceeds through multiple interactions between the Companion and the randomly generated artificial user.

Ultimately, this speech stress continuum will be expanded to a full emotional space, mapping Cichosz's reduced feature space (Cichosz and Slot, 2005), which defines a number of emotional speech indicators, to the emotional space defined by Wundt (ibid.). Over this space three things will operate, a partitioning of the space, and two points representing the user and Companion. A trained self-organising map (SOM) (ibid.) will partition the space, defining key 'emotional attractors', it is to these attractors that distinct emotionally driven actions will be attached reflecting and expressing the current emotional state of the Companion. Secondly, a representation of the user's emotional state will be defined from a number of multi-modal inputs (paralinguistic features and semantic content), these will form a compressed belief state for the Companion allowing the augmented MDP to proceed. Finally the Companion will maintain a representation of its own emotional state, this will move in response to the user attempting to manipulate them into a positive emotional state, Figure 4. Given this representation the Companion can now learn, through RL, how best to alter its own emotions and motivation.

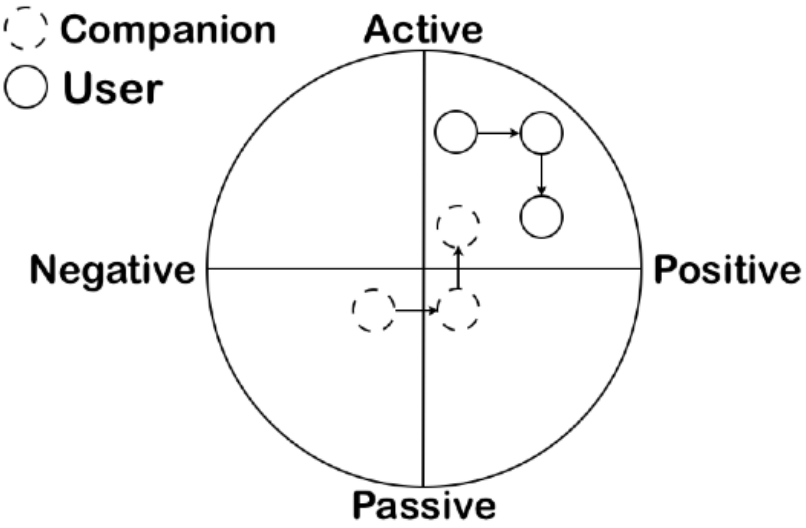


Figure 4. An example of a continuous emotional space

Movement through the continuous emotional space, an understanding of the users location in emotional space allows the companion to respond intelligently by modifying its own location.

By creating a SOM, capturing the variability of emotional speech, we can train the system on a wide range of emotional utterances forming distinct attractors within the defined emotional space. Having trained the system these attractors will be tied to a subset of dialogue acts, reflecting the agents current emotional state. Reinforcement learning can then proceed by allowing the Companion to learn how its own movement in this space affects the user. However, this is only possible if the Companion can place a representation of the state of the user into this space, accordingly we need to

ascertain the user's current emotional disposition. Clearly, as the model increases in complexity the representation of the emotional state of the user will become increasing subtle and uncertain.

We propose to reduce this uncertainty by extracting emotional information from a number of inputs. By combining the semantic content and acoustic information of the current utterance the uncertainty of the current emotional state can be reduced (Lauria, 2007; Lee et al., 2002) Having established this understanding the Companion will have a number of tools with which to manipulate the situation.

The Companion will converse with the user on a number of levels, the tone and paralinguistic nature of the utterance can be adjusted to convey emotional content, the structure of the dialogue act can be similarly adapted and the overall conversational strategy can be optimised. When exploiting these features it is important to maintain a consistent personality that accurately conveys the capabilities and limitations of the Companion, any gap between user expectation and reality would form an obvious source of dissatisfaction. As shown by previous WOZ studies (Moore and Morris, 1992) simple differences to tone of voice or appearance can cause dramatic differences in user behaviour. Accordingly, the Companion will present a unified personality (encompassing appearance, voice, personality and behaviour), which accurately manages the user's expectation, as we have demonstrated emotion lies at the heart of this approach.

3 Evaluation of the Senior Companion

The notion of companionship is not yet one with any agreed evaluation strategy or metric, though developing one is part of the main project itself. Again, there are established measures for the assessment of dialogue programs but they have all been developed for the standard task-based dialogues and the SC is not one of those: there is no specific task either in reminiscing conversations, nor in the elicitation of the content of photos that can be assessed in standard ways, since there is no clear point at which an informal dialogue need stop, having been completed. Conventional dialogue evaluations often use measures like "stickiness" to determine how much a user will *stay with* or *stick with* a dialogue system and not leave it, presumably because they are disappointed or find it lacking in some feature. But, of course, it is hard to separate that feature out from a task rapidly and effectively completed, where stickiness would be low not high. Traum (Traum et al., 2004) have developed a methodology for dialogue evaluation based on "appropriateness" of responses and the Companions project is developing a model of evaluation for the SC based on that (Webb et al., 2010).

The evaluation of phase I (years one and two) was conducted at Napier University (Benyon et al., 2008) and was principally concerned with initiating and testing the metric evaluation process. In June and July 2008 we evaluated three manifestations of the Companions concept – the Senior Companion (SC), the Health and Fitness Companion (HFC—this volume) and the mobile HFC. The purpose of the evaluations was as much to refine the evaluation protocol and data gathering method, as it was to evaluate the products. The evaluation also served as a preliminary feedback

mechanism for the development of the prototypes by providing initial interface feedback to the development teams A key intention of this first evaluation was to enable the prototypes to be as stable and usable as possible without altering fundamental underlying technologies. However, useful base-line data has been gathered during this process.

The mechanism for the first phase of evaluation was two-fold:

Metric-centric – The use of quantitative methods to determine values for speech metric data including word error rate (WER) and concept error rate (CER) of the various prototypes. Alongside these are a variety of dialogue metrics such as dialogue duration; number of turns; words per turn, etc.

User-centric - Qualitative methods used to acquire subjective opinions from the users of the Companions prototypes, including Likert based surveys, focus groups and interviews.

Eight people completed the whole protocol. All participants were native English speakers without strong accents ranging in age from 27 to 61. 2 were female, 6 were male. Each participant had to complete four distinct tasks; introductory tutorials, using prototypes, on-line surveys and interviews. The users of the SC had a voice training exercise with the "Dragon Natural Language" software before its associated introductory tutorial. This voice training exercise took five to seven minutes to complete. Each session began with an introductory tutorial. These ten to sixteen slide presentations introduced the prototype, established its intentions, its limitations, what the prototype would say and do, how to use the prototype and give the user suggestions in how to respond.

Participants then used the SC for 10 - 15 minutes each, completing the on-line questionnaire after each session. Researchers were sitting in the background while the participants interacted with the prototypes, and participants were video-taped during their interaction. Researchers were able to intervene in case of catastrophic failure. Finally the researcher interviewed the participants.

Standard timing information was collected from each interaction – to establish baseline guides for the usability and 'stickiness' of each prototype. System utterance length is calculated from the moment when audio output started to the moment when it ended. Everything between those events is reported as delays. For SC, the system itself has calculated average system and user turn durations. Vocabulary sizes and utterance lengths (in words) are available both based on ASR results and on transcriptions. Word error rate (WER) has been calculated using the standard formula (Deletion Errors + Insertion Errors + Substitution Errors) / (number of words actually uttered by user). Regular dynamic programming string alignment has been used to calculate the errors.

Dialogues with SC had between 100 and 160 dialogue turns (sum of both user and system turns). Dialogue durations were between 9 minutes 20 seconds and 15 minutes 15 seconds. Average length of user utterances varied between participants from 2.9 and 6.8 words for SC. There are significant differences in how verbose different people are. Comparing the actual utterance lengths with ASR results, ASR in the SC recognises fairly closely the same amount of words as uttered. Average system

utterance length for SC is around 14 words. User vocabulary size with the SC ranged between 33 and 131 words. The average of these is 70. Word error rates for SC range between 12% and 37%. Many of the errors are small insertion errors, but there are cases where larger segments are completely misrecognised.

Traditional dialogue systems' evaluations place a high reliance on user feedback. Measures of how people related to the Companions were collected through on-line questionnaires. The SC consisted of forty questions that were answered on a 5- point Likert scale (strongly agree, agree, undecided, disagree, strongly disagree). The last ten questions were concerned with gathering feedback about the aesthetics of the interface, in order to inform subsequent designs. The first thirty were aimed at validating a model of companions and at establishing a base line for further developments. Twenty-seven responses were collected. The questions were organised around six themes:

A	The behaviour of the Companion and what it looked like
B	The utility of the Companion
C	The nature of the relationship between participant and Companion
D	The emotion demonstrated by the Companion
E	The personality of the Companion
F	The social attitudes of the Companion

Table 2: Six themes for the questionnaire

The Likert scales asked people to indicate whether they agreed or not with statements such as those shown in table 3:

The dialogue between the Companion and me felt natural
I thought the dialogue was appropriate
Over time I think I would build up a relationship with the Companion

Table 3: Examples of questions asked with Likert scale

The answers were scored as 1 for strongly agree through to 5 for strongly disagree. Much of the data was biased to the 'undecided' option, partly because the prototypes are still in their early stages and are not yet achieving the higher-level ambitions of Companions. However, some very strong opinions were elicited. Twenty-seven responses were received with respect to the SC. Nine of those scored an average of over 3.0 and one scored an average of 1.8. All the others scored between 2 and 3. They agreed (average score 1.8) with the statement: The Companion was polite.

3.1 *Initial reinforcement learning for emotional utterances*

To test the fundamentals of the system proposed in section 2.5 an initial model will be constructed and analysed. The implemented system will consist of a number of components. The agent itself will contain two stages, a perceptual stage, section 3.1.1, which determines the emotional content of a speech signal and a reinforcement learning stage, section 3.1.2, which learns how to maintain the users positive emotional state given this information. The 'user' in this system is entirely artificial

and consists of two mappings, one which maps from a current emotional state to an utterance and the other which maps from the agents response to an adjustment of its own emotional state. To successfully maintain the user's positive emotional state the agent needs to acquire these two mappings.

3.1.1 Perception of emotion from speech

To provide the artificial users utterances training and test data was obtained from a database of German emotional speech (Burkhardt et al., 2005) consisting of 10 native German speakers (5 female and 5 male) simulating 6 basic emotions with 10 utterances per emotion. To simplify our training task we selected 2 basic emotions (stress and joy) and an individual neural network was trained on the F0 values (Kim et al., 2007) extracted from each utterance.

The F0 values corresponds to the central frequency of speech and is believed, among other things, to convey paralinguistic qualities of speech (Crystal, 1980), it has been hypothesised that this region contains a large amount of emotional content in speech. By training each network on a range of utterances from an individual speaker and stripping out the linguistic content of the utterance we are able to reduce the variance of the task and focus upon its emotional content.

The first stage of learning then proceeds through Levenberg- Marqundt (Marqundt, 1963) backpropogation, for each utterance 204 F0 (ibid.) values where extracted from 500ms of speech. After supervised training, figure 5(a), on the acted emotional speech the agent proceeds to the second stage, where it learns to exploit this information through conversation with the artificial user.

3.1.2 Use of emotion in dialogue

Having acquired a connection between speech and emotion the agent will try to manipulate the user into a positive emotional state. The user responds by moving along an abstract emotional continuum ranging from upset to happy (−1 to 1). Attractors along this scale at intervals of 0.2 capture how a users current emotional state influence their response to utterances, i.e., a certain phrase will not produce the same response in all scenarios. These attractors capture the behaviour of the converged SOM that will be present in the final system. Accordingly, to successfully maintain the users positive state the system needs to learn the mapping between speech and emotional consequence at each attractor point.

The artificial user conveys its current emotional state by selecting an appropriate speech signal from the German speech database. It then listens to the systems response and adjusts its emotional state according to its mapping between response and emotion, modified by a sensitivity parameter, σ. This parameter represents a proportion between the user's current state and the new state proposed by its mapping.

The agent's understanding of the users selected response forms a belief, Bs, about the users current emotional state. This belief is then modified by the agents learning

rate, λ, to update the agents understanding of the user's mapping from perceived utterance to emotional state. Here, λ represents a proportion between the systems current understating and Bs. This cycle of user utterance and agent reply continues until it is clear that the users emotional state has converged. Each round of utterance and reply is described as one time step in section 3.1.4.

During reinforcement learning we have modified the standard exploration approach α (Young, ibid.) buy adding the variable ρ, as shown in equation 1. This variable is the Euclidean distance to the desired emotional state of the user. Consequently, as the user is perceived to approach this state the agent feels increasingly at liberty to explore the space of possible utterances. As defined in equation 1, when combined with a variable decay rate, δ, the system avoids premature convergence on a local optimum and balances exploration/exploitation in response to the user, i.e. when the user is unhappy it will focus on making them happy, when they are happy it can indulge in exploration of the space of possible mappings.

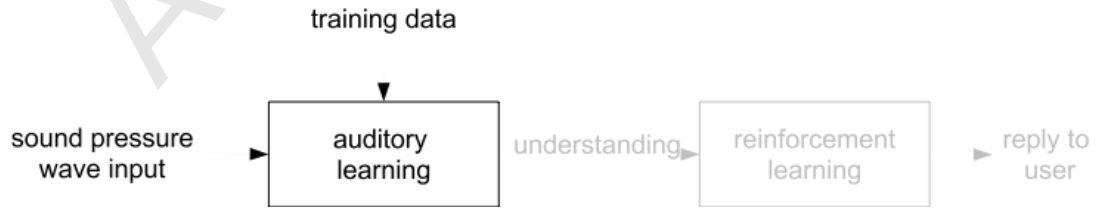
$$\alpha = E\left(-\frac{\rho+1}{\delta}x\right) \qquad (1)$$

In equation 1 α defines a proportion of the subset of possible utterances, attached to an emotional state, of which one is selected at random, allowing for exploration around the optimum utterance.

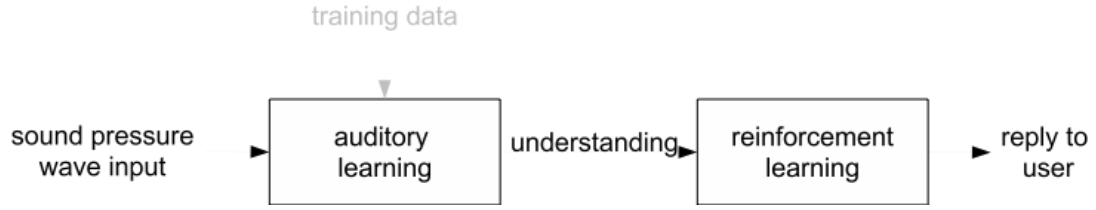
3.1.3 Experimental work

Having developed this model we will now investigate whether it can learn the users emotional responses and maintain the users positive emotional state. We will also investigate the systems robustness to parameter variation (δ,λ,σ) and the effect of a variable balance between exploration and exploitation, captured by ρ.

In all experiments we will record the results from 5 different German speakers (2 female, 3 male) averaged over 20 runs. In all cases training takes place on 80% of the data and testing on the remaining 20%, the number of emotional utterances varies according to the selected speaker. We will be recording the changing emotional state of the user and the average error for the system's understanding of the user's signal to emotion mapping.



(a) Stage one training, supervised learning of a mapping between speech and emotion.



(b) Figure 5: Stage two training, reinforcement learning proceeds through dialogue.

The two training stages of the agent as it attempts to learn the mapping between its own utterances and the user's emotional state and attempts to manipulate the user into a favorable emotional state.

3.1.4 Emotion Results

During supervised training, figure 6(a), the average classification error (stressed speech = −1/happy speech = 1) over 100 trials was 0.0049 with a standard deviation of 0.018. This suggests that for our limited purposes the change in F0 is sufficient to distinguish between these two types of emotional speech.

The results of stage 2 show that the system is highly robust to parameter variation: after 800 time steps the system's understanding had converged to a solution and the users emotional state was maintained. As can be seen in figure 6(b) without any learning-taking place, the control experiment, the agent fails to establish a positive emotional state in the user, as it cannot correct its initially random signal to emotion mapping.

Under a reasonable parameter set (δ = 500, λ = 0.5 and σ = 0.2) without variable exploration the system reduces the absolute average error to 0.81 with variable exploration the error is reduced to 0.67. Averaged over 100 runs (20x5 speakers) and between 95 to 125 signals to emotion mappings, depending on the chosen German speaker, the resulting p-value is less than $1E^{-51}$ establishing the clear significance of exploration.

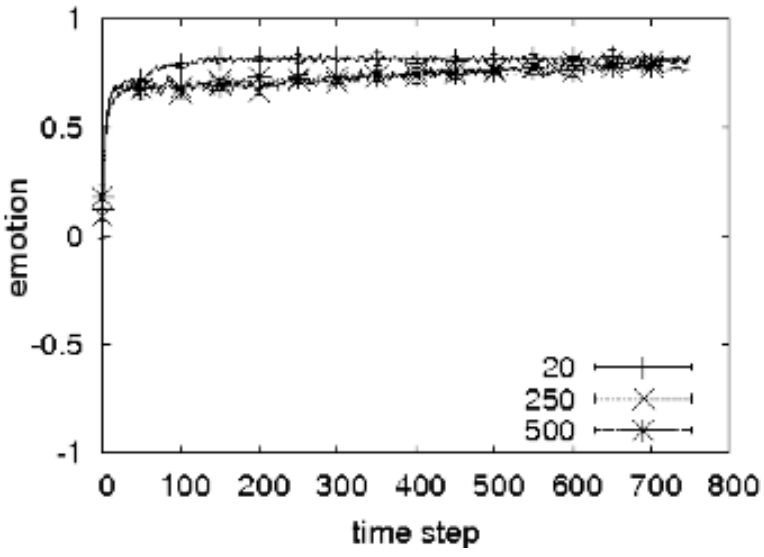


Figure 6 (a) Robustness to a varying decay parameter, δ,

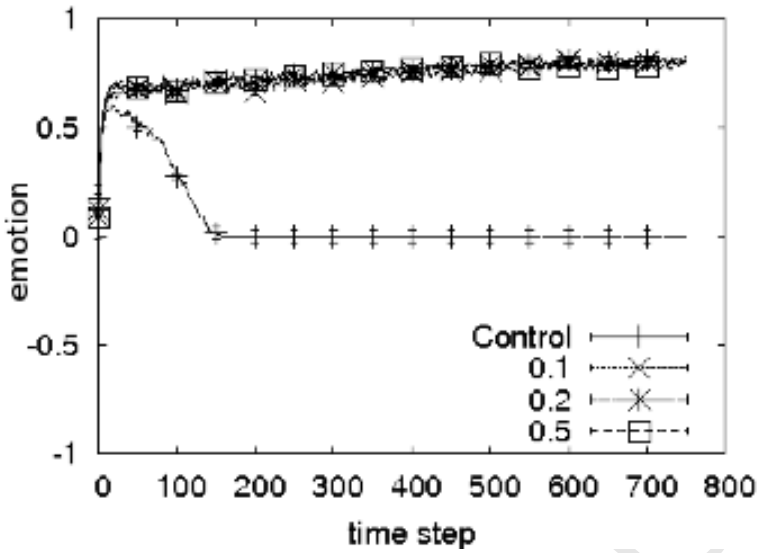


Figure 6(b) Robustness to a varying learning parameter, λ. Without learning the system fails.

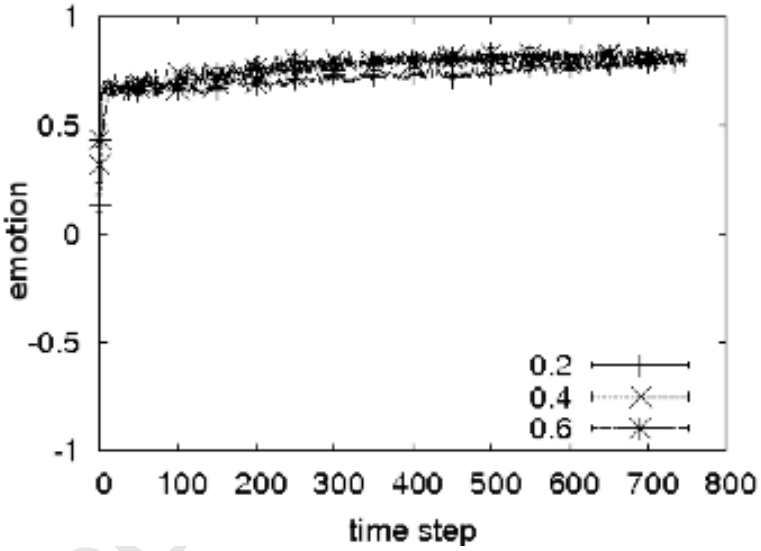


Figure 6(c) Robustness of the system to varying user sensitivity, σ

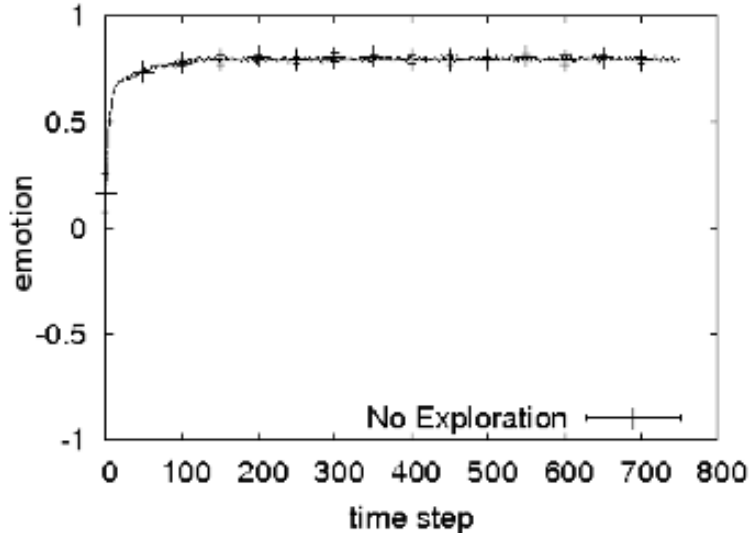


Figure 6(d) Without flexible exploration the system can still maintain a positive emotional state.

Figure 6. Demonstrating robustness to parameter variation, and the abstract nature of the user, the system is able to maintain the users positive emotional state (y-axis> 0) under a variety of scenarios. Results obtained from an average of 5 speakers with 20 runs per speaker, error bars have been plotted but are too small to show.

4 Conclusions

This preliminary work allows us, in common with the work of others, to draw those key correlations between user feedback and our observable metrics. It should be remembered that the principle point of the first Companion demonstrators was to provide a usable technology base from which we could explore the nature and content of companionable dialogues. Even as a lower baseline, the evaluation conducted in phase one provides us with useful information. However, the evaluation is a success from the following standpoints. The developers were able to deliver stable initial prototypes, and the evaluation mechanism – from recruiting participants to acquiring the necessary metrics – has been established.

Meanwhile, a wide set of Wizard of Oz (WoZ) experiments has been developed to deal with special issues that arise in the evaluation of Companions (Webb et al., 2010), which will allow for testing the metrics developed and mentioned above on a defined set of dialogues and interactions, and tuning the parameters to maximize over. The application of these evaluation criteria will help further refine the initial evaluation paradigm for Companion technologies. For example, to investigate the impact of the length, content or timing of a Companion utterance on users will be manageable in a WoZ scenario, where the wizard has a series of guidelines. There are several such explorations that will be best served by this methodology, including dialogue initiation as well as the proactivity and the personality of the Companion.

4.1 *Discussion on ASR and emotion*

Section 2 has presented an implementation of an open-ended dialogue strategy, through the implementation of the variable exploration/ exploitation balance the agent can continue to learn and function in response to the users emotional state. However, it is important to emphasise the abstract nature of this initial model, even though we use real speech our user is entirely artificial and, as figure 5 shows, it can be argued that it doesn't present enough of a challenge for our system. However, we believe that this model establishes the plausibility of this approach and, as shown in figure 6(b), without learning the system fails.

Due to the abstract nature of the user it becomes difficult to demonstrate the advantages of the variable exploration parameter, ρ. As shown by figure 6(d) the emotional state of the user does not suffer when this parameter is removed. However, the resulting difference in the average learning error demonstrates the importance of this parameter and we believe that as increasingly complex and realistic models are constructed its role will become increasingly important, as over the course of

prolonged interactions the system will have to know when it is appropriate to discover the emotional subtleties of its user.

In previous work (Oudeyer, 2003: Greasley ibid.) it has been shown that natural speech, with emotional ambiguity and few incidents of 'full-blown' emotion, is significantly harder to learn than acted speech. This has implications for our initial learning stage, Figure 5(a), which is further complicated by the fact that the Companion needs to make sense of emotional speech from the outset. In typical use a comprehensive, annotated, dataset of emotional speech tailored to each user will be unavailable, making a supervised learning approach impossible. Additionally, attempts to build a general emotional speech recogniser were unsuccessful as the variability between speakers was too great for the neural network, obscuring the relevant emotional features. Clearly, an advanced unsupervised approach will have to be developed to recognise real emotion in speech.

Although the results remain robust at no point do they reach an optimum emotional state, we believe that this is due to the perceptual error inherent when working with real speech. As a result misclassification errors occur preventing the system from achieving an optimum result. Additionally, as the system moves towards an optimum it is at increasing liberty to explore a range of utterances, accordingly the system trades a perfect emotional state for the opportunity to acquire a range of possible mappings. Clearly, a number of challenges remain but this initial model has served its purpose. We have demonstrated that emotional speech can motivate an open-ended dialogue and developed a method of balancing the twin goals of maintaining a positive emotional state while learning a users response to emotional utterances. We believe that as the technical challenges, highlighted in this section, are overcome this approach will become increasingly important, eventually reaching the point where it can converse with real users.

In this section 2 we have shown two things: an implemented proof of concept model demonstrating how the emotion of a dialogue can enable the application of reinforcement learning in an open-ended conversational system and a plan detailing how this work can be expanded to capture and exploit a wide range of uncertain human emotions. In future work we will begin to address this uncertainty by proposing the exploitation of multiple sources of information, integrating speech, language and the environment, which we take, as in section1, to include open-ended internet knowledge sources. Ultimately, we hope to present a unified Companion, one that can accurately assess and respond to the user's emotional state.

Acknowledgements

This work was funded by the COMPANIONS project sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434. Companions demonstrators can be seen at: http://www.dcs.shef.ac.uk/~roberta/companions/Web/

References

Akkaya, C., Wiebe, J., and Mihalcea, R.,. 2009. Subjectivity Word Sense Disambiguation, In: EMNLP 2009.

Barreto, J., et al., 2007. Non-intrusive physiological monitoring for automated stress detection in human-computer interaction, Human Computer Interaction, 4796, pp. 29–38.

Benyon, D., P. Hansen and N. Webb 2008. Evaluating Human-Computer Conversation in Companions In: Proceedings of the 4th International Workshop on Human-Computer Conversation, Bellagio, Italy, October.

Burkhardt, F., Paeschke, A., Rolfes, M., and Sendlmeier, M. 2005. A database of German emotional speech In: Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal.

Casale, S., Russo, A., and Serrano, S. 2007. Multistyle classification of speech under stress using feature subset selection based on genetic algorithms. Speech Communication, 49(10-11), pp. 801–810.

Catizone, R., Setzer, A., and Webb, N. 2002. Scaling-up Information Extraction, In: Proceedings of the Workshop on Event Modelling for Multilingual Document Linking, Language Resources and Evaluation Conference (LREC 2002), Las Palmas, Canary Islands.

Catizone, R., Setzer, A., and Wilks, Y. 2003. Multimodal Dialogue Management in the COMIC Project, In: Proc. Workshop on Dialogue Systems: interaction, adaptation and styles of management, European Chapter of the Association for Computational Linguistics (EACL), Budapest, Hungary.

Cichosz, J. and K Slot, 2005. Low-dimensional feature space derivation for emotion Recognition. In: Proc. Ninth European Conference on Speech Communication and Technology, Lisbon, Portugal.

Cowie, R., Douglas-Cowie, E., Savvidou, E. and McMahon, E. 2000.`'feeltrace': An instrument for recording perceived emotion in real time', in ISCA Tutorial and Research Workshop on Speech and Emotion, Newcastle, Northern Ireland, UK.

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

Cunningham, H., Humphreys, K., Gaizauskas, R., and Wilks, Y. 1997. GATE -- a TIPSTER based General Architecture for Text Engineering. In: Proceedings of the TIPSTER Text Program (Phase III) 6 Month Workshop. DARPA, Morgan Kaufmann, CA.

Crystal, D. 1980. A first dictionary of lingusitics and phonetics, Oxford: Wiley Blackwell.

Ekman, P., 1999. Basic emotions, In: The Handbook of Cognition and Emotion, Wiley: New York

Greasley, P., Sherrard, C., and Waterman, 'M. 2000. Emotion in language and speech: Methodological issues in naturalistic approaches, Language and Speech, 43(4), pp. 355–375.

Hopkins, RJ., Ratley, DS., Benincasa, R. and Grieco, JJ. 2005. Evaluation of voice stress analysis technology, In: Proc. 38th Annual Hawaii International Conference on System Sciences, pp. 1–10.

Kaelbling, LP., Littman, ML., and Cassandra, MR. 1998. Planning and acting in partially observable stochastic domains, Artificial Intelligence, 101(1- 2), 99–134.

Kim, S., Georgiou, PG., Lee, S. and Narayanan, S. 2007. Real-time emotion detection system using speech: Multi-modal fusion of different timescale features, In: Proc. IEEE 9th Workshop on Multimedia Signal Processing, pp. 48–51.

Kohonen, T. 1982. Self-organized formation of topologically correct feature maps. Biological Cybernetics, 43 pp.59-69.

Lauria, P. 2007. Talking to machines: Introducing robot perception to resolve speech recognition uncertainties, Circuits Systems Signal Processing, 26(4), pp.513–526.

Lee, C., Narayanan, S.m and Pieraccini, R. 2002. Combining acoustic and language information for emotion recognition, In: Proc. Seventh International Conference on Spoken Language Processing, Denver, CO.

Lin YL., and Wei, G. 2005. Speech emotion recognition based on hmm and svm, International Conference on Machine Learning and Cybernetics, volume 8.

Marquardt, DM., 1963. An algorithm for least squares estimation of nonlinear Parameters, In: Journal of the Society of Industrial and applied Mathematics, 11(2), pp. 431--441.

Moore, R. and Morris, A., 1992. Experiences collecting genuine spoken enquiries using woz techniques, In: Proc. Fifth DARPA Workshop on Speech & Natural Language, Narriman, NY.

Paeschke, A. 2004. Global trend of fundamental frequency in emotional speech, In: Speech Prosody, Nara, Japan.

Oudeyer, P-Y. 2003. The production and recognition of emotions in speech: features and algorithms', International Journal Of Human-Computer Studies.

Park, CH., and Sim, KB. 2003. Emotion recognition and acoustic analysis from speech signal. In: International Joint Conference on Neural Networks.

Roy, N., Pineau, J., and Thrun, S. 2000. Spoken dialogue management using probabilistic reasoning, In: Proc. 38th Annual Meeting of Association for Computational Linguistics.

Scherer, S., Schwenker, F. and Palm, G. 2008. Emotion recognition from speech using multi-classifier systems and rbf-ensembles, in Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, pp. 49–70, Springer: Berlin.

Schuller, B., Rigoll, G. and Lang, M. 2003. Hidden Markov model-based speech emotion recognition' In: International Conference on Multimedia and Expo, volume 2, pp. 1–4.

Traum, DR., Robinson, S., and Stephan, J. 2004. Evaluation of multi-party virtual reality dialogue interaction, In: Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC 2004), pp.1699-1702

Webb, N., Liu, T., Hepple, M., and Wilks, Y. 2008. Cross Domain Dialogue Act Tagging. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC-08). Marrakech, Morocco.

Webb, N., Benyon, D., Hansen P. and Mival, O. (submitted) Wizard of Oz Experiments for a Companion Dialogue System: Eliciting Companionable Conversation. Submitted to LREC 2010.

Wundt, W., 1913. Grundriss der Psychologie, A. Kroner: Berlin.

Wilks, Y. 2007. Has there been progress on talking sensibly to computers ? Science, Volume: 318.

Wilks, Y. 2008. The Semantic Web and the apotheosis of annotation. In: Proc. IEEE Intelligent Systems. (May/June)

Wilks, Y., Catizone, R., and Mival, O. 2008. The Companions paradigm as a method for eliciting and organising life data, In: Proc. Workshop on Memories for Life, British Computer Society, London, March.

Young, S. Y. 2002. Talking to machines (statistically speaking), In: Proc. Seventh International Conference on Spoken Language Processing, Denver, CO.