

Different ways of ending human-machine dialogues

Loredana Cerrato

TMH-CTT,

KTH Royal Institute of Technology,

GSLT- Graduate School of Language Technology
Sweden

loce@speech.kth.se

Susanne Ekeklint

School of Mathematics and Systems Engineering,
Växjö University,

GSLT- Graduate School of Language Technology

Sweden

sek@msi.vxu.se

ABSTRACT

More and more dialogue systems are developed to provide public services to inexperienced users. This makes the competition harder among developers and increases the interest in finding methods to evaluate and improve the overall performance of dialogue systems. We believe that analysing the way in which users end their interactions with dialogue systems could provide some new metrics to evaluate users' satisfaction. To prove our hypothesis we carried out a preliminary analysis of the way in which users of two experimental Swedish multi-modal conversational dialogue systems (namely August and Adapt) ended their interactions. The results indicate that prosodic cues, such as pitch contour and intensity, can be used as an indication for users' satisfaction.

1. INTRODUCTION

Evaluation plays a crucial role in speech and natural language processing, both for system developers and for technology users. Developers need to evaluate their product at different stages: before starting the actual development (Initial Evaluation), during the development (Evaluation in Progress) and at the end of the development (Final Evaluation).

Initial Evaluation: Developers are interested in finding the modules that are most likely to give the best dialogue model. Different modules are often separately validated for this purpose. Evaluation of what is adequate for the system to handle often considers parameters such as cost and performance according to the task under consideration [8].

Work in Progress Evaluation: During the development of systems, different kinds of "Diagnostic Evaluations" are carried out. For example different generations of the system can be compared according to specific parameters that are particularly interesting for different modules or sometimes for the overall performance of the system.

Final Evaluation: In the very last stage of a system development it is necessary to determine the adequacy of a system for a specific purpose. This evaluation can be carried out both by developers and by users, since it can be used typically to compare similar systems (i.e. alternative implementations of the same system or successive generations of the same implementation). The

developer can look at parameters for the overall performance of the system by counting for example elapsed time on a particular task or number of speech turns to complete a particular task.

During all these stages in the development there is the need to look at the performance of the separate modules. There are several evaluation methods that propose evaluation measures for individual components. For example the standard way to evaluate speech recognition and language understanding modules is to (given a certain input) compare the actual output to the desired output [9]. Many systems have some kind of confidence scoring of the recognized utterances [5]. Another module that might be evaluated separately is the dialogue manager. As suggested by Danieli and Gerbino [4] the total score for the dialogue systems robustness can be measured in terms of the dialogue manager's ability to perform both implicit and explicit recovery when the speech recognition or the language understanding unit fails.

Evaluating the overall performance of a system is also necessary. One way to do this kind of evaluation is to look at the users' satisfaction. One of the more recent tools for evaluation of spoken dialogue systems is PARADISE [13]. This method uses various parameters to calculate an estimation of the user satisfaction. The PARADISE paradigm breaks down the term user satisfaction into costs and success – the goal is to maximize the success and minimize the costs. PARADISE takes into account several parameters when calculating the user satisfaction. For example, counting number of rejects, cancels, time-outs and mean recognition score. Other things taken into account are number of turns, requests for help, barge-ins, elapsed time, etc. Rejects, for example, are the number of times that the recognizer cannot produce a result with enough confidence.

Being a relatively new field, the development of conversational dialogue systems lacks both evaluation tools and established sets of evaluation criteria. Very few empirical tests have been carried out to learn about the benefits of embodied agents on different aspects such as: entertainment, mental load and system efficiency [10]. The more modalities that are used in a system the more complex the evaluation of the individual components will be, since each component will need to collaborate with a larger number of components. The objectivity when looking at separate modules may also be influenced by other modules since one module's weakness may very well be saved by another module

[3]. For example: if the speech recognizer is not so good, a well working dialogue manager may fix some of the problems of the speech recognizer.

The different methods that have been proposed so far to evaluate dialogue systems are based mainly on modular evaluation and very little focus is put on how users interact with different systems. It is of course possible to ask users for a subjective judgement of a system. This can be done by asking the users specific questions about the way in which their interaction with the system went on. However it is not always practical to use forms or interviews to evaluate dialogue systems, since the judgement of the users can be influenced by many factors and further more there are the issues of costs, time and users' integrity. Moreover there are particular questions, which are difficult to formulate such as whether the interface has influenced the users' feelings and expectations during the interactions.

2. MOTIVATION

Finding new parameters and metrics to evaluate complex dialogue systems is a challenging experience. We believe that the prosody of users' utterances could be considered as one of these parameters that we are searching for.

Humans are able to understand if a speaker is happy, sad, worried, disappointed; the cues for the interpretation of these emotions are in the utterances spoken by the speakers, therefore we aim at finding out if it is possible to exploit prosodic information for the evaluation.

Our method does not aim at providing an overall evaluation for dialogue systems, but it proposes to exploit prosodic information as a complement to other traditional subjective and modular evaluation methods. The acoustic measurement can be used as an indication to whether or not the users are satisfied with the way the interaction with the system went on. The reasons for the degree of satisfaction will not be given by this measurement. It will however point out when something did not go as well as expected. Even though the method does not cover all the aspects of users' satisfaction it has several positive sides:

- It can be used as an indication of whether or not the system is working as well as the user expected it to.
- It is a method that can be implemented at low-cost, since it is possible to do the tests automatically.
- It can be used as a trigger to start certain modules, for example help modules.
- It might work as a parameter to decide when to save data from dialogues for further investigation.

3. MATERIAL

To carry out our analysis we used material collected by means of two experimental Swedish multimodal conversational dialogue systems: *August* and *Adapt*.

The August database counts more than 10.000 utterances produced by 2685 users, all visitors at the Stockholm Cultural Centre, where the dialogue system was displayed as part of the *Cultural Capital of Europe '98 program*. The August system was developed for research purposes [6]. It was endowed with a talking head with a "distinctive personality" (resembling the author August Strindberg), which invited users from the public to try the system and somehow induced them to socialize with it, rather than just seek for straightforward information. The users received no instructions about how to interact with the system.

The system was able to give information about three different domains, the Royal Institute of Technology, the city of Stockholm and the life and work of Strindberg. The material of the August database is quite heterogeneous, since it contains recordings from users of different ages, languages and background [2]. Moreover the interactions with the systems are not always carried out by a single user. In fact sometimes groups of users tried to interact with the system overlapping or alternating each other in the turns. Many of the recorded dialogues are quite short (less than 5 turns). For these reasons we selected a sub-corpus of 274 interactions among those that presented more than 5 turns in which the users did not overlap with each other. We will refer to this material as the *August corpus*.

The Adapt database was collected in 1999 by means of the Wizard of Oz technique. It includes a total of 50 dialogues produced by 33 users. Adapt is an experimental Swedish conversational multi-modal dialogue system [7]. The system is able to provide information about apartments on sale in Stockholm. The dialogues in the Adapt database can be described, in our opinion, as "information seeking" or "task oriented" since the users were given the specific task of finding apartments in Stockholm that fulfilled certain criteria. From this database we selected 32 dialogues; we will refer to this material as the *Adapt corpus*.

4. ANALYSIS

We carried out an analysis of users' last turns in the two corpora. For the August corpus we started by categorized the final utterances using the typology proposed by Bell and Gustafson [1]. They suggest a grouping of the utterances in the August database in two main categories: socializing and information seeking. To these two main categories they add further subcategories as reported in table 1.

Table 1: Utterance typology proposed by Bell and Gustafson (1999a)

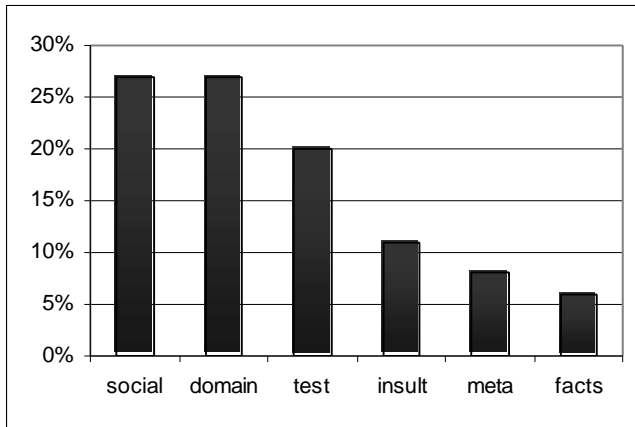
Socializing	Information seeking
Social	Domain
Insult	Meta
Tests	Facts

Bell and Gustafson divided the socializing type of utterances in social, insult and test. They explain that the social category includes greetings and remarks of a social kind while the insult category consists of negative comments and swearwords. The test

category on the other hand consists of utterances produced with intention of testing the system.

The results of the distribution of the last users' turns in our August corpus are reported in figure 1. The majority of final utterances have been categorised as social. Also in the whole August database a great number of utterances were categorised as social and Bell and Gustafson explained this result by the existence of the animated agent, an explanation that we support.

Figure 2: Distribution of the last utterance in the August corpus, according to Bell and Gustafson typology



As we can see from the graph the number of final utterances that fall under the categorisation of socializing (social, test, insult) is quite high. The sub-categorisation of socializing utterances proposed by Bell and Gustafson however does not completely serve our purpose of analysing the final utterance, since it was intended to describe all the utterances in the database. For a categorisation of final utterances we thought it could be appropriate to take into account the notion of conventional closures [11].

A conventional closure is expected to appear at the end of a spoken interaction between humans. A goodbye, a thank you or some other kind of courtesy expressions such as "it has been nice talking to you", can be considered conventional closures.

It can be supposed that humans interacting with dialogue systems (and in particular with dialogue systems featuring embodied conversational agents) also would end their interactions with conventional closures. This has been supposed for instance in the Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems-DATE- [14]. DATE counts ten categories to label speech acts. Among these, the opening/closing speech act category is intended for the categorisation of utterances that open and close the dialogues.

Analysing the August database we noticed that the dialogues did not always end with a conventional closure. In fact 35% of final utterances consisted of negative comments or even swearwords, which cannot be considered, in our opinion, as conventional closures. To group all those final utterances that cannot be categorised as conventional closures, such as negative comments

and swearwords, we introduced the category of "non-courtesy expressions".

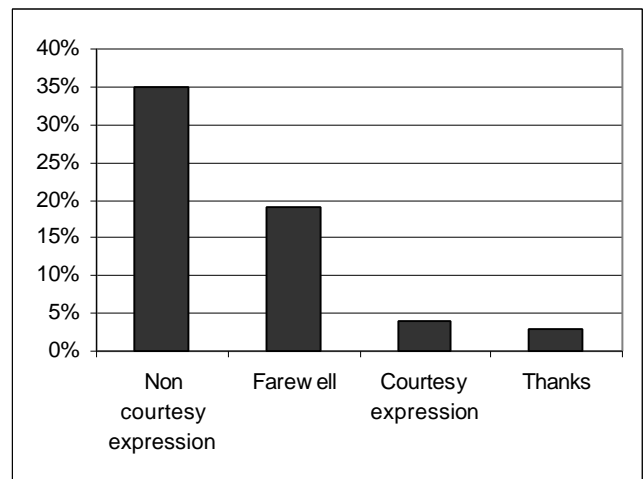
The typology we suggest to categorise the final utterances in the August corpus is reported in table 2.

Table 2: Categorisations of final utterances in closures

Closures
Farewell
Thanks
Courtesy expression
Non courtesy expression

In figure 2 the distribution of the socializing final utterances in the August corpus is reported according to the sub-categories in table 2.

Figure 2: Distribution of the subcategories of the socializing final utterances in the August corpus



A great deal of dialogues (39%) in the August corpus do not end with an utterance that we can consider a closure. We cannot judge if the turns categorised as non-closures depend on the user's choice of concluding the interaction without a closure or whether they are due to the alternation of users in the interactions or to some problems of recording.

This categorisation of final utterances was carried out on the basis of lexical analysis. By this we simply mean that we categorised final utterances by looking at the words without taking eventual suprasegmental information given by the prosodical cues into account. For example when an utterance included the word "goodbye" we categorised this as a "farewell".

However, it is well known that by means of prosody, a different meaning can be given to an expression depending on the attitudinal intentions and the emotional state of the speaker. For instance a "goodbye" uttered with an ironic tone may imply that the user wishes to express his/her dissatisfaction with the interaction. Because of this we believe that more detailed analysis

of the final utterances is needed to find the acoustical correlates that can be used for a deeper semantic interpretation.

We carried out a preliminary deep semantic analysis of some of the words included in the final utterances by looking at their pitch contour. We used the software package Wavesurfer [12].

We selected words from the farewell (“hej” “hej då”) and thanks (“tack”) groups. We noticed the following trends:

- a farewell or a thank with a rising pitch contour is typical when the user has had a successful interaction with the system;
- a farewell or a thank with a falling pitch contour is typical when the user has not had a successful interaction with the system, due to problem of reciprocal understanding or because the user was asking questions outside the domains;
- a thank you or a greeting with a higher intensity and longer duration (respect to the rest of the words in the utterance) were produced by some users who had not had a successful interaction with the system. These users, in our opinion sounded “ironic”.

When using the Bell and Gustafson typology to categorise the final utterances in the Adapt corpus we found that the utterances could be assigned to two categories only: social (63%) and domain (9%).

As in the August corpus, the majority of final utterances are categorized as socializing, but in the Adapt corpus there are no final turns consisting of insults. This probably depends on the fact that the majority of the interactions with the Adapt system went on in a positive way, that is, without problems in reciprocal understanding, while in the August corpus many interactions went on in a problematic way due to problems of reciprocal understanding. 44% of the users of the Adapt system positively completed their task (i.e. finding the apartment they were instructed to look for).

28% of the users could not find the apartment they were looking for (because there were no similar apartments in the system database), but anyway they managed to have an interaction without problems with the system. The fact that the users of the Adapt system never insulted the system at the end of the interaction can also be an indication of the fact that the interactions went on without problems of understanding.

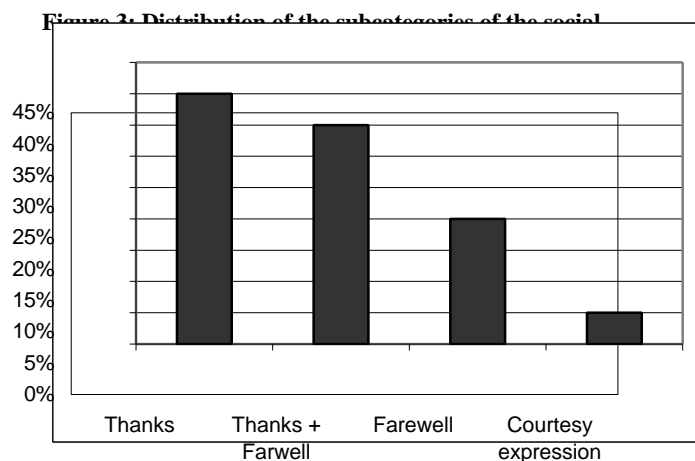
For the remaining 28% of the users the final turns consist of an empty sound file. This can both depend on the personal choice of the users of not concluding the interaction using a conventional closing utterance or it might as well depend on some problems with the recordings.

When we tried to use the closure typology proposed in Table 2 to sub-categorize the socializing final utterances in the Adapt corpus, we realized that some final utterances could be assigned to more than one category, for instance an utterance like: “thank you and good bye” can be assigned both to the “farewell” category and to

the “thanks” category. For these cases we need to add a “joined category” such as “Thanks+Farewell”.

Moreover in the Adapt corpus there are no final utterances that can be categorised as “non-courtesy expressions”. This might depend on the fact that the users of the Adapt system managed to carry out interactions without great problems of mutual understanding.

The sub-categorisation of the socialising final utterances in the Adapt corpus is reported in Figure 3.



An acoustic analysis of the expression “*adjö*” in the farewell category, produced by 7 different users, showed a rising pitch contour for those 3 users that successfully completed their task and a falling pitch contour for those 4 users who did not get the information they required.

5. CONCLUSIONS AND FURTHER INVESTIGATIONS

The results here reported and the results reported by Bell and Gustafson [1,2] support the idea that the human-like appearance of the agent induces users to have a more “social” behaviour towards the system they are interacting with. This is an important observation to support the idea that special evaluation methods are needed for multimodal conversational dialogue system.

The aim of our study was to test our preliminary hypothesis that the way in which users end their interactions with dialogue systems could provide some new metrics to evaluate users’ satisfaction.

To prove our hypothesis we carried out a preliminary analysis of the way in which users of two Swedish experimental multi-modal conversational dialogue systems (namely August and Adapt) ended their interactions.

Even if it was difficult to propose a general typology for the categorization of final utterances, the results of our analysis showed that there is a tendency among the users to end their interaction with the system with a conventional closure when the

interaction proceed without great problems of mutual understanding.

Moreover the results of the preliminary acoustic analysis of farewells and thanks indicate that prosodic cues, such as pitch contour and intensity, can be used as an indication for users' satisfaction.

More thorough investigation is needed to better support our results. In particular to test our hypothesis we need a larger amount of data collected by means of conversational dialogue systems and to validate our hypothesis we need the possibility to compare our results with the results of a more traditional evaluation method (like Paradise).

REFERENCES

- [1] Bell, L. & Gustafson, J. (1999a). Utterance types in the August System. In *Proceedings of the Third Swedish Symposium on Multimodal Communication*. url in May 2002: http://www.speech.kth.se/ctt/publications/papers/ids99_augutt.html
- [2] Bell, L. & Gustafson, J. (1999b). Interacting with an animated agent: an analysis of a Swedish database of spontaneous computer directed speech. In *Proceedings of Eurospeech 1999*, 1143-1146, Budapest, Hungary.
- [3] Carlson R & Granström B (1996). The Waxholm spoken dialogue system. In Palková Z, ed. *Phonetica Pragensia IX. Charisteria viro doctissimo Premysl Janota oblata*. Acta Universitatis Carolinae Philologica 1; 39-52
- [4] Danieli, M. & Gerbino, E. (1995). Metrics for evaluating dialog strategies in a spoken language system. In *Working Notes, AAAI Spring Symposium Series*, Stanford University, 34-39.
- [5] Glass, J. R. (1999), *Challenges for spoken dialog systems*. Spoken Language Systems Group, MIT Laboratory for computer science, Cambridge.
- [6] Gustafson, J., Lindberg N. & Lundeberg M. (1999). The August spoken dialogue system. In *Proceedings of Eurospeech 1999*, 1151-1154, Budapest, Hungary.
- [7] Gustafson J., Bell L., Beskow J., Boye J., Carlson R., Edlund J., Granström B., House D., Wirén M. (2000) AdApt- a Multimodal conversational dialogue system in an apartment domain. In *Proceedings of ICSLP 2000* (2), 134-137, Beijing, China.
- [8] Hirschman, L. & Thompson, H. S. (1996). Overview of Evaluation in Speech and Natural Language Processing, Introduction to ch. 13 In Cole R, Mariani J, Liberman M, Uszkoreit H, Zaenen A, Zue V (eds.) *Evaluation in the Joint EC-US Survey of the State of the Art in Human Language Technology*, (url in May 2002: <http://cslu.cse.ogi.edu/HLTsurvey>)
- [9] Lippmann, R. (1997). Speech recognition by machines and humans. In *Speech Communication*, 22 Elsevier Science B.V., 1-15.
- [10] Sanders, G. A. & Scholtz, J. (2000). Measurements and Evaluation of Embodied conversational agents. In Cassel, J., Sullivan, J., Prevost, S., Churchill, E., *Embodied conversational agents*, 346-373, MIT press.
- [11] Schegloff E.A., Sacks H., (1977) , Opening Up Closings. In *Semiotica*, 8: 298-327,
- [12] Sjölander, K., Beskow J. (2000). WaveSurfer - an Open Source Speech Tool. In *Proceedings of ICSLP 2000*, 464-467, Beijing, China.
- [13] Walker, M. A., Kamm, C. A., Litman, D. J. (1998). Towards Developing General Models of Usability with PARADISE. In *Computer Speech and Language*, 12-3, (url in May 2002: <http://www.research.att.com/~walker/>)
- [14] Walker, M., Passonneau R., DATE: A Dialogue Act Tagging Scheme for Evaluation of Spoken Dialogue Systems. In *Proceedings of Human Language Technology Conference*, San Diego, March, 2001. (url in May 2002: <http://www.research.att.com/~walker/dtag6.pdf>)