

# SPEECH INTERACTION FOR A CHILDREN'S TOY REFLECTION REPORT

This reflection report describes the design process of creating a simple-to-use speech interface for a children's toy. The ephemeral quality and linear progression of audio proved two great challenges for this design, as well as the unconscious expectation amongst humans for spoken interfaces to express human speaking behavior. The final product uses a natural, conversational, dialog, and gives contextually appropriate instructions to lead the user through the interaction.

Jonas Ohlsson – 42411754

*Team Point and Ask*

18/06/2012

# TABLE OF CONTENTS

## Table of Contents

|   |    |
|---|----|
| INTRODUCTION _____                              | 3  |
| BACKGROUND _____                                | 4  |
| PROBLEMS WITH NON-GRAPHICAL INTERFACES.....     | 4  |
| SPEECH INTERFACE DESIGN PRINCIPLES .....        | 5  |
| DESIGNING OUR SPEECH INTERFACE _____            | 8  |
| CREATING A NATURAL CONVERSATION .....           | 8  |
| MAKING OUR SPEECH ACTIONS PERCEPTIBLE.....      | 9  |
| INDUCING HUMAN HELPFULNESS TO THE PENGUIN ..... | 10 |
| EVALUATION_____                                 | 11 |
| CONCLUSION _____                                | 13 |
| REFERENCES _____                                | 16 |

# SPEECH INTERACTION FOR A CHILDREN'S TOY

*Team Point and Ask*

Jonas Ohlsson - 42411754

Physical Computing & Interaction Design Studio Reflective Report

18/06/2012

## Introduction

The 'help the penguin' prototype is a small physical penguin toy, designed for children to support independent exploration of nearby objects. It uses speech interaction to communicate with the user, who carries it around. The purpose of the interaction is for the kid to figure out and find items that the penguin is looking for, which is done by verbally asking it for clues, and finally, presenting found objects to the penguin. During our design process, we came up with three success criteria for the toy: for the interaction to be simple to understand, for the toy to be mobile, and for the toy to be fun to use.

This report focuses on the speech interface of the toy, and the design challenges encountered in relation to the first core requirements: making it easy to understand and use. I also spent a lot of time and effort into other parts of the design of the toy, especially in regards to making the interaction fun and engaging, but these aspects of my design process will not be covered in this report.

We chose to use speech interaction because we wanted to use interaction methods kids are already familiar with. As we learnt during this project, however, this familiarity with the method itself does not at all guarantee the interaction will be simple. Benefits and disadvantages of speech interaction over other methods will not be covered further in this report. I believe that instead looking at the problems we faced in our project, and how these problems are handled in the literature, proves both a more interesting report for the reader, as well as a greater lesson for me as a designer.

In the next section, I present a brief survey of the relevant, covering first some problems associated with non-graphical interfaces, followed by suggested design principles for speech interaction. In the main body of this report I describe my efforts on making the speech interaction natural, easy to figure

out, and helpful. In the last section, I present my final thoughts about this project, mainly in regards to how simple the final speech interaction was to use, but also in terms of well I think our final toy as a whole met our success criteria.

## Background

In this background section I first cover problems mentioned in the literature with using non-graphical interfaces in general, and speech only interfaces in particular. In the second section, I give an overview of solutions to these problems, as well as general design principles, for speech only interfaces found in the literature.

### PROBLEMS WITH NON-GRAPHICAL INTERFACES

The most significant problem for products lacking a graphical interface is that the user is given no visible clues of possible actions (Norman, 2010; Rosenfeld et al., 2000; Larsen).

Gaver (1991) provides a good framework for affordances that help illustrates this problem. In his framework, he distinguishes between the actual existence of affordances, ways in which tools supports being used, and whether the user has any perceptually information about these affordances (See Figure 1). Only *perceptible affordances* are helpful for users. Speech actions, on the other hand, only have *hidden affordances*, as the user cannot perceive them.

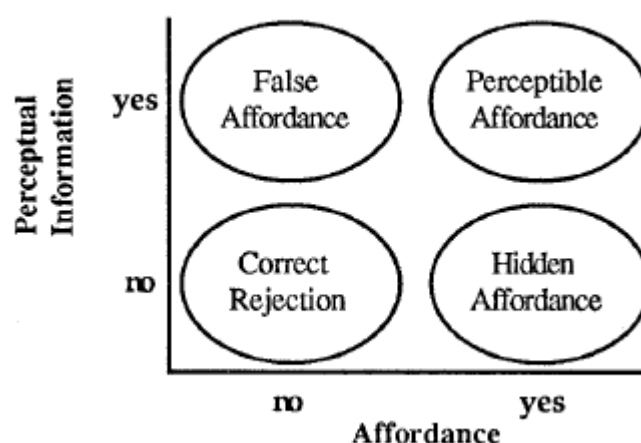


Figure 1. Affordances and perceptual information. From Gaver(1991)

It is important to note that Gaver (1991) supports the idea of perceptible affordances in other mediums than vision. An affordance could be mediated to the user using audio; however, using audio inherently imposes certain prob-

lems. Because audio input is naturally linear (Larsen et al., 2007), only one given affordance can be mediated at the time. In addition, since audio is ephemeral, mediating information using audio places a greater cognitive strain on users (Larsen et al., 2007), and the information is at greater risk of being missed and forgotten (Norman, 2010). This stressed the need when giving audio information to adhere to the general design principle of separating instructions into manageable chunks (Wickens et al., 2002).

Another problem using a natural interaction such as speech is of accidental activation (Norman, 2010). Because the interaction method is natural, the user might be using it for other means than interacting with the system, but for the system to be able to distinguish when the user is talking to the system and when the user is talking to a nearby person, is not easy (Yankelovich, 1997). Another type of accidental activation in speech interaction is incorrectly recognized speech, which also makes the system perform an action the user did not intend.

The biggest problem with accidental activation for non-graphical interfaces is the difficulty of providing feedback (Norman, 2010). As soon as the interaction takes a turn the user did not expect, it is very hard to for the user to understand why (Norman, 2010).

A problem specifically for speech interaction comes from that speech is a familiar, and almost by definition, *human* activity. This seems to lead users to, at least unconsciously, expect the same kind of dialog when speaking to a computer as to when speaking to a human (Yankelovich, 1997). This expectation is not only about format, but also about what kind of input the interaction can handle, that is, what kind of commands the user can use. Returning to the terminology of Gaven (See Figure 1), merely the associations of speech puts *false affordances* in the system, in that users expect and try to do what is not possible.

## SPEECH INTERFACE DESIGN PRINCIPLES

The literature provides a range of answers to the problems posed above, both in terms of general design principles, and some more directly practical problem solutions. I present these recommendations grouped under three topics: making the dialog natural, making possible actions perceptible, and leading the user through the interaction.

## MAKING THE DIALOG NATURAL

The most overarching principle recommended in the literature is to make the speech dialog *conversational* (Larsen et al., 2007; Yankelovich, 1997); however some authors debate to what extent (Rosenfeld et al., 2000). This suggestion provides one answer to the above stated problem of people expecting human-like dialog from speech interaction. The literature also very strongly suggests never translating graphical interfaces directly to speech interfaces (Larsen et al., 2007; Yankelovich, 1997), as graphical interfaces are based on the direct manipulation paradigm which speech interaction does not support (Larsen et al., 2007).

Yankelovich (1997) describes two separate ways to make the speech dialog natural. Firstly, the grammar of recognized input should be based on natural speech dialog. This author suggests using *natural dialog studies*, a kind of field studies of the language used in the context of use of the proposed application, for collecting this data. In practice, this comes down to the system supporting the language the users expect to use. Whereas it is important to allow for many different ways of saying the same thing (Yankelovich, 1997), the number of unique commands should preferably be kept low, as this helps keeping speech recognition accuracy up (Rosenfeld et al., 2000).

The second way to keep the dialog natural suggested by Yankelovich (1997) is to make the system cooperate with the user. Two practical examples of such behavior given by Yankelovich (1997) is for the system to try to make sense of speech references in user speech, and to shorten output that the system repeats more than once. In addition, she suggests that the user should be able to interrupt and correct the system.

## MAKING POSSIBLE ACTIONS PERCEPTIBLE

Unfortunately, even in making the dialog sensitive to the most natural user language, users will not know what actions are available, and most likely not how to perform them all either.

The easiest, but perhaps least elegant solution is to provide initial instructions when the dialog initiates (Rosenfeld et al., 2000), telling the user how to use the speech interaction.

A more elegant way of giving the same instructions to the user is to use what Rosenfeld and colleagues (2000) refers to as *lexical entrainment*. In the answers the speech interaction returns, the text is phrased in a way that instructs users how to proceed, using the same words that the user has to use in

their command. Ideally these instructions form a natural next step of the dialog, conforming to the cooperation principle, however they can also be more explicit instructions, of the kind "Say X, Y or C" (Yankelovich, 1997).

Finally, one way to get away of this whole problem is to just not use a speech only interface, but only using speech as input (Larsen et al., 2007). However, using different modalities between task input and output has been shown to be less efficient (Wickens & Liu, 1988)

#### *LEADING THE USER THROUGH THE INTERACTION*

The final set of principles the literature suggests for improving speech interaction is to in different ways help the user by letting them know what the system is doing, what they can and should do next.

The easiest help to give in relation to this is to let users know what is going on, by letting them know what their speech was recognized as (Larsen et al., 2007; Yankelovich, 1997). Naturally, if the application has a graphical interface, this information can easily be shown there, without much of a problem, however in a speech only application it gets more complicated. When dealing with sequential, choice-driven speech input, Yankelovich (1997) recommends repeating back what was recognized, to allow the user to make corrections. However, to keep with the cooperation principle, the system should quickly continue with the next step after repeating the information, to get going with what the user wants to do.

Another suggested principle is to keep track of the overall state of the speech dialog (Larsen et al., 2007; Yankelovich, 1997). By doing so, the system does not only increase its ability to handle referential speech input, but more importantly can give more contextual help and instructions to the user, when needed (Rosenfeld et al., 2000).

A final suggestion for speech interaction given in the literature is to try to interpret non-recognized speech input (Yankelovich, 1997). Even if the input was not recognized, given the context it might still make sense for the system to choose a very likely interpretation, especially if the effect of doing so can be easily reversible, (for example, interrupted), by the user (Yankelovich, 1997). In addition, an invalid recognition can trigger context sensitive help (Rosenfeld et al., 2000).

## Designing our speech interface

In this section I describe my efforts in the designing process in relation to the three design principles mentioned in the previous section: making our speech dialog natural, improving the perceptibility of our speech actions, and leading the user through the interaction. Even though this information was not all known to me during the process of the design, I refer back to the literature to point out where we, in hindsight, followed the recommendations made in the literature.

### CREATING A NATURAL CONVERSATION

From the first day that I was writing the dialog for the toy; his greeting, the clues he would give, and the way he would congratulate the user when he found something for him, I made sure to make the dialog sound as natural as possible. We had chosen a speech interaction to make use the knowledge kids already possess of this interaction method, so naturally we tried to stick to the natural mode of conversation. The literature also recommends this way of speech interaction (Larsen et al., 2007; Yankelovich, 1997).

Rather than requiring of the kids to learn specific command, or having the penguin give unnatural instructions such as do or say x, I always tried to accept the way a kid would naturally express an request, and would in the same way always try to have the penguin to respond naturally.

The biggest obstacle I faced to making the dialog natural was to avoid having the penguin sound robotic and repetitive. What I did was that I wrote a couple of ways of saying the same thing for every utterance the penguin would ever respond with, and then picked what response to use randomly. I found this especially important for the responses the penguin gives when the kid finds something, as this is kind of the goal of the interaction, where we reward the kid and also try to encourage them to continue to play.

However, I always drew the line for how far to take my efforts at trying to code in any kind of understanding of linguistics into the application, as I thought this would be too ambitious for the project, and could not really be justified in our time budget. This meant that whenever I was combining two sources of text into one spoken output the penguin would give, the first part would either have to have a known ending, or the last one a known beginning, in order to ensure that the combination of them both does not result in an ungrammatical phrase. One effect of this limitation is that all the clues the penguin ever gives start with "This is...".



## MAKING OUR SPEECH ACTIONS PERCEPTIBLE

The design aspect of our speech interface that I spent most time on during our project was making it understandable to our users. As we designed it for children to use on their own, the toy had to explain itself, or to use Gaven's (1991) terminology, have *perceptible affordances*. Naturally, because of the inherent ephemeral and cognitively taxing properties of audio (Larsen et al., 2007), this was a challenging task for me to handle.

My first attempts included expanding the grammar by allowing more and more accepted ways of issuing the same command, as recommended in the literature (Yankelovich, 1997). This however did not really solve the problem. The number of different ways in which people expressed the same action when given no guidance, for example when asking for a clue, were too many for our grammar to handle without losing speech recognition accuracy. After each time I tested the speech interaction on new users I fiddled with the grammar back and forth, adding one keyword and removing another, but after a while I realized I was not going to solve the problem this way. Using Gaven's (1991) framework the reason why this would never have worked becomes obvious: no matter how many more affordances for performing an action I added, or how well I made them fit the users natural language, they were still *hidden*, and that was the problem I needed to address.

I finally realized we had to give our users some kind of instructions on how to use the device. I took this as a loss, as I felt that explaining the interface because it was too complicated to *explain itself* was not in line with the core requirement of simple interaction that we had set up for the toy. I challenged myself to include the instructions inside the speech interaction in a natural way, so that the user would not feel like they were explicitly being told how to use it, but rather that these utterances were part of the experience.

The question was when to give the user these instructions. Giving them when the dialog starts (Rosenfeld et al., 2000) was never really an option, both because it did not fit with my requirement of a natural way of communicating the instructions, and because it would not have been practical. We did not want to have to reset the toy between each user, as this took too long, so we could not possibly know when a new interaction started. In addition, resorting to just explaining the interface in one piece would have resulted in a rather long explanation, going against the recommendations of chunking auditory instructions in smaller steps (Wickens et al., 2002).

The instructions therefor had to be separated into smaller steps, and be given at different times. The user needed to be given three pieces of information, namely the following:

- Explaining the purpose of the interaction: the penguin is looking for something he has lost and he wants the kid to help him.
- Instructions on how to ask the penguin for a clue of the thing to look for
- Instructions on how to ask the penguin if a found object is the one

After having received a suggestion from one tutor to have the penguin toy say thing by itself sometimes, I started giving a random line of instruction to the user after a certain period of no given recognized speech. I made the assumption that a user who was not producing any correct speech commands did not know what to say, and could therefor make use of instructions. This is technique is also mentioned in the literature (Yankelovich, 1997). To not make these unasked for instructions to intrusive, I made sure they only appeared after a longer period of silence, and that the user could interrupt these comments by issuing another command, also recommended in the literature (Yankelovic, 1997).

#### INDUCING HUMAN HELPFULNESS TO THE PENGUIN

This was a start, but not good enough. A lucky user could receive just the instruction they needed, but an unlucky one would instead only hear an already heard instruction repeated over and over again. As stated before, we never knew when the interaction with a new user started, so somehow iterating through the different parts of instruction was not an option either. We needed to know which instruction the user needed, to have the penguin offer just the right instruction, naturally, just as human participant would have done, obeying the conversational law of cooperation (Yankelovich, 1997).

To do this, I started to track how well the user seemed to understand the speech interaction, which is what the literature recommends (Larsen et al., 2007; Yankelovich, 1997). I started with two variables. The first one kept track on if the user had asked for any clues of the object the penguin was currently looking for. Until they did so, I would occasionally tell them how to ask for a clue, to get them going. The second variable kept track on how much time had passed since the user last asked the penguin whether an object was the one he was looking for. After the user had asked for one or more clues of an

object, if a longer period of time passed without them asking for if an object was what the penguin was looking for, I gave them an instruction on how to ask the penguin this.

The third instruction, the purpose of the interaction, that the penguin was looking for something thing, was not explicitly given in the same way as the other two pieces of instruction. This information was instead given whenever a user said hello to the penguin, or asked who he was, and was embedded throughout the dialog. The penguin would say things such as (“This is a water bottle – I am looking for something else”, or “When you think you have found what I am looking for, let me smell it!”). Smelling things was our final implementation of how to let the penguin tell users whether they had found the right thing, which he would automatically do when users lifted something close to his nose. The way the information of what the penguin wanted flowed through the interaction, I thought we were safe in regards to users understanding what the penguin wanted them to do.

In our final testing of the toy, however, I found that people still did not always understand the purpose of the interaction. Instead they would just let the penguin smell one thing after another. If they would have done this just to explore it would have been okay - I do not want to say that a device can only do what designers have imagined it to - but when asked these users gave no indication to having understood that the penguin was explicitly looking for something specific. This led me to start tracking one more thing. Whenever the user had let the penguin smell two or more things in a row that was not what the penguin was looking for, I would make him say something more explicit about what he wanted the user to do, for example to ask for another clue and stop just guessing what he was looking for.

## Evaluation

During an one-day exhibition on the Edge, a digital culture centre, in Brisbane, Australia, we collected questionnaire responses from users who had tested our penguin, in order to determine whether our final product met the success criteria we had define: simple interaction, full mobility, and fun to use.

Before presenting the results of the evaluation, a very important thing to note is that these responses were not collected from our user target group – children – and that the results might therefor not be generalizable. As our evaluation was opportunistic, in that we encouraged everyone who had used the penguin to fill in the questionnaire, this was the best evaluation we

could get. We still only got 30 responses, which only gives an indication of how well we succeeded. The results of the evaluation are presented below.

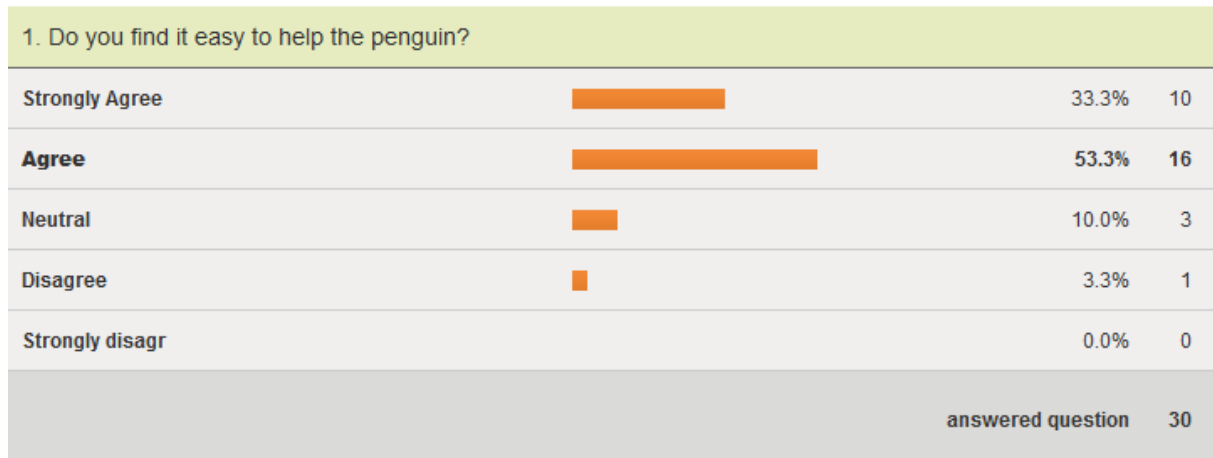


Figure 2. Simple interaction – questionnaire result

86% of our users agreed or strongly agreed that it was easy to help the penguin (find his lost things), and only one user, and only one person disagreed (See Figure 2).

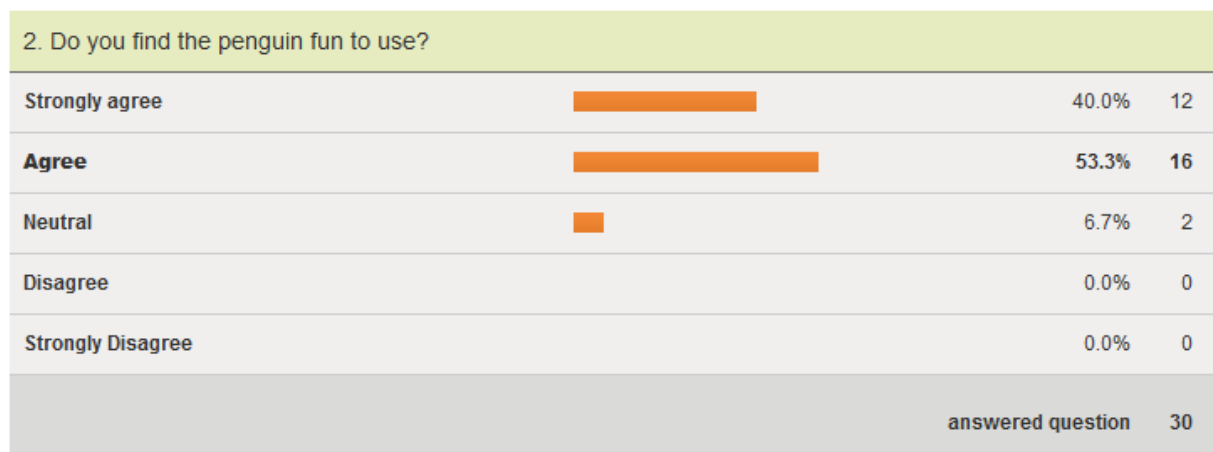


Figure 3. Fun to use – questionnaire result

93.3% of our users agreed or strongly agreed that they found the penguin fun to use, and no one disagreed (See Figure 3).






| 3. Did feel like the penguin limited your ability to move around? |   |              |           |
|---|---|--------------|-----------|
| Strongly agree  |  | 6.7%         | 2         |
| Agree   |  | 3.3%         | 1         |
| Neutral   |  | 26.7%        | 8         |
| <b>Disagree</b>   |  | <b>33.3%</b> | <b>10</b> |
| Strongly disagree   |  | 30.0%        | 9         |
| answered question   |   |              | 30        |

Figure 4. Mobility – questionnaire result

63.3% of our users disagreed or strongly disagreed to that the penguin limited their ability to move around. 26.7% percent of users gave a neutral answer, while 10% agreed or strongly agreed to the statement (See Figure 4).

## Conclusion

In this report I mainly covered how I worked on improving the speech interface to comply with our success criteria of the toy being simple to use, and this is what my reflection here will mainly focus on. I will I will also reflect on what lessons I learn in the process of creating this toy. First, however, I will reflect on the general success of our toy in regards to all our three success criteria – simplicity, mobility, fun.

As can be seen in the previous section, the results of our evaluation points to that, if anything, we met our success criteria (See Figure 2-4). I would like to make two points here. As stated before, the respondents in this survey were not at all representative of our user group, and it is therefore questionable whether these results generalizes to kids.

Secondly, ignoring the low validity of this survey, the results regarding mobility being less confirming than the results to the other questions looks really suspicious to me. Our toy is fully functional when moved within a range of at least 75m (~250 feet) away from a base station, which completely satisfies my criteria for mobility for this kind of toy for children. I suspect that users' responses might be due to that we never showed them this mobility, but instead only had them test our toy within our test booth in the exhibition, which was not very large. In fact, during the later parts of the evening, our booth got so crowded that our users could hardly move around at all. It is also possible that

we expressed this question (the only reverse-phrased question) influenced the answers to this question.

Speaking of my personal opinion of the outcome of this project, specifically in regards to my design efforts with the speech interaction, I would say that what I created is in some ways a success, in some ways a failure. It is successfully simple in that anyone who pays attention to what the penguin says will soon enough understand how to handle the interaction, which is what the project came to be about for me. On the other hand, this was not at all the kind of simple I imagined at the start of this project. I thought more in terms of being self-explanatory, or as I have now repeated many times, having perceptible affordances.

In hindsight, I think I have to accept that I was naïve and did not fully consider the practical limits of speech interaction when we started the project. I have now gotten a much better practical understanding these things. On another note, it is very interesting to me that even though I never did the kind of research covered in this report before the project, the design decisions I took still almost completely agrees with the recommendations made in the literature. This might partially be because of some kind of bias when I was looking for articles related to speech interaction, but to some extent I also hope and believe that my past experience in design, and my background in psychology, got me here. If anything, I feel stupid for not thinking about these things earlier.

The way I seem to have reinvented the wheel during my design process in this project might serve to teach me a lesson. Surely it would have been more time efficient if I would have done this research at first, to inform my design. One thing I can for this ignorance of research is that we all think we know what 'speech' is – we think of human speech. We think of conversation where both parts have an understanding of what is going on, and cooperate to reach wherever the conversation is heading. To replicate anything like this behavior in a computer system is anything but easy, but for it is worth, I cut my costs I did the best I could in this project.

Something I really wonder is how we came to conclude that speech interaction was a simple kind of interaction, suitable for kids. Surely, the false associations mentioned above are part of the explanation, but still - merely the load on working memory that auditory stimuli puts on listeners suggests it to be a terrible fit for delivering complex instructions to kids. We probably never conceived that our toy would need complex instructions. Not to sound to

negative however, the user reviews we got of our toy were better than anyone could have expected. We might just have been lucky enough to hit the perfect balance between requiring enough attention to provide an engaging experience for users, and requiring too much to make it unpleasant. Figure 5 depicts a kid concentrating hard on listening to the clue the penguin is giving.



Figure 5. User, holding the toy penguin, listening carefully

## References

Gaver, W. W. (1991). Technology affordances. *Proceedings of the SIGCHI conference on Human factors in computing systems: Reaching through technology (CHI '91)*

Larsen, L. B., Jensen, K. L., Larsen, S. & Rasmussen & M. H. (2007). Affordance in mobile speech-based user interaction. *Proceedings of the 9th international Conference on Human Computer interaction with Mobile Devices and Services.*

Norman, D. (2010) Natural user interfaces are not natural, *ACM Interactions*, May/June.

Rosenfeld, R., Zhu, X., Toth, A., Shriver, S., Lenzo, K. & Black, A. (2000). Towards a universal speech interface. In *Proceedings of ISCLP*, Beijing, China.

Wickens, C. & Liu, Y. (1988). Codes and modalities in multiple resources: a success and a qualification. *Human Factors: The Journal of the Human Factors and Ergonomics Society*

Wickens, C., Gordon S. & Liu Y. (2002). *An introduction to human factor engineering*. New York: Harper Collins

Yankelovich, N. (1997). Using natural dialogs as the basis for speech interface design. *Automated Spoken Dialog Systems*. Cambridge, MA: MIT Press.