

Prosodic Features for Speech User Interfaces

Jaakko Hakulinen

Human-Computer Interaction Group
Department of Computer Science
University of Tampere
FIN-33101 Tampere, Finland
+358-3-2158558
jh@cs.uta.fi

Markku Turunen

Human-Computer Interaction Group
Department of Computer Science
University of Tampere
FIN-33101 Tampere, Finland
+358-3-2158559
mturunen@cs.uta.fi

ABSTRACT

The designing of system utterances is a very crucial part of speech user interfaces, especially if speech synthesis is used. Although speech synthesis is quite intelligible in well formed and simple sentences it is very difficult for users to understand when complex structural elements like tables are spoken. Furthermore, most users do not like the way synthesizers use prosody. Most of the previous research has focused on what information should be presented to the user. Recent research has also brought up the question of how this information should be presented. In order to improve intelligibility and naturalness of synthetic speech we arranged an experiment to find new ways to use prosody.

In our experiment subjects listened to three human readers and a speech synthesizer reading system utterances from our e-mail system. Questions were asked to measure how well the utterances were understood. Subjective evaluations of the voices were also collected. We used the results to find those prosodic elements that help users to better understand what they are hearing. Pauses were found to make a significant difference in comprehension. Good variation in pitch and speed seem to make a voice more pleasant to listen to but have only minor positive effect on comprehension. We will transfer these elements to our e-mail system taking advantage of its conceptual information presentation capabilities.

Keywords

Speech user interfaces, prosody, dialogue management, spoken language systems, synthetic speech, e-mail

INTRODUCTION

Speech output is widely used in many computer applications today. Telephony applications, mainly interactive voice response systems (IVR's), have been very successful and today we use them as a part of our daily life. The expansion of multimedia has brought speech also to desktop applications like games and educational programs. These applications use mainly real speech recorded by professional speakers. In most cases the use of stored voice is a very sophisticated and efficient way to express information.

However, it is not always possible to use prerecorded speech. For example, it is impossible to build large-scale query systems like library information systems using only prerecorded prompts. An alternative way to use speech in human-computer interaction is the use of synthetic speech.

It is commonly argued that although current speech synthesizers can produce very understandable sentences, most people do not like the way in which they are expressed. In general people feel that synthetic speech is unnatural and they do not like to listen to it if not necessary. This restricts the usefulness of synthetic speech in many areas.

Synthetic speech sounds very monotonous when compared to normal human speech. This is one of the most frequent complaints against synthetic speech. One of the main reasons for this is the fact that synthesizers lack prosodic information, which makes human speech sound lively. In everyday communication we all make heavy use of prosodic elements like pitch and volume to accent what we say.

Prosodic information conveys also information that cannot be obtained from anywhere else. For example, when we change emphasis from one word to another, it is possible that the meaning of a sentence changes dramatically. This is especially true in complex sentences where misinterpretations are likely if adequate prosodic elements are not present. In spoken language systems various elements like lists, tables and addresses are often presented to the user. These elements tend to be very problematic, even in human-to-human communication.

Prosodic features in human speech

The most important prosodic features found in human speech are *pitch*, *volume*, *speed* and *pauses*. Current speech synthesizers allow decent control of those parameters. Therefore these prosodic features could be utilized in user interfaces. Next we describe in detail what we mean by prosodic features.

Pitch (or fundamental frequency from the viewpoint of sound production) is a basic method to introduce new topics and emphasize important sections in a spoken language. By varying pitch we could also add more liveliness to the synthesizer's output, but this should be done care-

fully since incorrect use or overuse makes spoken output incoherent. Since everyday communication skills have influenced the way we use prosody, it is important to follow common practices in human-computer communication.

Volume could be used much like pitch to emphasize certain important sections in speech. However, volume is not as flexible, since it is impossible to lower or raise volume very much.

Speed of speech (measured usually by words per minute) is a very crucial factor in spoken language systems. Slow speed makes speech easier to understand but could be very annoying since listeners cannot usually skip sections. On the other hand, too fast speed could make it impossible to understand speech. The speed of speech also affects how much of the message is recalled. However, experience increases the ability to listen to the faster speech.

In addition to pitch, pauses are also a very important factor in the use of prosody. Pauses could be used to emphasize essential issues in the same way as pitch and volume. Pauses affect also the overall rhythm of speech along with speed. In general, people use pauses heavily in human-to-human communication. When used in human-computer interaction, they could be problematic since pauses are ambiguous (they suggest turn takings, user may feel that computer has failure etc.) – this feature is very crucial in dialogue management.

Because prosodic information is such an important factor for successful speech user interfaces it seems natural to support prosody in as low level as possible. If we add prosodic features to speech synthesizer engines, prosodic elements are always available and they are used automatically. Unfortunately, it is impossible to add meaningful prosodic information to unconstrained sentences because this leads us to an unsolved problem of natural language understanding. That is why we should bring prosodic information to the application level. In this paper we will introduce how we added support for prosodic features to our spoken language system.

To utilize prosodic features on the application level we could use existing guidelines like the ones described by Dobroth [1]. Unfortunately these commonly known rules are not enough for complex situations. It is reasonable to believe that we still have a lot to learn from human-to-human communication.

In order to find new ways to use prosody we arranged an experiment in which human speakers recorded a set of utterances. A group of listeners heard those utterances and answered questions about them. We found that some prosodic features seemed to increase intelligibility of speech while others made speech more pleasant to listen. We believe that by bringing these features to synthetic speech we could increase both its intelligibility and pleasantness.

In the rest of this paper we first will introduce how prosody could be supported in speech applications. Second we describe the experiment and its results. Next, conclusions from the experiment are drawn and ideas for future work are presented.

SUPPORTING PROSODY IN SPOKEN LANGUAGE APPLICATIONS

We have been building telephone based speech user interface for electronic mail. E-mail has been a very popular application for speech user interface research since 1980's [6, 4]. E-mail has many features which make it both suitable and challenging for research purposes. First of all, telephone based interaction is a very natural way for most of us. Secondly, the richness and broad scale of e-mail messages provides a great challenge for expressing information.

Previous research has focused mainly on prompt design [7], dialogue management issues like navigation in lists and menus [5] and on dialogue management strategies [8]. In general, most of the previous research in speech user interfaces has focused either on input issues (i.e. speech recognition) or in the case of output on *what* information should be expressed in system utterances. However, there is not much research done about *how* this information should be expressed.

We wanted to examine how system utterances could be expressed more efficiently by using prosodic features. Next we explain how prosodic support works in our system.

In our e-mail application there are three kind of utterances expressed to the user: system utterances that are part of dialogue management, views of messages and the messages themselves. The first case is the easiest one because we know in advance what these utterances are. In the second case things get more complicated since we have to deal with information that is not known in advance. However, the structure of information is still fixed. The third case is the most challenging one since information is totally unconstrained.

To improve the way in which speech output is expressed by using prosodic features we could add control codes in messages when messages or their structure are known in advance. In this way messages could be fine-tuned by hand. However, this approach is not possible in all cases. Furthermore, expression style should not be static in order to be natural and efficient. Instead it should be dynamically constructed and based on current context and dialog history. This in turn means that information should be represented in conceptual form as opposite to static string-based representation.

Figure 1 presents the way in which our system handles input and output messages. The dialogue management component receives input events from the input handler and produces conceptual output messages. Output mes-

sages are then sent to the person manager. Based on the message's category, the current state of the dialog and user preferences, a part of the system called *person manager* selects one of the many *persons* ('readers') available and directs the message to that person. It is the selected person's duty to interpret the message, form a sentence level expression and add the necessary prosodic features to the message. Finally the message is sent to the output handler which in turn presents it to the user.

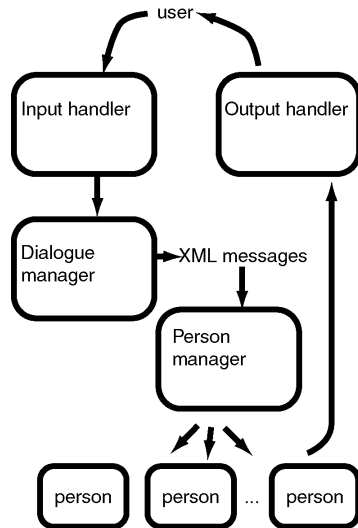


Figure 1. System structure.

All information flow in our system is based on XML [2] messages. Messages form a hierarchy of XML classes that are fine-tuned for special purposes. Typically one person is specialized to handle only a few types of messages. For example, mail readers are good at dealing with mail messages but don't know how to process error messages. For error messages we have a different set of persons.

Every person has a set of attributes, which describes its personality and behavior. Attributes are important when choosing a right person from a group of persons, which have the same abilities to process messages. For example, if a user prefers to get brief explanations or this is beneficial for other reasons the person manager selects a person who produces brief output messages. Selecting the right persons is not a trivial task and it is a very interesting research issue when we implement more persons for handling the same messages.

Persons produce the prosodic information found in the utterances. Since they are aware of the situation in which the message appears and have access to knowledge about past events and dialogue flow, they could produce highly customized and context sensitive utterances with rich prosodic features. Even in the case of unconstrained mail messages some structured elements like lists, addresses and tables can be found and prosodic cues added.

Since existing rules and guidelines are not enough to produce efficient and natural prosodic elements for speech synthesis we arranged an experiment to examine how human speakers use prosody. Based on the results of this experiment we hope that we could find the most suitable methods to be used by the persons in our system.

EXPERIMENT

We arranged an experiment where subjects listened some short utterances containing various interesting structural elements. The subjects then answered some questions about the content of the utterances to see if the message was understood. At the end subjects were also asked to answer some questions where they considered how good the reading voice was.

Material

We assembled a set of fifteen utterances for our experiment. These utterances were fictional system utterances from our e-mail system. The utterances formed an entire session with the system. The first utterances were login messages telling the user about the system he/she had logged into and the mailbox that was opened. In later utterances the user was presented a list and descriptions of folders into which the mail messages are organised. Later there were utterances that listed mails in those folders. Some mails were then read to the user.

Three human readers and a synthesiser each read all fifteen utterances. The synthesiser that we are using in our e-mail system is Infovox 230 [3] using the "Finnish Male" voice. Because the synthesiser's voice is male we used only male readers. The three human readers had some differences in their background with using their voice. One reader does not use his voice much in his work. In this paper he is referred as the non-professional speaker. The second reader does lecturing at the University of Tampere. The third reader is studying radio work in the same university and has some experience working as a radio jockey.

We gave strict instructions for readers on how to read the utterances, so that for example dates were read exactly the same way. Only in one place we gave the reader some freedom to choose how to read the material. This was a table in one utterance. Readers were free to express the table seen in Figure 2 in any order they wanted. They were not allowed to add any information to or interpret it when reading, but they could read elements in any order they wanted and read any element several times. The synthesiser read the table in the normal, left to right, top down reading order. All human readers read the table by first reading both row and column headings for data and then the data.

From the viewpoint of prosodic elements the other interesting elements in the utterances were dates and times, a web-address, and a telephone number. Most of the words were in Finnish as we are developing our e-mail system in

Finnish. Still, there were some short phrases in English and also some English names. These were read in English as we are building multilingual support for our system.

	January	February
Production	30000	20000
Orders	50000	56000

Figure 2. The table used in the experiment.

Subjects

We used sixteen subjects to listen to the utterances. Each subject listened to one reader. Therefore each reader got four listeners. The actual amount of subjects we used was eighteen. One was used in a pilot test, which gave us some ideas on how to improve the test. One of the actual tests had to be discarded because of some serious problems with the test situation. From the sixteen subjects in the final tests, nine were men and seven women. Our only requirement for subjects was that they had to be native Finnish speakers.

Experiment setup

Experiments took place in two working rooms. The subjects were seated next to a table. The experimenter explained them what was about to happen. The pile of papers, containing questions about the utterances, was placed on the table and the subject was given a pencil and an eraser. The experimenter gave a headset to the subject. Some test utterances were played to the subject through the headphones so that he/she could set the volume level.

There were two questions about each utterance. The test subject was allowed to read the first question before he/she heard the utterance. After answering the first question, the subject was allowed to see the second question about the utterance just heard and to answer it. After answering that question, the subject read the first question about the next utterance. Then the subject signaled the experimenter that he/she was ready to listen to the next utterance. One utterance exceptionally did not have the second question because the utterance did not have enough information for two questions and answering the first question took a long time.

Each question was on its own piece of paper and all papers were in a pile or file in front of the subject so that turning the page revealed the next question. It was the subject's task to turn the pages. During the experiments, the experimenter's only task was to play the next utterance, when the subject signaled that he/she was ready for it.

Each paper with a question had also another question asking if the subject had problems answering the first question. If user had had problems then she/he could describe what kind of problems occurred (did not hear, did not remember etc.).

There were different types of questions about the utterances. Some asked specific details and others required more understanding on what was heard.

After all questions about the utterances were answered, there was one more paper for the test subjects to fill in. This questionnaire asked how the reader's voice was like, giving us speech profiles. There were scales asking for such parameters as pitch, speed and volume and their variations. Scales ranged from too little to too much. The amount and the length of pauses and overall emphasis were scaled. At the end of the paper there were two open questions. The first question asking if listening to the speech required concentration and the second asking about possible problems with the voice.

RESULTS

We got two kinds of data from the experiment. First data set consisted of the answers to the questions about the content of the utterances. Second set of data contained speaker profiles. Latter one corresponded to the questions about the readers' voices.

As we had 16 subjects, each of whom answered to 29 questions, we got a total of 464 answers to the questions about the content. As all questions were open, it was not appropriate to require exactly correct answers to all questions. For example some names, especially foreign ones are impossible to spell correctly without good luck. Answers may also be partially correct, for example the table may have the correct form but the subject may have forgotten some exact digits.

18 of the 29 answers could be analyzed using numerical rightness scale. We gave each of these 18 answers a score between zero and one, based on the accuracy of the answer. For the rest, it is more important to look at the actual answers rather than their correctness.

Next we explain the results from the eighteen scored answers, then from the remaining questions and finally we introduce the speaker profiles.

Numerical results

The eighteen scored questions were divided into two groups: those that were presented *before* the spoken utterances were heard and those presented *after* hearing the utterances. There were 9 of both types of questions. As one might expect, the questions in the first group ("before") had much better answers. There were also differences between readers and listeners.

Table 1 gives a summary of answers from the viewpoint of readers. All values are average scores. As Table 1 shows, there is only a little difference between readers in questions that are presented before utterances was heard. The only difference was that listeners answered slightly better when the reader was the lecturer.

By comparing questions that were presented after the reading of utterances, much higher variations could be

found. The listeners got only one question right out of four when the non-professional reader's utterances were presented. The synthesizer's and the radio reader's utterances were understood and remembered much better and answers to questions that the lecturer presented were much higher still.

Figure 3 presents another view of same results. It shows how answers differ from the average for each question class. As it can be seen, answers to the lecturer's utterances are much more accurate than the rest. It also shows that the listeners had the most difficulties to answer questions about the sentences spoken by the non-professional reader.

	Before	After	Total
Synthesis	3,60	2,17	5,77
Non-pro	3,60	1,06	4,66
Lecturer	3,94	2,94	6,89
Radio	3,61	2,14	5,75
Total	3,69	2,08	5,77
Max	4	4	8

Table 1. Numerical results.

We analyzed also individual differences between the sixteen listeners. Results fit surprisingly well to groups formed by readers. Only one anomaly was found. We plan to arrange a replacement experiment since this anomaly is huge and it effects the results about the radio reader. If this anomaly is removed, answers to the radio reader's utterances are more like answers to the lecturer's utterances.

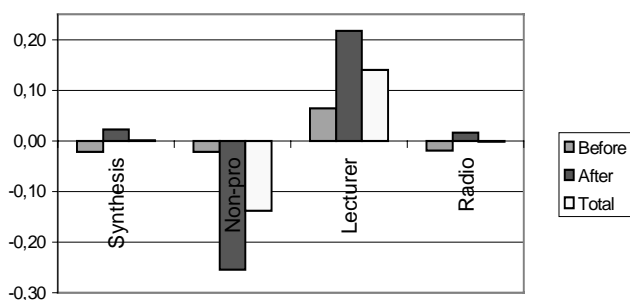


Figure 3. Numerical results (b).

Other results

In addition to the scored answers, we found other interesting issues from the answers. First, there exists a great difference in answers when a simple table is expressed. With all other sentences, we gave strict instructions to readers about how to read them. With the table, the readers were able to choose how to read it. However, since it turned out that all human readers used exactly the same words, all human readers' presentations are consistent.

Since we asked the listeners to reproduce the table from what they heard, the results varied much. Although it is not possible to evaluate the results the same way as we did with most of the questions, the results showed that answers to sentences spoken by the lecturer are the most accurate and complete. Again, results to the non-professional speaker's expression were the worst.

We asked the listeners to pick up some names from the spoken utterances. Some of these names were Finnish and some were foreign. As expected, the names were often transformed when reproduced by the listeners. Especially foreign names were difficult, although we hand-tuned the synthesizer's expressions. A similar effect occurred when we asked the listeners to pick up some abbreviations.

The third notable detail occurred when an Internet address was spoken and a telephone number followed it. In answers to the non-professional reader's sentences some listeners concatenated these together. It shows the importance of the correct lengths of pauses between sentences. It is also interesting that with this question listeners had the least difficulties with the utterances spoken by the synthesizer.

Speaker profiles

We also gathered speaker profiles for each speaker. A profile consisted of ten scales indicating problems with different aspects of speech and a few questions for possible open comments. The scales ranged from too little to too much and an optimal value in the middle. For example with speed of speech left side of the scale meant too slow speed and right side too fast. It is possible, that different subjects could give inconsistent comments, for example one saying speed was too slow and other that it was too fast. Of total 40 values (10 values per subject), there were eight such cases. There was no inconsistency within the non-professional speaker's profiles, one with the lecturer, three with the radio voice and four with the synthesiser.

When summing the absolute problem values for each speaker, the radio voice has clearly the lowest value of 17. Surprisingly, the synthesiser got into the second place with 24. The third place goes to the lecturer with the score of 25 and the non-professional speaker has the biggest problem score of 25,5. The order is similar, no matter how the inconsistent values are handled. Letting positive and negative values cancel each other, the radio voice gets 11, the synthesiser 17, the lecturer 23 and the non-professional 25,5. Here the differences are somewhat bigger. Also, we get a similar order, if we only look at the last scale in the profiles, which is the overall feeling about the speaker's use of prosody. The radio voice is clearly the best with value -1 and the synthesiser is second with -3. The non-professional speaker gets the third place with value -5 and a small margin before the lecturer who has -6. For this last scale, every speaker got values telling that their speech is too monotonic. Absolute values of these

numbers are visualised in Figure 4. The lack of intonation can also be seen in the open comments. With the exception of the radio voice, everybody got negative comments about too monotonic speech. However, there was one subject who said that the synthetic voice should be even more monotonic: then it would be easier to understand, but also more boring.

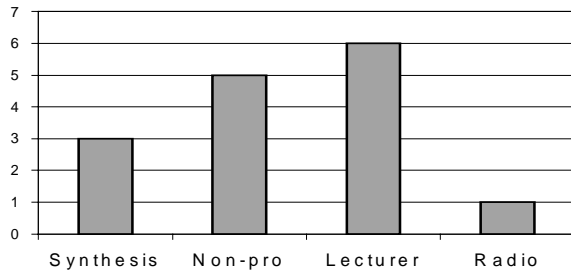


Figure 4: Overall scores from speaker profiles.

Unclear voice and missing phonemes were reported for the non-professional speaker and surprisingly also for the synthesiser. Also, not so surprisingly, problems with foreign names were noted on several comments for all readers.

Parameter	Synth		Non-pro		Lecturer		Radio	
	Min	Max	Min	Max	min	Max	min	max
Pitch	-1,5	0,0	-2,0	0,0	-2,0	0,0	0,0	0,0
Pitch variation	-2,5	1,0	-3,5	0,0	-2,5	0,0	-1,0	1,0
Use of volume	0,0	0,0	0,0	0,0	-1,0	0,0	0,0	1,0
Volume variation	-0,5	2,0	-0,5	0,0	-1,5	0,0	-1,0	0,0
Speed	0,0	2,5	0,0	1,5	-2,0	0,0	0,0	2,0
Speed variation	-2,5	1,0	0,0	1,0	-2,0	0,0	0,0	0,0
Pauses (amount)	-1,5	0,0	-4,5	0,0	0,0	1,0	-4,0	0,0
Length of pauses	-2,5	0,0	-4,5	0,0	-1,0	2,0	-2,0	1,0
Emphasis	-2,5	1,0	-3,0	0,0	-4,0	0,0	-2,0	1,0
Overall amount	-3,0	0,0	-5,0	0,0	-6,0	0,0	-1,0	0,0

Table 2. Speaker profiles.

Table 2 lists the problems of the voices from the speech profiles. For each voice there are two values, the sum of negative and positive values given for each voice parameter in the questionnaire. The optimal value for each parameter is zero, negative values mean that there is too little or too low amount of the feature, for example too low pitch or too little pitch variation or too slow speed or too short pauses. Positive values are also problems in the opposite side like too many pauses, or too much speed variations. If there was inconsistent inputs between subjects, i.e. both positive and negative values were given, then both Min and Max columns have values other than zero.

The interesting parts in Table 2 include both rows describing the use of pauses. We can see that the lecturer had good values there, just a slightly too many pauses and some inconsistency in the length of pauses. However all others were told to have too few and too short. Also we can notice that the radio voice scored well on almost all other scales. His use of pitch, volume and speed was liked more than any other one's. Everybody else was considered to be too monotonic.

The overall view seems to be that the lecturer had too monotonic and slow style of speaking. However, he is the only one who had pauses that are long enough, in some cases even too long. The radio voice seemed to be most liked, the other people were considered too monotonic. The synthesiser actually got rather good scores, far better than the non-professional speaker. However, according to these results, it seems that people would like the synthesiser more if we would add some more and longer pauses to it and try to get a bit more emphasis to its voice like the radio voice had in it.

DISCUSSION

First of all, there exist some notable differences between the speakers in the numerical results. As Table 1 and Figure 3 show, answers to questions about utterances spoken by the lecturer are most accurate. It could also be clearly seen that listeners had most difficulties in answering questions about the non-professional speaker's utterances.

There exist also great differences in speaker profiles. As it could be seen from Table 2, the problem score of the radio voice is clearly the lowest. Other speakers have similar overall scores but there exist variations between individual factors. The only factor where the radio voice's value is not excellent is the amount of pauses.

As one could guess from the results of the speaker profiles, expressions of individual speakers' differ a lot from one another. Since we haven't yet analyzed them formally, only general assumptions based on hearing experiences can be presented here. First of all, the lecturer used a lot of pauses. On the contrary, the non-professional speaker used only a minimal amount of pauses. The other two speakers (synthesizer and radio voice) fit somewhere between these. The radio voice sounded most lively, although it was not the clearest one.

When we compare the speaker profiles to numerical results it seems likely that pauses are an important factor of intelligibility. Since the lecture's scores are very high on both "before" and "after" questions and because he used pauses more than others did, it seems likely that pauses help listeners to interpret what they hear. We find more support to this phenomenon as we compare the non-professional speaker's use of pauses and his numerical results. We find a very similar but negative effect. In order to get clearer picture about this we are planning formal

studies where we will use synthesizer to read same utterances both with and without additional pauses.

The main reasons for wrong answers are forgetting, difficulties to understand the heard words and difficulties to understand whole spoken utterances. The latter two correspond to segmental intelligibility and comprehension. Since most of the “before” questions do not need memorizing or comprehension we could say that wrong answers in the “before” questions are because of poor segmental intelligibility of speech.

The case of the “after” questions is more complex. Since answers to these questions needed usually both memorizing and understanding of the content of the message it is impossible to say exactly what caused errors. Still, it is obvious that the “after” questions needed more comprehension than the “before” questions.

When “before” and “after” results are examined we find that difference in accuracy was more dramatic in the “after” questions. This might influence that besides affecting segmental intelligibility, pausing could have an even stronger effect on comprehension. We are planning further studies to examine this issue.

However, in spite of being the most intelligible, the lecturer’s speech was also described to be very monotonous. Instead, the listeners liked most the radio person’s voice. In general, the listeners did not complain about any other problems with him except his use of pauses. It is a surprise that listeners did not answer the questions better when he was the speaker. It might be that his use of pauses cancelled out the positive effect from the use of pitch. It might also be that even though the use of pitch is very pleasant, it does not bring much intelligibility to speech.

Finally, we were surprised about how well synthetic speech managed. It was at least average in every respect and its problem score was the second lowest. This was the most unexpected result. It might be that listeners judged the synthesizer more lightly than human speakers.

FUTURE WORK

We will continue our research by analysing the utterances the speakers have read. This analysis will include pitch extraction and energy extraction so that it is possible to make analyses about the use of pitch and volume. Also word boundaries will be marked. This gives us possibilities to find and measure pauses and their lengths and also the speed of speech in different utterances and even in single words. As such analysis will give us large amounts of data, these operations will be done only to interesting utterances. From this data we should be able to gather knowledge about how the speakers used prosodic elements. Also, as we will analyse the synthesiser output, we should see differences between humans and the synthesiser and find situations where the synthesiser is missing

prosodic elements. We should also be able to do some comparisons with the results of this research.

CONCLUSIONS

Intelligibility and pleasantness of a message vary a lot depending on how it is spoken. This is especially true when complex structural elements are presented. By using prosody we could improve greatly both intelligibility and naturalness of speech output. However, we need to know more about how prosody could be utilized in human-computer interaction. We believe that we could borrow a lot from professional human speakers. Furthermore, speech applications should be built in a way that makes it possible to use prosodic features efficiently.

Our listening experiment has given us information about the effects of prosodic elements in speech. Pauses seem to be able to significantly improve understandability of speech. In our experiment, the use of pitch and volume had positive effect on pleasantness of speech but only little effect on comprehension. We are about to do careful work to find out exactly how these prosodic elements can be used with speech synthesis in our e-mail system to make it easier to understand.

REFERENCES

1. Dobroth, K. It’s both what you say and how you say it: The role of prosody in effective prompt design. *In Proceedings of AVIOS ’98 The 17th Annual International Voice Technologies Application Conference*, 1998: 213-220.
2. Extensible Markup Language (XML) 1.0. W3C Recommendation 10-February-1998. Editors: Bray, T., Paoli, J., Sperberg-McQueen, C.
3. Infovox 230 Text-to-Speech system version 1.1. Telia Promotor.
[<http://www.promotor.telia.se/infovox/230.htm>]
4. Ly, E. Chatter: A Conversational Telephone Agent. MIT Master’s Thesis, Program in Media Arts and Sciences, 1993.
5. Marx, M., Schmandt, C. MailCall: Message Presentation and Navigation in a Nonvisual Environment. *In Proceedings of ACM CHI 96 Conference on Human Factors in Computing Systems*, 1996: 165-172.
6. Schmandt, C. Speech Synthesis Gives Voiced Access to an Electronic Mail System. *Speech Technology*, August/September 1982: 66-68.
7. Yankelovich, N. How Do Users Know What to Say? *ACM Interactions*, Volume 3, Number 6, November/December 1996.
8. Walker, M., Fromer, J., Fabrizio, G., Mestel, C., Hindle, D. What can I say?: Evaluating a spoken language interface to Email. *In Proceedings of ACM CHI 98 Conference on Human Factors in Computing Systems*, 1998: 582-589.