

THE USE OF PROSODIC FEATURES TO HELP USERS EXTRACT INFORMATION FROM STRUCTURED ELEMENTS IN SPOKEN DIALOGUE SYSTEMS

Jaakko Hakulinen, Markku Turunen, and Kari-Jouko Rähkä

Human-Computer Interaction Group, Department of Computer Science, University of Tampere,
P.O. Box 607, FIN-33101 Tampere, Finland.

Tel. +358 3 2156952, FAX +358 3 2158557, E-mail: {jh, mturunen, kjr}@cs.uta.fi

ABSTRACT

Most of the previous research on speech user interfaces has focused on what information should be presented to the user. Equally important is the question of *how* this information should be presented. Although speech synthesis is quite intelligible in well-formed and simple sentences, it may be very difficult to understand when complex structural elements, like tables or URLs, are spoken. We arranged a controlled experiment to identify the prosodic features that affect the intelligibility and pleasantness of synthetic speech. Pauses were found to make a significant difference in comprehension. Good variation in pitch and rate seem to make a voice more pleasant to listen to but have only minor positive effect on comprehension. We analyzed the exact ways in which human readers used prosodic elements so that we could construct unique and human like computer 'persons' for spoken dialogue applications.

1. INTRODUCTION

Speech output is widely used in many computer applications. Telephone applications, mainly interactive voice response systems (IVRs), have been very successful. These applications mainly use real speech recorded by professional speakers.

However, it is not always possible to use prerecorded speech. An alternative is to use synthetic speech. However, it is often argued that although current speech synthesizers can produce very understandable sentences, most people do not like the way in which they are expressed.

Indeed, synthetic speech sounds very monotonous when compared to normal human speech. Synthesized speech usually lacks the prosodic features that make human speech sound lively. In everyday communication we all make considerable use of prosodic elements like pitch and volume for emphasis.

Prosodic information also conveys information that cannot be obtained in any other way. For example, when we change emphasis from one word to another, it is possible that the meaning of a sentence changes dramatically. Furthermore, if complex elements like lists, tables and addresses are produced using speech synthesis, the use of prosody is essential. Such verbal representations may be very hard to understand, even in human-to-human communication.

The most important prosodic features found in human speech are *pitch*, *volume*, *rate* and *pauses*. Current speech synthesizers allow reasonable control of these parameters. Therefore these

prosodic features could be utilized in speech user interfaces [1]. In our case, the motivation was a speech interface (in Finnish) to an e-mail client that we have been developing.

In order to find new ways to use prosody we arranged an experiment in which human speakers recorded a set of utterances. The same sentences were also produced using a speech synthesizer. 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 to. We analyzed the exact ways in which human readers used their voices. We believe that by bringing these methods to synthetic speech we could increase both its intelligibility and pleasantness.

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

2. SUPPORTING PROSODY IN SPOKEN LANGUAGE APPLICATIONS

Previous research in speech user interfaces has focused mainly on prompt design [6], dialogue management issues like navigation in lists and menus [4] and on dialogue management strategies [7]. In general, most of the previous research on speech output has studied *what* information should be expressed in system utterances. We wanted to examine *how* system utterances could be expressed more efficiently by adding prosodic features.

In our e-mail application, three kinds of utterances are spoken to the user: system utterances that are part of dialogue management, descriptions of sets of messages ("views"), 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, since information is totally unconstrained.

To improve the quality of speech output 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, in order to be natural and efficient the style of speech should not be static. Instead it

should be dynamically constructed and based on current context and dialogue history.

In our system, the utterances are produced by a collection of software modules called “persons”. Every person has a set of attributes describing its personality and behavior. The user can select the person with the properties that s/he prefers. For example, if a user prefers to get brief explanations, s/he (or the system) can select a person who produces brief output messages.

Persons produce the prosodic information found in the utterances. Since they are aware of the context 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.

Prosodic features found in current speech synthesizers are not very sophisticated. A lot of research work has been done and several concept-to-speech synthesizers with added prosodic elements have been introduced [3, 5]. These systems usually want to bring prosodic elements to be part of the synthesizer at a low level. We are interested in bringing prosodic elements to application level and especially in finding ways to use prosody in complex structural elements. We also want to construct unique, human kind ‘persons’ with different prosodic features.

3. EXPERIMENT

We arranged an experiment where the subjects listened to short utterances containing various structural elements. The subjects then answered questions about the content of the utterances. The answers revealed whether the messages were understood or not. At the end of the experiment the subjects were also asked to give their personal opinions regarding the quality of the reading voice that they had listened to.

3.1 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 s/he had logged into and the mailbox that had been opened. In later utterances the user was presented with a list describing the folders into which mail messages were organized. These were followed by utterances that listed some mail messages in those folders. Finally, some mail messages were read to the user.

Three human readers and a synthesizer each read all fifteen utterances. The synthesizer that we are using in our e-mail system is Infovox 230 [2] using the “Finnish Male” voice. Because the synthesizer’s voice is male we used only male human readers. The three readers had different characteristics. One reader does not use his voice much in his work. In this paper he is referred to 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 of working as a radio voice.

We gave precise instructions to the readers on how to read the utterances, so that dates, for example, were read in exactly the same way. There was only one place, a 3 by 2 table, where the reader had some freedom in choosing how to read the material. Readers were not allowed to add any information to it or interpret it when reading, but they could read elements in any order they wanted and also repeat elements. The synthesizer read the table in the normal, left to right, top down reading order. All the human readers read the table by repeating column headings for each row.

From the perspective 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 a multilingual support to our system.

3.2 Subjects

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

3.3 Experiment Setup

The subjects were seated at an office desk. The experimenter explained to them what was about to happen. A 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 s/he could set the volume level.

The questions were of different types. Some asked for specific details and others required more understanding of what was heard. There were two questions about each utterance. The test subject was allowed to read the first question before s/he heard the utterance. After answering the first question, the subject was shown the second question about the utterance just heard and asked to answer it. After answering that question, the subject read the first question about the next utterance. Then the subject signaled the experimenter that s/he was ready to listen to the next utterance.

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

After all the questions about the utterances had been answered, there was a questionnaire for the test subjects to fill in. The users were asked to evaluate such parameters as pitch, rate and volume and their variations in reader's voice. Scales ranged from too little to too much. The amount and the length of pauses and overall emphasis were scaled as well. The answers enabled us to create "speech profiles", i.e. characteristics of each speaker. Finally, at the end of the questionnaire there were two open questions. The first question asked if listening to the speech required concentration and the second asked about possible problems with the voice.

4. RESULTS

We obtained two kinds of data from the experiment. The first data set consisted of the answers to the questions about the content of the utterances. The second set of data contained speaker profiles. The latter corresponded to the questions about the readers' voices.

As we had 16 subjects, each of whom answered 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 absolutely 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 been into correct form but the subject may have forgotten some exact digits.

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

Next we explain the results from the 18 scored questions, then from the remaining questions and after that we introduce the speaker profiles. Finally we describe findings from the analysis of the readers' voice samples.

4.1 Numerical Results

The 18 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 each type of question. As one might expect, the questions in the first group ("before") had much better answers. There were also differences between readers and listeners. The differences between the listeners were small.

Table 1 gives a summary of the answers from the viewpoint of readers. All values are sums of average scores. As Table 1 shows, there is only a slight difference between the readers in questions presented before the utterances were heard.

By comparing questions that were presented after hearing the utterances, much greater variations were found. The lecturer ranked clearly ahead of the others, and the non-professional reader's utterances were found to be the hardest for producing correct answers.

	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	3,92	2,69	6,62
Average	3,77	2,22	5,98
Max	4	4	8

Table 1. Numerical results.

4.2 Observations

In addition to the answers scores, we found other interesting issues from the answers. First, there exists a great difference in the answers when a simple table is expressed. With all other sentences, we gave precise instructions to readers about how to read them. With the table, the readers were free 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 considerably. Although it is not possible to evaluate the results in 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 from the non-professional speaker's expression were the worst.

We asked the listeners to pick out 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. Foreign names in particular were difficult, although we hand-tuned the synthesizer's expressions. A similar effect occurred when we asked the listeners to pick out 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. This shows the importance of appropriate pauses between sentences and phrases. It is also interesting that with this question the listeners had the least difficulties with the utterances spoken by the synthesizer.

4.3 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 with an optimal value in the middle. For example, in the scale of rate of speech the left side of the scale meant excessively slow speech and the right side excessively fast speech. It is possible that different subjects could give inconsistent comments, for example one saying that the rate was too slow and another that it was too fast. Out of 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 synthesizer.

When summing the absolute problem values for each speaker, the radio voice has the lowest and therefore best value of 22. Surprisingly, the synthesizer got into the second place with 24. The third place goes to the lecturer with a score of 25 and the non-professional speaker had the biggest problem score of 25,5. The order is similar, no matter how the inconsistent values are handled. Also, we got a similar order, if we only looked at the last scale in the profiles, which was the overall feeling about the speaker's use of prosody. The radio voice was clearly the best with -1 and the synthesizer came second with -3. The non-professional speaker took third place with -5 and a small margin before the lecturer who had -6. For this last scale, every speaker got values telling that their speech was too monotonous. 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 monotonous speech. However, there was one subject who said that the synthetic voice should be even more monotonous: then it would be easier to understand, but also more boring.

Unclear voice and missing phonemes were reported for the non-professional speaker and surprisingly also for the synthesizer.

Table 2 lists 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 an amount of the feature, for example, too low pitch or too little pitch variation or too slow a rate or too short pauses. Positive values are also problems in the opposite direction, such as too many pauses, or too much rate variation. If there were inconsistent inputs between subjects, i.e., both positive and negative values were given, then both 'min' and 'max' columns have values other than zero.

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	-2,0	1,0
Use of volume	0,0	0,0	0,0	0,0	-1,0	0,0	0,0	2,0
Volume variation	-0,5	2,0	-0,5	0,0	-1,5	0,0	-1,0	0,0
Rate	0,0	2,5	0,0	1,5	-2,0	0,0	0,0	2,0
Rate variation	-2,5	1,0	0,0	1,0	-2,0	0,0	0,0	1,0
Pauses (amount)	-1,5	0,0	-4,5	0,0	0,0	1,0	-5,0	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	-3,0	1,0
Overall amount	-3,0	0,0	-5,0	0,0	-6,0	0,0	-1,0	0,0

Table 2. Speaker profiles.

The interesting parts in Table 2 include the rows describing the use of pauses. We can see that the lecturer had good values there, just slightly too many pauses and some inconsistency in the length of pauses. However, all others had too few and too

short pauses. We can also see that the radio voice scored well on almost all other scales. His use of pitch and volume and rate was preferred any other. All others were considered to be too monotonous.

The overall view seems to be that the lecturer had a too monotonous 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 the best liked, the other readers were considered too monotonous. The synthesizer actually got fairly good scores, far better than the non-professional speaker. However, according to these results, it seems that people would like the synthesizer more if we add more and longer pauses to it and try to get more emphasis to its voice such as the radio voice had.

4.4 Analysis of the voice samples

The four most interesting utterances in the test material were selected for careful analysis. As four readers had read each utterance, a total of 16 samples were analyzed. Word boundaries were marked in these files. This gives us exact data on the use of pauses by the readers. The pause lengths were compared by including all pauses and by excluding and analyzing pauses between sentences separately. We also did pitch extraction and energy extraction using a Kay CSL machine so that it was possible to analyze the use of pitch and volume. The use of prosodic features varied a lot, as speaker profiles suggested.

Pauses. All readers used paragraph pauses, but there were huge differences within sentences. The non-professional speaker used pauses very seldom. The radio voice used pauses in some places and just as the listeners reported, the lecturer had the longest pauses and also the greatest number of them.

The lecturer's pauses between sentences were about 1900 milliseconds on average. The average length of pauses inside sentences was 276 milliseconds. Both values were more than twice as long as any other readers' pauses were. Pauses were made all over the sentences. The radio voice had the second longest pauses inside sentences with an average length of 117 ms. The average length of sentence pauses was 1213 ms. The radio voice used lots of pauses when reading an URL and had long pauses when reading a normal address and a telephone number. With the exception of sentence pauses, he used very few pauses elsewhere.

The non-professional reader had the least pauses of all human readers and his sentence pauses were also shorter than those of the synthesizer. His average pause length inside sentences was 111 ms and sentence pauses were just 596 ms long on the average. The only place where pauses were found inside sentences was in the address. The telephone number had very short pauses and the URL was read almost without pauses. The synthesizer's use of pauses was straightforward. There were about 700 millisecond pauses between sentences and shorter pauses (53 ms on average) in the places of commas and before numbers. No other pauses were occurred.

If we look at only pauses inside sentences that are longer than 100 milliseconds, which should all be intentional pauses, we got

the following results. The synthesizer had only two such pauses in the data, caused by periods in the address and in an abbreviation. The non-professional speaker had 20 of these pauses and the radio voice 26. The lecturer had a total of 59 such pauses.

The lecturer used pauses differently from other speakers. First, he used three different kinds of pauses: very long between paragraphs, moderately long to split sentences into meaningful parts and quite short to separate individual words (for example when reading list elements). The second case is the most interesting since other readers did not use pauses in this way. Also, he used pauses after emphasizing point – the radio voice used pauses almost invariably before emphasizing point.

The use of pitch and energy. Table 3 shows pitch and volume emphasizing counts for all human readers. As can be seen, the radio-voice used a lot of lowered pitch and energy levels to de-emphasize certain unimportant points. The non-professional speakers use of lowered energy was usually because of missing syllables at the end of sentences. The lecturer had more positive pitch variations than other speakers, although some of these could be because of his use of pauses which tend to rise pitch.

Reader	Pitch +	Pitch -	Energy +	Energy -
Non-pro	18	9	7	12
Lecturer	22	8	5	6
Radio	12	12	10	12

Table 3. The use of pitch and energy.

The radio voice’s use of lowered energy to de-emphasize certain words is noteworthy. For example, when an http- address was read this kind of de-emphasizing achieved good results. He de-emphasized characters like ‘:’ and ‘/’, thus emphasizing the importance of more meaningful parts of the URL.

Readers’ prosodic rates were as follows: the synthesizer’s average pitch level was 99 Hz and the radio voice’s was 101 Hz. The lecturer’s value was 84 Hz and the non-professional’s pitch level was 147 Hz. The non-professional speaker’s pitch level was much higher than other speakers’ and the lecturer’s a little lower than others. These may also have affected the results to some extent and it would be very interesting issue to study how much different pitch levels could affect the understanding and pleasantness of a synthetic voice.

5. DISCUSSION

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

There also exist great differences in speaker profiles. As could be seen from Table 2, the problem score of the radio voice is the

lowest. Other speakers had similar overall scores but there exist variations between individual factors.

The individual speaker profiles differ a lot from one another and so also did the actual uses of voices. In general, speaker profiles and the actual use of voices matched very well. First of all, the lecturer used a lot of pauses. By contrast, the non-professional speaker and synthesizer used only a minimal number of pauses. The listeners complained most about the non-professional speaker’s use of pauses. The radio voice used pauses sometimes, but in general listeners liked him to use them too seldom. The radio voice sounded most lively, although it was not the clearest one.

When we compare the speaker profiles and actual uses of voices to numerical results it seems clear that pauses are the most important factor for intelligibility, especially when listener comprehension and memorability are needed. Since the lecturer’s scores are very high on both “before” and “after” questions and because he used pauses more than the others did, it seems likely that pauses help listeners to interpret what they hear. We find more support to this phenomenon when 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 confirmation for this we are planning a formal study where we will use the synthesizer to read the same utterances both with and without additional pauses.

The main reasons for wrong answers are forgetting, difficulties in understanding the words heard and difficulties in understanding 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 usually needed both memorizing and understanding of the content of the message, it is impossible to say exactly what caused the errors. Still, it is obvious that the “after” questions needed more capability to comprehend than the “before” questions.

When “before” and “after” results are examined, we find that the difference in accuracy was more dramatic in the “after” questions. For example, in one utterance where meeting times were read the non-professional speaker and the radio voice got very low scores and the lecturer and the synthesizer got very good scores. The first two used only a few pauses and rest two used a lot of them. It is noteworthy that it was built into the synthesizer to add short pauses before numbers – even with this kind of straightforward method the synthesizer achieved surprisingly good results. This might indicate that besides affecting segmental intelligibility, pausing could have an even stronger effect on comprehension. It seems especially likely that the lecturer’s use of long pauses to divide sentences into meaningful parts and shorter pauses for other emphasizing is a very powerful method. His way of adding pauses after emphasized words instead of putting them before those words is also interesting. We are planning further studies to examine these issues.

Many listeners complained that the lecturer overused pauses. Our analysis supports this, since sometimes the radio voice's use of much shorter pauses resulted as high scores as the lecture's way to use quite long pauses. This should be noted when his use of pauses is transferred to our e-mail system. The lecturer sometimes used pitch variations instead of pauses to emphasize words: usually this happened when he used pauses to emphasize longer word groups in one sentence. This could be a good alternative, since the overuse of pauses could increase usability problems in interactive situations.

In spite of being the most intelligible, the lecturer's speech was also described as very monotonous. Even though he used a quite lot of pitch variations, the listeners did not rank his voice as varying as the radio voice. In general, the listeners did not complain about any other problems with radio voice except his use of pauses. It seems likely that radio voice's way to use combination of both positive and negative pitch and energy variations is very pleasant to listeners. His way of lowering volume in unimportant sections is also interesting. However, the radio voice's 'after' scores were much lower than the lecturer's. It may be that his use of pauses cancelled out the positive effect from the use of pitch. It may also be that even though the use of pitch is very pleasant, it does not bring much comprehension to speech.

Finally, we were surprised at how well the synthetic speech was received. It was at least average in every respect and its problem score was the second lowest. This was a most unexpected result. Of course it may be that listeners judged the synthesizer more lightly than human speakers. We must also remember that there are other elements in speech than those considered here and they could greatly affect both the intelligibility and pleasantness of a voice.

6. FUTURE WORK

We will continue our research by instrumenting the speech synthesizer so that it uses prosody in the manner that the results presented here indicate to be the most profitable. We will then repeat the test and compare results obtained using the synthesizer with and without added prosody. If the test confirms our hypotheses, we will incorporate these features into the utterances produced by our e-mail system.

We do not only want to find general methods to make synthesis more intelligible and pleasant, but also to produce different kinds of persons that could use rich and varying set of prosodic elements in their speech. After we have built these persons, we shall arrange user tests in real interactive situations. These tests could yield more information and suggestions how prosodic elements could be used in spoken dialogue applications.

7. CONCLUSIONS

The 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 can greatly improve both the intelligibility and pleasantness 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. Good use of pauses seems to be able to improve speech intelligibility considerably. In our experiment, the use of pitch and volume had positive effect on pleasantness of speech but not so much effect on comprehension. As we have analyzed the exact ways in which the human readers used prosodic elements in their speech we are able to construct computer persons, a special kind of software modules, which have a rich and varying set of prosodic features in their speech. In this way we could bring unique, human kind speech output to spoken dialogue applications.

8. 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*: 213-220, 1998.
2. Infovox 230 Text-to-Speech system version 1.1. Telia Promotor.
[<http://www.promotor.telia.se/infovox/230.htm>]
3. Hitzeman, J., Black, A., Taylor, P., Mellish, C., and Oberlander, J. "On the Use of Automatically Generated Discourse-Level Information in a Concept-to-Speech Synthesis System", *In International Conference on Spoken Language Generation (ICSLP)*: 2763-2768, 1998.
4. Marx, M., and 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.
5. Pan, S., and McKeown, K. "Integrating language generation with speech synthesis in a concept to speech system", In Alter, K., Pirker, H., and Finkler W., editors, *Concept to Speech Generation Systems*: 23-28, Madrid, Spain, July. Association for Computational Linguistics, 1997.
6. Yankelovich, N. "How Do Users Know What to Say?", *ACM Interactions*, Volume 3, Number 6, November/December 1996.
7. Walker, M., Fromer, J., Fabrizio, G., Mestel, C., and 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*: 582-589, 1998.