

Online Collaborative Learning System Using Speech Technology

Sid-Ahmed. Selouani, Tang-Ho Lê, Chadia Moghrabi, Benoit Lanteigne, and Jean Roy

Abstract—A Web-based learning tool, the Learn IN Context (LINC) system, designed and being used in some institution's courses in mixed-mode learning, is presented in this paper. This mode combines face-to-face and distance approaches to education. LINC can achieve both collaborative and competitive learning. In order to provide both learners and tutors with a more natural way to interact with e-learning applications, a conversational interface has been included in LINC. Hence, the components and essential features of LINC+, the voice enhanced version of LINC, are described. We report evaluation experiments of LINC/LINC+ in a real use context of a computer programming course taught at the Université de Moncton (Canada). The findings show that when the learning material is delivered in the form of a collaborative and voice-enabled presentation, the majority of learners seem to be satisfied with this new media, and confirm that it does not negatively affect their cognitive load.

Keywords—E-learning, Knowledge Network, Speech recognition, Speech synthesis.

I. INTRODUCTION

RECENTLY in the context of rapidly growing network applications, an abiding vision consists of providing computer-based media support where no sophisticated training is required. Among these applications, e-learning systems are rapidly gaining in popularity. In fact, easier-to-use development tools, lower costs, availability of broadband channels and potentially higher returns, in the form of better learner productivity, have made e-learning technology attractive to a wider variety of institutional and individual users. Numerous studies including those of Najjar [1] and Alty [2] confirm the fact that the type of computer-based media incorporated in e-learning materials can have a significant impact on the amount of information retained, understood, and recalled by learners.

Several web-based techniques are used to develop the online collaborative learning (CL) systems. These systems

may integrate a form of chat window or forum through a public or private communication channel. To some extent, these features switch the system to an interactive and communication system, which may be separated from the underlying learning context (that justifies the development of our "in context" system as described below).

Moreover, to design a good e-Learning tool for CL, we must avoid some common issues rising from applying or developing them. For example, (1) the teachers fear to apply them in the classroom (because of the apparent loss of control in the classroom); (2) the students resist collaborating together (because of the lack of familiarity with CL techniques and class management), and (3) one of the obstacles for the implementation of collaborative activities is that students are accustomed to working competitively, not cooperatively [3]. With lessons learned from our past projects and the above experiments from others, we have developed a Web-based learning tool called LINC (Learn IN Context).

The ideal user environment has not yet been found, but individual interface technologies are sufficiently advanced to a point where it becomes feasible to design systems capable of making a positive impact on the e-learning experience, thanks to a high quality of human-computer interaction. Central to such systems is a conversational interaction using speech recognition and text-to-speech synthesis. Deng confirms in [4] that in recent years, automatic speech recognition (ASR) and text-to-speech (TTS) have become sufficiently mature technologies that allow their inclusion as effective modalities in both telephony and other multimodal interfaces and platforms. In this paper, we propose to include such technologies into an e-learning environment using mixed-mode learning. This mode combines face to face and distance approaches to education where an instructor meets with students in the classroom, and a resource-base of content material is made available to students through the web.

This paper is further organized as follows. Section II is concerned with both text-to-speech and automatic speech recognition background. Section III describes the main components and features of LINC and LINC+. Section V presents the results of experiments carried out to objectively and subjectively evaluate ASR and TTS modules of LINC+. Finally, in section IV we conclude and discuss future perspectives of this work.

II. SPEECH TECHNOLOGY BACKGROUND

Applications using a TTS module are numerous and quite diversified. They range from talking document browsers, to

Manuscript received August 5, 2006. This work was supported by the Natural Sciences and Engineering Research Council of Canada.

Sid-Ahmed Selouani is with the Information management department of Université de Moncton, campus of Shippagan, NB, E8S1P6, Canada (phone: 506-336-3625; fax: 506-336-3477; e-mail: selouani@umcs.ca).

Tang-Ho Lê is with the Computer science department of Université de Moncton, NB, E1A 3E9, Canada (phone: 506-858-4120; Fax: 505-858 4541; e-mail: letangho@UMoncton.CA).

Chadia Moghrabi is with the Computer science department of Université de Moncton, NB, E1A 3E9, Canada (phone: 506-858 4521; Fax: 505-858 4541; e-mail: moghrac@UMoncton.CA).

Benoit Lanteigne and Jean Roy are research assistants with the Computer science department of Université de Moncton.

personal computer-based agents, to voice-mail and unified messaging systems, to new telephone directory services.

The current challenge is to extend existing technology to include new modalities, more complex inputs, and more natural-sounding outputs. With current development in memory and CPU, it is no longer necessary to trade off speech quality for storage or calculation capabilities. Thus, it becomes possible to perform large segment concatenation where segments covering each sound might be recorded several times. Each record has its own prosodic features so that more complex methods of locating the appropriate segments and storing them compactly are needed [5].

The maturation of TTS speech interfaces enables the computer to generate any question necessary to clarify any spoken input processed by automatic speech recognition (ASR) system. Thus, new TTS-based architectures allow developers to create natural language dialogue systems that combine TTS with natural language speech recognition [6].

Speech recognition has also made enormous progress over the past 20 years. Advances in both computing devices and algorithm development have facilitated these historical changes. In general, automatic speech recognition (ASR) can be viewed as successive transformations of the acoustic micro-structure of the speech signal into its implicit phonetic macro-structure. The main objective of any ASR system is to achieve the mapping between these two structures. To reach this goal, it is necessary to suitably describe the phonetic macro-structure which is usually hidden behind the general knowledge of phonetic science, as studied by Allen in [7] and O'Shaughnessy in [8].

Generally speaking, ASR systems based on statistical models such as HMMs are able to automatically recognize speech sounds by comparing their acoustic features with those determined during the training. A Bayesian statistical framework underlies the HMM-speech recognizer [9]. The development of such a recognizer can be summarized as follows.

$$w' = \operatorname{argmax}_{w \in \Psi} p(w/o) = \operatorname{argmax}_{w \in \Psi} p(o/w) p(w), \quad (1)$$

where Ψ is the set of all possible phone sequences $p(w)$ is the prior probability determined by the language model that the speaker utters w , and $p(o/w)$ is the conditional probability that the acoustic channel produces the sequence o . Let Λ be the set of models used by the recognizer to decode acoustic parameters through the use of the MAP procedure. Then Equation 1 can be written as follows:

$$w' = \operatorname{argmax}_{w \in \Psi} p(w/o, \Lambda) p(w). \quad (2)$$

The mismatch between the training and testing environments induces a corresponding mismatch in the likelihood of o given Λ and consequently involves a breakdown of ASR systems. Decreasing this mismatch should increase the correct recognition rate.

Currently, the challenges for the ASR are the use of robust acoustic features and models in noisy and changing environments, the use of multiple word pronunciations and efficient constraints allowing one to deal with a very large vocabulary, the use of multiple and exhaustive language models capable of representing various types of situations, and the use of rich methods for extracting conceptual representations from word hypotheses, and automatic learning methods for extracting various types of semantic and pragmatic knowledge from corpora. Most ASR research has shifted to conversational and natural speech. The ultimate goal consists of making ASR indistinguishable from the human understanding system.

The e-learning system that we present in this paper, allows the use of dictation software or automatic speech recognition (ASR) as another option for text input. However, it is often assumed that dictation software will insert correct grammar and punctuation automatically; unfortunately this is not always the case as it may mistakenly transpose homonyms, i.e. wrong words that sound similar to the correct ones. Furthermore, dictating is a skill in its own right, and perfect results may not necessarily be achieved. These considerations underlie our choice to include in this first version of LINC+, a module which recognizes short vocal sentences instead of dealing with long-sentence dictation.

III. WEB-BASED LEARNING ENVIRONMENT WITH VOCAL USER INTERFACE

The Learn In Context Web-based learning environment (LINC) was effectively used for some of our courses in the fall 2004 semester. It is a mixed mode learning tool (both in class and on line at the same time). LINC+ is its new version with a vocal user interface. In the following subsections we describe the LINC system and the vocal features added to the LINC+ version.

A. The LINC Environment and its Components

The LINC system has three components with different functionalities and interfaces. The first component is a multi-user online authoring system, which allows at most six team members (instructional designers) to collectively create the lesson's contents that include demonstrations (multimedia files) and corresponding documents (an URL or documents preloaded on the server).

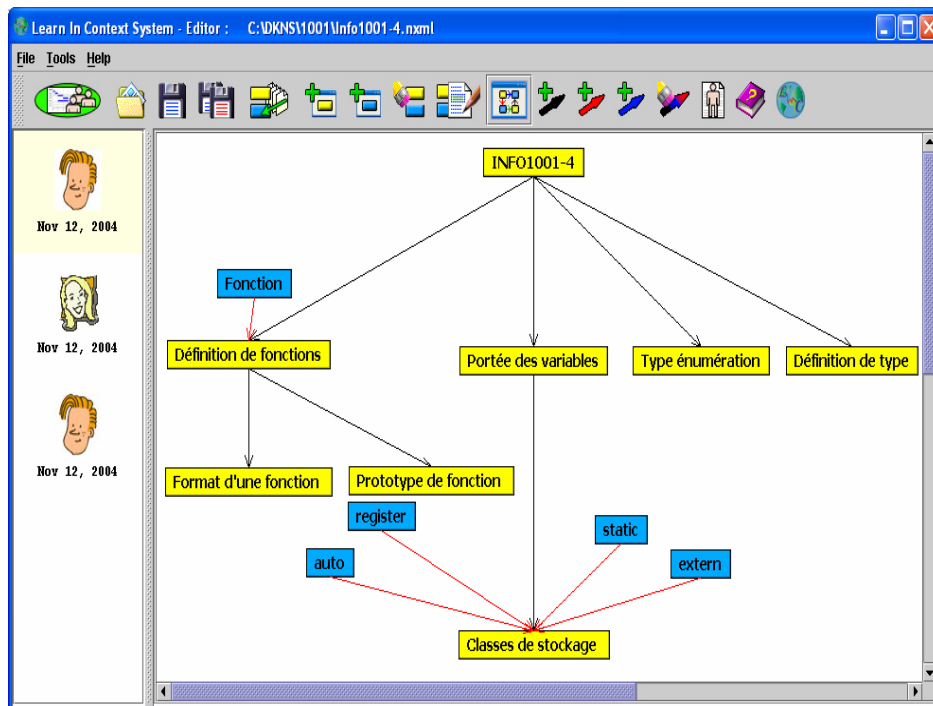


Fig. 1 Example of a Knowledge Network

This component is standalone software, which has been used in several educational projects. It allows easy structuring of the lessons in a task-oriented fashion and makes allowance for saving them in graphical knowledge networks (XML files). Each network (Fig. 1) includes several connected nodes, which can be selected (clicked) and expanded into a text frame with several slots. The last slot is special and is reserved for stimulating questions in which a good instructional designer chooses a suitable question for stimulating learner reflection about the underlying subject matter.

The second component is the Learning Management System, the platform in which the course content as well as associated files (teacher's photo, multimedia files) are assembled before online distribution. We describe some of these details below:

- The Designer's communication facility is used by the designers to communicate with each other about the course content and thus allows them to cooperatively build the content of the course by exchanging ideas about it. To access the designer's communication facility, the user simply has to click on his/her displayed picture, and a new frame appears with recent messages. To send a new message, the user types the message in the text field and sends it over the network. After a message is sent, the date under the picture is updated thus making it easier to verify newly transferred messages.

- The Designer's profile is an add-on functionality allowing the designers to modify their user profiles, where they can

change their passwords, their pictures, and much more.

- The Team leader's tasks are among the main aspects of this tool as they involve the creation and initialization of a project procedure.

The third component is the software that generates the Web pages from every Text Frame created by the first component. Its interface is a standard Web page with the index of the actual lesson on the left side of the screen, which corresponds to the Knowledge Network illustrated in Figure 1, and the topic's content displayed on the right.

In the LINC interface, one can see the hyperlinks for "Reference" and "Demo" below the instructor's photo to display the referenced document and to activate the multimedia demonstration (video clip, Flash file, etc.). There are also some highlighted words in the text; by passing the mouse over these words, a small yellow box will appear to explain the concept or to give a definition.

The forum for discussion between students can be opened by clicking on one of the six mailbox icons labeled from "Group 1" to "Group 6". The content of each mailbox is anchored to the corresponding topic of the lesson; in addition to the content messages, there is also an indication of their date and time.

Thus, researchers can later on trace the student's knowledge evolution. Finally, the three hand-icons are for the instructor's intervention when it is necessary. If the red Stop-hand icon is highlighted, it means that the discussion has been going on too long for the underlying topic; hence the instructor gives this

signal as a warning to the learners. If the yellow Attention-hand icon is highlighted, the discussion may be going in the wrong direction; therefore, the instructor gives some hints to help the learners. And when the green Good-hand is highlighted, the instructor encourages the learners to continue in the same right way or to promise a bonus, an award, etc. We note that his/her role in this course is more suitable as animator, facilitator, and questioner, i.e. a less active role than the learner's role in such a learning process.

B. LINC+: A Vocal User Interface for E-Learners

With the addition of a synthesized voice and a facility for speech recognition to the LINC environment described above, the learner can listen to a voice reading the Web page's content as well as navigate within this learning environment by using his/her own voice instead of using a mouse. We call this new environment with a vocal interface LINC+. After accessing LINC+'s Home page, a simple click allows the user to see a list of courses; another click to choose a course. After this click a login window will be displayed asking the user for a username and a password. The system can recognize two types of users: a professor and a student. Different vocal commands are available depending on the type of user. Users access to the complete functionalities of the vocal interface, as illustrated in Fig. 2, after logging in and choosing a lesson.

Fig. 2 is labeled with sections from A to F. By clicking on one of the speaker icons, the text displayed in the corresponding section will be read by the TTS. Section A contains the Links list (lesson's topics); Section B, the page's header; Section C, the Situation; Section D, the Actions; Section E, the Comments (not indicated); and Section F contains all sections from C to F.

Some words, as the «Introduction» in the Actions section, are light blue, but are hotspots. If the user passes the cursor over this word, a tool tip will appear with a supplementary content in a yellow window (for a concept explanation or the definition of a term). Clicking on the word will cause the content to be read.

It is also possible to interact with the site by launching a vocal command. To enable this facility of speech recognition, the user must first click on the microphone icon (on the right in the first yellow field). This microphone button is a toggle; clicking on the microphone icon again will disable it. A small tooltip appears to show if this facility is enabled. Thus, if the user enables it and then exits LINC+ without disabling it, when he/she goes back again, it is still enabled. There are several commands that can be used to navigate within the site. Some of them are only available for the teacher, and some others are only available for the students.

-Commands available for all users:

Read Information, Read Situation, Read Action, Read Comments.

Read all content (i.e. the three sections above). Read all links (topics on the left of screen).

Read this link (the link that the cursor is actually on, do nothing if not).

Read tooltip names (read all the tooltips of a lesson, even if not on the current page).

Read tooltip X. For example, «Read tooltip Introduction», even if not on the current page.

Read this tooltip (the tooltip below the cursor of the mouse).

Read tooltip N. For example, «Read tooltip 1» read the first tooltip defined for this lesson.

Go to link X. For example, «Go to link MODULE 1».

Go to link N. For example: «Go to link 1» (the Introduction is link 1).

Go to this link (the link pointed by the cursor).

Go back to the lesson list.

Go to reference (link to the external Web page referenced by the actual page, if any).

Open demo (link to the demonstration appearing in the actual page, if any).

What state is it? (To give the actual state - a page may be in one of three possible states: green, yellow or red, only the professor can change a state).

Show commands (open a small window to show the available commands).

Hide commands (close this window).

-Commands only available for teachers:

Set state red (changes the page state to red and then opens the window allowing the professor to write a message to students). Set state green. Set state yellow.

Open group 1 (open the forum of discussion of the group 1, and so on for other groups).

-Commands only available for students:

Open teacher's message (open the window to show the message of the professor).

Open my group (open the forum window to show the email messages of all students in a specified group - 1, 2, 3, 4, 5 or 6). The professor cannot use this command because he/she is not in any group.

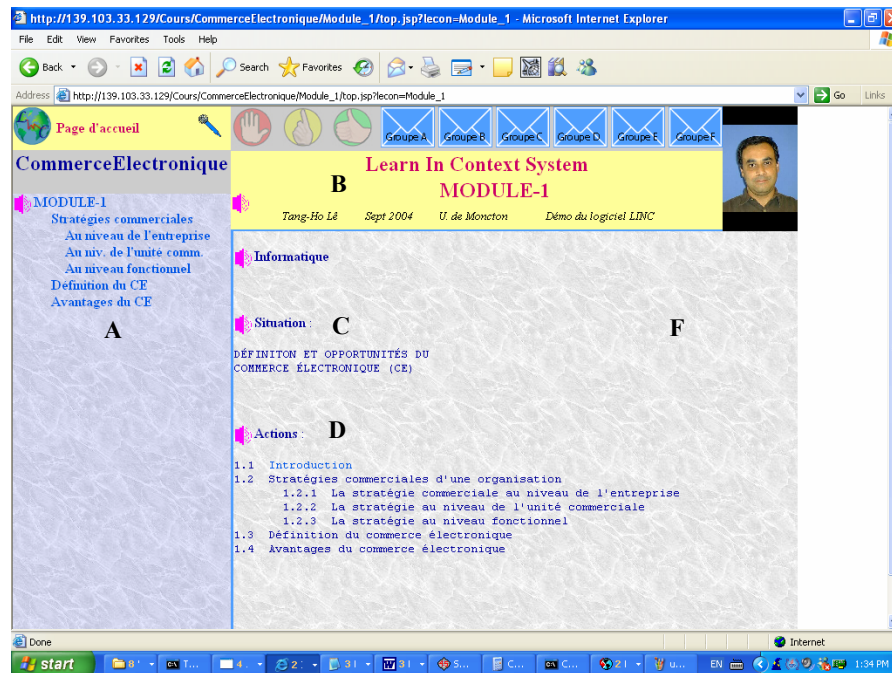


Fig. 2 The LINC+ Vocal Interface

IV. EVALUATION OF LINC+

We have used LINC+ e-Learning tool in a trimester programming course at our university. We have created ten lessons using the authoring system and fifteen Flash files for the demonstrations to accompany the lessons. Experiments have been carried out in order to perform objective and subjective evaluations of ASR and TTS modules of LINC+. The evaluation of LINC+ involved both speech recognition and speech synthesis with a total of fifty experiments. Twenty people were asked to test the speech recognition system, while thirty subjects were recruited to test the quality of the speech synthesis for a number of voices. The participants were mainly students and professors at our university who had never been involved in the project development. All participants were from a multicultural background. In order to collect data for testing *speech recognition*, a number of web navigation commands, usually used for the Learn in context system (LINC+), were tested. The idea was that all 20 test subjects would try each command three times in order to see how many attempts would be successful, and hence obtain a total of sixty trials per command.

As for the evaluation of *speech synthesis*, a simple program was provided in order to permit the user to control the repetition of each sentence until he/she understood it. The program calculates the time (in seconds) needed to understand each sentence, as well as the number of times each of them had to be repeated before it was understood. Thirteen meaningful sentences were used for this test. Some were very short while others were quite long. About half of the sentences were made of commonly used words, while the other half

used scientific terms. The choice of scientific terms intended to avoid the easy guess of a misunderstood word by a listener. Four voices were used for this testing. Two from the L&H TTS 3000 [10] speech synthesizer (Pierre and Veronique) as well as two from the IBM ViaVoice Outloud [11] speech synthesizer (Jacques and Jacqueline) were used for comparison purposes. Some people tested only one voice and some tested two voices for comparison, in a proportion of 50%. Twenty people were considered enough for the evaluation of the speech recognition engine. However, we needed some participants to *compare* voices for informal discussions. We created a new set of sentences for the Jacqueline voice and ten new participants tried both the Pierre and Jacqueline voices. The French sentences used for synthesis varied in length and in difficulty. For each participant, the percentage of word recognition rate per command and for all commands was calculated, as well as the standard deviations.

Table I summarizes the data for every command and for men and women. During the experiment, we realized that the total success rate for men was better for men than for women. Total results for men were around 74% while those of women were around 62%. On average, the recognizer understood 69.99% of the commands which is a little more than two thirds. Also, during testing it was noticed that the recognizer seemed to work better when the person spoke with an American accent.

TABLE I
PERCENTAGE OF WORD RECOGNITION RATE (OVERALL AND BY GENDER)
OBTAINED FOR 10 COMMANDS USED IN LINC+

| Commands | % word recognition rate (men & women) | Standard deviation | Women | Men |
|-------------------------------|---|-----------------------|-------|-------|
| Read information | 66.68 | 27.890 | 66.68 | 66.68 |
| Read situation | 63.33 | 29.649 | 49.95 | 69.06 |
| Read action | 83.34 | 22.360 | 61.12 | 92.86 |
| Read all links | 75.01 | 33.126 | 66.68 | 78.57 |
| Read all content | 70.06 | 31.444 | 61.27 | 73.82 |
| Go to this link | 80.01 | 28.675 | 77.78 | 80.96 |
| Read this tool tip | 65.01 | 26.831 | 66.68 | 64.29 |
| Set state yellow | 65.01 | 35.712 | 61.12 | 66.67 |
| Open group A | 76.67 | 36.667 | 61.12 | 83.34 |
| Go back to the lesson list | 51.67 | 30.699 | 38.88 | 57.15 |
| All | 69.99 | 13.073 | 61.65 | 73.8 |

The testing program used for speech synthesis calculated the time (in seconds) needed by the participant to understand a sentence as well as how often the sentence had to be repeated. The average and standard deviation were calculated for those data. Table II summarizes the results for each voice in L&H TTS 3000. Veronique and Pierre performed similarly. Veronique had slightly fewer repetitions on average, but with Pierre participants took less time to determine whether they had understood or not. According to the test results, the IBM ViaVoice Outloud engine performed better than the L&H TTS 3000 engine both in number of repetitions and time of completion. Of the two ViaVoice Outloud voices, Jacques had better results than Jacqueline. Generally speaking, there was satisfaction with almost every user. With this result as well as with the evident advantage for the instructor, we believe that the LINC/LINC+ system is very useful, to both learners and teachers.

V. CONCLUSION

Traditional e-learning suffered from a "boredom" factor. The introduction of conversational technology constitutes a solution to improve retention rates. This paper presented such a solution which proposes a convivial e-learning environment (LINC) dedicated to french computer science courses and provides augmented interaction modalities by incorporating ASR and TTS modules (LINC+). The ideal system we target is one where the trainee uses earphones and a microphone to interact with LINC+, and where the learning procedures are reconfigured depending on the system's interaction with the trainee. However, in incorporating speech technology into an e-learning tool, there are many technological challenges. The ASR must be sufficiently robust and flexible, whatever the acoustical environment.

TABLE II
EVALUATION OF TTS SYSTEMS AND SYNTHETIC VOICES INVOLVED IN
LINC+

| Voice synthesizer | Average number of repetitions | Standard deviation | Average time in seconds | Standard deviation |
|-----------------------------|-------------------------------------|-----------------------|----------------------------|-----------------------|
| L&H TTS 3000 | | | | |
| Veronique | 1.607 | 0.731 | 5.883 | 7.828 |
| Pierre | 1.623 | 0.844 | 3.809 | 4.641 |
| IBM ViaVoice Outloud | | | | |
| Jacques | 1.323 | 1.323 | 1.636 | 18.609 |
| Jacqueline | 1.415 | 0.429 | 3.107 | 4.639 |

In some situations, the TTS module must be bilingual, since in many technology oriented courses, French and English are used interchangeably. The major finding of our study is that when learning material is delivered in the form of a speech-based presentation, the majority of learners seem to be satisfied with this new media, and claim that it does not negatively affect their cognition load. Learners are faced with the task of reviewing their overall learning strategy in light of the opportunities provided by speech technology. Speech technologies can drastically enhance a student's ability to access e-learning, but it is vital to match the right tools with both the user and the tasks to be undertaken.

REFERENCES

- [1] Najjar, L.J. (1996). Multimedia information and learning. *Journal of Educational Multimedia and Hypermedia*, 5(1), pp.129-150.
- [1] Alty, J.L. (2002). Dual Coding Theory and Education: Some Media Experiments to Examine the Effects of Different Media on Learning, in the *Proceedings of EDMEDIA2002: World Conference on Educational Multimedia, Hypermedia & Telecommunications*, Denver, Colorado, 42-47, USA.
- [2] Bosworth, K. (1994). Developing collaborative skills in college students; *New Directions for Teaching and Learning*, 59, 25-31.
- [3] Deng, L. & Huang, X. (2004). Challenges in adopting speech recognition, *Communications of the ACM*. 47(1), 69-75.
- [4] Sproat, R.W. (1995). Text-to-speech synthesis, *AT&T Technical Journal*, No 74, pp. 35-44.
- [5] Fukada, T., Yoshimura, T., and Sagisaka, Y. (1999) Automatic generation of multiple pronunciations based on neural networks, *Speech Communication*, vol.27, pp. 63-73.
- [6] Allen, J.B. (1994). How do humans process and recognize speech? *IEEE Transactions on Speech and Audio Processing*. 2:4, 567-577.
- [7] O'Shaughnessy, D. (2001). *Speech communication: Human and machine*, IEEE Press.
- [8] Jelinek, F. (1997). *Statistical methods for speech recognition*, MIT Press.
- [9] L&H TTS3000, Nuance: trade mark: <http://www.nuance.com/>
- [10] IBM ViaVoice outloud, Text-To-Speech downloadable for many languages <http://www.306.ibm.com/software/voice/viavoce/dev/msagent.html>