# A Speaking Electronic Librarian

The design of a speech agent for automatic library services is presented in this paper. The proposed system will be based on speech recognition and synthesis technologies, applied to the library environment. The client of the library can have access to various automatic electronic services through a sophisticated interface, making use of the embedded technologies. The access to OPAC, the loaning process, the database access and the resource retrieval are some of the services that could be greatly facilitated through the use of the system. The speech interface is considered as a factor that contributes greatly to the global access initiative, giving equal opportunities to the individuals with special needs.

## Presentation of the speaking librarian system

The system is designed to consist of a speech synthesis and a speech recognition module co-operating with various library services concerning resource catalogues, bibliographic / full text databases and communication facilities (fax, e-mail etc.)

The system requirements are: a moderate PC system, e.g. a Pentium at 1.7 GHz with 256 MB of RAM, Windows environment and a standard compatible sound card. Under these specifications the system executes the various services in real time.

The components of the system are shown in Figure 1.

The user can give his command through the microphone or the keyboard. In case of speech input, the speech recognition module is required in order to transform the physical speech flow into a series of electronic commands. Then each command is passed to the library service, where it is processed, concluding either in the execution of the command or in the informing of the user about inadequate input. The user takes all the information either in speech or in a monitor display. The first step for the speech output is the definition of the language (Greek, English, etc.), next the transcription of text into phonetic items and lastly the dictation of the phonetic text in the selected language. Additionally, the system gives to the user the capability of help topics either presented in the monitor or dictated by synthetic speech.

Various speech synthesis systems used for facilitating the services of the citizen in the information society have been suggested [1].

More information about the speech recognition and the speech synthesis modules are presented in Appendices A and B correspondingly.

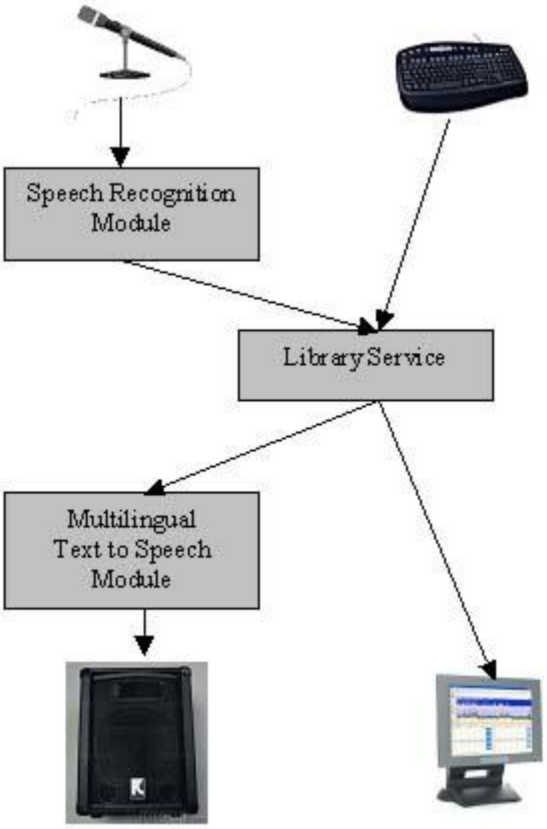

Fig. 1. The speaking librarian system

## Candidate library services to be embedded in the speaking librarian system

Some library services that seem to be suitable to be performed in coordination with the speech processing facilities are the access to the OPAC (On-Line Public Access Catalogue), the loaning process, the use of the inter-library loan system, the access to a bibliographic or full-text database, the selection or retrieval of a database resource and the selection or retrieval of a hard-copy or virtual resource of the library.

All these services are performed through the speech processing or the typical Input / Output interface.

The library services mentioned above, the content requested in the frame of these services and the response actions in each case are given below in a tabularised form :

| **Library service** | **Request content** | **Response actions** |
|---|---|---|
| 1. Access to OPAC | Subject category selection through microphone or keyboard | 1a. Transition to the requested category<br>1b. Speech confirmation that the action has been completed (necessary for visually impaired persons) |
| 2. Access to OPAC | Item selection through microphone or keyboard | 2a. Transition to the required item<br>2b. Dictation through a speech synthesizer of all the text information related to the item with parallel scrolling on the screen |
| 3. Access to inter-library loaning | Library selection through microphone or keyboard | 3a. Transition to the required library loaning service<br>3b. Speech confirmation that the action has been completed (necessary for visually impaired persons) |
| 4. Access to a database | Database selection through microphone or keyboard | 4a. Transition to the required database<br>4b. Speech confirmation that the action has been completed (necessary for visually impaired persons) |
| 5. Resource selection / retrieval in the frame of a certain database | Resource selection / retrieval through microphone or keyboard | 5a. Transition to the required resource<br>5b. Dictation of the text information content of the resource (in case of digital resource) |
| 6. Loaning process (hard copy resource retrieval service) | Resource selection through microphone or keyboard and filling in an application form with the information of the borrower and the mailing or faxing command | 6a1. Confirmation of the availability of the resource given through speech and written on the screen<br>6a2. Offer of an identity number for the monitoring of the process |

| | | |
|---|---|---|
| | | OR<br>6b. A negative message declaring the current unavailability of the resource |
| 7. Information about the retrieval status | Asking for the status of a request through microphone or keyboard | Dictation through speech synthesizer of the information concerning the related mailing or faxing process with parallel scrolling on the screen |
| 8. Virtual resource retrieval service | Speech command for downloading or selection of the 'download' button through the mouse | Periodical informing through synthetic speech that the process is in progress and declaration in the end that the process has been completed |

Table 1. Library services with the corresponding user requests and the system response

In all the above cases some words are isolated from the natural dialogue request through a word-spotting algorithm. The means for the word isolation is the identification of a natural dialogue word with a word in the catalogue menus or the dialogue-box buttons or the contents of file lists or in general in any text element of the working window. All of these words are available in the system, since they can be captured through Applications Program Interface (API) of the Windows environment. Any word contained in the natural dialogue having a distance below a certain threshold from some of the active words of the current window (from a pattern matching view) is considered as a selection command of the user.

## The state of the art in the area of the interfaces of visually impaired persons

A Section of Libraries for the Blind (SLB) was established in 1983 as a forum for libraries for the blind. SLB participates in the annual IFLA conference and also in a bi-annual pre-IFLA conference for the Section. In the 2001 pre-conference, which took place in Washington, many interesting articles were presented, focused on increased information choices through web-based technologies, future library services for blind students, digital delivery for the blind and mainstreaming library services for blind and print disable users [2].

Despite growing technological developments in the information and communications area, only a small percentage of documents are actually made available to the blind in accessible formats including speech output, braille output, tactile devices or even simple adjustments to a browser[3].

Integration of blind and visually impaired persons into schools, universities and training centers is being considered through projects such as BrailleNet [4]. BrailleNet concerns document delivery in the frame of Internet and aims to achieve integration through Web accessible assistive technologies and teaching materials. The delivery of these special books is further enabled through co-operation with publishers, adaptation centers and printing centers.

National Library Service for the Blind and Physically Handicapped, Library of Congress (NLS) is making use of the Internet to deliver a number of its services [5]. A continuously growing number of Web-Braille titles (3800 titles now) has been made available to 1500 users.

What must be stressed here is the great importance of the speaking librarian system for the visually impaired persons. Speech input / output through speech recognition / synthesis correspondingly, is the most user-friendly interface for offering information in the case of visual inefficiencies, contributing greatly to a more efficient interoperability of the previously mentioned initiatives.

## Conclusions

An agent based on speech processing technologies is present in this article to be applied to a library environment. Many library services can be greatly facilitated in the form of a virtual librarian, who serves and helps either the user visiting physically the library or the user having a remote access to the library from his home. Such a system is useful for all the user groups familiarised with the basic computer use and it is of great importance in the case of visually impaired persons, helping them to address their commands through the microphone and take the system response in synthetic speech. These sophisticated interfaces are coordinated with a great tendency for global access to the information and communication technologies (ICT).

## Appendix A. The speech recognizer

Large progress has been made in speech recognition technology over the last few years, especially in Large Vocabulary Recognition (LVR). Current systems are capable of transcribing continuous speech from any speaker with average word error rates of between 5% and 10%. The best policy is to allow speaker adaptation. In that case, after 2 or 3 minutes of speech, the error rate will drop under 5% for most speakers [6].

The speech recognition module to be used is a unit of continuous speech recognition based on acoustic and linguistic models. The acoustic models are related to the minimal phonetic units, which are modeled in the learning phase. These models are used next as reference patterns for the recognition of the phonetic units of the continuously introduced speech stream. The comparison of the unit to be recognised to the available patterns can be based on Hidden Markov Models (ÇÌÌ) methodology. The linguistic models are related to the structure of each written language and improve the efficiency of the recognition system by using context information. A key-factor of a language model is the number of words used in the examined word-strings, the so-called N-grams. The performance of the large vocabulary recognisers is greatly improved in case of speaker adaptation. Adaptation can be supervised or unsupervised, and it can be performed additively as the speaker is talking or off-line at the end of the session. Unsupervised additive adaptation is the least intrusive technique from the user view.

Another essential feature of the large vocabulary recognisers is their robustness in noisy environments. There are a variety of approaches to dealing with noise [7]. Noise can be removed at the front-end from the speech, alternatively noise-robust features can be used, the noise can be masked or the features can be mapped. The problem of noise presence still remains important since all the features derived for the compensation of noise are based on the global statistics of the noisy signal instead of the noise itself. The noise elimination is a necessary component of the total recogniser, since without any form of compensation the recognition performance drops dramatically.

A representative system of the current generation of recognition systems is the Cambridge University HTK system, which will be used in the speaking electronic librarian. The HTK error rate reduces with an increase of the used vocabulary and an increase of the length of the word-string N-gram. An error percentage of nearly 7% can be reduced more in case of speaker adaptation.

Two scenarios can be adopted in the case of the incorporated speech recogniser, the first one without speaker adaptation focused on the service of any client and the second one addressed for a certain client, increasing his performance as much as possible. In the first approach, a greater number of errors is allowed in expense of serving any new client. In the second approach the recognized are substantially eliminated but this superior performance needs an extra time for the adaptation process. The acoustic models of each customization case are stored and retrieved for use of the specific client.

After the recognition of the phrases of the speech stream, the location of the key-information words follows through word spotting technology. These located and isolated words are the commands introduced to the system for the execution of the various library services. Whether an error occurs there are two possibilities. The first one is that the false word has a distance from the active words of the window, which exceeds the threshold, thus it is not selected as a command word. The second possibility is that the false word is relatively near to one of the active words, leading in this way to an action not actually requested by the client. Such false alarms can be eliminated, if a voice or written confirmation message is inserted at the beginning of the service, avoiding a subsequent spending of time.

## Appendix B. The speech synthesizer

A number of commercial and laboratory prototype systems have been presented for text-to-speech synthesis. The majority of them are based on one of the three most popular paradigms:

· Rule-Based Speech Synthesis [8].

· Speech Synthesis based on Time-Domain Techniques [9], [10].

· Speech Synthesis based on Articulatory Models of the human speech production system.

Each of these methods possesses quite different characteristics, which render it more suitable for specific application areas. In case that speed of execution is mainly concerned, a time-domain technique is the prime candidate. On the other hand, for memory-sensitive application environments, formant-based techniques present a distinct advantage.

Modelling the human speech production system is a very demanding task, since the incorporated articulatory models require intense calculations. This fact severely inhibits the implementation of articulatory models into real-world commercial applications.

Time-domain text-to-speech conversion (TD-TtS) relies on a large database of pre-recorded natural speech segments, which are then appropriately concatenated to obtain the speech transcription of arbitrary text [11]. By employing sophisticated algorithms for seaming the discrete segments one can achieve synthetic speech of high naturalness [12], [13], [14]. Rule-based text-to-speech conversion (RB-TtS), on the other hand, models the human speech production system more closely, requiring a more profound examination and a direct modelling of all the phenomena involved. A number of high-quality state-of-the-art systems based on RB-TtS have been presented confirming the value of this method.

Synthetic speech quality, especially naturalness, is largely dependent on the sophistication of the prosodic modelling and prosodic rules employed. On the other hand, detailed prosodic implementation increases substantially the intelligibility of the system even at segmental level.

The majority of TtS systems are based on sentence-level prosody, which provides various degrees of intelligibility but hardly any quasi-natural output from a prosodic and thus a phonetic point of view. Thus, the main directions for improving the naturalness of synthetic speech involve studying (i) the synthetic signal quality as well as (ii) the prosodic modelling of natural speech. Both aspects are the subject of intense research activity for improving the naturalness of synthetic

speech [15].

Porting an existing speech synthesiser to a different language is a task requiring language-specific resources. Focusing to a TD-TtS approach, the creation of a high- quality speech synthesiser consists of developing an ensemble of modules for the target language. These may be divided into the areas of (i) linguistic processing and (ii) digital signal processing, and are briefly described as follows:

· Text-to-Phoneme Module: Converting written character complexes into phonemes to be dictated;

· Segment Database: Creating a database of segments (and associated corpus) that covers sufficiently the target language;

· Deriving an algorithm for decomposing text into segments;

· Prosodic Modelling: Creating a prosody generator for the target language that provides the desired synthetic speech quality;

· Speech Corpus: Obtaining an adequate corpus of pre-recorded utterances, which will provide the basis for defining speech segments in various environments to be concatenated during synthesis;

· Synthesis Algorithms: Designing the algorithms that join the segments so as to generate the synthetic speech signal.

· Unit Selection: To improve the speech quality, multiple instances of each segment, possessing different prosodic properties, may be provided in the database. An algorithm is then used to select the unit that most closely resembles the prosodic characteristics dictated by the model, thus minimising the audible mismatches [16], [17], [18].

## References

Raptis, S., Malliopoulos, C., Bakamidis, S. and Stainhaouer, G., “A Speech Agent for Remote E-Mail Access”, *The 4th IEEE workshop on Interactive Voice Technology for Telecommunications Applications and ESCA Tutorial and Research Workshop on Applications of Speech Technology in Telecommunications*, September 29-30, Torino, Italy 1998

Jenny Craven, “The development of digital libraries for blind and visually impaired people”, athttp://www.ariadne.ac.uk/issue30/ifla/ (access date 27/2/2002)

Miesenberg, K., “Future library services : developing research skills among blind students” in *Digital libraries of the blind and the culture of learning in the information age*, Conference proceedings of the IFLA SLB Pre-Conference, Washington DC, USA, Aug. 13-15, 2001, IFLA/SLB, 2001

Burger, D., “BrailleNet: digital document delivery for the blind in France” (as in Ref. 2)

Sung, C., “The future of lifelong learning in the next generation of library services” (as in Ref. 2)

Steve Young, “Large Vocabulary Speech Recognition : A Review”, *IEEE Signal Processing Magazine* 13(5): 45-57, 1996

Haton J-P, “Automatic Recognition of Speech in Adverse Conditions : a Review”, *IEEE Speech and Audio*, SAP 276, 1994

8. Conkie, A & Isard, S, “Optimal Coupling of Diphones”, in *Progress in Speech Synthesis*, Van Santen, J., Sproat, R, Olive, J. & Hirschberg, J. (eds.), pp. 279-282, Springer-Verlag, New York, 1997.

9. Conkie, A et al, 1997, op.cit., pp.279-282.

10. Moulines, E., & Charpentier, F., “Pitch Synchronous Waveform Processing Techniques for Text-to-speech Using Diphones”, *Speech Communication*, Vol. 9, No. 5, pp. 453-370, 1990
11. Dutoit, T., *An Introduction to Text-to-Speech Synthesis*. Kluwer Academic Press, Dordrecht, 1997.
12. Dutoit, T. & Leich, H., “Text-to-Speech Synthesis Based on a MBE Resynthesis”, *Speech Communication*, Vol. 13, pp. 435-440, 1993
13. Stylianou, Y., *Harmonic plus Noise Models for Speech, combined with Statistical Methods, for Speech and Speaker Modification*, Ph.D. Thesis, Ecole Nationale Superieure des Telecommunications, 1996
14. Stylianou, Y., “Removing Linear Phase Mismatches in Concatenative Speech Synthesis”, *IEEE Transactions on Speech and Audio Processing*, Vol. 9, No.3, pp. 232-239, 2001.
15. Keller, E., Bailly, G. Monaghan, A, Terken, J. & Huckvale, M., “Improvements in Speech Synthesis” *COST 258: The Naturalness of Synthetic Speech,* John Wiley & sons Ltd., Chichester, England, 2002.
16. Conkie, A et al, 1997, op. cit. pp. 279-282.
17. Founda, M, Tambouratzis, G, Chalamandaris, A. & Carayannis, G, *Proceedings of the Eurospeech-2001 Conference*, Vol. 2, pp. 837-840, 2001.
18. Klabbers, E.A.M & Veldhuis, R. (1998), “On the Reduction of Concatenation Artifacts in Diphone Synthesis”, *Proceedings of the ICSLP 98 Conference*, Vol. 5, pp. 1983-1986.

## Author Details

*Markos Dendrinos*
Ass. Professor of the Dept. of Library Studies in Technological Educational Institution of Athens (TEI-A)
e-mail : mdendr@teiath.gr
Researcher of Speech Technology Dept of Institute for Language and Speech Processing (ILSP)
e-mail : mark@ilsp.gr
Web site : http://www.ilsp.gr/homepages/dendrinos_eng.html
Athens
Greece

***Article Title***: *"A speaking electronic librarian"*
***Author***: *Markos Dendrinos*
***Publication Date***: *10-Oct-2002*
***Publication***: *Ariadne Issue 33*
***Originating URL***: *http://www.ariadne.ac.uk/issue33/dendrinos/*