Romanian Speech Enabled Human Computer Interaction System Applied for Information Retrieval

Mircea Giurgiu
Technical University of Cluj-Napoca,
Telecommunications Department
26 Baritiu Str., 400027 Cluj-Napoca
Romania
mircea.giurgiu@com.utcluj.ro

Abstract – **This work presents several results on the implementation and testing of a human computer interaction system using continuous speech recognition. The main task of the system is to allow students to access and to retrieve information from a database via a speech recognition engine based on hidden Markov models. The system's architecture is composed by a database server with students' information and the speech recognition engine that transforms the speech commands uttered in continuous manner into a query string for the database. For experimental purposes the dialogue was kept as simple as possible, so a very simple language model was incorporated in order to get the appropriate sequence of commands. The accomplished experiences provided an interesting point of view on the problems of real applications of speech technology and they open the doors for system integration with the telephone line.**

I. INTRODUCTION

An important progress has been carried out in recent years in the area of spoken language technologies and it is directed to the implementation of practical systems for dictation, telephone applications for switchboard, information retrieval systems, computer-telephony integration, interactive voice response systems, etc.

Research activities in the area of speech technology started at Technical University of Cluj-Napoca almost ten years ago by the investigation of Automatic Speech Recognition (ASR) of Romanian isolated digits. Since then, important results have been achieved in this field by practical implementation in laboratory of ASR systems from the beginning (speech acquisition) up to the end (evaluation of recognition accuracy) using different techniques: Dynamic Time Warping (DTW) [1][2], Hidden Markov Models (HMM) [3] or Artificial Neural Networks (ANN) [4].

During all these years the main target remained the implementation of an experimental system based on Continuous HMM (CHMM), where the observation probabilities are modelled by continuous Probability Density Functions (PDFs), aimed for speech communication with a computer [6]. The system has been implemented in the frame of several projects and within a project with the Ministry of Education and Research in Romania in 2001. This system is our first trial in the application of Continuous Speech Recognition (CSR) for a practical purpose aimed to identify and solve problems related with the incorporation of language models in such an application. The paper gives only a general overview of the elements contained in the system (Fig. 1) as we are not going to present here theoretical fundaments, but more practical results instead.

II. RESULTS ON CONTINUOUS SPEECH RECOGNITION IN ROMANIAN

Before the implementation of the human interaction system for information retrieval using ASR, a series of experiments have been done in order to evaluate a more simple CSR system. This was aimed to explore the capacity of integration of acoustic models with the language models. The established task was to command a telephone with very simple commands spoken in a continuous manner, as follows: a) "Suna la <nume>" (in English: "call <name>") and b) "Telefoneaza <numar de telefon de sase cifre> (in English: "make a phone call at <number>") [7].

The adopted strategy was to use HMM acoustic models with three states per phoneme. A set of experiments have been accomplished with CHMMs to recognize Romanian digits, where the acoustic vectors were set at: a) 12 components of Mel Frequency Cepstral Coefficients (MFFCs), b) MFCC and Energy (MFCC+E), c) MFCC and the corresponding derivatives to see the effect of their correlation in time (MFCC+D), d) MFCC, energy and derivatives (MFCC+E+D). The recognition results are presented in Table I and the confusion matrix for the case of using only MFCC, in Table II. The database is composed by 110 speakers uttering the Romanian digits. The statistical analysis of the speech material is presented in Table III, Table IV and Table V.

Because of the nature of the application, the training and testing speakers were selected from the students in our faculty: 37 females and 73 males with ages between 20 – 25 years. The database was collected in an interval of three years, in each year two sessions for speech acquisition have been organized in order to cope with the speech variability issues.

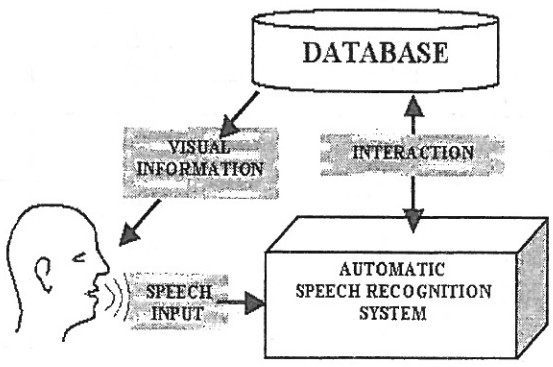


Fig.1. General scheme of the system

No speech deffects have been remarked for training materials. The result of speech variability in time is embeded into the speech features, so attention was given to this issues. As the system is speaker independent it is not important the interval between the training and experimentation. The recognition technique with HMMs is a well known one [6]. For more clarity we may explain that for each of the word to be recognised there is a HMM whose parameters (transition probabilities and PDFs for each state) are trained with the Baum-Welch algorithm. In the recognition step, the HMM that gives the maximum probability for an unknown speech features sequence is selected for the the recognised word. In our case, the situation is rather complex as the HMMs are three state models and they are trained at phoneme level. A chain of such micromodels gives a statistical acoustic macromodel at phrase level. Apart of these models there is the language model which defines the grammar rules and the linguistic rules. Acoustic models and linguistic models define the continuous recognition model.

TABLE I
RECOGNITION ACCURACY FOR CHMM IN DIGIT RECOGNITION TASK USING DIFFERENT SET OF ACOUSTIC VECTORS

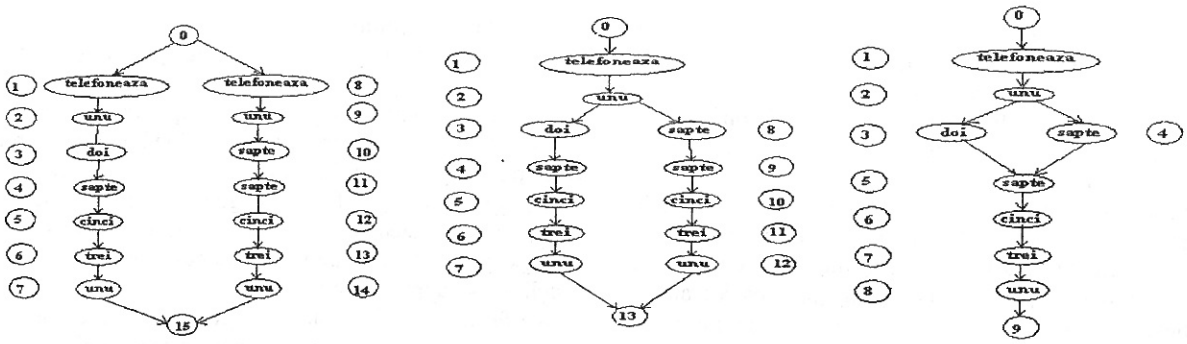
Parameters	Vector dimension	Recognition accuracy [%]		
		3 Gaussians	4 Gaussians	5 Gaussians
MFCC	12	97,53	98,08	98,43
MFCC + E	13	97,63	98,32	98,49
MFCC + D	14	98,15	98,47	99,47
MFCC + E + D	25	98,33	99,41	99,44

TABLE II
CONFUSION MATRIX FOR ROMANIAN DIGIT RECOGNITION (MFCC AND PDFS MODELLED WITH 3 GAUSSIANS) AND DIGIT ACCURACY [%]

	zero	unu	doi	trei	patru	cinci	sase	sapte	opt	noua	TOTAL
zero	159						1				97,50
unu		156			3	1					96,30
doi			162				1	1			98.80
trei			1	160			1	2			97,60
patru					162		1				99,40
cinci						163		1			99,40
sase							157	6			96,30
sapte							1	155			95,10
opt		1			1	1			157	1	97,50
noua	1	1			1		1			149	97,40

TABLE III
STATISTICAL ANALYSIS OF THE SPEECH CORPUS AT WORD LEVEL

No	Word (in Romanian)	No. of occurrences	Average length (ms)	Minimum length (ms)	Maximum length (ms)
1	suna	125	355.3	200	469.9
2	pe	125	169.5	110	310
3	adi	21	308	190	459.9
4	dan	22	353.7	180	480
5	marin	20	459.2	370	620
6	petre	20	492.1	370	630
7	sorin	21	468.7	360	598.2
8	stefan	21	567.6	420	803.9
9	telefoneaza	104	852.7	610	199.9
10	unu	48	308	170	577.2
11	zero	28	364	230	501
12	trei	84	318	168.6	430
13	doi	68	299.2	180	410
14	patru	75	467.1	300	690
15	sase	69	458.9	280	630
16	sapte	67	523.1	350	740
17	cinci	78	384.5	259.9	530
18	noua	58	384.5	180	570
19	opt	45	310.4	179.9	467.6



a) full representation requires 14 states, b) partial optimisation-12 states, c) complete optimisation -9 states

FIG. 2. AN EXAMPLE OF OPTIMISATION OF THE GRAMMAR SPACE

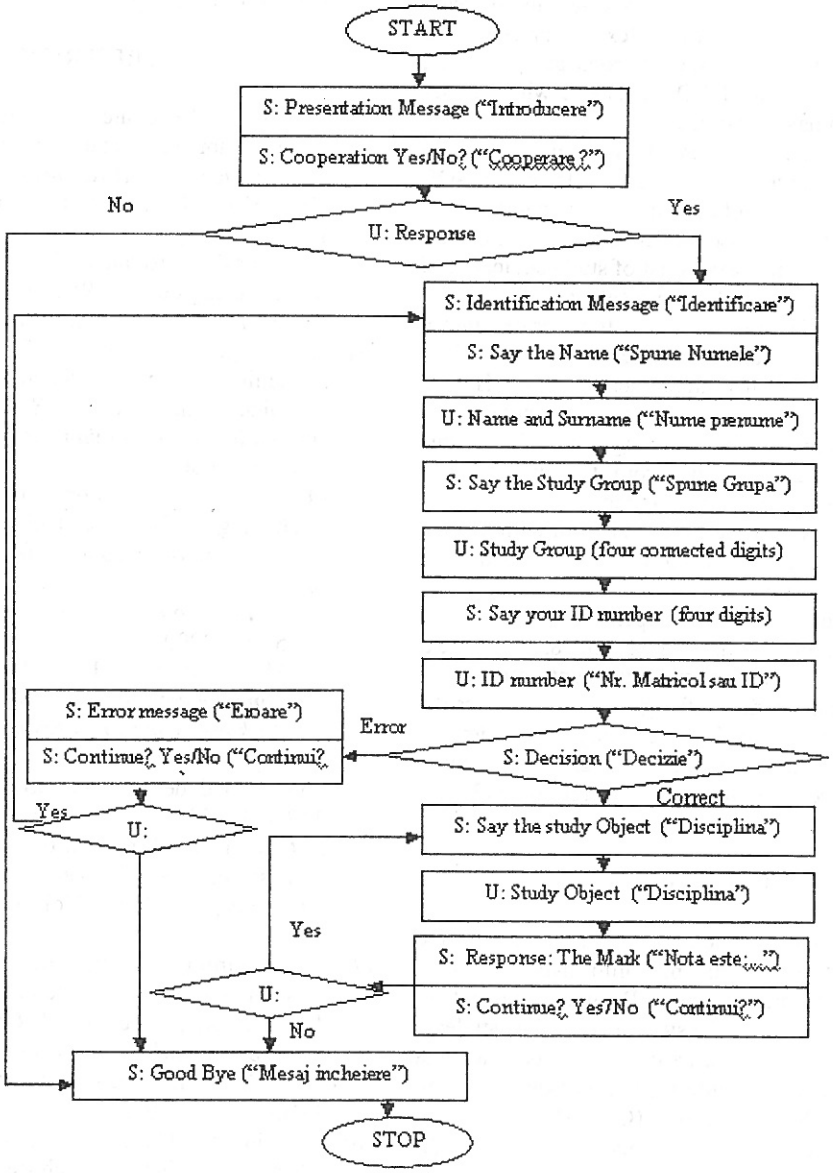


FIG. 3. THE SCHEME OF THE MAN-MACHINE DIALOGUE

One very important issue in language modelling was the optimisation of the grammar space at the phrase level. For

example, supposing we want to discriminate between the next phrases which are spoken in continuous way: a) "telefoneaza unu doi sapte cinci trei unu (call to number: one, two, seven, five, three, one)" and b) "telefoneaza unu sapte sapte cinci trei unu (call to number: one, seven, seven, five, three, one)".

The language model can be represented by a tree as in Figure 2. The tree can be simplified according to the language model and it normally expands according to the phoneme modelling by HMMs. Taking into considerations all these preliminary experiments, we have proposed to implement a more complex system able to offer information that is stored in a database.

III. THE SYSTEM INTERACTION BY SPEECH

The idea behind this system is to offer audio-visual information in the form of written text or pre-recorded message (not Text to Speech incorporated at this stage), which is retrieved from a MySQL database when a user asks a specific command by speech.

The general scheme is presented in Figure 1 and the structure of the dialogue in Figure 3. Inside the MySQL database, the tables store information on: students names and their passwords, addresses, telephone numbers, their classes in each day of the week, a list of study subjects and the students' marks at exams for these subjects.

The speech input is processed and transformed into a sequence of acoustic vectors that are fed at the input of the acoustic recogniser, which is implemented with CHMMs at phoneme level. A sequence of recognized phonemes is generated according to the language model. If this sequence is recognized as one from the dialog, it is converted into an SQL query in the database.

The MySQL server extracts needed information from the database and presents it on the computer screen or as pre-recorded audio messages.

The global recognition rates depend on the adopted language model (context dependent left 83%, context dependent right-86% or context dependent left-right – 93%). The context dependent left-right offers better results as the information provided for training the system is larger and the restriction imposed by the grammar recover possible acoustic recognition errors delivered by the HMM recogniser.

IV. CONCLUSIONS

It was presented here practical results concerning the implementation of an experimental information retrieval system based on Continuous Speech Recognition (CSR) in Romanian. The main task of the system is to allow students to access and to retrieve information from a database via an Automatic Speech Recognition (ASR) engine based on Continuous Hidden Markov Models (CHMMs).

The system's architecture is composed by a MySQL database with students' information and an ASR engine that transforms the speech commands uttered in continuous manner into a query string for the database. The feed-back from the machine is not yet a text to speech one, but we are studying this possibility for the future.

For experimental purposes the dialogue was kept as simple as possible, so a very simple language model was incorporated in order to get the appropriate sequence of commands. It is implemented also, a guided dialog in order to simplify the tasks in the language processing part of the system. In every stage of the dialogue, the ASR generates a string of recognized words, which is used by the system manager to create the query for database interogation and to deliver the appropriate response.

We faced several problems in grammar modelling and in generation of the correct string of words, which we to hope to solve in the future by cooperation with experts in this field.

Current research is focused on the collection of the speech from the telephone line using a Dialogic acquisition board and to test the application for such specific cases.

V. REFERENCES

[1] M. Giurgiu, "Rezultate privind recunoasterea automata a cuvintelor pronuntate izolat in limba romana", *Proceedings of Awareness day on Language Technology,* Bucharest, Romanian Academy, 29-30 January, 1996, pp.127-130.

[2] M. Giurgiu, "Isolated Word Speech Recognition System using both DTW and VQ", *2nd International Conference DMMI*, Bled, Slovenia, pp. 56-60, 1995

[3] M. Giurgiu, T. Muresan, T. Abrudan, "Dialogue Modelling in an Experimental Spoken-based Man Machine Communication System", *Proceedings of International Carpathian Conference,* Slovakia, pp. 197-200, 2000.

[4] M. Giurgiu, "Education and research in Speech Technology at Technical University of Cluj-Napoca", *Proceedings of European Workshop on Education and Research in Speech Communication Sciences in Associated Countries,* 27-28 October, Cluj-Napoca, pp. 57-64, 2000.

[5] A.M. Peinado, J.C. Segura, A.J. Rubio, *"Reconocimiento de Voz Mediante Modelos Ocultos de Markov – Seleccion y Estimacion de Parametros"*, Monografias del Dept. de Electronica No.31, Universidad de Granada, ISBN 84-7951-008-0, pp. 118-161, 1994.

[6] L.R. Rabiner, "A tutorial on Hidden Markov Models and selected applications in speech recognition", *Proceedings of IEEE*, Vol. 77, No. 2, pp. 257-286, 1989.

[7] M. Giurgiu, "Experimental Information Retrieval System based on Romanian Continuous Speech Recognition", *Proc. of 2nd Conference on "Speech Technology and Human-Computer Dialogue"*, 10-11 April, 2003, Editura Academiei Romane, pp161-166, ISBN: 973-27-0963-4.

[8] M. Giurgiu, "On the use of Semicontinuous Hidden Markov Models for Speech Recognition", *Proc. of 2nd Conference on "Speech Technology and Human-Computer Dialogue"*, 10-11 April 2003, Editura Academiei Romane p.187-192, ISBN: 973-27-0963-4.