24

Progress in Speech Recognition for Romanian Language

Corneliu-Octavian Dumitru and Inge Gavat
Faculty of Electronics Telecommunications and Information Technology,
University POLITEHNICA Bucharest
Romania

1. Introduction

In this chapter we will present the progress made in automatic speech recognition for Romanian language based on the ASRS_RL (Automatic Speech Recognition System for Romanian Language) research platform.

Speech recognition is a research domain with a long history, but despite this fact, still open for new investigations and answers to the not yet finally solved questions. This situation can be explained by the difficulty of the task, underlying on the fact that speech is a human product, with a high degree of correlation in content, but with a great variability in the formal manifestation as an acoustic signal. Great difficulties cause also the imperfection of the audio chain and the noises in the environment.

The best-known strategies for speech recognition are the statistical and the connectionist ones, but fuzzy sets can also play an important role.

Based on HMM's the statistical strategies have many advantages, among them being recalled: rich mathematical framework, powerful learning and decoding methods, good sequences handling capabilities, flexible topology for statistical phonology and syntax. The disadvantages lie in the poor discrimination between the models and in the unrealistic assumptions that must be made to construct the HMM's theory, namely the independence of the successive feature frames (input vectors) and the first order Markov process.

Based on artificial neural networks (ANNs), the connectionist strategies for speech recognition have the advantages of the massive parallelism, good adaptation, efficient algorithms for solving classification problems and intrinsic discriminative properties. However, the neural nets have difficulties in handling the temporal dependencies inherent in speech data.

The learning capabilities of the statistical and the neural models are very important, classifier built on such bases having the possibility to recognize new, unknown patterns with the experience obtained by training.

The introduction of fuzzy sets allows on one hand the so-called fuzzy decisions, on other hand the "fuzzyfication" of input data, often more suitable for recognition of pattern produced by human beings, by speaking, for example. In a fuzzy decision, the recognizer realizes the classification based on the degree of membership to a given class for the pattern to be classified, a pattern belonging in a certain measure to each of the possible classes. This

Source: Advances in Robotics, Automation and Control, Book edited by: Jesús Arámburo and Antonio Ramírez Treviño,
ISBN 78-953-7619-16-9, pp. 472, October 2008, I-Tech, Vienna, Austria

relaxation in decision leads to significant enhancements in recognition performances, situation that can also be obtained by looking in a fuzzy way to the input data
The learning capabilities offered by the statistical and the connectionist paradigms and also a "nuanced" inside in the reality of the input and output domains of the speech recognizers contribute to a kind of "human likely" behaviour of this automata.
These three main strategies were applied in our speech recognition experiments and the developed algorithms were incorporate in our research platform, the ASRS_RL. The statistical strategies played the most important role in the development of continuous speech recognition based on HMMs. The neural strategies are applied in form of multilayer perceptrons (MLP) and Kohonen maps (KM) for tasks like vowel or digit recognition. Algorithms for hybrid structures like fuzzy HMM, fuzzy MLP or MLP –HMM are also incorporated in our research platform.
The chapter will be structured as: **Section 2** will describe the "state of the art" of the techniques applied for Romanian language. The **Section 3** is dedicated to the capabilities of the ASRS_RL system. The data bases for training and testing, the feature extraction methods and the learning strategies incorporated in the system are briefly introduced. In **Section 4** are shown and commented the experimental results of the ASRS_RL platform in two tasks: first a task of continuous speech recognition, next a telephone dial experiment. In the first task are applied HMMs with context dependent modelling and the performance is expressed in word recognition rate (WRR) for the recognition experiments and in phrase recognition rate (PRR) for speech understanding experiments. For the second task are implemented HMMs without context dependent modelling, for monophones. **Section 5** presents results and comments for experiments done applying alternative strategies to HMM. These strategies are investigated for simple tasks like vowel and digit recognition. The error rate is evaluated for vowel recognized by using MLP, KM, SVM (Support Vector Machine), fuzzy – MLP and fuzzy–HMM strategies and for digits recognized by the HMM – MLP.
The last Section (**Section 6**) end the chapter with conclusions extracted from the experimental results and with perspectives for our future work.

2. Techniques applied for Romanian language

A brief history of the speech processing techniques for Romanian language was given by the president of the Romanian Academy, Prof. Dr. ing. Mihai Draganescu in the introductory speech to the third conference dedicated to man-machine communication by natural languages, SPED 2003, held at Bucharest: "*Almost 20 years ago, the Romanian Academy organized the first session on the analysis and synthesis of the speech signal. It was a moment of recognition of the activity of Romanian scientists in the domain of speech technology, with works beginning in 1963 (Edmon Nicolau, Inge Weber, Stefan Gavat), 1973 (Aurelian Lazaroiu), 1976 (Eugeniu Oancea, with a volume on analysis and synthesis of speech) and with new papers of Corneliu Burileanu, Horia Nicolai Teodorescu, Eugeniu Oancea, Grigore Stolojanu, Virgil Enatescu and others. Since then the domain evolved from Speech technology to spoken language technology. This had to be foreseen however from the very beginning involving the use of artificial intelligence, both for natural language processing and for acoustic-phonetic processes of the spoken language.*" (Draganescu, 2003).
Continuing this tradition, our team has experimented a set of Computational Intelligence algorithms for Automatic Speech Recognition (ASR) extended also for speech understanding in the frame of the three main paradigms mentioned in the introduction.

We started caring out experiments with neural recognizers for vowels (Grigore et al., 1996), (Grigore et al., 1998), for isolated words (Valsan et al., 1998-a) and word spotting (Valsan et al., 1998-b). Then we introduced the HMM modelling in our experiments, first for isolated word recognition, further for continuous speech. The refinement of the HMM based recognizers was our permanent goal, and today we can report a system with reasonable performances in continuous speech recognition. for Romanian language (Gavat et al. 2003, Dumitru et al,, 2006, Dumitru et al,2007, Gavat et al, 2008)
But also very attractive in order to improve recognition performance was the development of hybrid systems. One of the firsts was a neuro-statistic hybrid, in different ways proposed by many authors (Bourlard et al., 1990), (Richard et al., 1991), (Lippmann et al., 1993), (Bourlard et al., 1994). We realized such a system for Romanian language as a variant consisting in a HMM assisted by a MLP as *a posteriori* probability estimator (Gavat et al., 1996-a). This combination, realized between components with complementary properties (i.e. the good integrating time successions HMM and the high discriminative MLP), leads to an increase in recognition rates that exceeds 2%. Furthermore, as suggested in (Juang et al., 1992), (Reichl et al., 1994) we refined our hybrid structure by a supplementary discriminative training (Gavat et al., 1998), obtaining a further improvement with around 3.5%. We applied also fuzzy variants for neural and hard classifiers. By applying fuzzy decisions, we have obtained improved performances in classical algorithms like k-NN or ISODATA (Gavat et al., 1996-a) and also in neural recognizers realized with MLP or self-organizing feature maps (Gavat et al., 1997), (Grigore et al., 1999). To apply fuzzy concepts to HMMs was the next natural step to be followed in our studies, experimenting in speech recognition the hybrid called Fuzzy-HMM (FHMM). (Gavat et al., 2001-a), (Gavat et al., 2001-b).This developed hybrid systems were applied until now only in tasks for vowel and isolated word recognition, the implementation for continuous speech recognition remaining as a future desiderate.

3. ASRS_RL system

Speech is a communication modality, so how the communication is accomplished is a question of high interest. Traditionally (Juanhg et al., 2000) it can be admitted that the speech communication chain is organized in four stages: detection of acoustic-phonetic cues to form words, syntactic and grammatical analysis to build sentences, semantic evaluation to determine the possible meanings of a sentence and pragmatic evaluation in order to select the convenient meaning.
An ASR (Automatic Speech Recognition) system is a machine that performs in two stages (the first two) and an ASRU (automatic speech recognition and understanding) system is a machine for which all stage are necessary. But, a lot of work must still be done to acquire the whole knowledge necessary to entirely realize this last task.

3.1 Capabilities of the ASRS_RL

Automatic Speech Recognition System for Romanian Language (ASRS_RL) is a viable structure, offering multiple options for the tasks that can be performed, for the features that can be used to characterize speech frames, for the learning strategies applied, for the speech databases necessary in training and testing stage. It allows also multilingual (Romanian, English, German and can be extended to other languages) and multimodal (speech and image entries) experiments. One window of the MATLAB interface to choose the working options is shown in Fig.1

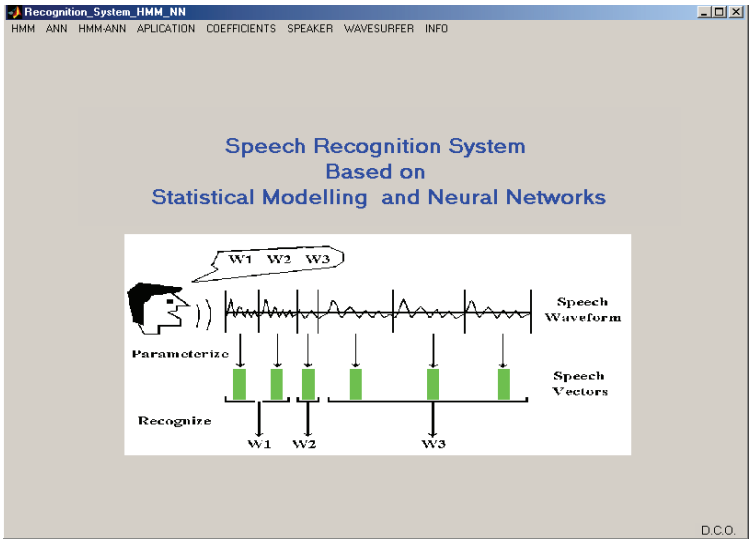


Fig. 1. The recognition system ASRS_RL

a. The possible tasks that can be performed with this system are: continuous speech recognition; digit recognition and vowel recognition.
b. The possible databases for training/testing are: CDRL (Continuous Database for Romanian Language); DDRL (Digit Database for Romanian Language); DDEL (Digit Database for English Language); VDRL (Vowel Database for Romanian Language). The data are sampled by 16 kHz, quantified with 16 bits, and recorded in a laboratory environment (for CDRL, DDRL and VDRL) (Dumitru et al., 2007). For DDEL the database used is the AMP database (www.amp.ece. cmu.edu/_download /Intel/feature_data.html). The DDRL and DDEL are available for three different SNR: 19dB, 25dB and 30dB.
c. The phonetic transcription for each database was created based on the SAMPA standard (Sampa).
d. The features vectors to characterize speech frames (and the number of coefficients for each type) can be constituted (Huang at al., 2001) (Hermansky, 1990) by: formants, obtained from cepstral analysis; mel frequency cepstral coefficients (MFCC) with or without first (D–delta) and second (A–acceleration) order variations, obtained from perceptual cepstral analysis; linear prediction coefficients (LPC) obtained by linear prediction; perceptual linear prediction (PLP) coefficients obtained by perceptual linear prediction. To this features frame energy (E), zero crossing rate (Z_0), and other information can be added.
e. The possible learning strategies (Goronzy, S. (2002) that can be chosen in the system in order to build acoustical models are: hidden Markov models (HMM) for monophones (phonemes without context) and triphones (phonemes with left and right context) (Young, 1992); Support Vector Machines (SVM); artificial neural networks in form of multilayer perceptrons (MLP) and Kohonen maps (KM); hybrid structures like the neuro–statistical hybrid (HMM–MLP), or the fuzzy variants for the MLP or the HMM.

3.2 Databases

In this sub-section we present our databases for Romanian language with the principal characteristics, as well as a "state of the art" on the most important international databases.
Based on the investigation of the existing standardized databases for different languages, the most important databases are: for English - TIMIT, WSJO, AMP, etc; for German – Vermobil; for French - NEOLOGOS, for Spanish - Entropic Latino40; for Hungarian – OASIS; but there are also some multilingual databases like CSLU, SpeechDat, Appen , etc.
TIMIT Acoustic-Phonetic Continuous Speech Corpus was created by DARPA-ISTO (Defense Advanced Research Projects Agency - Information Science and Technology Office), and the text of this corpus was realized by MIT (Massachusetts Institute of Technology), SRI (Stanford Research Institute) and TI (Texas Instruments).
The database contains 630 speakers from 8 region of the United States of America, 438 male speakers and 192 female speakers, each of them spoken 10 phrases (Timit).
WSJO (Wall Street JOurnal) is the biggest and variat corpus for English, created by ARPA (Advanced Research Projects Agency). The phrases are from *Wall street Journal*. The initial number of words was 37 million, but finally reduced to 64000 words, the most frequent words in the journal. There are two corpora:J0 (84 speakers, 42 male speakers and 42 female speakers, 7193 phrases) and WSJ1 (200 speakers, 100 male speakers and 100 female speakers, 29320 phrases) (Douglas et al., 1992); WSJCAM0 (for British English created at the Cambridge University in 1995; the corpus was selected from the *Wall street Journal* between 1987-1989; the speakers are: between 18-23 years old, 46, between 24-28 years old, 30, between 29-40 years old, 7, more then 40 years old, 9; each of them uttered 110 phrases) (Robinson et al., 1995).
AMP (Advanced Multimedia Processing) is a database created at the Carnegie Mellon University in Advanced Multimedia Processing Laboratory. The database is sampled by 44,1kHz and quantified with 16 bits, mono system (Amp).
The database contains 10 speakers (7 ale speakers and 4 female speakers) each spoken 78 words for 10 times.
CSLU (Centre of Spoken Language Understanding) is a database crated by Oregon Health & Science University. This database contains many corpora: *22 languages* (a database for 22 language); *Alphadigits* (3025 speakers uttered digits and stream of digits, 78044 audio files); *Kids Speech* (this database is created to help the children with hearing problems; about 100 kids); *Multilanguage Telephone Speech* (for 11 language and contains 2052 speakers, spoken fixed phrases or free continuous speech); etc (Cslu).
SpeechDat is created to develop systems for *voice driven teleservice* and *speech interfaces*.
SpeechDat(M) is a database for 7 languages from West Europe (SpeDatM), 1000 speakers; **SpeechDat(II)** is for 11 languages, official languages of the Europeen Union, each of this databases contains between 500-5000 speakers; **SpeechDat(E)** is for 5 languages from Est Europe, contains between 1000-2500 speakers (SpeDatE).
Appen is a collection of databases for 20 languages for different applications like: telephony, broadcast audio, desktop, pocket PC (Appen). The recorded files are realized in different location: car, studio, office, street or public place. We present the database for one application, telephony. The database is for spoken English in Australia, recorded in office environment. The first part of the database contains 500 speakers and immigrant speakers, each reading 165 phrases; the second part contains 1000 speakers (Australian speakers) reading 75 phrases.

Others databases are for spoken English in Canada (contains 49 speakers each uttered about 99 phrases in office environment) and for spoken French in Canada (contains 48 speakers each spoken about 100 phrases) in office, car, home environment and public place. The phrases are constructed by digits, phonemes, names places jobs, answers yes / no), commands.

Neologos was recorded on the French telephone network (Charlet et al., 2005) and two databases were obtained. The first database is IDIOLOGOS (with two sub-databases: *Bootstrap* – contains 1000 speakers (470 male speakers and 530 female speakers) with ages from 18 to 61 and up to 61, with a very good distribution, each of them uttered 45 phrases; *Eingenspeakers* – contains 200 speakers (97 male speakers and 103 female speakers) for different age, from 18 to 61 and up to 61, each of them spoken 45 phrases/appeal telephonic, 10 appeal are considered).

The second database PAIDIALOGOS is for the children between 7 to 16 years, with one child having more than 16 years. The number of the children is 1010 (510 male children and 500 female children), from different regions of France. The database contains: digits, numbers of credit cards, names of cities, etc. The data are sampled by 8 kHz, quantified with 8 bits.

Vermobil (Elra) was created between 1993-1996 by the research institutions and companies: Institut für Phonetik und digitale Sprachverarbeitung (IPDS) - Kiel, Institut für Kommunikation und Phonetik (IKP) – Bonn, Institut für Phonetik und sprachliche Kommunikation - München and Universität Karlsruhe. The database was divided in two databases: one for training (contains 12000 turns) and one for testing (1800 turns), with different dialects (12 dialects). One turn is equal with around 22.8 words.

Entropic Latino40 (Entro40) was recorded in 1994, in Palo Alto – California, the speakers are Spanish, from Latino America. Database contains 5000 phrases, spoken by 40 speakers (20 male speakers and 20 female speakers) each of them reading 125 phrases. The database is sampled by 16kHz, quantified with 16 bits and recorded in office environment. All the speakers are not specialized speakers and are between 18-59 years old.

OASIS (Kocsor et al., 1995) was created at Hungarian Academy of Sciences (Research Group on Artificial Intelligence). The database contains speech files from 26 speakers (16 male speakers and 10 female speakers) each of this speakers uttered the digits from 0-9, from 10-100, 1000, 1000000 and variants of composed numbers. The database is sampled by 22,05kHz, quantified with 16 bits and recorded in office environment.

Because in Romanian language no standard database is available, we create our database for our recognition tasks. The databases created are: OCDRL (Old Continuous Database for Romanian Language), CDRL (Continuous Database for Romanian Language), DDRL (Digit Database for Romanian Language) and VDRL (Vowel Database for Romanian Language). The data are sampled by 16 kHz, quantified with 16 bits, and recorded in a laboratory environment (Dumitru, 2006).

OCDRL: for continuous speech recognition, is constituted by two databases: the training database contains 500 phrases, spoken by 10 speakers (8 males and 2 females), each speaker reading 50 phrases; the testing database contains 350 phrases spoken by the same speakers. The phrases from the database are for an application of telephone dial.

A statistical investigation of the existing phonemes in the training database, based on phonetic transcription of the SAMPA standard (Sampa), is shown in the Table 1. We can see easy that the preponderant are the vowels (i, i_O, e, a, @, o, u, 1) and semi-vowels (j, e_X, w,

o_X). The distribution of phonemes is very similar to that given by linguists for Romanian language, so that we trusted the data base for the acoustical modelling.

a	**55**	l	21	p	9	b	3	ts	4
sp	102	**j**	**18**	**e_X**	**4**	S	2	g	3
tS	10	f	10	z	4	**1**	**5**	dZ	3
e	**45**	n	29	@	**17**	d	14	**o_X**	**3**
s	24	**o**	**36**	r	35	k_	3	g_	3
t	32	k	8	m	12	**i_O**	**4**	h	4
i	**27**	**u**	**23**	**w**	**6**	v	3	Z	3

Table 1. Number of phonemes existing in the training OCDRL database

CDRL: for continuous speech recognition, the database is constituted for training by 3300 phrases, spoken by 11 speakers (7 males and 4 females), each speaker reading 300 phrases, and for testing by 220 phrases spoken by the same speakers, each of them reading 20 phrases. The phrases are from: information, telecommunication, geography, history, and sports domain. The training database contains over 3200 distinct words; the testing database contains 900 distinct words.

In order to carry out our experiments the database was reorganized as follows: one database for male speakers (MS), one database for female speakers (FS) and one mixed database for male and female speakers (MS and FS). In all cases we have excluded one MS and one FS from the training and used for testing.

Conducting the same statistical investigation of the phonemes distribution, (Table 2), we obtained very similar results with that obtained for the OCDRL data base.

a	**85**	l	48	p	23	b	9	ts	9
sp	152	**j**	**18**	**e_X**	**7**	S	5	g	6
tS	14	f	13	z	10	**1**	**9**	dZ	6
e	**95**	n	41	@	**27**	d	23	**o_X**	**5**
s	38	**o**	**45**	r	64	k_	5	g_	5
t	51	k	18	m	29	**i_O**	**7**	h	7
i	**53**	**u**	**41**	**w**	**10**	v	6	Z	5

Table 2. Number of phonemes existing in the training CDRL database

DDRL: for digit recognition, the database contains speech data from 9 speakers (6 males and 3 females) each speaker reading 9 digits (unu, doi, trei, patru, cinci, şase, şapte, opt, nouă).

VDRL: for vowel recognition, the database contains speech data from 19 speakers (9 males and 10 females) each reading the same 5 vowels (a, e, i, o, u).

The database based on **formants** contains for the training 500 formant vectors, 100 for each vowel and for the testing 250 formant vectors, 50 for each vowel.

3.3 Speech analysis

First step in all recognition tasks is speech analysis, where the speech signal is processed in order to obtain important characteristics, further called features. By features extraction data compression is done, so that the amount of data used for comparisons is greatly reduced and thus, less computation and less time is needed for comparisons (Huang et al., 2001).

Our features extraction is based on perceptual cepstral coding and perceptual linear predictive coding, methods that will be presented further.
The block scheme of the acoustical processor is shown in Fig. 2. Few blocks are common in both linear prediction and cepstral coding (Gavat et al., 2003). The first block in the scheme is the frame blocking, used because speech is fundamentally a non-stationary signal, so we cut short fragments of the speech signal, which are called frames and the speech is approximated as a quasi-stationary random process during a frame. Then we passed each frame through a Hamming window. We can compute at this time the energy of each frame and we can use the energy set of coefficients in the recognition process for more accuracy.

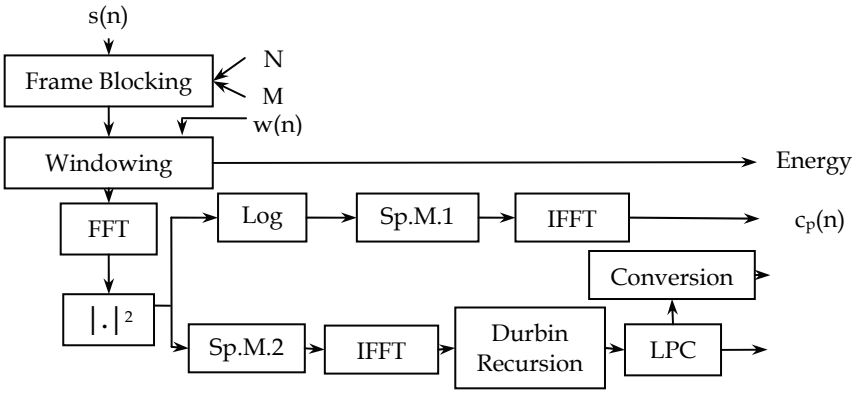


Fig. 2. The block scheme of the acoustical processor

Cepstral analysis is a very reliable method to speech analysis and it can be realized applying FFT (Fast Fourier Transform) to the blocked and windowed time discrete signal *s(n)*. After that, the modulus of the signal is calculated and the logarithm is taken, the result being proportional in fact to the power spectrum of the speech signal. Applying IFFT (Inverse Fast Fourier Transform) leads to the real cepstrum.
Cepstral coefficients, are obtained by re-sampling of the real cepstrum; alone or in addition with the energy *E*, and/or the first and second order differences constitutes a feature vector successfully applied in speech recognition
Mel-frequency cepstral coefficients are calculated from the power spectrum after a spectral manipulation Sp.M.1 in form of a filter bank that is a model for the critical band perception of the human cochlea. Often, in noisy environments the first (D–Delta) and second (A–Acceleration) order variations of the MFCC are applied to enhance the feature vector.
Formants can be calculated from the smoothed spectrum, resulted applying FFT to the windowed cepstrum near the origin.
Linear Predictive coding requires computation of the autocorrelation coefficients possible to be obtained according to the Wiener-Hinchin theorem, by applying IFFT to the power spectrum.
Linear Prediction coefficients are obtained by low order all-pole modelling. The Levinson–Durbin recursive algorithm is used to solve the Yule–Walker equations.

$$\begin{bmatrix} R(1) & R(2) & \dots & R(N) \\ R(2) & R(3) & \dots & R(N-1) \\ \dots & \dots & \dots & \dots \\ R(N) & R(N-1) & \dots & R(1) \end{bmatrix} \times \begin{bmatrix} A(1) \\ A(2) \\ \dots \\ A(N) \end{bmatrix} = \begin{bmatrix} -R(2) \\ -R(3) \\ \dots \\ -R(N+1) \end{bmatrix} \quad (1)$$

where $R(n)$ are the autocorrelation coefficients, and $A(n)$ are the all-pole model coefficients (the predictor), and $A(1) = 1$; through the conversion block the lasts can be converted in other types of coefficients like the reflection coefficients or the LPC cepstral coefficients.

Perceptual Linear Prediction (PLP) coefficients are obtained by spectral manipulation Sp.M.2, displayed in Fig.3. The PLP analysis method (Hermansky, 1990) is more adapted to human hearing, in comparison to the classic Linear Prediction Coding because of the critical band processing of the analysed signal. The LP all-pole model approximates power distribution equally well at all frequencies of the analysis band. This assumption is inconsistent with human hearing, because beyond 800 Hz, the spectral resolution of hearing decreases with frequency; hearing is also more sensitive in the middle frequency range of the audible spectrum (Dumitru et al., 2006).

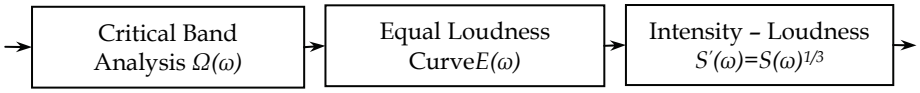


Fig. 3. Block representation for Sp. M. 2

The power spectrum is computed as follows:

$$P(\omega) = \mathrm{Re}(S(\omega))^2 + \mathrm{Im}(S(\omega))^2 \quad (2)$$

The first step is a conversion from frequency to bark, which is a better representation of the human hearing resolution in frequency. The bark frequency corresponding to an audio frequency is:

$$\Omega(\omega) = 6 \ln\left(\frac{\omega}{1200\pi} + \left(\left(\frac{\omega}{1200\pi} \right)^2 + 1 \right)^{0.5} \right) \quad (3)$$

The resulting warped spectrum is convoluted with the power spectrum of the critical band-masking curve, which acts like a bank of filters centred on Ω_i.

The spectrum is pre-emphasized by an equal loudness curve, which is an approximation to the non-equal sensitivity of human hearing at different frequencies, at about 40dB level. A filter having the following transfer function gives the curve:

$$E(\omega) = \frac{(\omega^2 + 56.8 \times 10^6)\omega^4}{(\omega^2 + 6.3 \times 10^6)^2 \times (\omega^2 + 0.38 \times 10^9)} \quad (4)$$

The last operation prior to the all-pole modelling is the cubic-root amplitude compression (Intensity – Loudness Conversion), which simulates the non-linear relation between the intensity of sound and its perceived loudness. Together with the psychophysical equal-

loudness pre-emphasis, this operation also reduces the spectral amplitude variation of the critical-band spectrum (Gavat et al., 2008).

3.4 Learning strategies

In the same way into which a human learns to perceive speech from examples by listening, the automatic speech recognizer does apply a learning strategy in order to decode the pronounced word sequence from a sequence of elementary speech units like phonemes, with or without context. By learning are created models for each elementary speech unit, serving in the comparisons to make a decision about the uttered speech unit. We will describe further the learning strategies implemented in our research platform: hidden Markov models (HMM), artificial neural networks (ANN) in form of multilayer perceptrons (MLP) or Kohonen maps (KM). Some hybrid learning strategies are also implemented in form of a HMM-ANN combination and a fuzzy perceptron or fuzzy HMM.

HMMs are finite automata, with a given number of states; passing from one state to another is made instantaneously at equally spaced time moments. At every pass from one state to another, the system generates observations, two processes taking place: the transparent one represented by the observations string (features sequence), and the hidden one, which cannot be observed, represented by the state string (Gavat et al., 2000).

In speech recognition, the left - right model (or the Bakis model) is considered the best choice. For each symbol, such a model is constructed; a word string is obtained by connecting corresponding HMMs together in sequence (Huang et al., 2001).

For limited vocabulary, word models are widely used, since they are accurate and trainable. In the situation of a specific and limited task they become valid if enough training data are available, but they are typically not generalizable. Usually for not very limited tasks are preferred phonetic models based on monophones (which are phonemes without context), because the phonemes are easy generalizable and of course also trainable.

Monophones constitute the foundation of any training method and we also started with them (for any language). But in real speech the words are not simple strings of independent phonemes; these phonemes are affected for the immediately neighboring phonemes by co-articulation. This monophone models are changed now with triphone models (which are phonemes with context) that became actually the state of the art in automatic speech recognition for the large vocabularies (Young, 1992).

A triphone model is a model that takes into consideration the left and the right context of the phonemes. Based on the SAMPA (Speech Assessment Methods Phonetic Alphabet) in Romanian language there are 34 phonemes; the number of necessary models (triphones) to be trained is about 40000, situation which is unacceptable.

In the continuous speech recognition task we modelled only internal – word triphones and we adopted the state tying procedure, conducting to a controllable situation. If triphones are used in place of monophones, the number of needed models increases and it may occur the problem of insufficient training data. To solve this problem, tying of acoustically similar states of the models built for triphones corresponding to each context is an efficient solution (Fig.4).

For example, in Fig. 4b four models are represented for different contexts of the phoneme *"a"*, namely the triphones *"k – a + S"*, *"g – a + z"*, *"n – a + j"*, *"m – a + j"*. In Fig. 4c and 4d the clusters formed with acoustically similar states of the corresponding HMMs are represented.

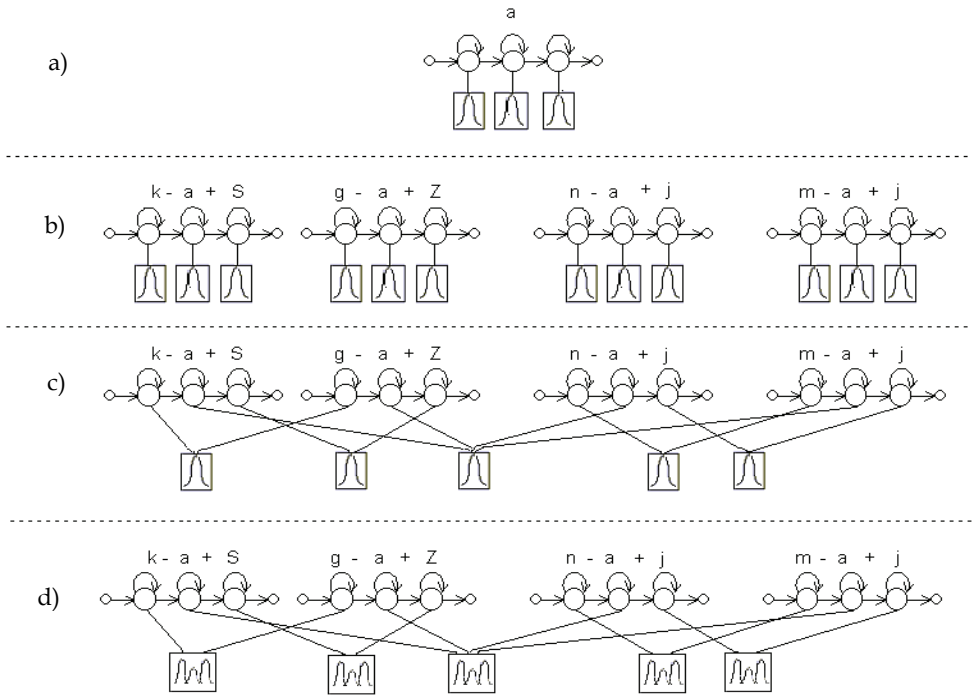


Fig. 4. Different models for triphones around the phoneme "a".

The choice of the states and the clustering in phonetic classes are achieved by mean of phonetic decision trees. A phonetic decision tree is a binary tree (Fig.5), where for each node of the tree questions are associated concerning the contexts of the phoneme (the number of the selected questions is 130 based on the knowledge about phonetic rules for our language).

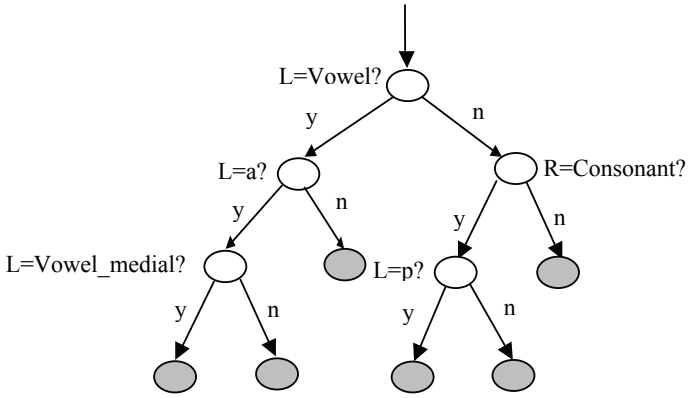


Fig. 5. Phonetic tree for phoneme "m" in state 2.

The questions are chosen in order to increase the log likelihood of the data after splitting. Splitting is stopped when increasing in log likelihood is less than an imposed threshold. In the leaf nodes are concentrated all states having the same answer to the question made along the corresponding path.

Fuzzy – HMM: The generalized model $\overline{\lambda} = (\overline{A}, \overline{B}, \overline{\pi})$ can be characterized by the same parameters (Huang et al., 2001) like the classical, well known model. The major difference in the fuzzy variant, is the interpretation of the probability densities for the classical HMM, as fuzzy densities. On this way the probabilistic similarity measure applied in the classical HMM is replaced by a more suitable fuzzy similarity measure.

The succession of feature vectors, called the observation sequence, *O*, produces the state sequence *S* of the model, and, visiting for example at the moment *t+1* the state $q_{t+1} = S_j$, the symbol b_j is generated. The corresponding symbol fuzzy density $b_j(O_t)$ measures the grade of certainty of the statement that we observed O_t given that we are visiting state S_j .To perform classification tasks, the fuzzy similarity measure must be calculated. Based on the fuzzy forward and backward variables, a fuzzy Viterbi algorithm is proposed in (Mahomed et al., 2000) for the case of the Choquet integral with respect to a fuzzy measure and multiplication as intersection operator.

The fuzzy formulation of the forward variable *a*, bring an important relaxation in the assumption of statistical independence.

The joint measure $\overline{\alpha}_{\Omega_Y}\left(\{O_1,...,O_t\}\times\{y_j\}\right)$ can be written as a combination of two measures defined on $O_1, O_2, ..., O_t$ and on the states respectively, no assumption about the decomposition of this measure being necessary, where $Y = \{y_1, y_2, ..., y_N\}$ represent the states at time *t+1* (Ω is the space of observation vectors).

For the standard HMM, the joint measure $P(O_1, O_2, ..., O_t, q_{t+1} = S_j)$ can be written as the product $P(O_1, O_2, ..., O_t) \cdot P(q_{t+1} = S_j)$, so that two assumptions of statistical independence must be made: the observation at time *t+1*, O_{t+1}, is independent of the previous observations $O_1, O_2, ..., O_t$ and the states at time *t+1* are independent of the same observations, $O_1, O_2, ..., O_t$.

These conditions find a poor match in case of speech signals and therefore we hope in improvements due to the relaxation permitted by the fuzzy measure.

Training of the generalized model can be performed with the re-estimation formulas also done in (Mahomed et al., 2000) for the Choquet integral. For each model we have trained with the re-estimation formulas the corresponding generalized models, GHMMs, with 3-5 states, analog to the classical case.

After the training, we have calculated the fuzzy measure $\overline{P}\left(O/\overline{\lambda}\right)$, with the fuzzy Viterby algorithm and made the decisions for recognition in the same manner like for the classical HMM: the correct decision corresponds to the model for which the calculated measure has a maximum.

MLP is the most common ANN architecture used for speech recognition. Typically, the classical MLPs have layered feed–forward architecture, with an input layer, one or more

intermediate (hidden) layers, and one output layer. The structure without hidden layer is called Boolean network and is a simple perceptron.
Each layer computes a set of linear discriminative functions, followed by a non-linear function, which is often a sigmoid function.
The numbers of neurons in the hidden layer was experimentally determined, trying to achieve an optimum between the following two opposite requirements: (a) lower computing volume and more rapid process of convergence in the learning period; (b) better performances from the correct classification of input patterns percentage.
In the learning phase are determined the optimum values for weights (Goronzy, 2002), (Valsan et al., 2002) connecting the pairs of neurons from the adjoint layers in the input–output direction using the Back-Propagation algorithm.
Introducing a fuzzy processing of the input layer of the MLP (**fuzzy-MLP**) is a solution to improve the MLP performances.
The structure of the fuzzy neural network (Fig. 6) like the classical one is composed from an input layer, a hidden layer and an output layer.

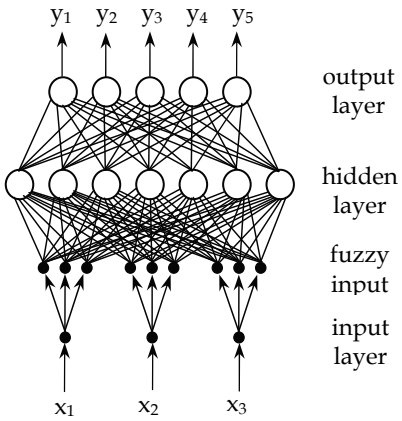


Fig. 6. Schematic diagram of fuzzy MLP

First, the input values are described through a combination of 3 membership values in the linguistic properties set: low, medium and high. For doing this the π membership function is used:

$$\pi(r,c,\lambda)=\begin{cases}2(1-\|r-c\|/\lambda)^2 & for\ 0\le\|r-c\|\le\lambda/2\\ 1-2(\|r-c\|/\lambda)^2 & for\ 0\le\|r-c\|\le\lambda/2\\ 0 & otherwise\end{cases} \quad (5)$$

where: λ>0 is the radius of the π function with c as the central point, || . || denotes the Euclidian norm.
For each component F_{ji} of the input vector F_j the parameters of the π membership function for each linguistic property: low (l), medium (m) and high (h) are computed using the relations:

$$\begin{aligned}
\lambda_{m(Fij)} &= (F_{ji\max} - F_{ji\min})/2 \\
c_{m(Fij)} &= F_{ji\min} + \lambda_{m(Fij)} \\
\lambda_{1(Fji)} &= (c_{m(Fji)} - F_{ji\min})/f_{dn} \\
c_{1(Fji)} &= c_{m(Fji)} - 0.5\lambda_{1(Fji)} \\
\lambda_{h(Fji)} &= (F_{ji\max} - c_{m(Fji)})/f_{dn} \\
c_{h(Fji)} &= c_{m(Fji)} + 0.5\lambda_{h(Fji)}
\end{aligned} \tag{6}$$

where $F_{ji\max}$, $F_{ji\min}$ denote the upper and lower bounds of the observed range of feature F_{ji} and f_{dd} is a parameter controlling the extent of overlapping.

After this, the structure of the fuzzy neural network, like the classical one, is composed from a hidden layer and an output layer .

The output vector is defined as the fuzzy class membership values. The membership value of the training $F_i = (F_{i1}\, F_{i2}\, \ldots F_{in})^t$ to class C_k is computed using:

$$\mu_k(F_i) = 1/(1 + z_{ik}/f_d))^{f_c} \tag{7}$$

where: f_d, f_c are constants controlling the amount of fuzziness in the class-membership set, z_{ik} is the weighted distance between the input vector F_i and the mean $O_k = (O_{k1}\, O_{k1} \ldots O_{k1})^t$ of the *k*-th class, defined as:

$$z_{ik} = \sqrt{\sum_{j=1} [(F_{ji} - O_{kj})/v_{kj}]^2} \tag{8}$$

where: v_{kj} is the standard deviation of the *j*-th vectors' component from the C_k class.

In the training stage, the Back-Propagation algorithm is used to determine the weights which minimized the mean square error (*mse*) between the real output d_j and the desired one y_j :

$$mse = \sum_{\substack{j=1 \\ F \in train}}^{n} (\sum (d_j - y_j)^2) \tag{9}$$

Kohonen maps are competitive neural networks with topological character. The setting up of the winner neurons at output is done with keeping the topological relations between the input vectors. That is the reason for which this neural network is successfully used in pattern recognition (Fig. 7).

It was applied the following structures to recognize vowels: (1) the input layer with 3 neurons, corresponding to the three formant frequencies; (2) three case for the output layer: one-dimensional with 25 neurons, bidimensional with 4x4 neurons, 5x5 neurons, 6x6 neurons, and toroidal with 25 neurons.

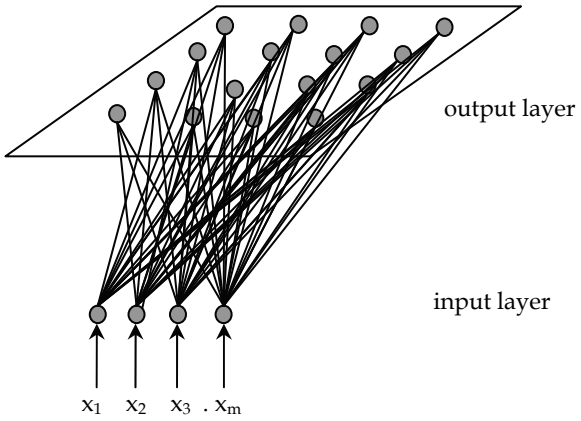


Fig. 7. Schematic diagram of Kohonen maps

In the learning phase (Gold et al., 2002) the structures are trained and the weights of the networks are established in two steps: (a) the determination of the winner neurons; (b) the adaptation of the weights for the winner neurons and for the neurons existing in a certain neighbourhood. In this step important are: the neighbourhood dimension *r(t)* decreasing during the training, and the learning rate $\eta(t)$ following in our experiments one of the laws:

$$\eta(t) = t^{-1} \text{ or } \eta(t) = t^{-1/2} \tag{10}$$

Hybrid systems: the HMM-based speech recognition methods make use of a probability estimator, in order to approximate emission probabilities $p(x_n/q_k)$, where x_n represents the observed data feature, and q_k is the hypothesized HMM state (Rabiner, 1989). These probabilities are used by the basic HMM equations, and because the HMM is based on a strict formalism, when the HMM is modified, there is a great risk of losing the theoretical foundations or the efficiency of the training and recognition algorithms. Fortunately, a proper use of the MLPs can lead to obtain probabilities that are related with the HMM emission probabilities.

In particular, MLPs can be trained to produce the *a posteriori* probability $p(x_n/q_k)$, that is, the *a posteriori* probability of the HMM state given the acoustic data, when each MLP output is associated with a specific HMM state. Many authors have shown that the outputs of an ANN used as described above can be interpreted as estimates of *a posteriori* probabilities of output classes conditioned by the input, so we will not insist on this matter, but we will mention an important condition, useful for finding an acceptable connectionist probability estimator: the system must contains enough parameters to be trained to a good approximation of the mapping function between the input and the output classes.

Thus, the *a posteriori* probabilities that are estimated by MLPs can be converted in emission probabilities by applying Bayes' rule (11) to the MLP outputs:

$$\frac{p(x_n / q_k)}{p(x_n)} = \frac{p(q_k / x_n)}{p(q_k)} \tag{11}$$

That is, the emission probabilities are obtained by dividing the *a posteriori* estimations from the MLP outputs by estimations of the frequencies of each class, while the scaling factor $p(x_n)$ is considered a constant for all classes, and will not modify the classification.

This was the idea that leads to hybrid neuro-statistical methods, that is, hybrid MLP - HMM methods (Fig. 8), applied for solving the speech recognition problem.

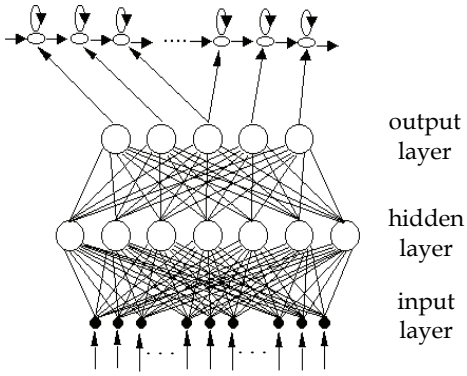


Fig. 8. Schematic diagram of HMM-MLP

4. Speech recognition using ASRS_RL system based on statistical strategies

4.1 Continuous speech recognition

To assess the performance of our ASRS_RL system we initiated comparative tests on the OCDRL and CDRL database for the word recognition rate (WRR) under different conditions for feature extraction, like perceptive cepstral analysis (MFCC_D_A–mel frequency cepstral coefficients with the corresponding first and second order variations –> having 36 coefficients), and linear prediction (LPC–Linear Prediction Coding –> having 12 coefficients). The evaluations are made in the case of acoustical model based on HMM with monophones (phonemes without context) (Dumitru, 2006).

In order to carry out our experiments about speaker independence, the CDRL database was reorganized as follows: one database for male speakers (MS), one database for female speakers (FS) and one database for male and female speakers (MS&FS). In all cases we have excluded one MS and one FS from the training and used for testing.

The words for both databases (OCDRL and CDRL) are chosen taking into account the characteristics of the Romanian language so that the phrase contains enough phonemes to create the models.

The results obtained with the two databases are summarized in Table 3. After that, chosen the second database (CDRL) other evaluations are made.

	OCDRL		CDRL	
	MFCC_D_A	LPC	MFCC_D_A	LPC
Monophone	96.10%	73.80%	57.44%	26.10%

Table 3. WRR for the two cases of databases, OCDRL and CDRL

Trying to create a multi-domain recognition system, we extended the numbers of speakers and of spoken phrases for each speaker (extending OCDRL to CDRL) introducing vocabulary from other domains. In the same time we introduced for WRR evaluation new feature extraction methods, like PLP and new acoustical models based on the HMM with triphones (phonemes with context) (Young et al., 1994).

Table 4 shows the comparative results (average WRR) for three feature extraction methods, MFCC_D_A, LPC and PLP and for two modelling situations, based on monophones and triphones.

Database	Type	MFCC_D_A	LPC	PLP
CDRL	Monophone	57.44%	26.10%	47.00%
	Triphone	78.24%	51.50%	70.11%

Table 4. WRR for monophone and triphone modelling and for three feature extraction methods

The next experiments performed on a Romanian language corpus prove that context-dependent models perform better than context-independent models. The recognition system was trained with 3000 phrases collected from ten speakers (more than 3000 distinct words). Gaussian output probability distribution was assumed for the 36 mel-frequency cepstrum coefficients (12 MFCC and first and second order variation).
Firstly, the 34 context-independent models (monophones) were trained, and the system was tested with an unenrolled speaker. The testing utterances contained distinct words and a loop-grammar was assumed, i.e. any word could occur after any word, anytime.
The core of our experiments is the construction of the decision tree for each state of the triphones derived from the same monophone. The monophones were cloned initially, and the resulted triphones were trained by embedded Baum-Welch procedure. Then, the decision tree was build for different thresholds (TL) in terms of log-likelihood resulting different size systems (example of the evolution of the log-probability values – in the training phase is presented in Fig. 9). The results are presented in Table 5. For a small threshold (TL) of 300, the trees are big and the system is large having 2954 tied states with a huge number of parameters. For a big threshold of 6000, the trees are much smaller, implying a great reduction in the system size, from 7521 triphone states to 416 states, (5.5% remained size) while the performance is degrading with less than 1%. In Table 5, are given also the word recognition rate (WRR), the accuracy.

TL	Initial states / final states	Remained size	WRR
300	7521 / 2954	39.30%	90.14%
900	7521 / 1448	19.30%	89.60%
1200	7521 / 1164	15.50%	90.31%
1800	7521 / 908	12.10%	90.51%
2400	7521 / 747	9.90%	90.02%
3000	7521 / 643	8.50%	89.97%
3600	7521 / 573	7.60%	90.07%
4200	7521 / 522	6.90%	89.79%
4800	7521 / 480	6.40%	89.60%
5400	7521 / 446	5.90%	88.85%
6000	7521 / 416	5.50%	88.75%

Table 5. The results obtained for different thresholds for constructing the phonetic trees

Now, we evaluate the WRR for the CDRL database taking into account the following situations: triphone modelling/monophone modelling; gender based training/mixed training and LPC and PLP/MFCC coefficients.
The speech files from these databases were analysed in order to extract the interesting features. The feature extraction methods used are based on LPC, PLP and MFCC.

The results obtained in the experiments realized under these conditions are summarized in the next three Tables.

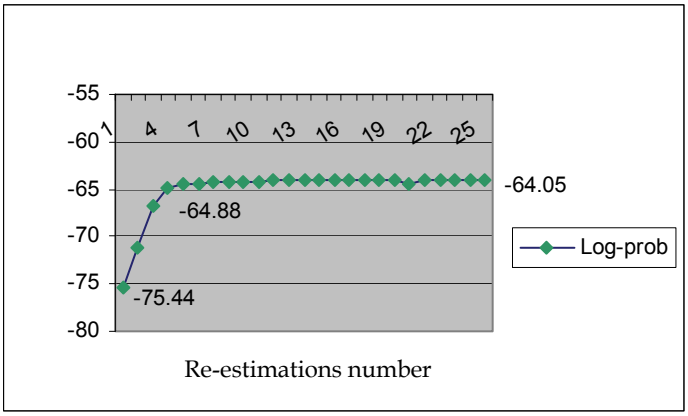


Fig. 9. Log-probability evolutions

Training MS	Type	WRR		
		MFCC_D_A	LPC	PLP
Testing MS	Monophone	56.33%	30.85%	34.02%
	Triphone	81.02%	49.73%	68.10%
Testing FS	Monophone	40.98%	23.23%	25.12%
	Triphone	72.86%	47.68%	59.00%

Table 6. WRR training MS and testing MS and FS

Training FS	Type	WRR		
		MFCC_D_A	LPC	PLP
Testing MS	Monophone	53.56%	26.72%	23.78%
	Triphone	69.23%	49.73%	53.02%
Testing FS	Monophone	56.67%	31.11%	34.22%
	Triphone	78.43%	61.15%	58.55%

Table 7. WRR training FS and testing MS and FS

Training MS and FS	Type	WRR		
		MFCC_D_A	LPC	PLP
Testing MS	Monophone	57.44%	26.10%	47.00%
	Triphone	78.24%	51.50%	70.11%
Testing FS	Monophone	49.89%	24.06%	41.22%
	Triphone	74.95%	50.49%	69.65%

Table 8. WRR training MS&FS and testing MS and FS

The following comments can be made:

a. for 12 LPC coefficients the word recognition rates are low: 30.85% (monophone) training and testing with MS and 49.73% (triphone); 31.11% (monophone) training and testing with FS and 61.15% (triphone); 26.10% (monophone) training MS and FS and testing with MS and 51.5% (triphone);
b. for 5 PLP coefficients the obtained results are very promising, giving word recognition rates about 58.55% (triphone training and testing FS), 68.10% (triphone training and testing MS) and 70.11% (triphone training MS and FS and testing MS);
c. for 36 MFCC_D_A coefficients (mel-cepstral coefficients with first and second order variation) we obtained the best results, as we expected: monophone 56.33% and triphone 81.02%, training and testing with MS; monophone 56.67% and triphone 78.43%, training and testing with FS; monophone 57.44% and triphone 78.24%, training MS and FS and testing with MS.

In the following part (ASRU system) (Juanhg et al., 200) are displayed the results obtained for PRR (Phrase Recognition Rate) in the same condition as well as evaluate the WRR: the CDRL database; the MFCC with the first and second order variation, or the PLP or the LPC feature vector; the acoustical model without and with context.

The experimental results for PRR are shown in the next Tables, using for the word recognition order a loop grammar.

Training MS	Type	PRR		
		MFCC_D_A	LPC	PLP
Testing MS	Monophone	20.00%	5.00%	10.00%
	Triphone	66.25%	11.25%	37.50%
Testing FS	Monophone	21.25%	6.25%	16.25%
	Triphone	48.75%	12.50%	32.50%

Table 9. PRR training MS and testing MS and FS

Training FS	Type	PRR		
		MFCC_D_A	LPC	PLP
Testing MS	Monophone	21.25%	3.75%	6.26%
	Triphone	37.50%	11.25%	21.25%
Testing FS	Monophone	30.00%	12.50%	22.50%
	Triphone	57.50%	33.75%	30.00%

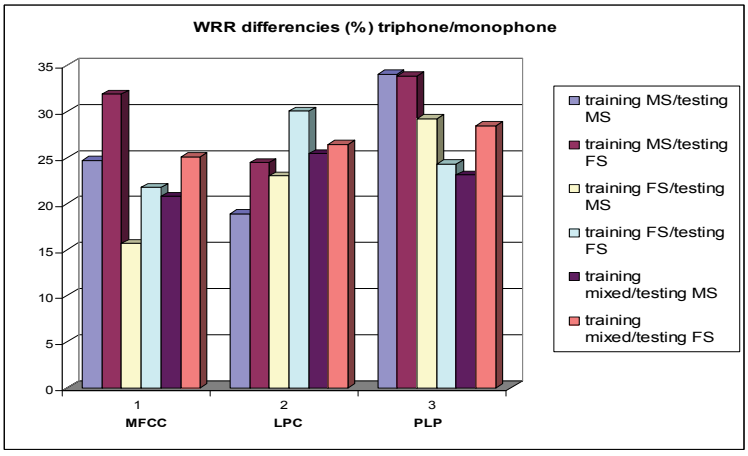
Table 10. PRR training FS and testing MS and FS

Training MS and FS	Type	PRR		
		MFCC_D_A	LPC	PLP
Testing MS	Monophone	23.75%	3.75%	18.75%
	Triphone	60.00%	10.00%	36.25%
Testing FS	Monophone	31.25%	6.25%	20.00%
	Triphone	55.00%	11.25%	46.25%

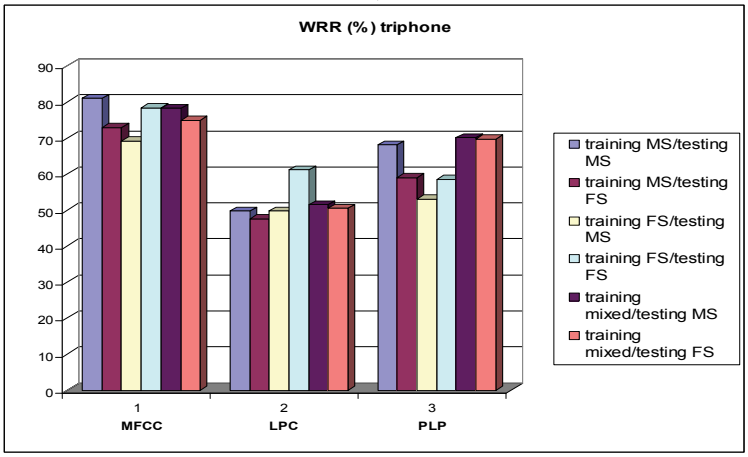
Table 11. PRR training MS&FS and testing MS and FS

After the experiments made on continuous speech recognition (Dumitru, 2006) we have following observations:

a. the triphone modelling is effective, conducting to increasing in WRR between 15% and 30% versus the monophone modelling. The maximal enhancement exceeds 30% for training MS and testing FS for MFCC_D_A (Fig. 10a).
b. A gender based training is conducting to good result for tests made with speakers from the same gender (training MS / testing MS: 81.02%, testing FS: 72.86%; training FS/testing FS: 78.43%, testing MS: 69.23%); changing gender in testing versus training leads to a decrease in WRR around 10%. For a mixed trained data base changing gender determines only variations around 5% in WRR (Fig. 10b).
c. The PRR variation follows the WRR line, but the results are drastically lower when we compare with WRR.



a)



b)

Fig. 10. WRR triphone *vs.* monophone

4.2 Telephone dial application

Other investigations using HMM are done for telephone dial domain. For this application the HMM without context modelling and MFCC coefficients are considered.

Our system was developed for a telephone numbers dialling application (Fig.11). The training data are constituted by the 850 phrases (OCDRL database), spoken by ten speakers, and the testing data are constituted by the 30 phrases, spoken by in training enrolled and also unenrolled speakers. The data are in the typical "wave" format, sampled by 16 kHz, quantified with 16 bits, recorded in a laboratory environment.

Examples of training phrases are:

Formeaza opt sapte sase cinci sase patru doi trei;

Telefoneaza Octavian;

Formeaza opt sapte cinci patru trei doi unu.

Examples of testing phrases are:

Formeaza sapte doi opt cinci unu sase patru cinci unu unu;

Telefoneaza Octavian Dumitru.

In the first experiments the speech signal was parameterized with 39 coefficients (MFCC_D_A): 13 mel cepstral coefficients; 13 delta coefficients; 13 accelerations coefficients. The recognition rates are summarized in Table 12.

Speaker	No. test words	No. recognised words	Recognition rate
Enrolled	103	98	95.1%
Unenrolled	103	93	90.3%

Table 12. Recognition rate

Fig. 11. Telephone dial application

For the enrolled persons the recognition rate was 95.1%, and for the unenrolled speakers the recognition rate was lower as in the first case, but still quite satisfactory (Dumitru, 2006).

Our next experiments were conducted for different choices in speech parameterization. The experiments made in the further described conditions lead to the recognition rates showed for the corresponding parameters in Fig. 12. The chosen parameter sets are the following: MFCC_0 (13) - 13 mel cepstral coefficients; MFCC_0_E (14) - 13 mel cepstral coefficients with log energy; MFCC_0_D (26) -13 mel cepstral coefficients and 13 delta coefficients; MFCC_0_D_A (39) -13 mel cepstral coefficients, 13 delta coefficients and 13 acceleration coefficients; MFCC_D_A_E (39) -12 mel cepstral coefficients, 12 delta coefficients, 12 acceleration coefficients, log energy, delta energy, acceleration energy.
The best rate, 96%, was obtained with 13mel-cepstral coefficients and 13 delta coefficients (MFCC_0_D).
Now, our interest is in medical domain to help the dentists in their job creating a system for completing the examination and treatment dental charts (a project between our laboratory and the University of Medicine and Pharmacy of Bucharest).

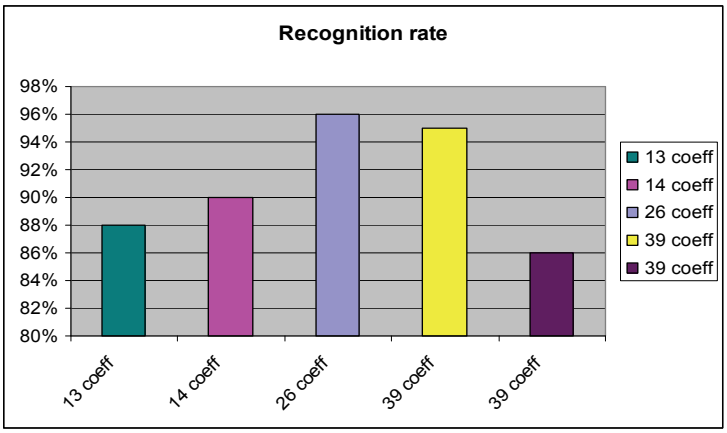


Fig. 12. Recognition rate for different number of coefficients

5. Digit and vowel recognition using ASRS_RL system based on neural and hybrid strategies

5.1 Digit recognition using neuro-statistical strategies

In the first digit recognition experiment, we compare the performance of two kinds of classifiers, namely the SVM (only to compare with other structure) and HMM, the performance being appreciated by their recognition rate and by their generalization capacity. For that, we performed two types of tests on the DDRL database: first, with enrolled speakers, which mean that the speakers were involved both in training and testing. We used five speakers for training and four for testing. For the second type of tests, with unenrolled speakers we used the leave-one-out method: for each word, we trained the classifier with 8 speakers and tested with the 9th repeating the procedure for each speaker.
The conditions for feature extraction are: perceptive cepstral analysis giving a 13 - dimensional vector having as components 13 MFCCs and perceptual linear prediction giving a 5 - dimensional feature vector having as components 5 PLP coefficients.

In the second digit recognition experiment, we evaluate the performance obtained with HMM and with the hybrid neuro-statistical system (HMM - MLP) for unenrolled and enrolled speaker (Dumitru, 2006).
The digit parameters were extracted by cepstral analysis, in form of 12 mel-frequency cepstral coefficients.

a. HMM: for each digit we constructed a left-right hidden Markov model with 3 states;
b. HMM - MLP: the system consists of 9 hybrid models corresponding to 9 digits. Each hybrid model is made of 5 states, each state being associated with one output node of the MLP. The MLP has one hidden layer (100 nodes), and the input layer consisting of 12 nodes (Dumitru et al., 2007).

The results for digit recognition are presented in Table 13.
The performance of the HMM, SVM and hybrid system are validate using other databases, for example Advanced Multimedia Laboratory from the Carnegie Mellon University. The results show the same behaviour, but the WRRs are slightly higher for the English database, recorded in a studio with specialized speakers.

Type	Features extraction	Enrolled	Unenrolled
SVM	MFCC	97.70%	91.70%
HMM	MFCC	98.00%	97.50%
HMM-MLP	MFCC	98.50%	98.30%
SVM	PLP	91.70%	84.70%
HMM	PLP	95.10%	94.20%

Table 13. WRR for digit recognition using SVM, HMM and HMM-MLP

Based on the experimental results obtained, the following conclusions can be extracted:

a. It is to seen that SVM performance is slightly after that of the HMM, but is really promising, taking into account that the HMM has the benefit of a so long refinement time.
b. Trying to reduce these limitation effects of HMMs (the model training is not discriminative), chosen an alternative approaches can be a good solution. This approaches combine the HMM with MLP into a hybrid system. The results for HMM are lower than the results for the hybrid system.

5.2 Vowel recognition using neural and fuzzy strategies

The learning strategies applied in our recognition experiments are: the Kohonen maps, the MLP, the fuzzy-MLP and the fuzzy-HMM.
There are three experiments can be made:
(A) In the first experiment, the vowel recognition rate using the VDRL database and the MFCC coefficients (in form of 12 mel-frequency cepstral coefficients) was determined; the results obtained with MLP and HMM as learning strategies are comparatively presented in Table 14.
The VDRL database was organized as follows: one database for male speakers (MS), one database for female speakers (FS). In booth cases one male speaker (MS) and one female speaker (FS) was excluded from the training database and used their data for the testing.
The experimented MLP is a two-layer perceptron trained with Back-Propagation algorithm, having in the output layer 5 nodes corresponding to the 5 vowels to be classified and 100

nodes in the hidden layer (experimentally chosen). The number of the input nodes is equal to the number of features (12 MFCC).
The HMMs chosen for comparison are Bakis (or left-right) structures with five states and for each vowel one model is created (Gavat et al., 2000).
In Table 14, are displayed only the results in the case of training MS and testing with MS and FS. Similarly results were obtained for the training with FS (Dumitru et al., 2007).

Vowel	HMM		MLP	
	MS	FS	MS	FS
a	100.00%	80.71%	100.00%	82.28%
e	85.81%	43.25%	94.82%	50.67%
i	85.71%	85.71%	95.15%	92.41%
o	90.90%	51.33%	97.00%	52.78%
u	88.88%	71.42%	94.83%	77.85%
Mean	*91.26%*	*66.48%*	*96.36%*	*71.20%*

Table 14. Vowel recognition rate in the case of training with MS and testing with MS and FS.

(B) In the second experiment, the vowels were described by three formant frequencies and the error rates obtained with different learning strategies are given in Table 15.
The database for formants contains 500 formant vectors, 100 for each vowel for the training and 250 formant vectors, 50 for each vowel for the testing.
The learning structures applied for these investigations are: (a) KM with the input layer with 3 neurons, corresponding to the three formant frequencies and three variants for the output layer: unidimensional with 25 neurons, bidimensional with 5x5 neurons, and toroidal with 25 neurons; (b) MLP with 3 layers organized as it follows: (1) the input layer with 3 neurons, corresponding to the three formant frequencies; (2) the hidden layer with 0 (Boolean network) or 4 neurons, (3) the output layer with 5 neurons corresponding each to a processed class (in our case the vowels *a, e, i, o, u*); (c) fuzzy-MLP.

Vowel	KM 1-dim	KM 2-dim	KM toroidal	Boolean	MLP	Fuzzy MLP
a	2.60%	1.20%	2.00%	4.00%	0.00%	1.50%
e	3.20%	2.40%	2.40%	6.00%	2.50%	1.00%
i	2.20%	1.60%	1.20%	4.00%	1.00%	0.50%
o	1.80%	1.20%	1.20%	10.00%	1.00%	1.00%
u	2.20%	2.10%	1.80%	10.00%	1.00%	0.00%
Mean	*2.40%*	*1.70%*	*1.72%*	*6.80%*	*1.10%*	*0.80%*

Table 15. Error rates in for different learning strategies for the case of formant

(C) In the third experiment the parameterization is realized with the mel cepstral coefficients and the first and second order differences of these coefficients deduced from homomorfic filtering. The error rates obtained in the vowel recognition tests are given in Table 16, comparatively for the generalized HMM (fuzzy-HMM) and the classical HMM.

Vowel	Fuzzy - HMM	Classical HMM
a	5.10%	6.90%
e	2.40%	4.80%
i	3.80%	7.30%
o	2.50%	5.90%
u	0.70%	3.90%
Mean	*2.90%*	*5.78%*

Table 16. Error rates for generalized and for classical HMMs

For vowel recognition in Romanian language the following conclusion can be reported:
(A) The recognition rates in the case of MLP are higher than in the case of HMM. A possible explanation can be the fact that the model training is discriminative, while in the case of HMM the training is not discriminative, which represents a disadvantage of HMM utilization.
(B) The KM 2-dimensional structures and the toroidal have the same performance, weaker is the performance of the 1-dimensional structure. The best balanced situation corresponds to the 2-dimensional map 5x5, in which all neurons are associated to a vowel to be recognized. The performance obtained in the case of the Boolean network is unacceptable, but the MLP acts well.
Using fuzzy-MLP structure it is an improvement with a mean value of 0.30% comparative with the non-fuzzy structure.
(C) A mean decreasing of nearly 3% is realized in the error rate by adopting the fuzzy-HMM instead of the probabilistic one.

6. Conclusions

In this chapter we presented the work done until now to implement a tool for research in speech recognition for Romanian language. What we realized can be summarized as follows:

- We have implemented basic algorithms for feature extraction for perceptual and simple cepstral analysis and perceptual and simple linear prediction
- We have implemented learning strategies based on HMMs, ANNs, and hybrid strategies like fuzzy-MLP, fuzzy HMM or HMM-MLP in the framework of the statistical, connectionist and fuzzy paradigms of Computational Intelligence
- We have constructed databases to train the models and to test our recognizers that obey the same statistical rules as given from linguists for Romanian language.
- We have trained the models and tested the recognizers on this data bases

What we achieved and what is further to be done is sketched bellow:

- The obtained results in ASR for WRR are acceptable, what can not be said for ASRU experiments, so that much work must be done to enhance PRR by extending the knowledge resources to the semantic and pragmatic levels. Our resources for the moment cover the acoustic and phonetic level by the trained models and the dictionary and in a very primitive manner the syntactic level, by a loop grammar. A more complex grammar and a language model are our next objectives

- For validation of our algorithms, a professional data base for Romanian language could be very helpful. Some universities in Romania try to do that by cooperation and we hope to succeed in our attempt.
- To extend the possibilities of our platform we are planning to enhance the feature extraction methods and also try to develop neural and hybrid strategies suitable for continuous speech recognition experiments

7. References

AMP Advanced Multimedia Processing Laboratory: http://amp.ece.cmu.edu/_download/Intel/feature_data.html

Appen Appen Database http://www.appen.com.au/services/

Bourlard, H. and Wellekens, C.J. (1990). Links between Markov models and multilayer perceptrons. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12, pp. 1167-1178.

Bourlard, H. and Morgan, N. (1994). *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Press.

Charlet, D., Krstulovic, S,. Bimbot, F., Boëffard, O., & others. (2005). Neologos : an optimized database for the development of new speech processing algorithms, *Proceedings Interspeech'05*, pp. 1549-1552, Lisbonne, Portugal.

Cslu Center of Spoken Language Understanding: http://cslu.cse.ogi.edu/

Douglas, B.P., Janet, M. B. (1992). The design for the wall street journal-based CSR corpus, *Proceedings ICSLP*, pp. 899-902.

Draganescu, M., (2003). Spoken language Technology, *Proceedings of Speech Technology and Human-Computer-Dialog (SPED2003)*, pp. 11-12, Bucharest, Romania.

Dumitru, C.O., Gavat, I. (2006). A Comparative Study of Features Extraction Methods Applied for Continuous Speech Recognition in Romanian Language, *Proceedings the 48th International Symposium ELMAR 2006*, pp. 115-118, Zadar, Croatia.

Dumitru, O. (2006). *Modele neurale si statistice pentru recunoasterea vorbirii*, Ph.D. thesis.

Dumitru, C.O., Gavat, I. (2007). Vowel, Digit and Continuous Speech Recognition Based on Statistical, Neural and Hybrid Modelling by Using ASRS_RL, *Proceedings EUROCON 2007*, pp.856-863, Warsaw, Poland.

Elra European Language Ressources Association, http://www.elra.info/

Entro40 Entropic Latino40 Speech Database http://noble.gs.washington.edu/~noble/latino40.html

Gavat, I., Zirra, M. and Enescu, V. (1996-a). A hybrid NN-HMM system for connected digit recognition over telephone in Romanian language, *Proceedings IVTTA '96,* pp. 37-40, Basking Ridge, N.J.

Gavat, I. and Zirra, M. (1996-b). Fuzzy models in vowel recognition for Romanian language, *Proceedings Fuzzy-IEEE '96*, pp. 1318-1326, New Orleans.

Gavat, I., Grigore, O., Zirra, M. and Cula, O. (1997). Fuzzy variants of hard classification rules, *Proceedings NAFIPS'97*, pp. 172-176, New York.

Gavat, I., Zirra, M. and Cula, O. (1998). Hybrid speech recognition system with discriminative training applied for Romanian language, *Proceedings MELECON '98*, pp. 11-15, Tel Aviv, Israel.

Gavat, I., & all. (2000). *Elemente de sinteza si recunoasterea vorbirii*, Ed. Printech, Bucharest.

Gavat, I., Valsan, Z., Sabac, B., Grigore, O. and Militaru, D. (2001-a). Fuzzy similarity measures - alternative to improve discriminative capabilities of HMM speech recognizers, *Proceedings ICA 2001*, pp. 2316-2317, Rome, Italy.

Gavat, I., Valsan, Z. and Grigore, O. (2001-b). Fuzzy-variants of hidden Markov models applied in speech recognition, *Proceedings SCI 2001, Invited Session: Computational Intelligence In Signal And Image Processing*, pp. 126-130, Orlando, Florida.

Gavat, I., Dumitru, C.O., Costache, G., Militaru, D. (2003). Continuous Speech Recognition Based on Statistical Methods, *Proceedings of Speech Technology and Human-Computer-Dialog (SPED2003)*, pp. 115-126, Bucharest.

Gavat,I., Dumitru, C.O. (2008). The ASRS_RL - a Research Platform, for Spoken Language Recognition and Understanding Experiments, *Lecture Notes in Computer Science (LNCS)*, Vol. 5073, Part II, pp.1142-1157.

Gold, B., Morgan, N. (2002). *Speech and audio signal processing*, John Wiley&Sons, N. Y.

Goronzy, S. (2002). *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*, Springer – Verlag, Berlin.

Grigore, M. and Gavat, I. (1996). Vowel recognition with nonlinear perceptron, *Proceedings CAS '96* , pp. 155-158, Sinaia, Romania.

Grigore, O., Gavat, I. and Zirra, M. (1998). Neural network vowel recognition in Romanian language, *Proceedings CONTI '98*, pp. 165-172, Timisoara, Romania.

Grigore, O. and Gavat, I. (1999). Neuro-fuzzy models for speech pattern recognition in Romanian language, *Proceedings ESIT'99*, pp. 98-103, Rhodos, Greece.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech, *Journal Acoustic Soc. America*, Vol. 87, No. 4, pp. 1738-1752.

Huang, X., Acero, A., Hon, H.W. (2001). *Spoken Language Processing–A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

Juang, B.H. and Katagiri, S. (1992). Discriminative learning for minimum error classification. *IEEE Transactions on Signal Processing*, Vol. 12, pp. 3043-3054.

Juanhg, B.H., Furui, S. (2000). Automatic Recognition and Understanding of Spoken Language–A First Step Toward Natural Human–Machine Communication, *Proceedings IEEE*, Vol. 88, No. 8, pp. 1142-1165.

Kocsor, A., Kovacs, K., Kuba Jr., A., Toth, L. (1999). An Overviey of the OASIS Speech Recognition Project, *Proceedings 4th International Conference on Applied Informatics*, Eger-Noszvajm Hungary, 30 August – 3 Sepetember.

Lippmann, R. and Singer, E. (1993). Hybrid neural network / HMM approaches to word spotting, *Proceedings ICASSP '93*, pp. 565-568, Minneapolis.

Mahomed, M., Gader, P. (2000). Generalized Hidden Markov Models, *IEEE Transactions on Fuzzy Systems*, Vol. 2, pp. 67-93.

Rabiner, L.R. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition , *Proceedings of the IEEE*, Vol. 77, No. 8, pp. 257-286 .

Richard, M. and Lippmann, R. (1991). Neural network classifiers estimate Bayesian a posteriori probabilities, *Neural Computation*, Vol. 4, pp. 461-483.

Reichl, W., Caspary, P. and Ruske, G. (1994). A new model-discriminant training algorithm for hybrid NN-HMM systems, *Proceedings ICASSP '94*, pp. 677-680, Adelaide, Australia.

Robinson, T., Fransen, J., Pye, D., Foote, J., Renals, S. (1995°. WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition, *Proceedings Int. Conf. Acoustics, Speech and Signal Processing*, pp. 81-84, Detroit.

Sampa. http://www.phon.ucl.ac.uk/home/sampa

SpeDatM The SpeechDat projects home page: http://www.speechdat.org/

SpeDatE SpeEastern European Speech Databases for Creation of Voice Driven Teleservices, http://www.fee.vutbr.cz/SPEECHDAT-E/

Timit DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus http://www.ldc.upenn.edu/Catalog/readme_files/timit.readme.html

Valsan, Z., Sabac, B. and Gavat, I. (1998-a). Combining self organizing feature map and multilayer perceptron in a neural system for fast key-word spotting. Proceedings SPECOM '98, pp. 303-308, St. Petersburg, Russia.

Valsan, Z., Sabac, B., Gavat, I. and Zamfirescu, D. (1998-b). Combining self-organizing map and multilayer perceptron in a neural system for improved isolated word recognition, *Proceedings Communications '98*, pp. 245-251, Bucharest, Romania.

Valsan, Z., Gavat, I., Sabac, B., Cula, O., Grigore, O., Militaru, D., Dumitru, C.O., (2002). Statistical and Hybrid Methods for Speech Recognition in Romanian, *International Journal of Speech Technology*, Vol. 5, No. 3, pp. 259-268.

Young, S.J. (1992). The general use of tying in phoneme-based HMM speech recognizers, *Proceedings ICASSP'92*, Vol. 1, pp. 569-572, San Francisco.

Young, S.J., Odell, J.J., Woodland, P.C. (1994). Tree based state tying for high accuracy modelling, *ARPA Workshop on Human Language Technology*, Princeton.

Advances in Robotics, Automation and Control

Edited by Jesus Aramburo and Antonio Ramirez Trevino

ISBN 978-953-7619-16-9
Hard cover, 472 pages
Publisher InTech
Published online 01, October, 2008
Published in print edition October, 2008

The book presents an excellent overview of the recent developments in the different areas of Robotics, Automation and Control. Through its 24 chapters, this book presents topics related to control and robot design; it also introduces new mathematical tools and techniques devoted to improve the system modeling and control. An important point is the use of rational agents and heuristic techniques to cope with the computational complexity required for controlling complex systems. Through this book, we also find navigation and vision algorithms, automatic handwritten comprehension and speech recognition systems that will be included in the next generation of productive systems developed by man.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Corneliu-Octavian Dumitru and Inge Gavat (2008). Progress in Speech Recognition for Romanian Language, Advances in Robotics, Automation and Control, Jesus Aramburo and Antonio Ramirez Trevino (Ed.), ISBN: 978-953-7619-16-9, InTech, Available from: http://www.intechopen.com/books/advances_in_robotics_automation_and_control/progress_in_speech_recognition_for_romanian_language

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821