We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities



WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com



13

Knowledge Resources in Automatic Speech Recognition and Understanding for Romanian Language

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

1. Introduction

In this chapter are presented the results obtained in automatic speech recognition and understanding (ASRU) experiments made for Romanian language in the statistical framework, concerning the performance enhancement of two important knowledge resources, namely the acoustical models and the language model. If the ASRU process is for simplicity seen as a two stage process, in the first stage automatic speech recognition (ASR) is done and in the second stage the understanding is accomplished. The acoustical models incorporate knowledge about features statistic in different speech units composing the words and are mostly responsible for the performance of the recognition stage, judged after the WRR (word recognition rate). The language models incorporate knowledge about the word statistic in the phrase and determine mostly the performance of the understanding stage, judged after the PRR (phrase recognition rate). The two considered stages are interrelated and the named performance criteria are interdependent, enhanced WRR leads to PRR enhancement too. In this chapter are exposed methods to enhance the WRR, based on introducing of contextual models like triphones instead monophones or building of gender specialized models (for men, women and children) instead of global models. The methods applied to enhance the PRR are based on introducing of a restrictive finite state grammar instead the permissive word loop grammar or a bigram based language model.

1.1 Short history

Speech recognition and understanding has in Romania also a long history and begins with recognition and synthesis of vowels, done in the University Politehnica from Bucharest around 1963 (Draganescu, 2003). Digit recognizers were built around 1970 in hardware form and in 1976 as software models and the first recognition experiments for continuous speech were successful around 1980 in the Institut for Linguistics of the Romanian Academy. The researches in this new domain of speech technology were extended after 1980 also in other universities and technical universities in cities like Iasi, Cluj – Napoca and Timisoara. To bring researchers together, starting with the year 1999 each two years an international conference namely SPED is organized under the aegis of the Romanian Academy. It is also to be mentioned participation of Romanian researchers in international research programs,

Source: Speech Recognition, Technologies and Applications, Book edited by: France Mihelič and Janez Žibert,
ISBN 978-953-7619-29-9, pp. 550, November 2008, I-Tech, Vienna, Austria

Open Access Database www.intechweb.org

in international conferences, in bilateral cooperations. In 2002 a special issue for Romanian contributions was dedicated by the International Journal of Speech Technology.
Our research group comes from the University Politehnica Bucharest, Chair of Applied Electronics and Information Engineering, Faculty of Electronics, Telecommunications and Information Technology. Our research interests concern mainly ASRU, but also other topics like speech recognition based on neural networks (Valsan et al., 2002), (Gavat & Dumitru, 2002-1), (Gavat et al., 2002-2), (Dumitru & Gavat, 2008) on fuzzy technics (Gavat & Zirra, 1996-2), (Gavat et al., 1997), (Gavat et al., 2001-1), (Gavat et al., 2001-2), or on Support Vector Machines (SVM) (Gavat et al., 2005-2), (Gavat & Dumitru, 2008), speech synthesis, TTS systems, speech and music retrieval (Gavat et al., 2005-1), speech prosody, multimodal systems, could be mentioned. Our main realization is the Automatic Speech Recognition System for Romanian Language, ASRS_RL (Dumitru, 2006), a research platform in order to implement and enhance different methods for ASRU in Romanian language (Gavat and Dumitru, 2008). Recognizers for phonemes, digits and continuous speech acting under the statistic paradigm based on hidden Markov models, under the connectionist paradigm of artificial neural networks or under fuzzy principles, but also under combined principles were experimented.
The system presented in this chapter has as main objective refinement of the statistic paradigm for ASRU in Romanian language, by enhancement of two important aspects, acoustical modeling and language modeling.

1.2 The proposed system

In the statistical approach, for the mathematical formulation of the problem, the recognition process can be modeled as a communication system, depicted in Fig. 1, consisting in four stages: text generation, speech production, acoustic processing, and linguistic decoding.

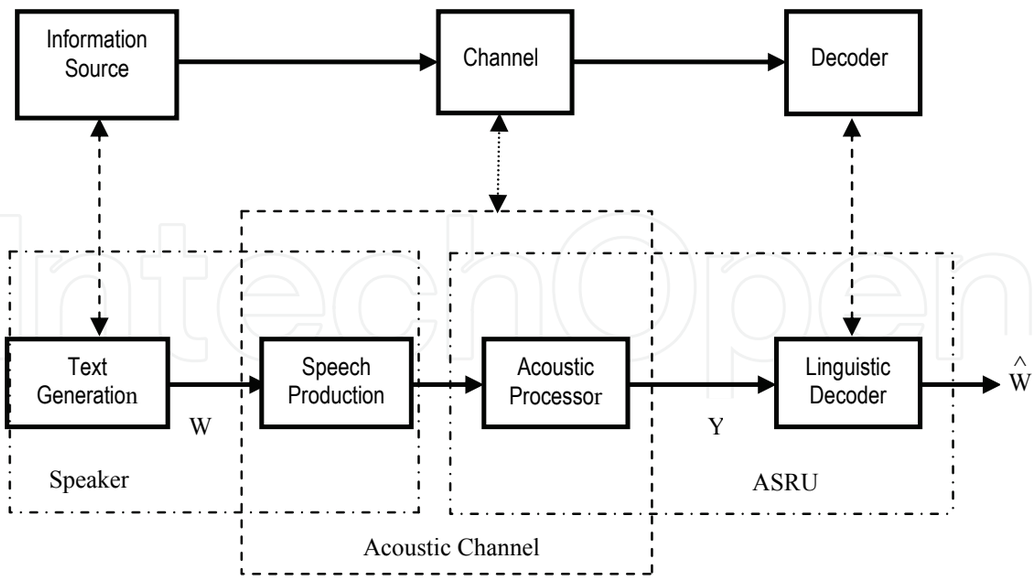


Fig. 1 Structure of continuous speech recognition system

A speaker is assumed a transducer that transforms into speech the text of thoughts to communicate. The delivered word sequence W is converted into an acoustic observation sequence *Y*, with probability *P(W, Y)*, through a noisy acoustical transmission channel, into an

acoustic observations sequence Y which is then decoded to an estimated sequence $\hat{W}$. The goal of recognition is then to decode the word string, based on the acoustic observation sequence, so that the decoded string has the maximum a posteriori probability (Huang et al., 2001):

$$\hat{W} = \arg\max_{W} P(W \mid Y) \tag{1}$$

Using Bayes' rule can be written as:

$$\hat{W} = \arg\max_{W} P(Y \mid W) * P(W) / P(Y) \tag{2}$$

Since *P (Y)* is independent of *W*, the maximum a posteriori decoding rule is:

$$\hat{W} = \arg\max_{W} P(Y \mid W) * P(W) \tag{3}$$

The term *P(Y|W)* is generally called the acoustic model as it estimates the probability of sequence of acoustic observations conditioned on the word string (Rabiner, 1989).
The term *P (Y)* is generally called the language model since it describes the probability associated with a postulated sequence of words. Such language models can incorporate both syntactic and semantic constraints. When only syntactic constraints are used, the language model is called a grammar.
The block diagram of the system, based on the pattern recognition paradigm, and applied in a continuous speech recognition task is presented in Fig. 2. The speech signal is analysed resulting sequence of feature vectors grouped in linguistic unit patterns. Each obtained pattern is compared with reference patterns, pre-trained and stored with class identities.

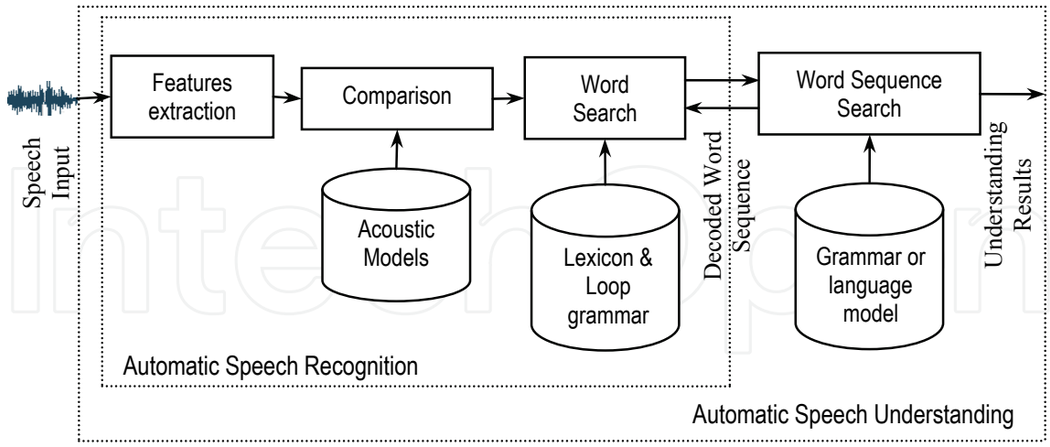


Fig. 2. Block diagram for an automatic speech recognition and understanding system.

These pre-trained patterns, obtained in a learning process are in our system the acoustical models for phonemes, with or without context, and represent a first knowledge source for the word sequence search.
Further, based on the dictionary the words are recognized and a simple loop grammar leads to the estimated word sequence. The important outcome of this stage is the word

recognition rate, the phrase recognition rate having low levels. To enhance the phrase recognition rate, a restrictive grammar or a language model must be applied. Like it is represented in Fig. 1, it is possible to separate the part of automatic speech recognition (ASR) as a first stage in the automatic speech recognition and understanding (ASRU) process (Juang et Furui, 2000).

1.3 Chapter structure

The remainder of this paper will be structured as follows. Section two will be dedicated to the acoustic models as first knowledge source in the ASRU process. In the third section is presented the second knowledge resource, in form of language models and restrictive grammars. Comments and discussions on the experimental results presented in the previous two sections will be made in section four. The final section will be dedicated to conclusions about the work done and to plan future activities.

2. Acoustic models

The acoustic models developed in our experiments are the hidden Markov models (HMM), basic entities in the statistical framework.

2.1 Hidden Markov models

2.1.a Basics monophones and triphones

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 (as for any language). But in real speech the words are not simple strings of independent phonemes, because each phoneme is affected through the immediately neighboring phonemes by co-articulation. Therefor to monophones context was added leading for example to triphones like monophones with left and right context, that became actually the state of the art in automatic speech recognition and understanding for the large vocabularies (Young, 1992).

Based on the SAMPA (Speech Assessment Methods Phonetic Alphabet) in Romanian language there are 34 phonemes and for each a model is to be trained. For triphones the situation is more complicated because the number of them is large, around 40000, and the control of the training could be lost. To solve this problem, tying of acoustically similar states of the models built for triphones corresponding to each context is an efficient solution.

In the realized continuous speech recognition and understanding task we modelled intra-word triphones and also cross- words triphones. We adopted the state tying procedure, conducting to a controllable situation.

2.1.b HMM Types

The hidden Markov model incorporates the knowledge about feature constellation corresponding to each of the distinct phonetic units to be recognized. In our experiments we used continuous and semi-continuous models.

To describe HMMs, we start for simplicity reasons with the discrete model (Gold, Morgan, 2002).

Discrete HMMs

A discrete HMM is presented in Fig.3 in a Bakis form. The basic parameters of the model are:

- N –The number of states $S = \{s_1, s_2, ...s_N\}$; a state to a certain time is denominated as q_t, $(q_t \in S)$.
- M – The number of distinct symbols observable in each state. The observations are $V = \{v_1, v_2, ...v_M\}$; one element O_t from V is a symbol observed at moment *t*.
- A – The transition matrix containing the probabilities a_{ij} of the transition from state i in state j:

$$a_{ij} = A(i,j) = P\left(q_{t+1} = s_j \middle| q_t = s_i\right) \quad 1 \le i, j \le N \text{ , } t \in [1,T] \text{, } a_{ij} \ge 0 \text{ , } \sum a_{ij} = 1 \tag{4}$$

- B – Matrix of observed symbols in each state of the model: $b_j(k)$ represents the probability to observe a symbol v_k in state j:

$$b_j(k) = P\left(O_t = v_k \middle| q_t = s_j\right) \quad 1 \le j \le N \text{ , } 1 \le k \le M \text{ , } t \in [1,T] \text{, } b_j(k) \ge 0 \text{ , } \sum b_j(k) = 1 \tag{5}$$

- Π – The matrix of initial probabilities

$$\pi_i = P(q_1 = s_i) \text{, } \pi_i \ge 0 \text{ , } \sum \pi_i = 1 \tag{6}$$

In a compact mode a discrete HMM can be symbolized with $\lambda = (\Pi, A, B)$.

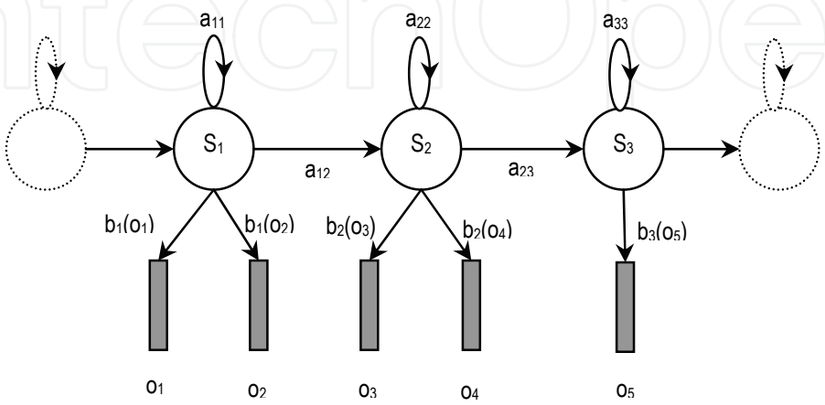


Fig 3. Bakis model with three states

Continuous densities hidden Markov models (CDHMM)

În the HMMs defined as $\lambda = (\Pi, A, B)$ the observations at a certain moment present a continuous probability density function, usually a Gaussian density or a mixture of Gaussian densities. For this last case we have:

$$b_i(O_t) = \sum_{m=1}^{M} c_{im} b_{im}(O_t), \; i = \overline{1, N} \tag{7}$$

c_{im} obey the restrictions: $c_{im} \geq 0$, $\sum_{m=1}^{M} c_{im} = 1$.

$b_{im}(O_t)$ is a K-dimensional Gaussian density with covariance matrix σ_{im} and mean μ_{im}:

$$b_{im}(O_t) = \frac{1}{\sqrt{(2\pi)^K |\sigma_{im}|}} \exp\left[-\frac{1}{2} (O_t - \mu_{im})^T \frac{1}{\sigma_{im}} (O_t - \mu_{im}) \right] \tag{8}$$

Semicontinuous hidden Markov models (SCHMM)

SCHMM are an intermediate form between the discrete and the continuous density hidden Markov models. An acoustical observation is described by a weighted combination of a number of probabilities, so that:

$$b_i(O_t) = \sum_{k=1}^{M} b_i(k) f_k(O_t) \text{ for } i = \overline{1, N} \tag{9}$$

$f_k(O_t)$ is a Gaussian density, with the covariance matrix Σ_k and mean vector μ_k.

Because speech is a signal with a high degree of variability, the most appropriate model capable to capture its complicated dependencies is the continuous one. But often also semicontinuous or discrete models are applied in simpler speech recognition tasks.

2.1.c Problems that can be solved with HMMs

Based on HMM's the statistical strategies has 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 (Goronzy, 2002).

The algorithms developed in the statistical framework to use HMM are rich and powerful, situation that can explain well the fact that today, hidden Markov models are the widest used in practice to implement speech recognition and understanding systems.

The main problems that can be solved with HMMs are:

- The evaluation problem, in which given the model the probability to generate a observation sequence is calculated. This probability is the similarity measure used in recognition (decoding) to assign a speech segment to the model of highest probability. This problem can be solved with the forward or the backward algorithm
- The training problem, in which given a set of data, models for this data must be developed. It is a learning process, during which the parameters of the model are

estimated to fit the data. For each phonetic unit a model can be developed, such phonetic units can be monophones or intra- word or inter-word triphones. Utterances result through concatenation of these phonetic units. Training of the models is achieved with the Baum-Welch algorithm.

- The evaluation of the probability of the optimal observation sequence that can be generated by the model. This problem can be solved with the Viterbi algorithm. Often this algorithm is used instead the forward or backward procedure, because it is acting faster and decode easier the uttered sequence.

2.2 Recognition experiments based on monophones and triphones models

First we will define the conditions under that our experiments were conducted and further the obtained experimental results will be displayed.

To solve the problem of continuous speech recognition we used the specialized recognition tool based on hidden Markov models (HMMs), namely the HTK-Toolkit (Young et al., 2006).

2.2.1 Experiments conditions

To conduct this experiments, we choose the speech material contained in the first self-made in the university data base, called OCDRL, meaning Old Continuous Database for Romanian Language. The OCDRL 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 speech material was recorded in a laboratory environment, sampled with 16 kHz and quantized with 16 bits, the speakers were students, not professionals (Gavat et al., 2003), (Dumitru, 2006).

As speech data, the utterances of the data base were processed by phonetical transcription after the SAMPA standard (Sampa), conducting to the phonetic dictionary of the system. Each word is decomposed in constituent monophones (34 for Romanian language) or triphones (34^3 for Romanian language) and for each monophone or triphone a model must be trained. Of course for monophones the number of necessary models is small, there are sufficient training data, so that the models will be good trained in a short time. For triphones the situation is more complicated because there number is huge and the training data become insufficient (Young, 1994). Therefore tying procedure must be adopted, combining in a model similar triphones. Beam searching is the solution adopted in our experiments to realize the tying.

The digitizated data are further analysed in order to extract characteristic cues, called features. By short term analysis a set of features is obtained for each speech frame, extracted by a windowing process. The frame duration is chosen making a compromise between a long time (20-40 ms) imposed in order to detect the periodic parts of speech and the short time during which the speech can be considered a stationary random process (around 20 ms.). The experimental results further presented are obtained with a Hamming window, with duration 25 ms and the overlapping factor of the windows ½.

The features that can be obtained in our system to characterize speech segments are from two kinds: static features obtained for each window and dynamic features, calculated over a number of windows and representing derivatives of first, second and third degree. The static features type (Gavat et al., 2003), (Dumitru & Gavat, 2006) we have extracted to use in our experiments are:

- Perceptive linear prediction coefficients (PLP)
- Mel-frequency cepstral coefficients (MFCC)
- Linear prediction coefficients (LPC)
- Linear prediction reflexion coefficients (LPREFC)
- Linear prediction cepstral coefficients (LPCEPC)

All this features are 12-dimensional.
Energy and zero crossing rate were also calculated for each frame and are one-dimensional.
To each of this kind of features we can add the first, second and third degree derivatives, in order to capture the dynamic of speech.
To prepare features for training of the models, we perform normalization applying two algorithms:

- Cepstral mean normalization (CMN)
- Global variance normalization (GVN)

The sequences of features vectors obtained from the OCDRL training database are used to train the acoustical models, like monophones and triphones (intra-word and inter-word or cross word).
Further, we will evaluate the efficiency of the trained models by the word recognition rate (WRR), the accuracy and the phrase recognition rate (PRR) in a task of continuous speech recognition for Romanian language using for that the OCDRL test database.
At the end of the first tests we compared the word recognition rates, and could establish a first ranking of the best feature set for recognition. The results are displayed in Table 1.

FEM	PLP	MFCC	LPC	LPREFC	LPCEPC
WRR (%)	58.96	54.98	39.04	39.57	47.81
Accuracy (%)	56.97	52.59	36.25	36.17	45.82

Table 1. WRR and accuracy for the basic feature extraction methods (FEM).

It is to be seen that the best scores were obtained with the PLP coefficients (Hermansky, 1990), so that we will display bellow only results for feature vectors having as components PLP coefficients with first order derivatives (D) and second order ones (A) with or without energy (E).

2.2.2 Experimental results

We conducted recognition tests on the OCDRL test database, proving the generalization capability of the models in the following situations:

- Training of the models with global initialization (TGI)
- Retraining of the models with global initialization (RGI)
- Retraining of the models with individual initialization (RII)

Detailed is analyzed TGI. For RGI and RII some comparative results are given.
Training with global initialization (TGI)
We applied first the training procedure with global initialization, meaning that the training starts with all initial models having zero mean and unitary variance.
We have trained and than tested along our experiments the following types of continuous density models:

- Simple mixture monophones (SMM)
- Simple mixture intra- word triphones (SMIWT)

- Simple mixture cross- word triphones (SMCWT)
- Multiple mixtures monophones (MMM)
- Multiple mixtures intra-word triphones (MMIWT)
- Multiple mixtures cross-word triphones (MMCWT)

We have also trained and tested semicontinuous density models based on monophones and triphones.

The obtained results for continuous densities models are displayed in Table 2 for simple mixtures and in Table 3 for multiple mixtures. The results for semicontinuous densities are displayed in Table 4.

CDHMMs	PLP + E + D + A			PLP + D + A		
	WRR	Accuracy	PRR	WRR	Accuracy	PRR
SMM	84.47	83.16	40.00	84.74	84.74	44.00
SMM+CMN	87.37	87.37	42.00	87.89	87.89	52.00
SMIWT	97.37	97.37	84.00	98.16	98.16	88.00
SMIWT+CMN	97.63	97.63	88.00	96.32	96.32	80.00
SMCWT	91.84	91.58	52.00	91.32	90.79	50.00
SMCWT+CMN	89.21	88.42	38.00	90.79	90.53	48.00

Table 2. Recognition performance for singular mixture trained monophones and triphones in continuous density hidden Markov models (CDHMM).

CDHMMs/ number of mixtures		PLP + E + D + A			PLP + D + A		
		WRR	Accuracy	PRR	WRR	Accuracy	PRR
MMM	5	96.58	96.32	80.00	97.37	97.37	86.00
	10	97.37	97.37	86.00	97.37	97.37	88.00
	15	98.16	98.16	90.00	97.89	97.89	88.00
	20	98.16	98.16	90.00	98.42	98.42	90.00
MMIWT	2	98.68	98.68	92.00	98.42	98.42	90.00
	4	98.68	98.68	92.00	98.95	98.95	92.00
	6	98.68	98.68	92.00	99.21	99.21	94.00
	8	98.68	98.68	92.00	99.21	99.21	94.00
	10	98.95	98.95	94.00	98.95	98.95	92.00
	12	99.21	99.21	96.00	98.42	98.42	90.00
MMCWT	2	93.68	92.89	58.00	94.21	93.95	68.00
	4	93.42	92.63	56.00	95.26	95.00	70.00
	6	93.68	93.16	58.00	94.74	94.47	64.00
	8	95.00	94.21	62.00	94.74	94.47	62.00
	10	95.53	94.74	64.00	95.26	94.74	68.00
	12	94.74	93.95	62.00	94.74	94.21	62.00

Table 3. Recognition performance for multiple mixture trained monophones and triphones in continuous density hidden Markov models (CDHMM).

HMMs	PLP + E + D + A			PLP + D + A		
	WRR	Accuracy	PRR	WRR	Accuracy	PRR
monophones	96.58	95.26	76.00	97.11	97.11	82.00
monophones +CMN	96.51	96.24	79.59	97.11	98.58	84.00
triphones	97.89	97.63	88.00	98.42	98.42	88.00
triphones +CMN	98.42	97.89	88.00	98.68	98.68	92.00

Table 4. Recognition performance for semicontinuous hidden Markov models (SCHMMs).

Detailed discussions and comments concerning these results will be made in section 4.
Some global observations can be made:

- Triphones are in all cases more effective models than monophones
- Increasing the mixture number is helpful only below certain limits: for monophones this limit is around 20 mixtures, for inter-word triphones around 12 mixtures, for cross-word triphones around 10 mixtures
- Due to the poor applied grammar, WRR is always higher than PRR
- SCHMMs are slightly less more effective than CDHMMs
- CMN is slightly more effective for semicontinuous models, producing increases in the same time for WRR, accuracy and PRR
- In all cases, the best performance is obtained with the feature set (PLP + D + A)

For applications, not only the recognition rates, but also training and testing durations are important. Training of models is done off line, so that the training duration is not critical. The testing time is important to be maintained low, especially for real time applications.
Training and testing of the models were done on a standard PC with 1 GHZ Pentium IV processor and a dynamic memory of 1 GB. The obtained training and testing durations are detailed in Table 5 for different categories of models.

HMMs		Training duration (sec.)	Average testing duration/word (sec.)
CHMM	SMM	157	0.092
	30 MMM	2.323	0.291
	SMIWT	263	0.098
	12 MMIWT	1.087	0.219
	SMCWT	220	0.129
	12 MMCWT	1.106	0.223
SCHMM	Monophones	3.779	0.125
	Triphones	2.887	0.831

Table 5. Training and testing durations

As general observation we can say that the processing durations depend of the model complexity for both the training duration and testing duration. The training time takes values between 157s and 3.779s. The testing duration/word is less than 0.3s, so that real time applications are possible with this system.

Retraining with global initialization (RGI) and with individual initialization (RII)

The performance of the recognition system can be enhanced by retraining, with global initialization or individual initialization. In the first case, the prototypes of the retraining are to be globally initialized with the means and variances extracted from the data trained with

that monophone mixtures conducting to the best recognition results. In the second case, the initialization is to be done with individual means and variances, extracted from the trained data with a high mixtures number.
Bellow are displayed the comparative results of the recognition performance obtained by training with global initialisation (TGI), retraining with global initialization (RGI) and retraining with individual initialization (RII) for SMCWT in Table 6 and for MMCWT in Table 7.

Training type/ normalization		PLP + E + D + A			PLP + D + A		
		WRR	Accuracy	PRR	WRR	Acuracy	PRR
TGI	PS	97.37	97.37	84.00	98.16	98.16	88.00
	CNM	97.63	97.63	88.00	96.32	96.32	80.00
	RGV	96.84	96.84	82.00	95.79	95.79	76.00
RGI	PS	98.68	98.68	92.00	97.63	97.63	86.00
	CNM	98.68	98.68	92.00	98.68	98.68	92.00
	RGV	97.63	97.63	84.00	97.89	97.89	84.00
RII	PS	96.32	96.32	76.00	97.11	97.11	80.00
	CNM	98.68	98.68	92.00	98.16	98.16	88.00
	RGV	97.63	97.37	86.00	97.37	97.37	84.00

Table 6. Comparative results for TGI. RGI and RII for SMCWT

Training type/ number of mixtures		PLP + E + D + A			PLP + D + A		
		WRR	Accuracy	PRR	WRR	Acuracy	PRR
TGI	2	96.68	98.68	92.00	98.42	94.42	90.00
	4	98.68	98.68	92.00	98.95	98.95	92.00
	6	98.68	98.68	92.00	99.21	99.21	94.00
	8	98.68	98.68	92.00	99.37	97.37	88.00
	10	98.95	98.95	94.00	98.95	98.95	92.00
	12	99.21	99.21	96.00	98.42	98.42	90.00
RGI	2	98.68	98.42	90.00	98.68	98.68	92.00
	4	99.21	98.95	92.00	99.95	98.95	92.00
	6	98.68	98.68	92.00	99.21	99.21	94.00
	8	98.42	98.16	86.00	98.68	98.68	92.00
	10	98.42	98.16	86.00	98.68	98.68	9000
	12	99.21	99.21	96.00	98.42	98.42	90.00
RII	2	98.95	98.95	92.00	98.95	98.68	90.00
	4	99.21	98.95	94.00	98.68	98.68	90.00
	6	99.21	98.95	94.00	99.21	99.21	94.00
	8	99.47	98.95	94.00	99.21	99.21	94.00
	10	99.74	99.21	94.00	99.21	99.21	94.00
	12	99.47	98.95	92.00	98.95	98.95	92.00

Table 7. Comparative results for TGI. RGI and RII for MMCWT

The training durations are increasing by retraining. Some comparative results are presented in Table 8.

HMMs		TGI Training time (sec.)	TGI Average testing time/word (sec.)	RGI Training time (sec.)	RGI Average testing time/word (sec)	RII Training time (sec)	RII Average testing time/word (sec.)
CHMM	SMM	157	0.092	211	0.096	360	0.11

Table 8. Comparative results for training and testing durations for TGI, RGI and RII

As global remark, we can conclude that retraining procedures enhance the recognition performance, but also are raising the training durations.

2.4 Gender trained models

In speaker independent speech recognition for large vocabularies the training strategies for the acoustical models are very important: a well trained model has high generalization properties and leads to acceptable word and phrase recognition rates, even without special speaker adaptation procedures. This purpose can be realized in the simplest way by speaker selection in the training phase.

In our experiments made we have assessed the speech recognition performance configuring the training database in three manners: only with female speakers (FS), only with male speakers (MS), combining male and female speakers (MS and FS). In order to find out which training strategy ensures the highest generalization capacity, the tests were made with two kinds of databases: only with female speakers (FS), only with male speakers (MS).

For continuous speech recognition there are two databases namely CDRL (Continuous Database for Romanian Language) and SCDRL (Second Continuous Database for Romanian Language).

The characteristics (Dumitru, Gavat, 2007) for the first database CDRL are the following: the database is constituted for training by 3300 phrases, spoken by 11 speakers, 7 males and 4 female speakers, each speaker reading 300 phrases, and for testing by 880 phrases spoken by the same speakers, each of them reading 80 phrases. The training database contains over 3200 distinct words, while the testing database contains 1500 distinct words.

The second database, SCDRL, contain 2000 phrases, spoken by 5 males speakers and 5 females speakers; each of them reading 200 phrases for training and 100 phrases, 20 phrases spoken by 5 speakers (3 males speakers and 2 females speakers) for testing. The numbers of the distinct words are: 11000 words for training and 760 for testing.

The data are sampled for CDRL by 44.1 kHz and for SCDRL by 16 kHz, quantified with 16 bits, and recorded in a laboratory environment.

In order to carie out our experiments the 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 and FS). In the case of independent speaker we have excluded one MS and one FS from the training and we used for testing (Dumitru, 2006).

To assess the progresses made with our ASRS_RL system we initiated comparative tests for the performance expressed in word recognition rate (WRR) to establish the values under the

new conditions versus the starting ones. The comparison (CDRL *vs.* SCDRL) is made for the following situations (in Table 9):

- Gender based training/mixed training;
- MFCC_D_A (36 mel-frequency cepstral coefficients with the first and second order variation);
- HMM - monophone modeling.

Training	Testing	CDRL	SCDRL
Training MS	Testing MS	56.33	55.45
	Testing FS	40.98	50.72
Training FS	Testing MS	53.56	43.91
	Testing FS	56.67	64.18
Training MS & FS	Testing MS	57.44	53.53
	Testing FS	49.89	63.22

Table 9. Comparison between CDRL and SCDRL for the case of independent speaker.

Similar results are obtained in the case of dependent speaker, (tested speaker was used in the training too) for example in Table 10 is presented the results for SCDRL.

Training	Testing	SCDRL
Training MS	Testing MS	71.62
	Testing FS	53.58
Training FS	Testing MS	55.07
	Testing FS	73.30
Training MS & FS	Testing MS	68.58
	Testing FS	67.79

Table 10. WRR for SCDRL for the case of dependent speaker.

Newly, trying to improve the word recognition rate, we chose the triphone modeling and we extended the area of extracted parameters (features) from the speech signal to PLP. The results obtained for triphone using two parameterizations, MFCC_D_A with 36 coefficients and PLP with only 5 coefficients are displayed in Table 11 for CDRL database (Gavat, Dumitru, 2008).

Training	Testing	MFCC_D_A		PLP	
		Monophone	Triphone	Monophone	Triphone
Training MS	Testing MS	56.33	81.02%	34.02	68.10
	Testing FS	40.98	72.86	25.12	59.00
Training FS	Testing MS	53.56	69.23	23.78	53.02
	Testing FS	56.67	78.43	34.22	58.55
Training MS & FS	Testing MS	57.44	78.24	47.00	70.11
	Testing FS	49.89	74.95	41.22	69.65

Table 11. WRR (for CDRL) in the case of monophone *vs.* triphone for MFCC_D_A and PLP coefficients.

The obtained results show that gender training is effective only if testing is done for the same gender. Than the results are better as in the case of mixed training.

3. Language models

The language model is an important knowledge source, constraining the search for the word sequence that has produced the analyzed observations, in form of succession of feature vectors. In the language model are included syntactic and semantic constraints. If only syntactic constraints are expressed, the language model is reduced to a grammar. In the following we will present first some basic aspects concerning the language modelling and further experimental results obtained in ASRU experiments on a database of natural, spontaneous speech, constituted by broadcasted meteorological news.

3.1 Basics about language modeling

Language models intend to capture the interrelations between words in order to gain understanding of the phrases, their meaning. The language models best known are from two kinds: rule – based and statistical. Rule-based models conduct to a certain word sequence based on a set of rules. Statistical models determine the word sequence based on a statistical analysis of large amounts of data (Huang et al., 2001), (Juang et Furui, 2000).

3.1.a N-gram statistical models

From statistical point of view the language model is represented in relation (10) by the probability *P(W)*, that can be written in the form:

$$P(W) = P(w_1, w_2, ..., w_n) = P(w_1)P(w_2|w_1)\cdots P(w_n|w_1, w_2, ..., w_{n-1}) = \\ = \prod_{i=1}^{n} P(w_i|w_1, w_2, ..., w_{i-1}) \tag{10}$$

$P(w_i|w_1, w_2, ..., w_{i-1})$ is the probability that the word w_i follows after the word sequence w_1, w_2, $\ldots w_{i-1}$. The choice of w_i depends on the whole input history. For a vocabulary having as dimension v there are possible v^{i-1} different histories; it is a huge number, making practically impossible to estimate the probabilities even for not big i values.

To find a solution, shorter histories are considered and the most effective one is based of a history of two preceeding words, called the *trigram* model $P(w_i|w_{i-1}, w \quad i-2)$). In a similar way could be introduced the *unigram* ($P(w_i)$), or the *bigram* ($P(w_i|w_{i-1})$). Our language model is *bigram* based

3.2 Parsing technics

Parsing algorithms are applied to search the desired word sequence in an utterance, based on rules or based on statistics.

A parser based on rules is represented in Fig. 4, the statistical one is depicted in Fig.5.

Based on rules, parsing becomes dependent from linguists specialized knowledge in order to establish this rules. Based on learning to create from a training corpus the language model, the statistical parser is flexible and also more independent from specialized expertize. Efficiency of the statistical parsing can be enhanced by the so called boot-strapp training (Huang et al., 2001), consisting in developping a model for a part of the training corpus and refining this model successivelly in completing the whole trainig material.

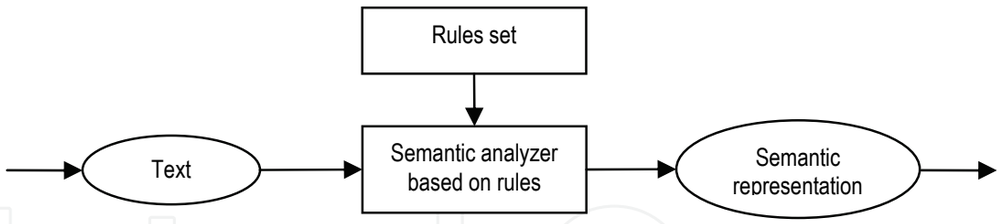


Fig. 4. Rule based parser

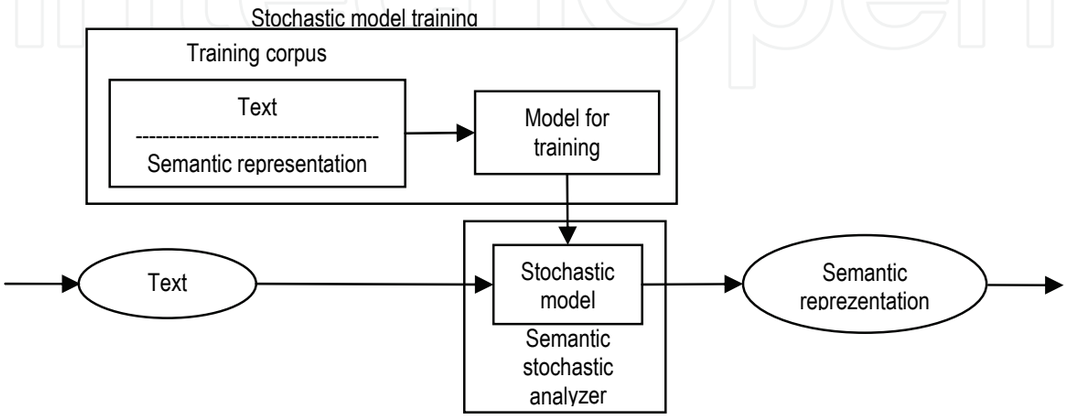


Fig. 5. Statistical parser

3.3 Experimental results

The language models applied in our research are:

- a simple word loop grammar, (WLG), permitting ever word sequence, without restrictions
- a restrictive finite state grammar (FSG), allowing only certain word sequences
- a bigram statistical model, extracting valid word sequences based on the bigram probabilities

The further displayed results were obtained on the MeteoRL database, constituted by registration of broadcasted meteorological news. The database contains 700 phrases, given a dictionary of 534 words. There are different speakers, male and females, not adnotated in our data.

The experiments had as objective to determine the influence of the language model on the ASRU performance, expressed in WRR, accuracy and PRR.

The conducted experiments were realized under the above listed conditions:

- Single mixtures for monophones, intra-word triphones and cross-word triphones (Table 12)
- Multiple mixtures for monophones, intra-word triphones and cross-word triphones (Table 13)
- Semi continuous models in form of monophones and triphones (Table 14)

Investigation of training duration and average testing duration/word were also done and the obtained results are displayed in Table 15. It is to notice the increase of the training but also of the average testing durations/word

HMMs/ number of mixtures		PLP + E + D + A			PLP + D + A		
		WRR	Accuracy	PRR	WRR	Acuracy	PRR
SMM	WLG	67.28	65.76	2.08	65.98	64.79	4.17
	FSG	95.89	95.11	55.32	96.33	95.67	53.19
	Bigram	97.40	96.97	72.92	98.70	98.37	81.25
SMIWT	WLG	87.87	81.15	12.50	89.38	85.70	14.58
	FSG	96.67	95.56	65.96	97.33	96.11	68.09
	Bigram	99.35	99.24	89.58	99.24	99.02	85.42
SMCWT	WLG	81.15	79.20	2.08	85.98	81.74	2.64
	FSG	96.11	95.56	59.57	97.67	96.56	61.70
	Bigram	99.35	99.24	87.50	99.13	98.92	85.42

Table 12. Comparative recognition performance for WLG, FSG and bigram model in the case of singular mixtures of monophones, intra-word triphones and cross-word triphones for continuous models.

HMMs/ number of mixtures		PLP + E + D + A			PLP + D + A		
		WRR	Accuracy	PRR	WRR	Acuracy	PRR
10 MMM	WLG	87.64	86.44	10.42	80.15	80.04	10.42
	FSG	97.22	96.44	65.96	96.89	96.00	59.57
	Bigram	98.70	98.59	83.33	98.37	98.16	83.33
6 MMIWT	WLG	87.87	81.15	12.50	89.11	87.11	17.02
	FSG	98.05	97.14	71.74	97.00	96.33	70.21
	Bigram	99.35	99.13	85.42	99.35	99.02	81.25
6 MMCWT	WLG	82.20	77.69	6.25	81.78	78.89	4.26
	FSG	97.71	97.37	69.57	96.67	96.22	68.09
	Bigram	99.44	99.44	89.36	99.00	99.00	87.23

Table 13. Comparative recognition performance for WLG, FSG and bigram model in the case of multiple mixtures of monophones, intra-word triphones and cross-word triphones for continuous models.

HMMs	PLP + E + D + A			PLP + D + A		
	WRR	Accuracy	PRR	WRR	Acuracy	PRR
Monophones						
WLG	83.42	80.82	6.25	78.01	77.03	4.17
FSG	97.00	96.22	61.70	96.89	96.22	59.57
Bigram	98.81	98.59	83.33	99.13	98.92	85.42
Triphones						
WLG	92.74	88.52	22.92	90.15	88.42	16.67
FSG	97.89	96.78	65.96	98.33	97.11	68.09
Bigram	98.81	98.37	77.08	98.22	98.00	76.60

Table 14. Comparative recognition performance for WLG, FSG and bigram model in the case of monophones and triphones of semicontinuous models.

HMMs		Training duration (sec.)	Average testing duration/word (sec.)		
			WLG	FSG	Bigram
CHMM	SMM	162	3.220	1.18	0.102
	10 MMM	672	3.930	1.43	0.130
	SMIWT	259	2.979	0.037	0.072
	6 MMIWT	361	1.250	0.069	0.097
	SMCWT	261	3.289	0.078	0.122
	6 MMCWT	652	7.390	0.143	0.205
SCHMM	Monophones	3.250	4.004	0.902	0.108
	Triphones	8.040	4.031	1.552	1.002

Table 15. Training duration and average testing duration/word for different language modeling and acoustical modeling techniques.

4. Discussion and comments of the experimental results

In sections 2 and 3 of this chapter we presented results obtained in continuous speech recognition and understanding experiments for Romanian language concerning the efficiency of:

- acoustical modeling based on monophones and triphones in continuous and semicontinuous models, with singular and multiple gaussian mixtures
- training with global initialization and retraining with global and individual initialization
- gender based training
- introduction of language models based on finite state grammars and bigram modeling

Some discussions and comments of this results could be usefull to conclude about the done work and future work directions.

4.1 Monophone and triphone models

All the experiments were carried out on the OCDRL database.

Comparing the results obtained for CDHMM models with singular mixtures (Table 2) it is obvious that triphone modeling enhance the recognition performance: WRR is increasing from 84.47% for monophones to 91.84% for CWT and to 97.37% for IWT, a maximum enhancement of more than 12%. Applying CMN the WRR marks again a slight increase.

The results obtained for CDHMM models with multiple mixtures (Table 3) show a WRR enhancement of around 12% for monophones with singular mixtures to monophones with five mixtures, with slight increase by increase of mixture numbers. For triphones, the WRR enhancement from single mixtures to multiple ones is not so spectacular: around 2% for IWT and 3% for CWT. Increasing the number of mixtures only slight increase in the WRR can be noticed. But because training time increases for multiple mixtures (Table 3) from 157s to 2323s for monophones, from 263s to 1087s for IWT and from 220s to 1106s for CWT it is better to not increase too much the mixtures number.

For SCHMM models, WRR increases of more than 1% can be remarqued by passing from monophones to triphones (Table 4).

4.2 Training and retraining

Training of the models was done with global initialization, retraining with global and individual initialization on the OCDRL database.
Of course an increase of the training and testing/word durations is to be noticed from 157s/0.092s for TGI to 211s/0.096s for RGI and 360s/0.11s for RII. (Table 8).
Increases, but not spectacular can be reported for WRR: for SMCWT for example, from 97.63% for TGI with CMN to 98.68% for RGI and RII with CMN (Table 6). Slightly higher increases can be reported for MMCWT for example: from 96.68% for TGI to 98.68% for RGI and 98.95% for RII for the case of two mixtures (Table 7).

4.3 Gender based training

The experiments were carried out on two databasis, CDRL and SCDRL for monophones and triphones, using as features MFCCs with first and second order variations and PLP coefficients, training with MS, FS and mixed and testing with MS and FS.
Gender based training and testing enhance the WRR. For example, training MS and testing MS leads to a WRR from 56.33% for the CDRL data base and 55.45% on the SCDRL data base for monophones, and 81.02% for triphones; testing with FS leads respectively to 40.98%, 50.72% and 72.86%, sensible lower values as for MS (Table9 and Table 11).
But it is to notice that in mixed training, the testing results are only slightly worser than for the gender based case: for training MS and testing MS WRR is 71.62%, training FS and testing FS WRR is 73.3%, but for mixed training WRR is 68.58% for testing MS and 67.79% testing FS (Table 11).

4.4 Language modelling

For the experiments a natural spontaneous spoken language database was used, namely MeteoRL. It is a way to explain why the results obtained on this database for WRR, accuracy and PRR are sensible lower than on the OCDRL database in which prompted, read text is used as speech material. For the SLG, in case of SMM for example, WRR, accuracy and PRR are respectively 84,47%, 83,16%, 40% for the OCDRL database (Table 2) and only 67,28%, 65,76%, 2,08% for the MeteoRL database (Table 12). Improving the language model, this data become respectively 95,89%, 95,11, 55,32% for FSG and 97,40%, 96,97%, 72,92% for the bigram model (Table 12), so that spectacular improving in ASRU performance is achieved, It is to notice that globally, the results obtained for the MeteoRL database follow the same trends as for the OCDRL data base. The known hierarchies are preserved: the WRR and PRR are higher for triphone models than for monophones, for multiple mixtures models than for single mixtures ones. For this experiments it is to highlite the improvement resulted by enhancing language modeling: starting for WLG with WRR, accuracy and PRR having the values of 82,20%, 77,69%, 6,25% for 6MMCWT, they became 97,71%, 97,37%, 69,57% for FSG and 99,44%, 99,44%, 89,36% for the bigram model (Table 13). Enhancements in ASRU performance can also be reported for semicontinuous models (Table 14)

5. Conclusions and future work

The done experiments helped us to obtain a deeper insight in the ASRU technics based on the statistical framework. The progress done in this work mainly consists in enhancing the language model applying for the first time in ASRU experiments for Romanian language

more elaborated language models as the simple WLG in form of the FSG and the bigram model. It is a work that in the future has to be further continued and improved.

Our major concern for future work is to obtain a standard database for Romanian language to validate the results obtained in ASRU experiments. The databases we have used were done in the laboratory of our university, carefully and with hard work, but still not fulfilling all standard requirements in audio quality and speech content.

6. References

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

Dumitru, C.O. and Gavat, I. (2006). Features Extraction and Training Strategies in Continuous Speech Recognition for Romanian Language, *International Conference on Informatics in Control, Automation & Robotics – ICINCO 2006*, Setubal, Portugal, pp. 114-121.

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

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

Dumitru, C.O. and Gavat, I. (2008). NN and Hybrid Strategies for Speech Recognition in Romanian Language, *ICINCO 2008 – Workshop ANNIIP*, Funchal–Portugal, pp. 51-60

Gavat, I., Zirra, M. and Enescu, V. (1996-1). A Hybrid NN-HMM System for Connected Digit Recognition over Telephone in Romanian Language. *IVTTA '96 Proceedings*, Basking Ridge, N.J., pp. 37-40.

Gavat, I. and Zirra, M. (1996-2). Fuzzy models in Vowel Recognition for Romanian Language, *Fuzzy-IEEE '96 Proceedings*, New Orleans, pp. 1318-1326.

Gavat, I., Grigore, O., Zirra, M. and Cula, O. (1997). Fuzzy Variants of Hard Classification Rules, *NAFIPS'97 Proceedings*, New York, pp. 172-176.

Gavat, I., Zirra, M. and Cula, O. (1998). Hybrid Speech Recognition System with Discriminative Training Applied for Romanian Language, *MELECON '98 Proceedings*, Tel Aviv, Israel, pp. 11-15.

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

Gavat, I., Valsan, Z., Sabac, B., Grigore, O. and Militaru, D. (2001-1). Fuzzy Similarity Measures - Alternative to Improve Discriminative Capabilities of HMM Speech Recognizers, *ICA 2001 Proceedings*, Rome, Italy, pp. 2316-2317.

Gavat, I., Valsan, Z. and Grigore, O. (2001-2). Fuzzy-Variants of Hidden Markov Models Applied in Speech Recognition, *SCI 2001 Proceedings, Invited Session: Computational Intelligence in Signal and Image Processing,* Orlando, Florida, pp. 126-130.

Gavat, I. and Dumitru, C.O. (2002-1). Continuous Speech Segmentation Algorithms Based on Artificial Neural Networks, *The 6th World Multiconference on Systemics, Cybernetics and Informations - SCI 2002*, Florida, SUA, Vol. XIV, pp. 111-114.

Gavat, I., Dumitru, C.O., Costache, G. (2002-2). Application of Neural Networks in Speech Processing for Romanian Language, *Sixth Seminar on Neural Network Applications in Electrical Engineering - Neurel 2002*, Belgrade, Yugoslavia, pp. 65-70.

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)*, Bucharest, pp. 115-126.

Gavat, I., Costache, G., Iancu, C., Dumitru, C.O. (2005-1). SVM-based Multimedia Classifier, *WSEAS Transactions on Information Science and Applications*, Issue 3, Vol. 2, pp. 305-310.

Gavat, I., Dumitru, C.O., Iancu, C., Costache, G. (2005-2). Learning Strategies in Speech Recognition, *The 47th International Symposium - ELMAR 2005*, Zadar, Croatia, pp. 237-240.

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.

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

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 .

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

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*, Kluwer Academic Publishers, Vol. 5, Number 3, pp. 259-268.

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

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.

Young, S., Kershaw, D., Woodland, P. (2006). *The HTK- Book*, U.K.



Speech Recognition

Edited by France Mihelic and Janez Zibert

ISBN 978-953-7619-29-9

Hard cover, 550 pages

Publisher InTech

Published online 01, November, 2008

Published in print edition November, 2008

Chapters in the first part of the book cover all the essential speech processing techniques for building robust, automatic speech recognition systems: the representation for speech signals and the methods for speech-features extraction, acoustic and language modeling, efficient algorithms for searching the hypothesis space, and multimodal approaches to speech recognition. The last part of the book is devoted to other speech processing applications that can use the information from automatic speech recognition for speaker identification and tracking, for prosody modeling in emotion-detection systems and in other speech processing applications that are able to operate in real-world environments, like mobile communication services and smart homes.

How to reference

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

Inge Gavat, Diana Mihaela Militaru and Corneliu Octavian Dumitru (2008). Knowledge Resources in Automatic Speech Recognition and Understanding for Romanian Language, Speech Recognition, France Mihelic and Janez Zibert (Ed.), ISBN: 978-953-7619-29-9, InTech, Available from: http://www.intechopen.com/books/speech_recognition/knowledge_resources_in_automatic_speech_recognition_and_understanding_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

© 2008 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License, which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen