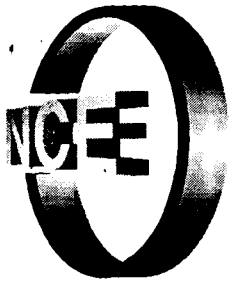

Relatório Técnico

Núcleo de Computação Eletrônica

Automatic Speech Recognition: A Study and Performance Evaluation on Neural Networks and Hidden Markov Models

Antonio G. Thomé
Sidney B. dos Santos
Suelaine S. Diniz

NCE - 15/99

Universidade Federal do Rio de Janeiro

Automatic Speech Recognition: A Study and Performance Evaluation on Neural Networks and Hidden Markov Models

Antonio G. Thomé (thome@nce.ufrj.br)
Federal University of Rio de Janeiro
Electronic Computational Center - IM
Computer Science Department
Rio de Janeiro, RJ - 20001-970 - Brazil

Sidney B. dos Santos (sidney@aquarius.ime.eb.br)
Military Institute of Engineering
Electrical Engineering Department
Rio de Janeiro, RJ - 22290-000 - Brazil

and

Suelaine S. Diniz (suelaine@fund.cepel.br)
Eletrobras Research Center
Rio de Janeiro, RJ - 20001-970 - Brazil

ABSTRACT

The main goal in this research is to find out possible ways to built hybrid systems, based on neural network (NN) and hidden Markov (HMM) models, for the task of automatic speech recognition. The investigation that has been conducted covers different types of neural network and hidden Markov models, and the combination of them into some hybrid models. The neural networks used were basically MLP and Radial Basis models. The hidden Markov models were basically different combinations of states and mixtures of the Continuous Density type of the Bakis model. A reduced set with ten words spoken in the Portuguese idiom, from Brazil, was carefully chosen to provide some pronounce and phonetic confusion. The results already obtained showed very positive, pointing toward to a high potentiality of such hybrid models.

Keywords: speech recognition, neural networks, hidden Markov models, hybrid systems.

1. Introduction

Automatic speech recognition is a very challenge and fertile area, with many fields still open and under intense study. The commercial interest on the subject is rapidly growing and is urging for scientific responses and solutions to the problems still present.

Among several different tasks to address within the area, we decided, for the sake of this investigation, to concentrate efforts on the framework of finding out possible ways to combine neural networks and hidden Markov models to face the isolated words recognition problem.

The research efforts in this field, have been concentrated on three major approaches: template, knowledge and stochastic based approximations. In the stochastic approach, the voice signal is seen as a random process which behavior can be reasonably learned and evaluated by a probabilistic model.

Hidden Markov and Neural Networks are the two major representative models within the stochastic class. Markov models, within the speech community, are far more well known than the neural models, that only recently have been started to be considered for the problem. The main difference between the two approaches is that Markov models are parametric and neural networks are non-parametric models. Either one presents its own set of characteristics, advantages and restrictions.[1,2]

It is known that any acoustic signal presents two major components: the temporal and the spectral components, that need to be considered on the construction of any approximation model. A third and not less important component for the recognition goal, is the phonetic context.

Hidden Markov, despite some peculiar restrictions, are able to address the first two components in a reasonable way. However, its independence hypothesis completely ignores any context information or relationship. The short dependence hypothesis on the first order Markov model, frequently causes difficulties to model coarticulation and duration of some phonetic elements. [3]

Feedforward neural networks on the other hand, despite not being able to address the temporal component, offer some important characteristics from the point of view of the recognition effort, such as: ability to learn complex and nonlinear mapping functions, ability to generalize from a set of given examples, parallelism, fault and noise tolerance.[2]

In this research we are investigating different types and configurations of neural network, hidden Markov and hybrid models. The hybrid models are from the type HMM-NN, where the Markov models can be seen as performing a nonlinear transformation on the input data set to be presented to the neural networks. In this combination, each HMM performs the temporal modeling and the NN does the spatial and the contextual modeling.

The neural networks in the experiments are basically different configurations of the MLP and the Radial Basis. For the hidden Markov models we only use Continuous Density models with different combinations of states and mixtures. Several simulations were done and the results are presented in this paper.

For all simulations we used a reduced set with ten words spoken in Portuguese from Brazil. The set of words was carefully chosen in order to provide some pronounce and phonetic confusion. The chosen words are: liga (turn on), pare (stop), grave (record), pausa (pause), avance (move), siga (proceed), volte (return), desliga (turn off), ejete (eject) and apague (erase). The amount of 1697 repetitions of this set of words were recorded from a group of 113 different male speakers, all among 20 and 30 years old, and coming from different geographic regions of Brazil.

In the following sections we describe the models used for the experiments, explain how the data set was formed, and discuss the results that were obtained.

2. Models and Configurations

The experiments were based on the construction of some hybrid models and the comparison of the performance obtained with them against those obtained with their single Markov and neural components.

The single HMM and neural models were those also used and described in [13].

2.1 Markov Models

Were restricted to the continuous density type and more specifically to the left-right strategy with delta equal 2, as in figure 1. Segmental Kmeans [6,7,8] and Viterbi [9] algorithms were used respectively for training and recognition.

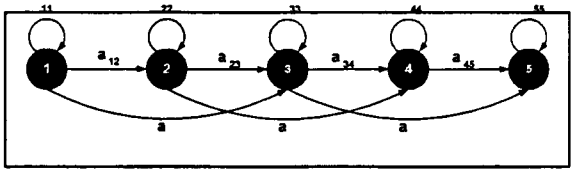

Figure 1 - Bakis Model

Six models were built using different combinations of states and gaussian mixtures: 8s5m (eight states and five mixtures); 8s10m, 10s5m, 10s10m, 15s5m and 15s10m.

2.2 Neural Models

Two models were investigated, as explained in [1 and 2]. One of them was the well known Multi-Layer Perceptron - MLP, trained with the backpropagation algorithm. The other model was of the Radial Basis Function type (RBF) where we changed the linear output layer by a nonlinear one, based on the sigmoid function. Both models were from the type feedforward.

The simulations, with the MLP model, were performed using different number of hidden layers and neurons. The best results were obtained with 2 hidden layers, 69 neurons in the first layer and 20 in the second (figure 2).

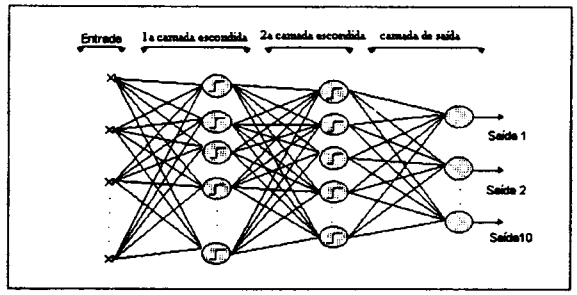

Figure 2 - MLP model with backpropagation

The best result with the RBF model was obtained with 10 gaussian nodes in the first layer, and 2 other layers with 40 and 20 sigmoid neurons. The gaussian parameters were defined through the k-means algorithm, and for the other layers it was used the backpropagation algorithm. For the output layer we used 10 neurons with sigmoid, one for each word to be recognized (figure 3).

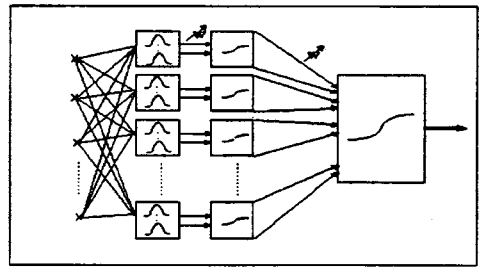

Figure 3 - RBF model with 3 layers, one gaussian and 2 sigmoid functions

2.3 Hybrid Models

Many possible ways to connect the two previous models into a third one were conceived. However, in all hybrid models we kept the same master approach, that is, the HMM was mainly used to catch the temporal component of the voice signal and the NN to catch the spatial and the contextual characteristics.

The recognition process using the Viterbi algorithm in a HMM, implies the search for the best segmentation of the sampled observations through the states of the model; the computation of the corresponding likelihood; and also the estimation of the probability function of the duration of each state. The state duration is measured directly from the training utterances and then, the mean and the standard deviation for the duration of each is calculated. The system likelihood is then estimated by the addition of the two previously found components, as can be seen below on the equation 1.

$$\log \hat{P}(q,O|\lambda) = \beta \log P(q,O|\lambda) + \alpha \sum_{j=1}^{N} \log[p(d_j)] \quad (1),$$

where α and β are scaling parameters.

With the expectation of improving the estimation of the system probability, that is, the estimation of the scaling factors, we decided to investigate the possibility to get any extra gain by adding a neural network to the output of the HMM model. The neural networks, this approach, have the advantage to receive the patterns of the response of each HMM to all words of the training set, what does not happen on the training of a Markov models.

In order to reduce the computational effort and the number of possible alternatives to investigate, we decided to work with only one architecture of HMM, the 8s5m, and one type of neural network, the feedforward MLP. They were not chosen because of the performance they achieved by themselves, but because they were simple and would demand less time and less computational effort for the training phase.

Four basic hybrid models were investigated: the first one, was built with the objective to verify the capability of a supervised training neural network (MLP) to improve the recognition performance by re-estimating the system likelihood based on a context information extracted from the conjunction of the likelihood provided by each one of the 10 HMM (figure 4). As showed below, a 2 layer MLP network with 10 input and 10 output, is trained taking as input the likelihood provided by each one of the Markov models.

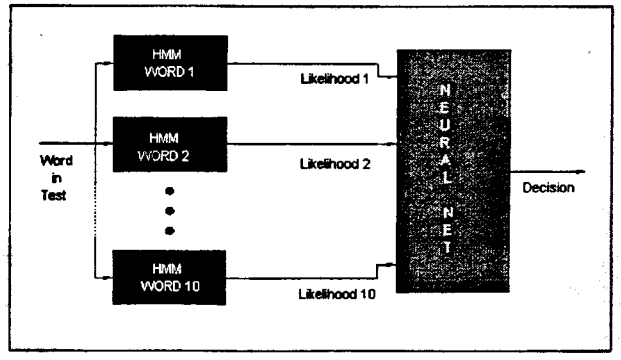

Figure 4 - Likelihood Hybrid Model

The second model was built considering as input not only the HMM likelihood but also the measured duration of its states. So, the network input dimension jumped from 10 to 90, that is, 10 likelihood plus 8 states duration from each HMM. The third model is similar to the second one, but instead of the duration of the states itself, it works with the estimated probability of the duration given by the Viterbi algorithm (figure 5).

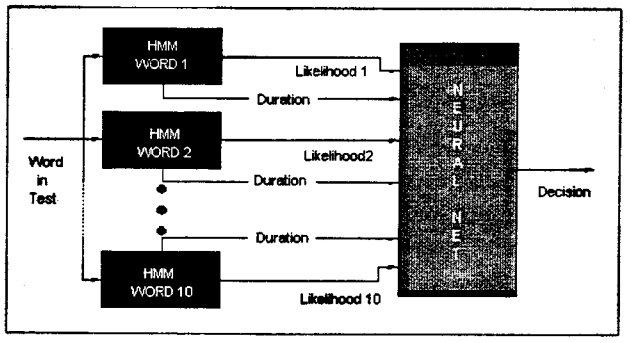

Figure 5 - Likelihood / Duration Hybrid Model

The fourth model, instead of working with the system likelihood provided by each HMM, receives 8 local, that is per state, likelihood from each model. The dimension of the network input now jumps to 160. The fifth and last model is also similar to the second one with respect to the type and number of inputs. The difference is that now there is a logic gate between the HMM and the NN. If the HMM output satisfies a defined confidence criterion, then its likelihood is assumed to be the likelihood of the hybrid model. Otherwise, a specific network is chosen from a set of candidates and activated to provide the hybrid likelihood. The network selection is based on the confusion group, that is, the subset of HMM which confidence factors are above certain limit.

The main goal with this fifth model was to investigate the capability of getting improve to the recognition giving special attention to those sets of words we noticed that the HMMs generated high level of confusion (figure 6).

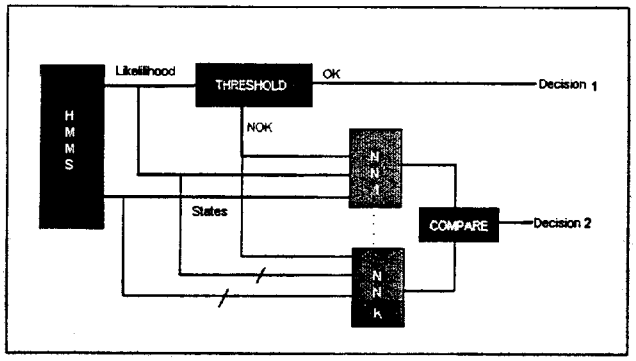

Figure 6- Switched Hybrid Model

The Experiment Context

Working with the same data set as in [1 and 2], we also split it into 3 groups: one for training and 2 for testing.

The data set consists of 1697 isolated words from a small vocabulary of 10 words and spoken by 116 different male speakers, all averaging 25 years old and coming from different regions of Brazil. All records were done with a common 16 bits sound blaster board, in a room without any acoustic protection or isolation, and using a simple directive microphone.

As showed in the table below, 1000 patterns were used for training (100 from each word), 110 for the first test group and 697 for the second test group. For the training set we used 46 different speakers, 8 new speakers for the first test group and other 48 new speakers for the second test group. Another important factor, responsible for the higher complexity of the second group of test, is that 288 of its patterns were recorded with a second microphone.

Table 1 - Training and Testing Sets

	training	testing #1	testing #2
patterns	1000	110	697
speakers	46	11	59

The set of words was chosen in such a way to provide some acoustic confusion, like those obtained with: liga, siga and desliga; pare, grave e apague; volte and ejete. In the next section it is showed that the major HMM confusion was between pare/apague, liga/desliga and pausa/apague.

In [x], since the neural networks were from the static type, that is, the number of inputs had to be fixed, it was chosen to segment the voice signal into 120 variable size window with 76% of superposition. There it was also decided to use a reduced number of features, 4 ($1^{\underline{o}}$ cepstrum, $1^{\underline{o}}$, $2^{\underline{o}}$ and $3^{\underline{o}}$ mel-cepstrum coeficients [5]), not to over increase the computational demand of the neural training. In [2] this same segmentation and number of features restriction was observed for the HMM investigation. This in order to be able to compare their performance against the neural models.

Here, with the objective to investigate possible gains provided by usage of hybrid models, we decided to adopt the same segmentation and features used in [1].

After the first battery of experiments we also investigated the performance adopting a larger number of features, 26 (12 mel-cepstrum, 12 delta mel-cepstrum, energy and delta energy [5]) and a segmentation based on a free number of windows with 50% of superposition. Delta parameters were used in order to give some contextual information to the HMM. These parameters are calculed through the following expression [4]:

$$o_{\Delta}(n) = o(n+i) - o(n-i) \qquad (2)$$

where:

$o_{\Delta}(n)$ is the delta mel-cepstrum vector at time stamp n;

$o(n+i)$ is the mel-cepstrum vector at time stamp n+i.

In this reserch we adopted i=2, what was found very efetive for speech recognition [3].

The results are presented and discussed in the next section.

The Experiments Results

As said before, the objective was to compare the results and the potentiality of hybrid models against their single Markov and neural components. The first simulation set was performed restricting the models to the usage of those 4 features mentioned in the previous section. Table 2 shows the results obtained with the Markov models.

Table 2 - Markov Recognition Performance [18]

model	8s5m		8s10m		15s10m	
window	G1	G2	G1	G2	G1	G2
120w76s	85.5	75.0	89.1	75.5	89.1	**78.2**
xw76s	89.1	73.6	**90.9**	73.7	**90.9**	77.2
xw50s	87.3	73.5	89.1	73.6	89.1	76.3

As can be seen, models 8s10m and 15s10m provided the best result for the test group 1, both using free number of windows and 76% of superposition (xw76s). The best result for the second and more complex test group was provided by the model 15s10m, with a fixed number of 120 windows and up to 76% of superposition (120w76s). From the results we confirmed some hypothesis: higher number of states may improve performance; free number of windows improves ability to approximate temporal variation; larger window size improves robustness; and larger superposition area among adjacent windows improves correlation.

Table 3 is relative to the neural models and shows the best results obtained from different sizes of the models MLP and Radial Basis. As seen, both networks provided similar accuracy behavior for the two test groups. The difference between them is that the training effort was much more expensive for the MLP model.

Table 3 - Neural Networks Recognition Performance

	G1	G2
MLP	95.0	79.0
RBF	98.0	80.5

Neural Network models outperformed Markov models in between 3 and 8% considering all cases. In fact, the difference was not that high, mainly if we remember that the Markov models were restricted to the small number features used for the neural models. In both cases we had a loss of performance, around 20%, between groups 1 and 2.

Table 4 summarizes the results obtained for each word of the vocabulary, considering test group 2, the best models and those selected for the hybrid investigations. Columns in italic highlight the models used in the construction of the hybrid ones.

Table 4 - Best Results per Word - Test Group 2

	#w	15s10m	*8s5m*	*MLP*	RBF
Liga	66	90.9	*86.4*	*86.4*	81.8
Pare	114	61.4	*54.4*	*86.0*	83.3
Grave	114	74.6	*65.8*	*83.3*	79.8
Pausa	114	74.6	*72.8*	*79.8*	90.4
Avance	114	70.2	*71.1*	*52.6*	60.5
Siga	16	100.0	*100.0*	*93.8*	100.0
Desliga	16	100.0	*100.0*	*62.5*	68.8
Volte	64	92.2	*92.2*	*87.5*	89.1
Ejete	16	93.8	*93.8*	*68.8*	68.8
Apague	63	93.7	*93.7*	*92.0*	85.7
Average	697	78.2	*75.0*	*79.0*	80.5

We noticed that some words were better recognized by the neural networks: pare, grave and pausa, and others by the Markov models: desliga, volte e ejete. Looking at the confusion matrix, like the one presented in table 5 for the 8s5m model, we also noticed that these are the words more frequently confused.

Table 5 - Confusion Matrix for the 8s5m HMM over the training set (100 repetitions of each word) and the testing set group 2 (table 4)

	Liga	Pare	Grave	Pausa	Avance
Liga	99/56				
Pare	0/2	95/54	0/4	0/4	1/4
Grave	0/3	1/1	99/65		0/3
Pausa				99/82	0/1
Avance			0/1	0/1	100/75
Siga					
Desliga	2/1				
Volte				0/2	
Ejete					
Apague		4/4	0/1	0/1	1/0

	Siga	Desliga	Volte	Ejete	Apague
Liga		1/10			
Pare	0/2				4/44
Grave	0/6	0/4		0/8	0/24
Pausa	0/1		1/11		0/19
Avance	0/1		0/15		0/21
Siga	99/16	1/0			
Desliga		98/15			
Volte			100/59		0/2
Ejete		0/1		100/15	
Apague	1/0				94/57

Notice that the majority of the confused words are: pare with apague; grave with apague; avance with apague and volte; and liga with desliga.

The first two hybrid models were built like the description in previous section, and the results obtained with them are summarized in the table below, where H1 refers to the first model, with likelihood only, and H2 to the second model, with likelihood and duration of states.

Table 5 - H1 and H2 Hybrid Models - Test Groups 1 / 2

	Test Group 1		Test Group 2	
	H1	H2	H1	H2
Liga	90.9	100.0	84.8	89.4
Pare	54.5	72.7	50.0	80.7
Grave	72.7	81.8	65.8	75.4
Pausa	100.0	100.0	69.3	76.3
Avance	90.9	100.0	69.3	82.5
Siga	100.0	100.0	100.0	100.0
Desliga	90.9	100.0	93.8	87.5
Volte	90.9	90.9	93.8	92.2
Ejete	100.0	100.0	100.0	100.0
Apague	81.8	90.9	95.2	88.9
Average	87.3	93.6	73.6	83.1

Second model outperformed the first model in about 8% and 11% respectively, for the first and second test groups. H2 provided 100% matching for 6 words in the first test group and both models provided 100% matching for the words siga and ejete in both tests.

We are still working on the simulations with the third and fourth models. Up to now they provided only a tinny improvement, around 4% and 7% respectively, on the results already obtained with H2, the second hybrid model. Our expectation, however, is that these gains will most likely be greater once we can solve some local problems.

The major obstacle we are facing with the third model refers to the amount of time required for the neural network training. They now are 160 input dimension against 10 and 90 respectively of the first and second hybrid models.

With the fourth hybrid model the difficulty refers to the confidence factors specification. Such values are very important because in this scheme all existing components of the model provide its own output for the given input and then, the output of the system is selected based on the confidence factor computed for each group: the Markov and neural agents, as in figure 6. We have being using ten HMM as before, one for each word of the vocabulary, and three NN, one for the whole set of words, one for the subset of the words "pare/grave /apague", and one for the words "siga/liga/desliga".

Conclusions

The main objective in this research is to search for different ways to build hybrid systems based on the combination of Markov and Neural models, applied to the problem of isolated word recognition in the context of independence of the speaker.

The vocabulary consists of a set of ten words, which were recorded from 113 male speakers speaking in the Portuguese idiom from Brazil. Four hybrid models were conceived, implemented and tested.

The comparison of the results provided by the hybrid models against those provided by the single Markov and Neural models, pointed toward a higher performance of the hybrid schemes.

The research is in progress and many simulations are still being performed, as well as other schemes to built hybrid systems are being tested. We are also working on the expansion of the vocabulary in order to include a more variety of speakers and words.

References

[1] RABINER, L. R. & JUANG, B. H.,"Fundamentals of Speech Recognition", Pretice Hall, Inc. Englewood Clifs, Nova Jersey, 1993.

[2] HAYKIN, S., "Neural Networks – A Compreehensive Foundation", Macmillan, 1994.

[3] DELLER Jr., J. R. et al., "Discrete-Time Processing of Speech Signals", Macmillan Publishing Company, New York, 1993.

[4] SANTOS, S. C. B.,"Continuous Speech Recognition for Portuguese Using Hidden Markov Models (in Portuguese), Ph. D. Thesis, Department of Electrical Engineering, Pontifical Catholic University of Rio de Janeiro, Brazil, 1997.

[5] DAVIS, S. B., MERMELSTEIN P., "Comparison of Parametric Representations for Monossyllabic Word Recognition in Contiuously Spoken Sentences", IEEE Trans. ASSP-28 (4), pp. 357-366, 1980.

[6] WILPON, J. G. and RABINER, L. R.,"A Modified K-Means Clustering Algorithm for use in Isolated Word Recognition", IEEE Trans. ASSP-33 (3), pp. 587-594, June 1985.

[7] JUANG, B. H., LEVINSON, S. E., SONDHI, M. M., "Maximum Likelihood Estimation for Multivariate Mixture Observations of Markov Chains", IEEE Trans. Information Theory, IT-32 (2), pp. 307-309, March 1986.

[8] RABINER, L. R., JUANG, B. H., LEVINSON, S. E. and SONDHI, M. M., "Recogniton of Isolated Digits Using Hidden Markov Models With Continuous Mixture Densities", AT&T Tech. J., 64 (6), pp. 1211-1234, July-Aug. 1985.

[9] FORNEY, G. D., "The Viterbi Algorithm", Proc. IEEE, 61, pp. 268-278, March 1973.

[10] LEE, K. F. "Automatic Speech Recognition — The Development of the SPHINX System", Kluwer Academic Publisher, Boston, 1989.

[11] PARANAGUÁ, E.D.S. "Reconhecimento de Locutores Utilizando Modelos de Markov Escondidos Contínuos", Tese de Mestrado, IME, 1997.

[12] DINIZ, S.S. "Uso de Técnicas Neurais para o Reconhecimento de Comandos à Voz", Tese de Mestrado, IME, 1997.

[13] COLE, R. A. "Survey of the State of the Art in Human Language Technology", Center For Spoken Languagem Understanding, Oregon Graduate Institute, Publicações Técnicas, Nov. 1995.

[14] DINIZ, S. S. , THOMÉ, A. C. G. "Uso de Técnica Neural para o Reconhecimento de Comandos à Voz", IV Simpósio Brasileiro de Redes Neurais, pp 23-26, Dez 1997.

[15] DUDA, O.R., HART, E., P., "Pattern Classification and Scene Analysis", Wiley- Intercience, 1973, pp. 114-118.

[16] RENALS, S., MORGAN, N., "Connectionist Probabilit Estimation in HMM Speech Recogation", International Computer Science Institute, Berkeley, Dec 1992.

[17] CHO, S-B, "Neural- Network Classifiers for Recognizing Totally Unconstrained Handwritten Numerals" , IEEE Transactions on Neural Networks, vol. 8, n.1, January 1997.

[18] SILVA, D.G. "Comparação entre os Modelos de Markov Escondidos Contínuos e as Redes Neurais Artificiais no Reconhecimento de Voz", Projeto de Fim de Curso, IME, 1997.

[19] LEE,C.H; SOONG F.K.; PALIWAL K.K. "Automatic Speech and Speaker Recognation -Advanced Topics", Kluwer Academic Publisher, Boston, 1996.