SPEECH RECOGNITION BY USING ASRS_RL

Inge Gavăt, Corneliu Octavian Dumitru

University Politehnica Bucharest, Faculty of Electronics Telecommunications,
313 Splaiul Independentei, Bucharest, Romania

igavat@alpha.imag.pub.ro (Inge Gavăt)

Abstract

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 and with a great variability in the formal manifestation as an acoustic signal.
Currently, the best ranked technology in speech recognition is based on hidden Markov models (HMM) as classifiers, of course there are other alternative like artificial neural networks (ANN) or support vector machine (SVM) in the classifier domain.
In the first part of the paper, a classifiers based on HMM is compared in the simple task of vowel recognition with a classifier based on the multilayer perceptron (MLP). In this situation, we have obtained better results for the last classifier, fact which highlights the advantage of the discriminative training of the perceptron versus the maximum likelihood training of the HMM. In the second part, a hybrid structure HMM/MLP is compared with the simple HMM in a digit recognition task. The hybrid structure improved with 2% the recognition rate. And in the last part, the continuous speech recognition experiments for Romanian language are describes by using HMM. The progresses concern enhancement of HMM by taking into account the context in form of triphones, improvement of speaker independence by applying a gender specific training and enlargement of the feature categories used to describe speech sequences.

Keywords: HMM, MLP, vowel, digit, continuous recognition.

Presenting Author's biography

Inge Gavăt. received the M.S. and Ph.D. degrees in Electronics and Telecommunications from the University "Politehnica" of Bucharest, Romania. She is currently a Professor at the Department of Applied Electronics and Information Engineering, Faculty for Electronics, Telecommunications and Information Technology and also at the German Branch of the Faculty of Engineering in Foreign Languages. She is teaching courses in Information and Estimation Theory, in Communication Theory and in Signal Processing and Artificial Intelligence for Man – Machine Communication. She is involved in advanced research projects and programs in pattern recognition with the Romanian Academy of Sciences, with the Military Technical Academy, with the Romanian Space Agency, with the Ministry for Education and Research. She is developing algorithms for pattern recognition with Markov models, neural networks and fuzzy systems.

ISBN 978-3-901608-32-2

1 Introduction

After years of research and development, the problem of automatic speech recognition and understanding is still an open issue. The end goal of translation into text, accurate and efficient, unaffected by speaker, environment or equipment used is very difficult to achieve and many challenges are to be faced.

Some factors which make difficult the problem of automatic speech recognition (ASR) are voice variations in context or in environment, syntax, the size of vocabulary. How this problems can easy be accommodated by humans, to continue studies in human speech perception can be very important to improve performance in speech recognition by machine. As alternative, human performance can be regarded as guide for ASRs and in this moment neither the best systems built for English language can not reach human performance.

The remainder of this paper will be also structured as follows: Section 2 illustrate the ASRS_RL interface. Section 3 present HMMs, MLP, database and experimental results for vowel recognition. Section 4 is dedicated to hybrid system, database and experimental results for digit recognition. Section 5 describes the context based acoustical modeling, realized with triphones, the database and experimental results for continuous speech recognition. Finally, conclusions are given in section 6.

2 ASRS_RL

We have made the experiments on our recognizer based on Automatic Speech Recognition System for Romanian Language (ASRS_RL), with multiple options for a large variety of speech recognition experiments.

The ASRS_RL has the following options, in order to choose:

- The recognition application, between vowel recognition, digit recognition and continuous speech recognition;
- The database for vowels recognition, for digit recognition and for continuous speech recognition;
- The database configuration for training and testing;
- The feature extraction method between LPC, PLP and MFCC;
- The modelling technique, between HMM-modelling, Neural Network-modelling, hybrid-modelling.

In order to easier handling the recognition experiments a MATLAB interface was designed. In Fig. 1 is presented the interface window for choosing our experiment conditions, for example: continuous speech recognition (phrases), Romanian language, HMM modelling.

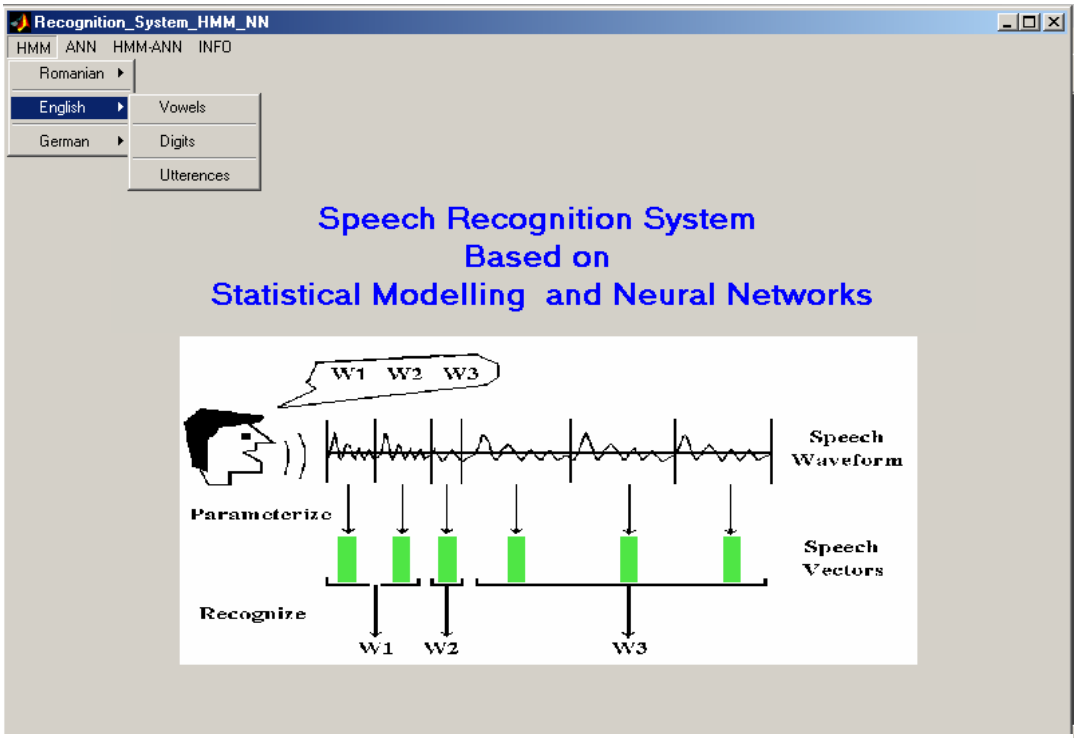

Fig. 1. Interface for ASRS_RL.

3 Vowel recognition

3.1 Hidden Markov Models (HMMs)

HMMs are finite automates, having 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 are taking place: the transparent one, represented by the observations string (feature sequence), and the hidden one, which cannot be observed, represented by the state string.

The left - right model (Bakis), which is considered the best choice for speech. For each symbol, such a model is constructed; a word string is obtained by connecting corresponding HMMs together in sequence [1].

Applying HMMs, there are three main problems:

- The first problem is the evaluation one, in the recognition phase, applied to the trained model. Given this model and the features sequence, we have to decide if the sequence is produced by the given model having as similarity measure the probability to produce this observation sequence with the given Markov model. This probability is calculated by the *"forward" or "backward"* algorithm.
- The second problem is about establishing the correct state sequence. The *"Viterby"* algorithm is one of the most used algorithms for this purpose and often only the probability of this state sequence is taken into account in the calculation of the similarity measure.
- The third problem is the dedicated to the training, the parameter optimization of the model to describe as good as possible the observation sequence. Training allows optimal adaptation of the model parameters to the training set of data by re-estimating them. The *"Baum-Welch"* algorithm is the most used parameter re-estimation algorithm.

3.2 Multilayer Perceptrons (MLPs)

In the following chapter we will discuss about MLP, that being the most common artificial neural network (ANN) architecture used for speech recognition.

Typically, MLP has a layered feed forward architecture, with an input layer, one or more intermediate (hidden) layers, and one output layer. Each layer computes a set of linear discriminative functions, followed by a non-linear function, which is often a sigmoid function.

One way of using the MLP for speech recognition consists of viewing the temporal acoustical vector as a spatial one, and processing it at once by the MLP. For the case of a vowel recognizer, each MLP output can be associated with one vowel.

In our case we have used a structure with three-layer MLP. An n-layer MLP consists of a layer of input nodes, n-1 hidden layer, and a layer of output node. The nodes from one layer are fully connected to the ones of the next higher layer by directed edges. The input features are propagated through the network from the input to the output nodes along the edges according to the optimization algorithm called *Back-Propagation* [2].

The *Back-Propagation* algorithm is guaranteed to converge to a local minimum, but it usually converges very slowly, and can be heuristically controlled by the parameter α called learning rate and by the momentum weight β, to be also practically chosen.

3.3 Database

The database contains speech data from 19 speakers (9 male speakers (MS) and 10 female speakers (FS)) for 5 vowels (a, e, i, o, u). We excluded one male speaker and one female speaker from the database and used their data for the testing.

The audio files contain signals sampled with 16KHz, 16 bit, mono, recorded with a desktop microphone, in a laboratory environment.

3.4 Experimental results

We compare the performance obtained for vowel with multilayer perceptrons (MLP) and hidden Markov models (HMM) for unrolled speaker.

The vowel parameters were extracted by cepstral analysis, in form of 12 mel-frequency cepstral coefficients (MFCCs).

- MLP: we used in our experiments 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. The number of the input nodes is equal to the number of features used to describe the vowels, namely the 12 MFC coefficients.
- HMM: for each vowel we constructed a hidden Markov model. The speech data were parameterized using cepstral analysis in form of 12 MFCCs.

For MLP we obtained recognition rates between 96.36% (training and testing with MS) and 77.85% (training with MS and testing with FS).

For HMM we obtained recognition rates between 91.26% (training and testing with MS) and 66.48% (training with MS and testing with FS) [6].

The recognition rates of the vowels in the case of MLP utilization are higher than in the case of HMM utilization (Fig. 2, Fig. 3). 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.

Trying to reduce this limitation effect of HMMs, we choose an alternative to combine the two approaches obtaining a hybrid system.

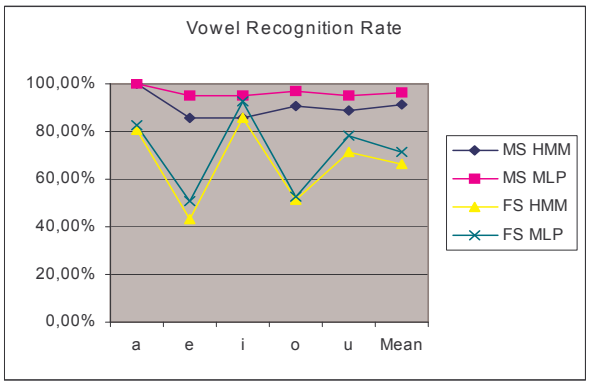

Fig. 3. Recognition rate: training FS and testing MS and FS.

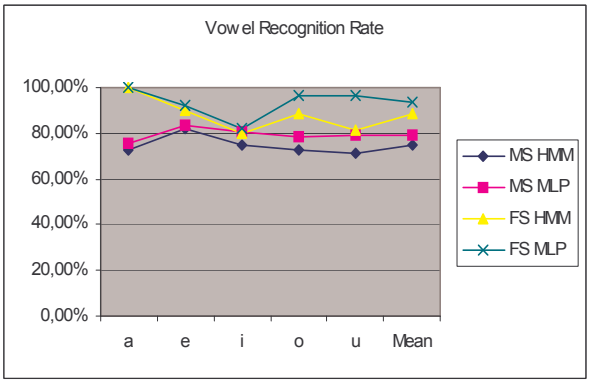

Fig. 4. Recognition rate: training MS and testing MS and FS.

4 Digit recognition

4.1 Hybrid system(HMM and MLP)

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. 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 $p(q_k / x_n)$, 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 contain 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 to the MLP outputs:

$$\frac{p(x_n \mid q_k)}{p(x_n)} = \frac{p(q_k \mid x_n)}{p(q_k)} \qquad (1)$$

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, applied for solving the speech recognition problem [4], [5].

4.2 Database

The database contains speech data from 9 speakers (6 male speakers (MS) and 3 female speakers (FS)) for 9 digits in Romanian language (unu, doi, trei, patru, cinci, şase, şapte, opt, nouă). We excluded the last 3 speaker (2 male speaker – speaker 9 (sp.9) and speaker 7 (sp.7)) and one female speaker – speaker 8 (sp.8)) from the database and used them for the testing.

The audio files contain signals sampled with 16KHz, 16 bit, mono, recorded with a desktop microphone, in a laboratory environment.

4.3 Experimental results

We compare the performance obtained in digit recognition task with hidden Markov models (HMM) and with the hybrid system (MLP-HMM) for unrolled and enrolled speaker.

The digit parameters were extracted by cepstral analysis, in form of 12 mel-frequency cepstral coefficients (MFCC).

- HMM: for each digit we constructed a hidden Markov model. The speech data was parameterized using cepstral analysis in form of 12 MFCCs.
- MLP-HMM: 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 [3].

The test results, for HMM and MLP-HMM for different training (for example: Tr=9 - number of speaker used for training = 9) and different system (HMM or MLP-HMM) are illustrated in Fig. 5.

The results obtained for unrolled speaker slightly lower than the results obtained for enrolled speaker.

Comparing the results obtained with hybrid system (MLP-HMM) and HMM we can see the results for HMM are slightly lower than the results for MLP-HMM with approximate 2.5% [6].

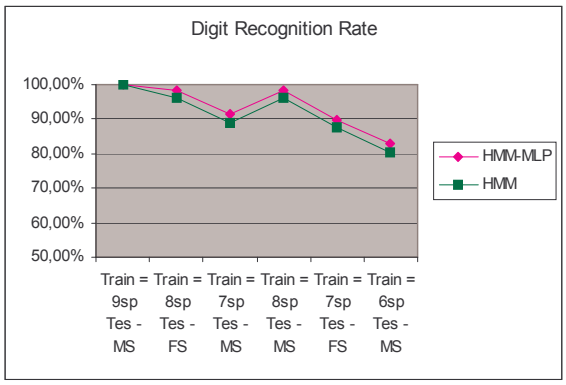

Fig. 5. Digit recognition rate.

5 Continuous speech recognition

5.1 Context-depending modeling

For small vocabulary recognition, 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. Therefore, usually for not very limited tasks are preferred phonetic models based on monophones, because the phones, as smallest linguistic units, are easy generalizable and of course also trainable. Monophones constitute the foundation of any training method, in any language, and we also started with them. But a refinement of this initial step was necessary because in real speech, the words are not simple strings of independent phonemes. The realization of a phoneme is strongly affected by its immediately neighboring phonemes by co-articulation. Because of this, monophone models have been changed in time with triphone models that became the actual state of the art in automatic speech recognition with large vocabularies [7], [8].

A triphone model is a phonetic model that takes into consideration the left and the right neighbouring phones. This immediate neighbour – phonemes are called respectively the left and the right context; a phoneme constitutes with the left and right context a triphone. For example in the SAMPA (Speech Assessment Methods Phonetic Alphabet) [9] transcription “m - a + j” of the Romanian word ”mai”, regarded as triphone, the phoneme “m” has as left context “a” and as right context “j”.

For each such a triphone a model must be trained: in Romanian that will give a number which equals 40,000 models, situation totally unacceptable for a real world system. In our speech recognition task we have modeled only internal – word triphones and the adopted state tying procedure has conducted to a controllable situation.

If triphones are used in place of monophonemes, the number of needed model 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.

The choice of the states and the clustering in phonetic classes are achieved by mean of phonetic decision trees. A phonetic decision tree is built as a binary tree and has in the root node all the training frames to be tied, in other words all the contexts of a phoneme. To each node of the tree, beginning with the parent – nodes, a question is associated concerning the contexts of the phoneme.

Possible questions are, for example: is the right context a consonant, is the left context a phoneme “a”; the first answer designates a large class of phonemes, the second only a single phonetic element. Depending on the answer, yes or no, child nodes are created and the frames are placed in them. New questions are further made for the child nodes, and the frames are divided again.

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 resulting a leaf node. In such leaf nodes are concentrated all states having the same answer to the question made along the path from the root node and therefore states reaching the same leaf node can be tied as regarded acoustically similar. For each leaf node pair the occupancy must be calculated in order to merge insufficient occupied leaf nodes.

A decision tree is built for each state of each phoneme. The sequential top down construction of the decision trees was realized automatically, with an algorithm selecting the questions to be answered from a large set of 130 questions, established after knowledge about phonetic rules for Romanian language.

5.2 Database

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

For continuous speech recognition, database for training is constituted by 3300 phrases, uttered by 11 speakers, 7 males and 4 females, each speaker reading 300 phrases [11].

The databases for testing contained 220 phrases uttered by 11 speakers, each of them reading 20 phrases.

The training database contains over 3200 distinct words; the testing database contains 900 distinct words.

In order to carry out our experiments about speaker independence, 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 all cases we have excluded one MS and one FS from the training and used for testing.

5.3 Experimental results

We initial compare tests for the performance expressed in word recognition rate (WRR) to establish the values under the new conditions versus the preceding ones. The comparison is made for the following situations:

- Triphone modeling/monophone modelling,
- Gender based training/mixed training,
- LPC and PLP/MFCC.

The speech files from these databases were analysed in order to extract the interesting features. The feature extraction methods used are based on linear predictive coding (LPC), perceptual linear prediction (PLP) [10] mel-frequency cepstral coefficients (MFCC).

The WRR are:

- 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).
- 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).
- 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 [12].

6 Conclusions

The recognition rates for vowels in the case of MLP utilization are higher than in the HMM utilization, because model training is discriminative, while in the case of HMM the training is not discriminative, and this represents a disadvantage of HMM utilization.

Trying to reduce these limitation effects of HMMs, we chose an alternative, combining the two approaches and we obtained a hybrid system.

This hybrid system has been successfully applied for digit recognition within MLP was used as an *a posteriori* probability estimator of the HMM states.

After the experiments made on continuous speech recognition we have following conclusions:

- The triphone modeling is effective, conducting to increasing in WRR between 15% and 30% versus the monophone modeling. The maximal enhancement exceeds 30% for training MS and testing FS for MFCC_D_A (depicted in Fig. 6).
- A gender based training conduct to good result for test 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 (depicted in Fig. 7).

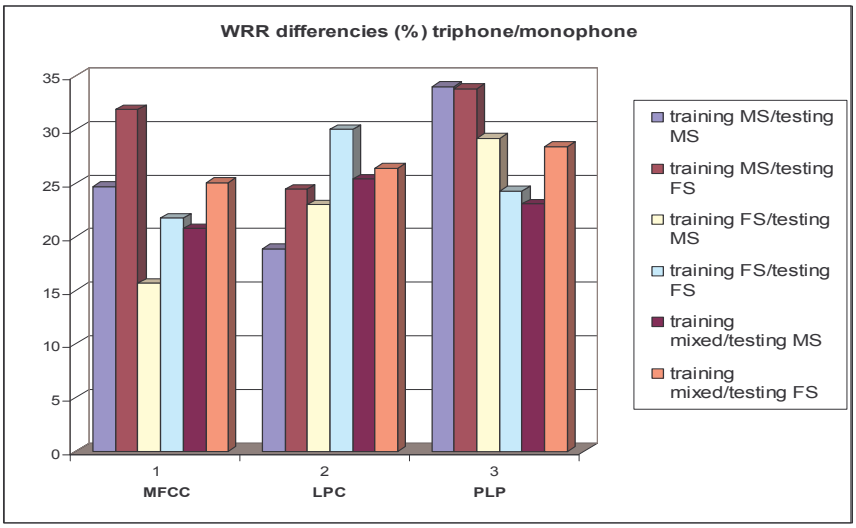

Fig. 6. Triphone modelling *versus* monophone modelling.

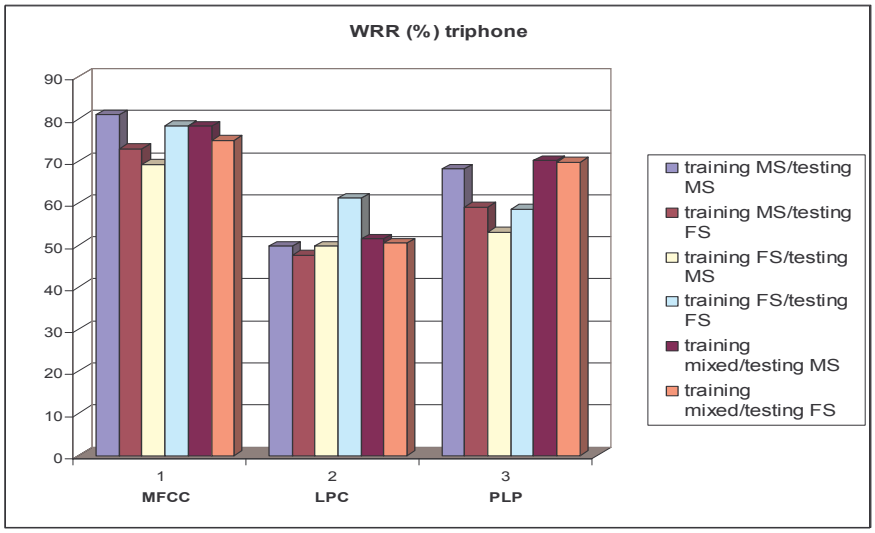

Fig. 7. WRR for triphone modeling for different case of training and testing speaker (MS or FS).

7 References

[1] I. Gavat, O. Dumitru, C. Iancu, G. Costache. Learning Strategies in Speech Recognition. *Proc. ELMAR 2005,* 237-240, 2005.

[2] S. Goronzy. Robust Adaptation to Non-Native Accents in Automatic Speech Recognition. Springer – Verlag Berlin Heidelberg, Germany. 2002.

[3] Z. Valsan, I. Gavat, B. Sabac, O. Cula, O. Grigore, D. Militaru, C.O. Dumitru, Statistical and Hybrid Methods for Speech Recognition in Romanian, *International Journal of Speech Technology*, 5(3): 259-268, 2002.

[4] I. Gavat & all. Elemente de sinteza si recunoasterea vorbirii. Printech, Bucharest. 2000.

[5] I. Gavat, M. Zirra. A Hybrid NN – HMM System for Connected Digit Recognition over Telephone in Romanian Language. *Proc. Third Workshop on Interactive Voice Transmission in Telecommunication Applications*, 37-40, 1996.

[6] C.O. Dumitru, I. Gavat. Statistical, Neural and Hybrid Methods for Speech Recognition in Romanian Language. *COMMUNICATIONS 2006*, 139-142, 2006.

[7] Young, S.J. The General Use of Tying in Phoneme-Based HMM Speech Recognizers. *Proceedings of ICASSP 1992*, 1: 569-572, 1992.

[8] S.J. Young, J.J. Odell, P.C. Woodland. Tree Based State Tying for High Accuracy Modeling. *ARPA Workshop on Human Language Technology*, 1994.

[9] http://www.phon.ucl.ac.uk/home/sampa/home.htm

[10]H. Hermansky. Perceptual Linear Predictive Analysis of Speech. *J. Acoust. Soc. America*, 87(4): 1738-1752, 1990.

[11] E. Oancea, I. Gavat, C.O. Dumitru, D. Munteanu. Continuous speech recognition for Romanian language based on context-dependent modeling. *Proceedings of COMMUNICATION 2004*, 221-224, 2004.

[12]C.O. Dumitru, I. Gavat. Features Extraction, Modeling and Training Strategies in Continuous Speech Recognition for Romanian Language. *Proc. EUROCON 2005*, 1425-1428, 2005.

