OPTIMIZATION METHODS FOR LARGE VOCABULARY, ISOLATED WORDS RECOGNITION IN ROMANIAN LANGUAGE

Horia CUCU[1], Andi BUZO[1], Corneliu BURILEANU[2]

Speech recognition algorithms in general and isolated word recognition algorithm in particular are computationally intensive processes that consume a lot of time in an automatic speech recognition (ASR) system. This issue is in fact the bottleneck of real time ASR systems. This paper approaches the optimization of the recognition process by presenting an innovative three-step recognition method that diminishes the processing time for isolated words recognition. The recognition rate improvement is also approached by an adaptive technique of combining speech units.

Recunoaşterea de vorbire în general şi recunoaşterea de cuvinte izolate în particular sunt procese intens computaţionale care consumă mult timp într-un sistem automat de recunoaştere de vorbire. Această problemă este de fapt gâtuirea sistemelor de recunoaştere de vorbire în timp real. Lucrare de faţă abordează optimizarea procesului de recunoaştere prezentând o metodă inovativă de recunoştere în trei paşi care diminuează timpul de procesare pentru recunoşterea de cuvinte izolate. Îmbunătăţirea ratei de recunoaştere este de asemenea vizată această lucrare prezentand şi o metodă adaptivă de combinare a unităţilor de vorbire.

Key words: isolated words recognition, automatic speech recognition, hidden Markov model, phones, triphones, token passing algorithm, Dynamic Time Warping (DTW)

1. Introduction

The field of speech recognition in the Romanian language has been approached by some studies that mostly focus on isolated words recognition. The first attempts [1, 2] report the usage of the Dynamic Time Warping algorithm, while the latest papers present the results obtained with the state-of-the-art algorithms in speech recognition: Hidden Markov Models (HMM) algorithms [3, 4, 5] and neural networks algorithms [6, 7]. Studies regarding continuous speech are also presented in [5], but, as in the other cases mentioned above, the speech

[1] Assist., Electronics, Telecommunications and Information Technology Faculty, University "Politehnica" of Bucharest, Romania

[2] Prof., Electronics, Telecommunications and Information Technology Faculty, University "Politehnica" of Bucharest, Romania

database used is relatively small (less than 5000 different words) and thus the results cannot be extended to real continuous speech applications.

Our latest study [8] presents an innovative training strategy for HMM based ASR systems and extensive experimental results obtained for isolated words recognition. The paper also shows how the HMM system can be tuned up to issue better recognition rates on a 50000 words (among which 10000 are different) database.

This paper reports the latest results obtained after tuning up the system and presents a new three-step recognition method that solves the large computational time problem for the recognition process. In addition to this, the paper presents an adaptive technique of combining speech units in order to improve the recognition rate. The task of mixing the results of different ASR systems in order to obtain better recognition rates has been tackled by other studies also [12, 13]. In [12] an implementation of a multistage Recognizer Output Voting Error Reduction (ROVER) framework is presented. This framework is based on the model sets respectively trained with different objective criteria which are MLE, Minimum Phone Error (MPE) and Boosted Maximum Mutual Information (BMMI). The last criterion is actually introduced in [13], which uses it for explicitly generating ASR systems that are complementary to each other.

The recognition process of an ideal ASR system should achieve the following goals: real-time recognition, 100% word accuracy, speaker independency, extensive dictionary (storing all the words of that specific language), noise independency. The results presented in Section 2 show that the second and third goal are nearly achieved by the usual recognition method. The first goal is far by being reached by this method and would be approached by the innovative three-step recognition method described in Section 4. The fourth goal is also far by being reached, but our 10000 different words dictionary is clearly the most consistent one to be used so far for Romanian. The fifth goal hasn’t been approached yet.

Section 3 analyzes the Token Passing algorithm [9] from an optimization point of view. Sections 5 and 6 introduce a novel technique of improving the recognition rate of the ASR systems (and thus help in achieving the second goal of an ideal ASR system), while in the end Section 7 draws some conclusion regarding the new recognition approach and states the next steps to be performed.

2. HMM based Automatic Speech Recognition system

The main steps in building an HMM based ASR system are as follows:

- Every speech unit (phoneme, triphone, syllable, word, etc) is modeled using a HMM.
- The set of HMMs is trained using a training speech database; this step aims to create HMMs that model all the various ways of speech units' pronunciations for different speakers, in different conditions.
- The set of trained HMMs is evaluated with a testing database; this step issues the recognition rates for the ASR system.

2.1. Speech database

As an ideal ASR system should be able to identify all the words that are intelligibly spoken by any person, in a specific language, our main goal is to create a continuous speech recognition system for the Romanian language. Romanian did not have a large enough speech database for this purpose therefore the first step in reaching our goal was to create it. Several types of speech databases were created to serve the training strategy, as presented in [8].

This study mainly focuses on exploiting the isolated words database. This database consists of 50000 speech audio clips and their corresponding label files that mark speech units that are uttered in each audio clip. These speech samples were recorded in laboratory environment by five speakers, two males and three females. Each speaker uttered 10000 different words (chosen to cover all the syllables in the Romanian language) resulting in 10000 speech clips.

The label files for these speech samples were programmatically generated and contain only the phones that compose the words and not the time boundaries between phones. The additional information regarding speaker identity and speech type are obvious as our team is the author of the recordings. More information regarding the acquisition process of this database, the benefits and drawbacks of recording isolated words, the phonemes histogram, etc. are presented in [8].

In order to obtain relevant and consistent recognition results the isolated words database was split into two parts: a training database comprising of 45000 speech audio clips (9000 clips per speaker) and a testing database consisting of the remaining 5000 speech audio clips (1000 clips per speaker). These being said and taking into account that the audio clips were recorded by five different speakers we assert that all the results presented further on are speaker independent recognition results and thus our ASR systems achieve the speaker independency goal stated in Section 1.

2.2. HMMs design and training

Choosing the most appropriate speech units to be modeled is one of the important factors in obtaining good recognition results. Up to this point we have experimented using phones and triphones and, as concluded in [8], triphones issue better recognition results. Our training strategy implies using both types of speech units and finally creating and ASR for triphones:

a) creating HMMs with phones as speech units (36 models, one for each of the 36 phones in Romanian)
b) training these phones HMMs using the training database
c) creating HMMs with triphones as speech units using the phones HMMs for initialization
d) training the triphones HMMs using the training database (this training process uses the tied-states technique)[8]

Both training processes (steps b and d) use the embedded training technique because the label files in the database don't have information about time boundaries between phones. Filling the label files with this information would waste lots of time as embedded training works well enough.

Another important aspect involved in designing the models (step a) implies choosing the model's parameters. As shown in [8] we've obtained the best recognition results using the following configuration for the HMMs:

- Number of states: more than 10 (see next section for updated results)
- Output parameters: twelve Mel Frequency Cepstral Coefficients (MFCC) with appended energy coefficient and first order derivates
- Output function: three Gaussian mixtures

2.3. The Speech Recognition process

The speech recognition process for an HMM based ASR system requires the following components:

- The set of trained HMMs resulted in the training process
- Language information including:
  - A dictionary containing all the possible output words
  - Grammar restrictions specifying the all the possible word combinations and their probability

The set of HMMs is chosen among the sets resulted after the training processes presented in the previous section in order to obtain the best recognition rates (usually triphones HMMs issue better results). The ideal dictionary would contain all the words in a specific language; in our case the dictionary contains the 10000 different words in the database. As the results presented further on will show we've experimented with different grammar restrictions types:

- all combinations of any number of phones have the same probability
- all combinations of any number of words have the same probability
- only one word "combinations" (only the succession <silence> <word> <silence> is allowed where all words have the same probability)

The best results obtained with the usual one step recognition method presented above are summarized in Table 1.

Table 1

Recognition results. Best recognition results using the usual recognition method

HMM configuration		Language information		Recognition algorithm		Rec. rate [%]	Average rec. time per file [s]
		Dictionary size	Grammar restrictions	Type	Steps description		
12 states 3 Gaussian mixtures MFCC_E_D	models for phones	10000 words	any combination of any words	one step recognition	recognize words	84.89	5.34
	models for triphones					88.59	2.92

Two important things should be noted about the results presented in Table 1. First, the triphones HMMs ASR system is faster and has a higher recognition rate than the phones ASR system. And second, both ASR systems are still far from achieving the real-time recognition goal (the average size of the testing speech audio clips is 1.65s).

In order to better understand these results and try to get closer to the real-time recognition goal, the recognition process algorithm has to be analyzed and enhanced.

3. Token Passing algorithm analysis

The recognition process is using the well known Token Passing algorithm [9] to decode an unknown speech utterance. Given the speech utterance this algorithm takes into account the language information to create all possible HMM combinations that have the output size equal to the length of the speech utterance. All these HMM combinations become recognition candidates. Going further, the algorithm uses the trained models to compute the probability that a given HMM combination could have been issued the speech utterance. The probability is computed for all candidates and, in the end, the candidate with the highest probability “wins the race”. The algorithm concludes that the particular HMM combination is the author of the speech utterance and outputs the corresponding phones or words.

Taking these into account it’s clear that the recognition time increases with the number of possible HMM combinations and this number depends on:

- the number models (usually a high number of models implies also a high number of possible HMM combinations)
- the size of the dictionary (more dictionary entries mean more possible HMM combinations)

- the complexity of the grammar restrictions (more language restrictions imply less possible HMM combinations).

Now we can better understand and explain why the triphones ASR system is faster than the phones ASR system. Although the number of HMMs is higher (7372 triphones models compared to 41 phones models) and apparently increases the number of combinations, the rigorous constructive restrictions applying to triphones drastically reduce the number of combinations. This happens because any triphone can be followed by only 41 different triphones. For example the triphone "c-a+b" (notation standing for central phone "a" preceded by "c" and succeeded by "b") can be followed by only 41 out of the 7372 triphones because its already implied that it would be a "b preceded by a" triphone (a-b+*).

Also, in the case of triphones HMMs 6 out of 10 emitting states are "tied". This means that HMMs combinations probability calculus is 60% common throughout all triphones groups that share the same central phone.

The algorithm analysis presented above unfolds a range of optimization possibilities. The first one that was exploited implied the introduction of some basic grammar restrictions. Taking into account that we approach isolated words recognition and thus we a priori know that the speech sample consists of a single word it's natural to allow only "one word combinations" recognition candidates. The results obtained by applying this type of grammar restrictions are presented in Table 2.

Table 2

Recognition results. Applying "only one word" grammar restrictions

HMM models	**Language information**		**Recognition algorithm**		**Rec. rate [%]**	**Average rec. time per file [s]**
	Dictionary size	**Grammar restrictions**	**Type**	**Steps description**		
phones	10000 words	only one word	one step recognition	recognize words	91.53	1.45
triphones					89.01	1.17

Two important things should be noted about the results shown in Table 2. First, and as expected, the average recognition time decreases a lot, being at this point lower than the average length of the testing speech audio clips. Second, the recognition rate for phones ASR system overcomes the recognition rate for the triphones ASR system.

The second optimization possibility implied decreasing the number of words in the dictionary. A first approach considered downsizing the testing dictionary from 10000 words to only 1000 words (the 1000 different words that composed the testing database).

Table 3

Recognition results. Using a short dictionary

HMM models	Language information		Recognition algorithm		Rec. rate [%]	Average rec. time per file [s]
	Dictionary size	Grammar restrictions	Type	Steps description		
phones	1000 words	any combination of any words	one step recognition	recognize words	96.24	0.31
triphones					92.92	0.25

Table 3 presents the results obtained applying a smaller dictionary. It's important to note that the ASR systems are more accurate because they have fewer words to choose from and also that the ASR systems are faster because the number of possible HMM combinations decreases.

Although the results in Table 3 are appealing we should not forget that the fourth goal of an ideal ASR system is to have an extensive dictionary. This means that if we're going to reduce the size of the dictionary we need to find a way to do it in an adaptive manner. That's the main focus of the three-step recognition method presented in Section 4.

Both optimizations, introduction of grammar and reduction of dictionary size are motivated by the existence of many applications where human-machine interaction is implemented through single commands. In this case each command is represented by one word and the number of commands cannot be very high.

4. Three-step speech recognition

Taking into account the results presented in the previous section and the conclusion that a small size word dictionary could improve both the accuracy and the recognition speed of an ASR system we've designed the following three step algorithm (these steps will be applied in real-time to the test speech audio clip):

Step 1. Unrestricted phone recognition is performed. It's supposed that this step will be very fast because the number of tokens in the network is smaller. The recognition result will be a set of ordered phones and their recognition probability.

Step 2. A small word dictionary is selected based on the recognition probability of the recognized phones (only the words that contain the phones with the highest recognition probability will be part of this dictionary).

Step 3. A regular word recognition (that uses the word dictionary selected at step 2) is performed.

This three-step recognition method raises several issues that will be presented and approached in the following subsections.

4.1. Step 1. Selecting the best recognized phones

A first thing that has been noticed is that the ASR system usually recognizes different phones with quite different recognition probabilities. The recognition probability depends on how well a model has been trained. Thus, a raw comparison between the recognition probabilities of phones "a" and "b", for example, is not suitable for our purpose.

In order to get more quantitative information about this aspect unrestricted phones recognition was performed on the whole training speech database. Then a DTW algorithm was used to align and compare the label files with the recognition files and select all the correctly recognized phones. The recognition probabilities for all these phones were utilized to compute the average recognition probability for each phone. Some of the results are presented in Table 4.

Table 4

Recognition probability average comparison

Phoneme	a	b	d	e
Logarithmic recognition probability average ($\overline{recProb}$ **)**	-54.54	-58.06	-60.74	-55.49

These results sustain the above mentioned note and are being used further on to complete step 1. The following measure, called recTrust, is proposed to be used to select the best recognized phones:

$$recTrust = recProb - \overline{recProb} \qquad (1)$$

$recProb$ is the current recognition probability for the phone and $\overline{recProb}$ is the average recognition probability for that particular phone.

Clearly, a positive and high recTrust means that the phone is recognized with a high probability relative to its recognition average and vice-versa, a negative recTrust means that the phone is recognized with a lower probability relative to its recognition average.

Using this measure, the output of step 1 would be a ranking of the best recognized phones within a given test speech audio clip.

It must be noted that the probabilities mentioned above are in fact log probabilities.

4.2. Step 2. Using the *recTrust* ranking to select a word dictionary

The process of selecting a word dictionary, even when the recTrust ranking is available, is not trivial as it will be shown further on. Using the following notations: ph1 – the top ranked phone, ph2 – the second ranked phone and ph3 – the third ranked phone, the words dictionary can be chosen to contain all the words that contain:

- phones ph1 and ph2 and ph3
- phones ph1 and ph2
- phone ph1
- phones ph1 or ph2
- phones ph1 or ph2 or ph3
- other phones logics

In order to evaluate the different types of dictionary selection we've experimented with some of them and we've summarized the results in Table 5. The average dictionary size is computed as a weighted average of all the dictionary sizes. The weights are the number of usages during the recognition process of the words in the testing database. We've measured the correct dictionary selection (CDS) rate as the percentage of correctly chosen dictionaries out of the total number of selected dictionaries. A dictionary is considered to be correctly selected if it contains the word that is actually uttered in the audio speech sample.

Table 5

Dictionary selection comparison

Dictionary selection logic	Average dictionary size [no. words]	Correct dictionary selection rate [%]
ph1 & ph2 & ph3	215	63.70
ph1 & ph2	n/a	74.14
ph1	3460	86.62
ph1 | ph2	5470	97.97
ph1 when ph1 is one of {a, e, i, i3, l, o, s1, u, z}, ph1 | ph2 otherwise	4756	96.84

As Table 5 shows, the dictionary size and the CDS rate increase as the selection logic becomes looser. One thing to note is that the average dictionary size selected by this method should be compared with the original dictionary size which is 10000 words. A second important aspect is that the CDS rate is only acceptable for the last two dictionary selection logics, because the recognition rate of the ASR system is now limited to this CDS value.

4.3. Experimental results

As the two main issues raised by the three-step recognition method have been addressed, and the third step of this method is actually regular word recognition that uses the word dictionary selected at step 2, the following table summarizes the results (the results presented in Table 2 are repeated for comparison).

Table 6

Recognition results. Comparison between one-step and three-steps methods

HMM models	**Language information**		**Recognition algorithm**		**Rec. rate [%]**	**Average rec. time per file [s]**
	Dictionary size	**Grammar restrictions**	**Type**	**Steps description**		
phones	10000 words	only one word	one step recognition	recognize words	91.53	1.45
triphones					89.01	1.17
phones	**Step 1:** 41 phones **Step 3:** average of 4756 words	**Step 1:** any combination of any phones **Step 3:** only one word	three step recognition	**Step 1:** recognize phones **Step 2:** select dictionary **Step 3:** recognize words	89.36	0.91
triphones					86.82	0.56

The results presented in Table 6 show significant recognition speed improvements (37% for the phones ASR system and 52% for the triphones ASR) with little (~2.18%) recognition rate drawbacks. The recognition rate was expected to degrade due to the imperfect dictionary selection.

Taking into account that the average recognition time per file includes the time needed for all the three steps and that this duration is now smaller that the average size of the testing speech audio clips (1.65s) we can assert that the real-time recognition goal has been achieved.

5. Combined ASR system

Exploring all the results obtained so far one can conclude that the triphones ASR systems are faster than the phones ASR systems and that the recognition rates differences vary (in some cases triphones models perform better – see Table 1, and in others phones models perform better – see Table 2, 3, 6).

More in-depth analysis showed that the gain in recognition rates for phones ASR systems or triphones ASR systems is not purely incremental. In other words, the triphones ASR systems fail to recognize some of the words that are recognized by the phones ASR systems and vice-versa. As an example, the set of words recognized (88.59% of the total words) by the triphones ASR system presented in Table 1 does not include all the words (84.89%) recognized by the phones ASR system presented in the same table. Figure 1 illustrates the actual intersection of the word sets.

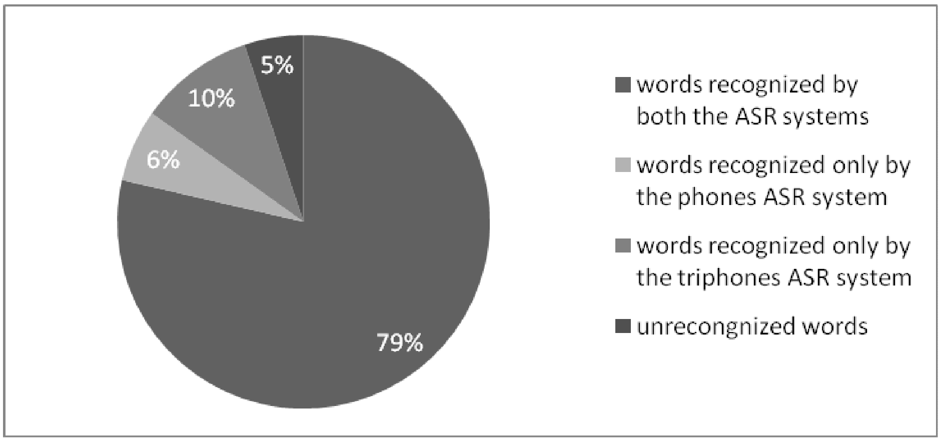

Fig 1. Recognized and unrecognized word sets for the ASR systems presented in Table 1

Figure 1 clearly shows that if the triphones ASR system would also be able to recognize the words recognized by the phones ASR system or vice-versa we would obtain a recognition rate of 95%.

The previous result triggers a new idea of a combined (phones + triphones) ASR system:

- the speech audio clip is inputted in the phones ASR system
- the speech audio clip is inputted in the triphones ASR system
- the results are compared and the better recognized word is selected

The selection of the better recognized word is in this case very simple because the two recognition probabilities outputted by the two systems can be directly compared.

The results obtained by this method are presented in Table 7 (the results in Table 1 are repeated for comparison).

Table 7

Recognition results. Combined phones and triphones recognition results

HMM models	**Language information**		**Recognition algorithm**		**Rec. rate [%]**	**Average rec. time per file [s]**
	Dictionary size	**Grammar restrictions**	**Type**	**Steps description**		
phones	10000 words	any combination of any words	one step recognition	recognize words	84.89	5.34
triphones					88.59	2.92
combined phones and triphones					92.72	8.36

The results show that the combined ASR system has a consistent recognition rate improvement thanks to the employment of both phones and triphones recognition. The downside of this system is that its average recognition time is the sum of the average recognition times for the composing systems incremented by an average phones/triphones decision time of 0.1s.

6. Three-step speech recognition in a combined ASR system

As seen in Section 4 the three-step recognition method significantly improves the speed performance of an ASR system offering the possibility of using it in a combined ASR system. The first two steps of the three-step recognition method would be entirely preserved, while the third step would be replaced be the steps of the phones-triphones combined recognition.

Table 8 presents the recognition results for a combined ASR system using the three-step method (some results presented in Table 6 are repeated for comparison).

Table 8

Recognition results. Three-step recognition method in a combined ASR system

HMM models	**Language information**		**Recognition algorithm**		**Rec. rate [%]**	**Average rec. time per file [s]**
	Dictionary size	**Grammar restrictions**	**Type**	**Steps description**		
phones	**Step 1:** 41 phones **Step 3:** average of 4756 words	**Step 1:** any combination of any phones **Step 3:** only one word	three step recognition	**Step 1:** recognize phones **Step 2:** select dictionary **Step 3:** recognize words	89.36	0.91
triphones					86.82	0.56
combined phones and triphones					91.95	1.56

The resulted ASR system has a 2.69% recognition rate improvement and still achieves the real-time recognition goal.

7. Conclusions and future work

This paper reports the first steps taken towards a continuous speech recognition system in Romanian language. As stated in [8] our approach to continuous speech involves designing and implementing an accurate, real-time isolated words recognition system. The best recognition results are being presented as a baseline while the paper focuses further on introducing two innovative recognition techniques that highly improve the performance of the ASR system in terms of both speed and recognition rate.

As proved by the results, the three-step recognition method enabled us to achieve the real-time recognition goal. A fast, unrestricted, phone recognition is needed first to select a small dictionary which is further on used in a regular word recognition. Although the method implies two successive recognition processes, the total time required to recognize one word is considerably lower than in the plain recognition case.

Conclusions about the most suitable speech units to be used in ASR systems can also be drawn by analyzing the results presented in this paper. Although triphones were considered to issue better results than phones [8] our current results show that this assertion is not always true. Moreover, even if a phones ASR system has lower recognition accuracy than the corresponding triphones ASR system, the phones ASR system would definitely be able to recognize some words that the other is not able to. These important experimental results motivated the design of a combined ASR system. The results turned out to be good and thus the employment of the phones/triphones recognition method improved considerably the recognition accuracy of the ASR system.

As continuous speech recognition is our final goal, a future plan is to implement the optimization method presented in this paper in a continuous speech recognition system. Both of the methods can still be improved to obtain better recognition accuracy in shorter time.

The first step of the three-step method can be improved by employing phones grammar restrictions. At the moment the phones recognition is unrestricted, but this can be changed to accommodate phones n-grams. The recognition rate will clearly be improved by this approach.

Also, the second step of the three-step method can be adapted by choosing a different type of dictionary selection logic. Some selection logics were presented in section 4.2, but many other are possible and some of them could possibly issue better results.

Finally, a big improvement in the combined recognition method would be to insert a decision block that could choose beforehand which models (phones or triphones) to use for the word recognition. This would save a lot of time because the ASR system would be needed to perform only one word recognition instead of two.

Acknowledgments. The research reported in this paper was funded by the Romanian Government, under the National Research Authority CNCSIS grant “IDEI” no. 114/2007, code ID_930 and by by the Sectoral Operational Program Human Resources Development 2007-2013 of the Romanian Ministry of Labor, Family and Social Protection through the Financial Agreement POSDRU/6/1.5/S/16.

REFERENCES

[1] *Mihaela Ioniță, C. Burileanu, M. Ioniță*, „DTW Algorithm with Associated Matrix for a Passworded Access System", Emerging techniques for communication terminals (ENSEEIHT), pp. 265 – 270, 1997.

[2] *B. Sabac, Z. Valsan, I. Gavat*, „Isolated Word Recognition Using the DTW Algorithm", in Proceedings of the International Conference COMMUNICATIONS '98, pp. 245 – 251, 1998.

[3] *C. Burileanu, V. Teodorescu, G. Stolojanu, C. Radu*, „Sistem cu logică programată pentru recunoaşterea cuvintelor izolate", independent de vorbitor, Inteligența artificială şi robotica, Editura Academiei Române, Bucureşti, vol. 1, pp. 266 – 274, 1983.

[4] *Diana Militaru, Inge Gavăt, Octavian Dumitru, Tiberiu Zaharia, Svetlana Segărceanu*, „Protologos, System for Romanian Language Automatic Speech Recognition and Understanding", Proceedings of Sped 2009, pp. 21 – 32, 2009

[5] *D. Munteanu*, "Contribuții la realizarea sistemelor de recunoaştere a vorbirii continue pentru limba română", Teză de doctorat, 2006.

[6] *D. Burileanu, M. Sima, C. Burileanu, V. Croitoru*, „A Neuronal Network-Based Speaker-Independent System for Word Recognition in Romanian Language", Text, Speech, Dialogue - Proc. of the First Workshop on Text, Speech, Dialogue, pp. 177 – 182, 1998.

[7] *J. Domokos*, „Contribuții la recunoaşterea vorbirii continue şi la procesarea limbajului natural", Teză de doctorat, 2009.

[8] *C. S. Petrea, A. Buzo, H. Cucu, M. Paşca, C. Burileanu*, „Speech Recognition Experimental Results for Romanian Language", Proceedings of ECIT 2010, 2010.

[9] *S.J. Young , N.H. Russell , J.H.S Thornton*, „Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems", 1989

[10] *S. Young, et al.*, „The HTK Book", Version 3.4, Cambridge University Engineering Department, 2006.

[11] HTK – http://htk.eng.cam.ac.uk

[12] H. Xu, J. Zhu, G. Wu, „An efficient multistage rover method for automatic speech recognition", Proceedings of ICME'09, pp. 894 – 897, 2009.

[13] C. Breslin, M. J. F. Gales, „Generating complementary systems for speech recognition", Proceedings of INTERSPEECH 2006, 2006.