BULETINUL Universităţii Petrol – Gaze din Ploieşti	Vol. LXVI No. 4/2014	81 - 90	Seria Tehnică

Automatic Speech Recognition: Methods and Applications for Romanian Language

Daniela Şchiopu

Universitatea Petrol-Gaze din Ploieşti, Bd. Bucureşti 39, Ploieşti
e-mail: daniela_schiopu@yahoo.com

Abstract

For many reasons, speech is a convenient mean for a human to interact with a machine. Computer speech recognition (or automatic speech recognition, ASR) is the process of converting the message uttered by a person into a sequence of words. In spite of all technological advances made in the field of ASR, due to many factors that influence the recognition rate, speech recognition is still a challenging research area. In this paper, the author presents two speech recognition systems for Romanian language, one based on statistical methods, hidden Markov models, and the other one based on dynamic time warping. The systems were tested for a set of voice commands applied to control electronic devices.

Key words: *automatic speech recognition, dynamic time warping, hidden Markov models, mel-frequency cepstral coefficients, statistical models.*

Introduction

Speech is the base of communication between people used to exchange information. Like any other technology, spoken language recognition was born from desire to simplify human life, from human-machine dictation applications to complex control systems or systems which require the communication between persons of different nationalities, who understand and speak only their mother tongue. Nowadays, international scientific communities are concerned with development of resources and methods for building high-performance systems in the most known languages, such as English, French, German, Mandarin, Japanese, etc. (see e.g. [8], [6]). These systems are embedded in smart-phones, intelligent cars, robots and so on.

Automatic speech recognition by machine (ASR) maps an acoustic signal that contains speech to a sequence of words. Essentially, ASR tries to solve the problem: what the speaker said.

This domain is very important for human life. The most important application area is human computer interfaces that involve for the most of the cases the use of speech recognition, mainly because speech has the potential to be a better interface than the keyboard for tasks where the natural language is helpful, or for tasks where the user has to manipulate different objects or to control certain equipment.

Specifically, the application fields where communication is essential are automatic call processing in telephone networks, translation and dictation, query based information systems, finding out travelling information, automatic medical transcriptions, etc.

Although the general problem of ASR by any speaker and in any environment is still far from completely resolved, in the last years some important realizations were made in this research area. A review of advances in ASR for Romanian language is presented in [7].

The history of Romanian speech recognition research spans over 50 years. Early research was limited by computational capabilities. After that, the research continued sporadically, thanks to enthusiasm of a small number of researchers (although due to the isolation from international scientific community, and the lack of an adequate material base), and was limited to phonemes recognition [2] or isolated words [4]. The scientific advance was made after 1989, by adopting modern techniques. Important achievements in the development of ASR systems, made for the Romanian language, can be found in the literature (e.g. [3], [13], [5]).

The study described in this paper is focused on building ASR systems for Romanian speech, applying different techniques.

The paper is organized as follows. Section 2 presents the field of ASR and reviews the main techniques used in this study. Section 3 describes how these techniques were implemented in the proposed speech recognition systems and the results obtained. The last section is dedicated to the conclusions.

Automatic Speech Recognition – an Overview

The Automatic speech recognition (ASR) or computer speech recognition is the process of converting a speech signal (a message uttered by a person) to a string of words. Although human listeners can determine pretty easy the speaker's message, this task is a difficult challenge for machine, due to many reasons: spoken message may come from several speakers (and, in this case, the ASR task is to identify the speakers and the messages); the speaker utters faster or slower, with a certain accent, pronunciation, articulation, nasality, speed; also health and emotions of the speaker can be reflected in his voice; acoustic environment can change; speech can be spontaneous or not. All these sources of variability make speech recognition a very complex issue.

Although ASR is very useful for hands-busy tasks (control of car navigation system while driving the vehicle, or control of different machineries in a factory), the main problems remain: sensitivity to speaking style (whisper, speaking loud); speech recognition affected by environmental noise.

Another problem in ASR is the difficulty of the speech recognition task. This can include the language, the size of the vocabulary and the linguistic domain. The size of the vocabulary is an important factor in ASR, but in certain situations, it is not a challenging task, for example in ASR control systems, where the speaker can utter a limited number of words.

Approaches for Speech Recognition

In ASR domain, there are three different approaches for solving speech recognition [1]:

- acoustic-phonetic approach;
- pattern recognition approach;
- artificial intelligence approach.

The acoustic-phonetic approach has the following steps:

- spectral analysis of the speech, converting the signal into a set of features that describe the acoustic properties;

- segmentation and labeling the speech signal, the result being a lattice characterization of the signal;
- validation a word (or string of words) from the phonetic label sequence.

The pattern recognition approach determines speech pattern representations, having a well mathematical framework. A speech pattern representation can be a speech template or a statistical model (as a hidden Markov model, HMM) and can be applied to a unit like phoneme, word or phrase. Other representations used for this approach are dynamic time warping (DTW) or vector quantization (VQ) which is applied for ASR specifically for data reduction.

The artificial intelligence approach combines the others two approaches. Here we can find knowledge-based approach that uses acoustic phonetic knowledge to develop classification rules for speech signal, or connectionist approach that uses artificial neural networks (ANN) to model learning and relationship between phonetic events.

Hidden Markov Models

Some of the most known and used strategies in speech recognition are the statistical methods [9].

The ASR aims to convert voice signal into text and this process can be formulated statistically as follows. Given a set of acoustic observations $O = (o_1, o_2, \dots, o_n)$ (sequence of speech vectors, where o_i is the speech vector observed at time i), which is the sequence of words $W = (w_1, w_2, \dots, w_n)$ that has the maximum probability:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} P(W|O) = \underset{W}{\operatorname{argmax}} \frac{P(W)P(O|W)}{P(O)} \quad (1)$$

Equation (1) specifies the most probable word sequence using Bayes rule, and $P(O)$ - the probability of the speech utterance - can be ignored, because it is independent of the sequence W. Thus, (1) becomes:

$$\widehat{W} = \underset{W}{\operatorname{argmax}} P(W)P(O|W) \quad (2)$$

Equation (2) contains two factors which can be directly estimated: the prior probability of the word sequence $P(W)$ and the probability of the acoustic data, given the word sequence $P(O|W)$. The first factor $P(W)$ can be estimated using only a language model, and the second factor can be computed on the basis of the acoustic model. The two models can be built independently, but they will be used together to recognize a spoken message.

Hidden Markov models represent the basis of the acoustical modeling.

An HMM is a stochastic finite state automaton, consisting of a set of states connected by transitions, in which the state sequence is hidden. The Markov process is considered to be "hidden" because the state sequence is not directly available to the observer, and instead of observing the state sequence, a sequence of speech vectors is observed, generated from a probability density function for each state.

An HMM representation as a probabilistic state automaton is shown in Figure 1. Thus, an HMM is characterized by the following parameters:

- a set of states $S = (s_1, s_2, \dots, s_N)$;
- transition probabilities between states: $A = (a_{11}, a_{12}, \dots, a_{NN})$, where each a_{ij} is the transition probability from state i to state j;
- observation probabilities $B = b_i(x_t)$, each defining the probability of an observation x_t being generated from the state i;

- the initial state distributions:
- $\Pi = \{\pi_i = P[s(0) = s_i], i = 1, \ldots, N\}$.

An HMM is complete if *A*, *B*, *Π* are specified, so the model is denoted by λ = (*A*, *B*, *Π*).

HMMs assume that the speech signal includes short time segments that are stationary. Because HMMs are able to handle very good variability of speech signal, they are very good models of speech. Although the definition of an HMM allows transitions from any state to another state, in speech recognition the models disallow transitions to earlier states. This HMM structure is called Bakis (or left-right) network. This type of HMM has the property that can easily models signals (such as speech signal) that changes their properties in time [12]. In a Bakis model, at ascending moments of time, the index of states is increasing.

An HMM can be a linguistic unit: a phoneme, a word, or a sentence. For continuous speech recognition, the linguistic unit is the phoneme. The word can be formed by concatenating phonemes and sentences can be obtained by enumeration of words. In an ASR system for command and control, with a limited vocabulary, the linguistic unit can be considered the word.

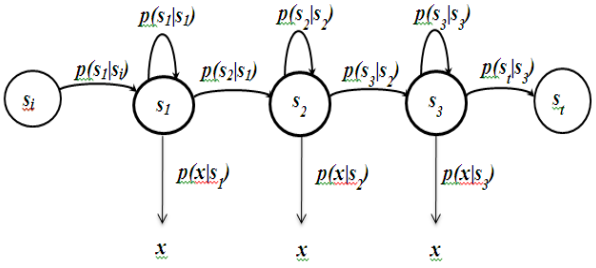

Fig. 1. HMM representation.

Dynamic Time Warping

Dynamic programming was introduced for nearly 5 decades, for solving the non-uniformity problem of time scale in speech [8]. This technique has numerous forms, like the Viterbi algorithm [14], used for finding the sequence of optimal states for an HMM, or dynamic time warping technique (DTW), used for measuring similarities between two sequences which may vary in time or speed [11].

DTW is a simple method that finds an optimal match between two time series, if one time series may be "warped" non-linearly by dragging it along its time axis. This warping between two time series can then be used to find equivalently regions among the two time series or to diagnose the similarity between them. This similarity is given by computing a minimum distance between the two time series (dynamic patterns).

If we suppose the two patterns are $P = (p_1, p_2, \ldots, p_i, \ldots, p_n)$ (the input signal) and $Q = (q_1, q_2, \ldots, q_j, \ldots, q_m)$ (the reference template signal) of length *n* and *m* respectively, DTW will computed an *n*-by-*m* matrix, where the element (*i*, *j*) is the distance between p_i and q_j. Then, using the Euclidean distance, the absolute distance between the two sequences is computed:

$$d(p_i, q_j) = (p_i - q_j)^2 \tag{3}$$

Each p_i and q_j is a vector of parameters (e.g. MFCC). The warping is equivalent to the problem of finding the minimum distance between *P* and *Q*.

For obtaining the optimal path between the starting point (1, 1) and the end point (*n*, *m*), we need to compute the optimal accumulated distance *D*(*n*, *m*):

$$D(i,j) = min[D(i-1,j-1), D(i-1,j), D(i,j-1)] + d(i,j) \quad (4)$$

The optimization process is performed using dynamic programming, which can reduce the amount of computation by avoiding accounting for all sequences that cannot possibly be optimal: once a sub-problem is solved, the result is retained and recalculation is not necessary.

Consequently, DTW has the following advantages:

- effectiveness for small-vocabulary speech recognition, even for isolated speech recognition applications;
- the method is suited to match sequences with missing information.

Also, DTW can successfully be used together with HMM in ASR systems (e.g. in [10] a unified view for HMM and DTW is presented).

The Experimental Systems

The general architecture of a HMM-based Isolated Speech Recognition System

We propose a general architecture of an HMM-based isolated word speech recognition system (IWSR) for the Romanian language. A Bakis HMM with 6 states (4 states are emissive and 2 states are not emissive, see fig. 3) is used in order to recognize a limited set of Romanian words for a specific domain of application (e.g. automatic control).

Figure 2 presents the general architecture of the HMM-based IWSR system.

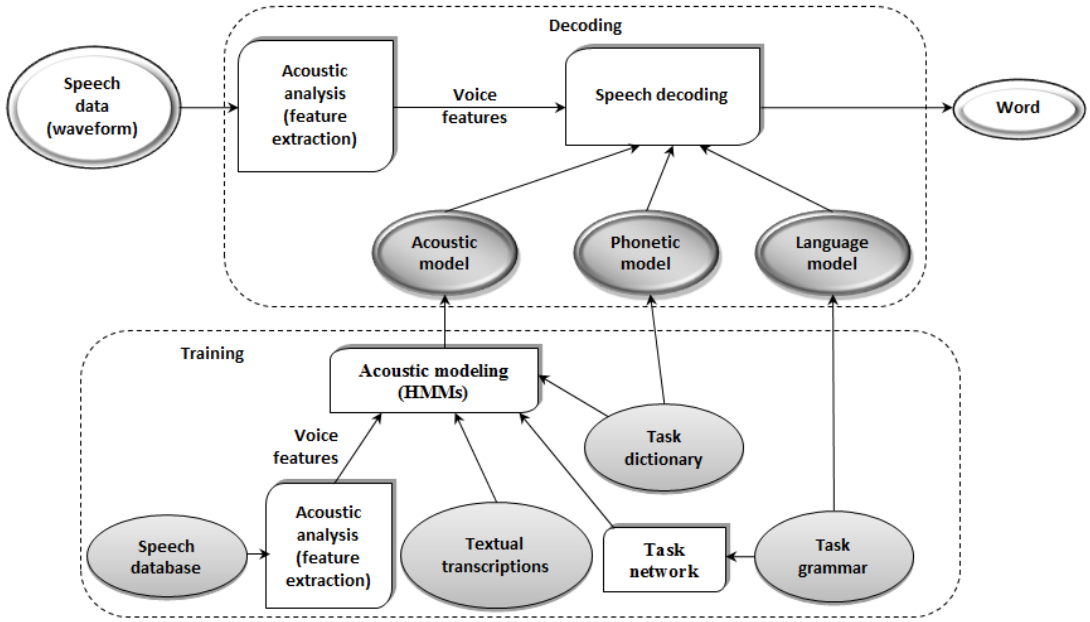

Fig. 2. The general architecture of the HMM-based IWSR system.

Speech recognition process includes two important parts: recognition is based on acoustic parameters extracted from the speech signal and not directly using spoken message; then some models are developed separately (acoustic, phonetic and language models).

HMMs don't work with the waveform of the speech signal. For this reason, in the ASR system a feature extraction block is necessary to determine speech vectors, which are further modeled by the acoustic model. The audio signal is converted by windowing into a sequence of windows in

time domain. From each of these windows are extracted spectral or cepstral parameters. For speech modeling, the most used parameters are the cepstral-perceptual parameters: mel-frequency cepstral coefficients (MFCC), perceptual linear prediction coefficients (PLP).

Besides each feature vector computed on a short frame of speech signal, the information embedded in the temporal dynamics of the features is also useful for recognition: velocity of the features (or delta features), which is determined by its average first-order temporal derivative and acceleration of the features (also known as delta-delta features), which is determined by its average second-order temporal derivative. Also, the total log energy has been computed to provide best results for speech recognition.

Acoustic model has to estimate the probability to utter a message, given a sequence of words. Language model decides whether a word (or a sentence) is valid in a certain language. Phonetic model aims to connect acoustic model (which works with phones) with the language model (which works with words). Phonetic model can be in our case a phonetic dictionary (task dictionary).

In the following section, we will present an experimental system for Romanian language with a particularization of the general architecture proposed in this section.

HMM_SR system

In this section, we present an ASR system based on HMMs. The system recognizes four Romanian isolated spoken words, *porneşte* (start), *opreşte* (stop), *creşte* (increase) and *scade* (decrease). The selection of these words was not random. Although the words *start* and *stop* exist in Romanian language too (as in English), we want to observe whether the behavior and acoustic similarities of these four words affect the recognition rate, especially when the application purpose is to command and control a device (or a robot). The command *porneşte* must start the device, *opreşte* must stop it, *creşte* can increase a certain indicator (e.g. temperature) and *scade* must reduce that indicator.

Romanian is a language that makes intensive use of the diacritics. Even though it uses only 5 diacritical characters (ă, â, î, ş, ţ), their occurrence frequency is very high: about 30% to 40% of the words in a general text contain at least one diacritical character.

Three of the chosen words for automatic recognition contain the diacritic *ş*. If these words can be used for controlling an electronic device, we were interested in the situation that a command is misunderstood, with some consequences.

First, we built the training corpus and the textual transcriptions (see fig. 2). For this purpose, we used voices of three speakers (two females and a male), each one uttering for ten times each word.

Then every speech data was converted into a set of feature vectors. For each frame, we extracted 39 coefficients: 12 MFCC parameters along with the first and the second derivatives and their log-energy.

For each of the five acoustic events (*porneşte*, *opreşte*, *creşte*, *scade* and silence – *sil*), we defined an HMM. The topology of the five HMMs (one model for each word, together with a model for the silence) is presented in Figure 3.

The grammar need to be defined according to syntactic rules. In our isolated word ASR system, the task grammar is defined as follows:

```
$cuvant=porneste|opreste|creste|scade;
({Start_sil} [$cuvant] {End_sil})
```

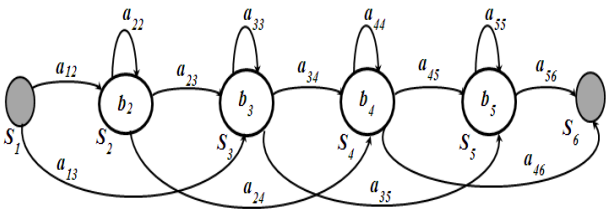

Fig. 3. HMM topology.

The variable $cuvant (word, in Romanian) can be replaced with one of the four Romanian words. The system can recognize each word or a segment of silence before and after the word.

The task dictionary contains information about the association between the word and its corresponding HMM.

The task grammar is then compiled and the result is a task network (fig. 4).

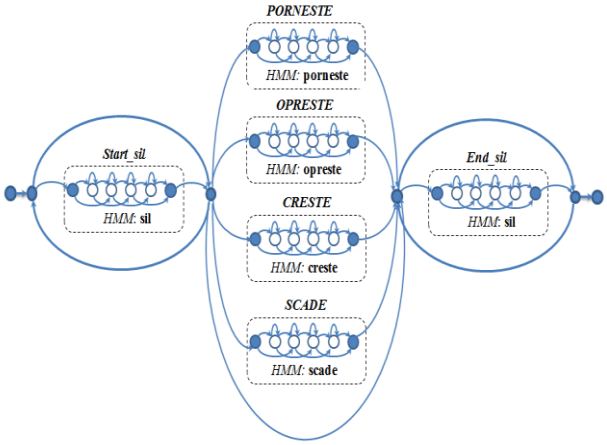

Fig. 4. HMMs network

Regarding the testing and evaluating the performance of the system, each speech signal from the test corpus is converted first into a set of acoustic vectors (fig. 2). The result is then processed with Viterbi algorithm [14] which compares the signal with the HMMs in the recognition module. The overall results are provided by HTK (Hidden Markov Model Toolkit) [16].

Word recognition rate for this experimental system was 83.33%.

DTW_SR system

In this section, we present an ASR system based on DTW. The system recognizes the same four Romanian isolated spoken words: *porneşte* (start), *opreşte* (stop), *creşte* (increase) and *scade* (decrease). We used the same corpus but different methods, in order to compare the results obtained in the both cases.

The voice recognition process consists of two steps:

- building the training corpus (when the voice of the speakers has to be recorded in order to build the reference template model), and
- the testing phase (when the recognition is made, based on the stored reference template model).

We used a simple speech recognition scheme with DTW method, as it is shown in Figure 5.

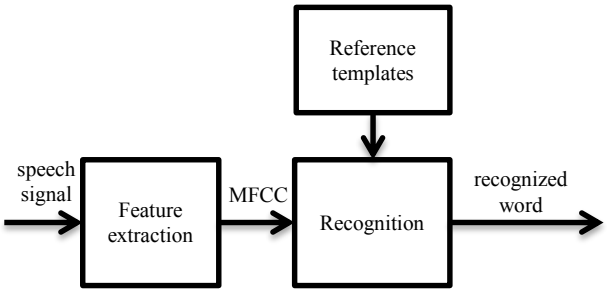

Fig. 5. ASR scheme using DTW

The same training corpus (as in the HMM-based system) was used, but the implementation was made in Matlab® [17]. Likewise, the feature vectors extracted from the voice signals were MFCC parameters. The system has to recognize one of the four classes, each one for every word. The word error rate for this system was 11.1%.

The characteristics and the results obtained (for the both systems) are synthetized in Table 1.

Table 1. Comparative analysis of the two systems

The system	Vocabulary type	Speech analysis technique	No. of parameters	System accuracy
HMM-based system	isolated-word	Mel-cepstral	12 MFCCs + energy + Δ + ΔΔ	83.33%
DTW-based system	isolated-word	Mel-cepstral	12 MFCCs + energy + Δ + ΔΔ	88.89%

Conclusions and Future Work

This paper evaluated how pattern recognition methods can be used in automatic speech recognition. We made two experiments, one by using HMM and the other by using DTW for isolated word speech recognition.

The first system estimates the parameters of a set of HMMs using a training speech database and the associated transcription of the speech data (manual labeling of the files containing speech). The second system uses reference templates for warping them with the input signals. The MFCC parameters were used for both systems.

The recognition rate for HMM-based ASR system was 83.33%, while for DTW-based ASR system was 88.89%.

Also, due to the fact that the words *porneşte* and *opreşte* are acoustically very similar, the word *opreşte* was more difficult to recognize in the case of the HMM-based system (it was replaced with *porneşte*), while *creşte* and *scade* were recognized 100%. At some tests for the first system, the word *scade* was replaced with *cade* (falls) and the two words have been confounded, because in the pronunciation of the word *scade* [s k a d e], the fricative consonant (phoneme *s*) is barely audible.

In the case of the DTW-based system, the only major problem was confounding the word *creşte* with *opreşte*, at some tests.

The results demonstrate the efficacy of the both proposed system, although DTW was more effective than HMM, mainly because the vocabulary was reduced and the words were not connected.

Considering these results, as future work, we intend to add other Romanian words for voice commands in the vocabulary, increase the speech database, and extend the present system to an ASR system for controlling a robot.

References

1. Anusuya, M.A., Katti, S. K. – Speech recognition by machine: A review. *International Journal of Computer Science and Information Security*, Vol. 6, No. 3, 2009, pp. 181-205.
2. Boldea, M., Bărbulescu, C. – Vowel Recognition Using a Microsystem, *"Tehnic 2000" Bulletin*, Timişoara, 1984, pp. 274-277.
3. Chivu, C. – Romanian Continuous Speech Recognition Applied to Automatic Controlled Systems. *Annals of the Oradea University, Fascicle of Management and Technological Engineering*, Vol. VI (XVI), 2007, pp. 728-735.
4. Constantinescu, M., Cristescu, D. – System for Analysis and Automatic Speech Recognition, in: Drăgănescu, M., Burileanu, C. (editors), *Analysis and Synthesis of the Speech Signal*, Romanian Academy Publishing House, Bucharest, 1986, pp. 210-220.
5. Cucu, H., Buzo, A., Petrică, L., Burileanu, D. – Recent Improvements of the SpeeD Romanian LVCSR System. *Proceedings International Conference on Communications (COMM)*, Bucharest, Romania, 2014.
6. Desai, N., Dhameliya, K., Desai, V. – Feature Extraction and Classification Techniques for Speech Recognition: A Review. *International Journal of Emerging Technology and Advanced Engineering*, Vol. 3, No. 12, pp. 367-371.
7. Dumitru, C.-O., Gavăt, I. – Progress in Speech Recognition for Romanian Language. *Advances in Robotics, Automation and Control*, Vienna, Austria, 2008, p. 472.
8. Furui, S. – 50 years of progress in speech and speaker recognition. *Proceedings SPECOM* 2005, pp. 1-9.
9. Jelinek, F. – *Statistical Methods for Speech Recognition*. Cambridge, MA, MIT Press, 1998.
10. Juang, B. H. – On the hidden Markov Model and Dynamic Time Warping for Speech Recognition - A Unified View. *AT&T Technical Journal*, Vol. 63, No. 7, 1984, pp. 1213-1243.
11. Muda, L. – Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. *Journal of Computing*, Vol. 2, No. 3, 2010.
12. Rabiner, L. – A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, 1989, pp. 257-286.
13. Teodorescu, H.-N. – AI Tools for Speech Analysis Applied to the Romanian Language. *Proceedings of the 4th European Computing Conference*, 2010, pp. 272-279.
14. Viterbi, A.J. – Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, Vol. 13, No. 2, 1967, pp. 260-269.
15. Zaharia, T., Segărceanu, S., Cotescu, M., Spătaru, A. – Quantized Dynamic Time Warping (DTW) algorithm. *Proceedings of the 8th International Conference on Communications (COMM)*, 2010.
16. * * * – *Hidden Markov Model Toolkit* (available at http://htk.eng.cam.ac.uk/).
17. * * * – *Matlab R2012a*, Mathworks Inc. (available at http://www.mathworks.com/).

Recunoaşterea automată a vorbirii: metode şi aplicaţii pentru limba română

Rezumat

Vorbirea este un mijloc convenabil de interacţiune omului cu o maşină. Recunoaşterea limbii vorbite de către un computer (sau recunoaşterea automată a vorbirii, RAV) este definită ca fiind procesul de transformare a unui mesaj rostit de o persoană într-o secvenţă de cuvinte. În ciuda tuturor progreselor realizate în acest domeniu, datorită multor factori care influenţează rata de recunoaştere, RAV rămâne încă un domeniu de cercetare complex şi provocator. În acest articol, autoarea prezintă două sisteme de recunoaştere a vorbirii pentru limba română, unul bazat pe reţele Markov ascunse, iar celălalt bazat pe metoda de aliniere temporală. Sistemele au fost experimentate pentru comenzi vocale care pot controla dispozitive electronice.