Spontaneous Speech Database for Romanian Language

Vladimir Popescu, Cristina Petrea, Diana Haneş,
Andi Buzo, Corneliu Burileanu

Faculty of Electronics and Telecommunications
"Politehnica" University of Bucharest
cburileanu@mesnet.pub.ro

Abstract. *The spontaneous speech recognition for Romanian language is an opened domain, not yet explored as other fields of the speech recognition area. The goal is to achieve performance in order to obtain a stat-of-the-art with worthy results. This paper presents the research work in the field of spontaneous speech recognition. This work begins with the new Romanian corpus, built from scratch with words and triphones which led to important statistical results regarding the linguistic structures in the Romanian language. Conventions e used in order to reach high standards of performance with aplicability to further research. This paper sumarizes the passed phases, the statistical results and the achievements obtained in the corpus building phase, on the challenging way to a new spontaneous speech recognition tool for Romanian language.*

Keywords: *Speech Recognition, Spontaneous Speech, Romanian Word Corpus, Romanian Triphones Corpus, Romanian Speech Database*

1. Introduction

Fields of speech recognition like isolated-word speech recognition or continuous speech recognition are knowledge territories that have been crossed over and over again with considerable results.

For the international languages (English, French) there are complete human-computer dialogue systems. In other languages such as Romanian, considered "under-resourced, from a software point of view" according to recent studies, [8], [9], the development of spoken dialogue systems is a long-term process.

However, reusing language independent components in the architectures designed in the developed countries, with whom Romanian researchers are cooperating (e.g. France) and creating, in Romanian language, only the language-specific components, human-computer dialogue applications in Romanian language become feasible in the mid-term.

The issue of Romanian spontaneous speech recognition in dialogue is of high importance, as spontaneous speech recognition it's a hardly crossed territory with poorer results. That's because there are major differences between speech read from a written script and real time mind made and freely expressed spontaneously speech.

In spontaneous speech the speaker doesn't preserve rules; he talks free, not necessarily grammatically correct. He can pronounce words in slang or in short forms, he can come up with unexpected interjections, he can make pauses or he can speak too fast, he may stammer or he may deeply breathe, he may hesitate, he may be incoherent. He may or may not be aware that his words are not always grammatically correct and his language contains disfluencies.

The speaker's mood has a big influence on his spontaneously spoken behavior as he may yawl, he may whisper, he may get confused and begin to stammer, he may laugh or cry and all his emotions will get reflected into the way he is expressing.

The issues that appear with the spontaneous speech are: false starts, filled pauses, incoherence of the speaker with possible ungrammatical constructions and other similar behaviors. These mentioned problems make the spontaneous speech more difficult to deal with, compared to the read speech, when it comes to recognition. The state-of-the-art in this field of speech recognition reveals the poor accuracy for the freely spoken spontaneous speech. As the spontaneous speech makes use of the acoustic and linguistic models that have been constructed especially for speech read from scripts, the results are far away from the desired ones.

The corpora used for the freely spoken speech has to be wide in order to permit a knowledge improvement on the structure of spontaneous speech, as for the moment it is limited.

Speech recognition systems, based on statistical approaches, are available in the research community (SPHINX system - Carnegie Mellon University, HTK toolkit - Cambridge University, RAPHAEL system - Laboratoire d'Informatique de Grenoble, etc.), as well as in the commercial domain.

At the same time, for Romanian language, the first continuous speech recognition system has been developed in 2006 at Military Technical Academy in Bucharest (D. Munteanu).

The concern of developing a spontaneous speech recognition module for Romanian language must be, on the one hand, adapted to the spontaneous nature of speech, and, on the other hand, to be able to interact with other modules in the dialogue systems, namely the semantic analyzer and the dialogue manager.

2. Recognition System Architecture

Building a speech recognizer from scratch involves in the first place a very good task and sub-task determination and separation.

Figures 1 and 2 represent the approach proposal for the spontaneous speech recognition system.

A typical speech recognition system works in two regimes **training** and **recognition**. At the end of these processing steps, the output of the recognition is represented by a number of alternative word sequences, for an utterance. The choice of the most relevant alternative, in a specified context, is the responsibility of other components in the dialogue system.

In general terms speech recognition involves finding a word sequence, using a set of determined models, acquired in a prior training phase, and matching those models to the input speech signal. For small vocabularies (a few tens of words), these models can capture word properties, but sounds units are generally modeled (such as phonemes or triphones, which represent phonemes in the context of right and left neighboring phonemes). The most successful approaches nowadays consider this model matching as a probabilistic process that has to account for the *temporal* variability (due to different sound durations), as well as for the *acoustic* variability (due to linguistic, subjective and channel-related factors, emphasized above).

The recognition system is based on trained hidden Markov models (HMMs) for each triphone. Figures 1 and 2 show the proposed architecture for both training and testing phases.

The training phase

Recorded voice files (wave files) were used as input in the training phase.

The wave files labeling was most time consuming. The phonetic transcription using SAMPA conventions [6] was done in order to build the phonetic dictionary. After these steps, the labeling passed at triphones level.

One important step in the training phase was the definition of Hidden Markov Model (HMM) prototype. This step was done manually. In the first place one prototype was chosen for all triphones and then several parameters of the prototype were varied in order to obtain a higher score.

HMMs were initialized with a default matrix of transitions and observations. These steps were realized by using HInit and HCompV, very helpful HTK tools.

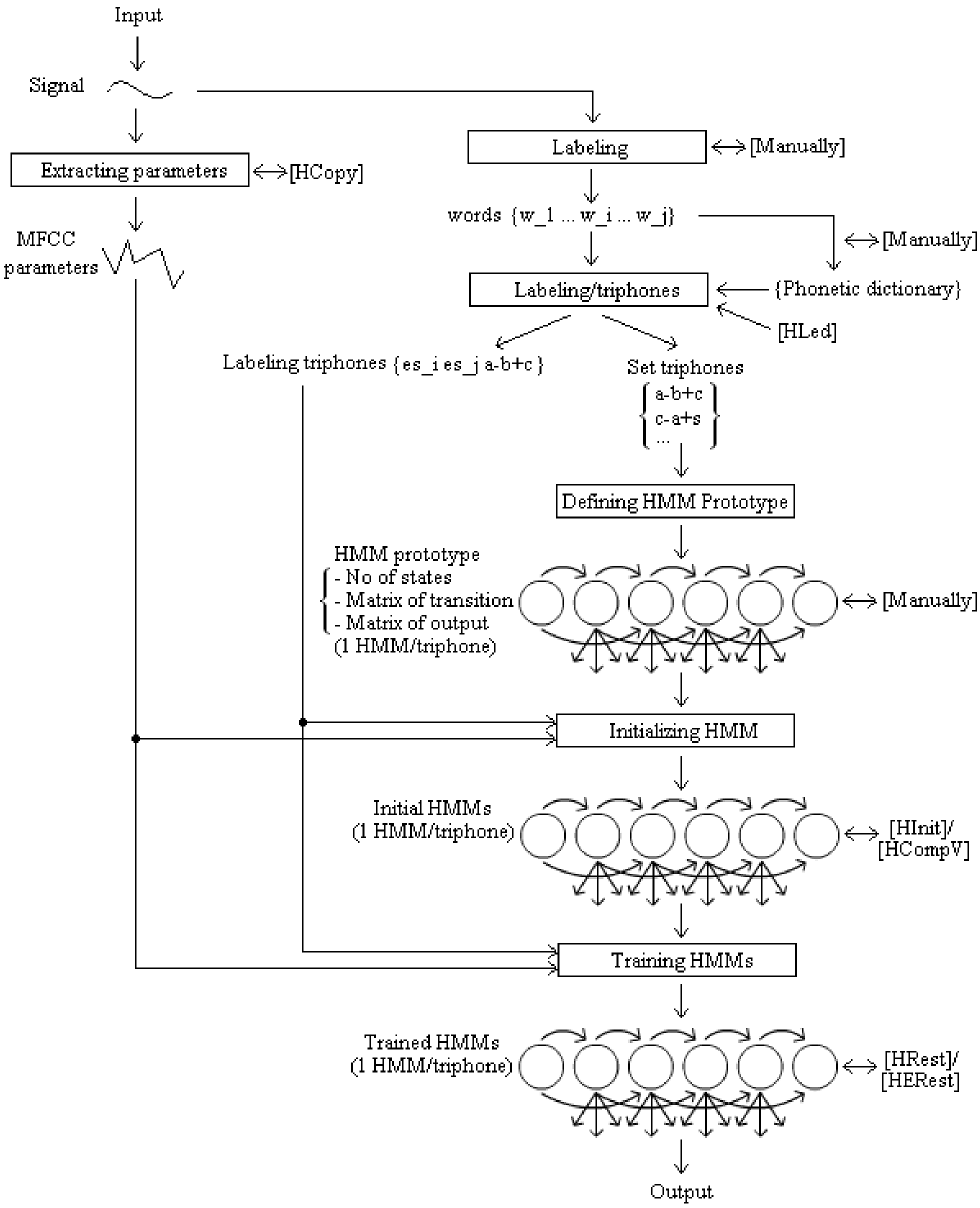

Fig. 1. The system architecture considered for the HMMs training phase

Before starting the training phase, the voice signal parameters were needed. The used parameters are MFCC. The training of the HMMs for each triphone is performed using the MFCC parameters and the labeled triphones. An iterative Viterbi alignment for the hidden Markov models will be used in order to obtain the maximum probability for certain HMM to represent the corresponding triphone. This step will be realized using HRest and HERest HTK tools [5].

The testing phase

The testing phase will use all the output obtained in the training phase.

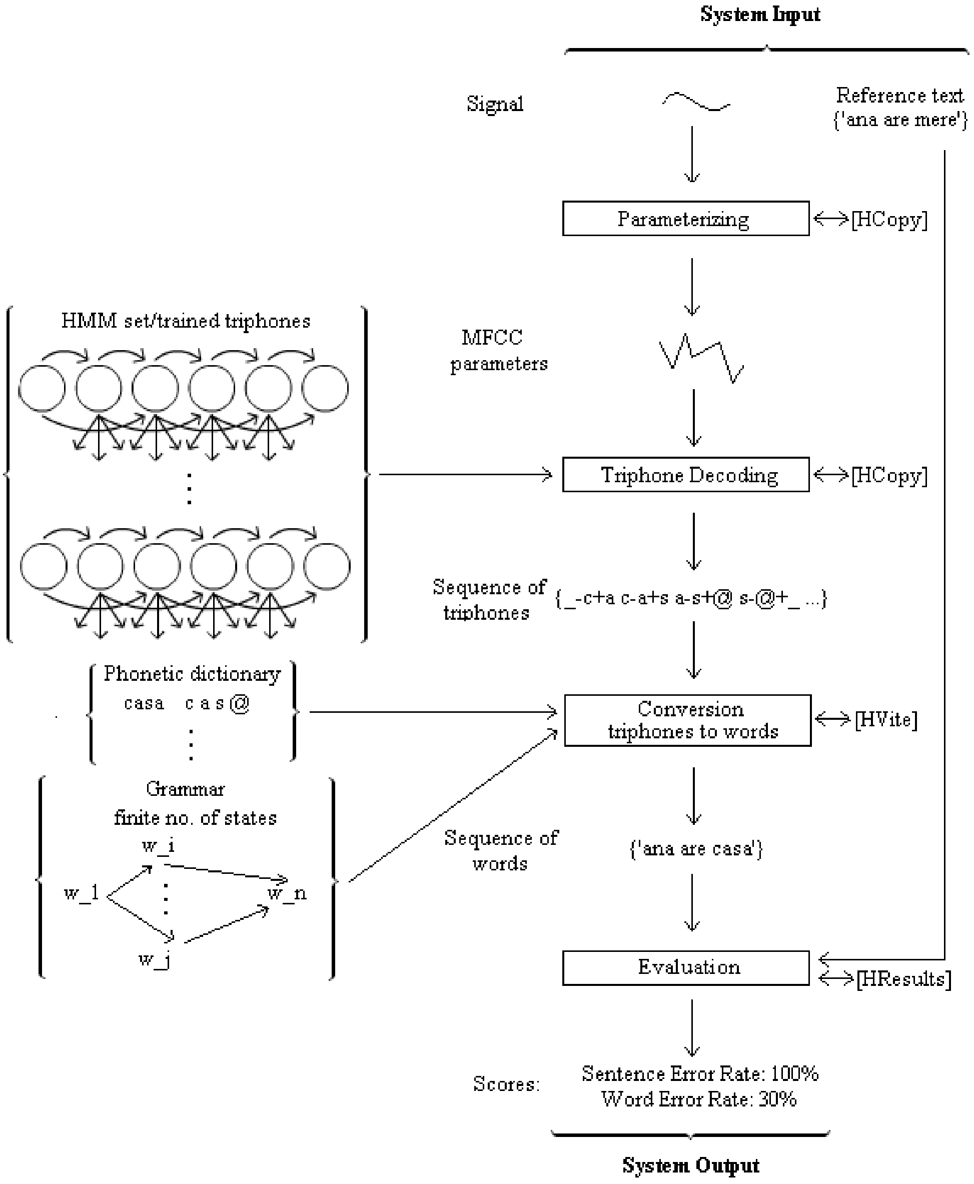

Fig. 2. The system architecture considered for the testing phase

Testing sentences will be used as input and the voice signal will be chosen from the already existent data base.

The wave files used as input were first parameterized and the MFCCs (Mel Frequency Cepstrum Coefficients) parameters were extracted using the HCopy HTK tool [5].

The MFCC acoustic parameters will be used as input for the next step – the triphone decoding phase. At this step the acoustic parameters will be decoded using the HMMs trained for each of the existent triphones and a sequence of triphones will be obtained.

At the next step this sequence of triphones will get converted into a word sequence. This is done by also considering a grammar with a finite number of states.

Having the reference text, the results can be evaluated. Reference text will be compared to the text obtained by the system and spontaneous speech recognition tool performance can be determined with two kinds of evaluation scores: Sentence Error Rate and Word Error Rate.

3. Spontaneous Speech Data Base Building

The corpora used in speech recognition should have the property of being wide. As wide as a corpus is as more useful it gets in spontaneous speech recognition domain.

The corpus was created from. The intentions are to keep the scalability property for the corpus in order to add as many words as the further investigations and research will need.

Efforts have been concentrated in order to create conventions and well followed steps for future use in database creation and enlargement.

3.1. Issues in Building Spontaneous Speech Data Bases

A series of problems and specific elements have been identified related to construction of databases for continuous speech recognition.

Speech recognition may be considered as a process of form recognition, and this can be achieved usually [1], based on rules or based on statistical methods. The last option is at the moment the preferred one because of good performances obtained under acceptable production costs [2]. Such step assumes usage of input data for training purposes so that the system may create automatically information to use at later stage.

For speech recognition process the specific aspects of input data used for training depend on the type of application that has been considered initially: speech recognition or speaker recognition. Databases corresponding to these two types of applications have common parts but also specific parts. [3]

When performing speech recognition the important observations are related to system type (speaker dependent or speaker independent), vocabulary size (small vocabulary: 10-100 words, medium vocabulary: 100-1000 words, large vocabulary: 10000-100000 words).

Databases used for speech recognition must: ensure good coverage of given vocabulary and of the most significant acoustic units (phonemes, phonemes in certain context); ensure inter-phoneme separation as good as possible; be speaker voice independent (for speaker independent recognition purposes) or reflect the voice of the speaker (for speaker dependent recognition purposes).

Options available for database creation*: direct recordings* may cause specific problems: premises selection for recording purposes (studio), microphone selection (unidirectional/multidirectional, with/without active filter); *data acquisition from radio or TV shows that have been already transmitted over Internet*, may have some specific problems: recordings are performed under different conditions (open air/studio, movies, etc), coding type uniformity issue (A - PCM, μ- PCM, etc.), usage of different sampling frequencies (4, 8, 16 kHz,); *direct acquisition from TV and radio channels* may have specific problems: digitization of acquired analog signal, homogeneity of recording conditions, control of possible power outages or jammed transmission.

Fields needed in a vocal signal database for spontaneous speech recognition: vocal signal files, in various formats – WAV, OGG, AIFF, RAW etc; characteristics of the vocal signal files: acquisition moment, recoding length, speaker identity, speech type – read, spontaneous etc; label files, indicating the words of phonemes corresponding to each segment of the vocal signal recording; acoustic parameters files, synthetically representing the vocal signal: linear prediction coefficients (LPC), cepstral coefficients (that may be filtered on a MEL frequency scale) etc; an acoustic parameter file is associated to each vocal signal file.

Table 1 shows the main characteristics for the newly created data base.

TABLE 1. Data base characteristics

Collecting procedure	Recording Internet broadcasted Romanian TV shows
Used language	Spoken Romanian
Recordings duration	~4 hours with vocal signal
Speakers	12
Females	8
Males	4
Sessions per speaker	3-20
Time between recording sessions	One day to two weeks
Words total occurrences	37604
Words unique occurrences	8068
Speech type	Oral, spontaneous
Recording environment	TV studio
Vocal recorded signal sampling frequency	8kHz

Based on these elements the following set of essential problems was determined: vocal signal recording segmentation – for the ease and reliability of vocal signal processing by the human expert, the preferred length of the vocal signal files corresponds to a recording between 60s and 180s; vocal signal labeling – labeling can be performed on word level (time consuming, manual process) or on phoneme or triphone (semi-automated, bootstrapping process starting from an initial manual labeling [5]); the latest lacks reliability, due to the statistical algorithm process, e.g. forced Viterbi alignment of hidden Markov models [6]; vocal signal parameterization – specific criteria need to be fulfilled. In speech recognition, the criteria are maximizing inter-phoneme dispersion and minimize intra-phoneme dispersion. In other words, the acoustic parameters must outline the maximum variation between acoustical production of a phoneme and the acoustical production of a different phoneme, and a minimum variation between the same phoneme productions. Implementing these criteria is quite difficult; manual process is relatively not reliable, automated process does not scale robustly to database extensions. Hence, the designer's experience is essential for an efficient and scalable database definition and population.

3.2. Raw Data Acquisition and Structuring

In order to create a proper data base for spontaneous speech recognition in Romanian language, the chosen particularities are presented in table 1.

The recordings gather radio news, stories, TV shows, medical discussions, financial discussions, weather forecasts and other kind of information.

The speaker variability is easy observed as the audio data base contains twelve speakers taken from different domains, with different styles of life, knowledge, experience and speaking habits. They speak fluently as they are presenting different kinds of information.

For each speaker, there are between five and thirty eight wave files with all kinds of news.

3.3 Data Annotation Protocol

During the mentioned bellow phases spell check had to be performed several times.

Audio-to-text translation: the translation of the audio files into text files was done manually in the following manner: listening to each wave file, writing all the heard words from a wave file into a text file.

As it's spontaneous speech and the speakers are stammering or using incomplete grammatical word, in the translated text file the exact pronunciation was reproduced in writing.

One step was jumped by directly writing the words with the phonetic translations for the diacritics: ă -> @; â -> i_; î -> i_; ş -> S; ţ -> ts

Foreign names, acronyms and all other words were written the way they are pronounced in Romanian. As a convention for the research it was decided that long vocals should be written double. For instance, "uuuuu" was translated just "uu".

The number of text files obtained is the same as the number of wave files.

Gathering the files in a single text file: all the text files were gathered in a unique text file. This file contains all the unsorted words from all the speakers. It contains multiple occurrences of the spoken words. The histograms created for this file in order to obtain the number of occurrences for each word show a total number of words of 37604 as in table 1. Alphabetical sorting the file and taking out the multiple occurrences of the words generated another file with a total number of 8068 words with single occurrences. As figure 3 states, in the multiple word occurrences file, there are 12147 words with more than 100 occurrences each.

Creating the phonetic dictionary: the file containing single occurrences of each word used by the speakers is the basis of the phonetic dictionary.

The phoneme translation was made using the SAMPA conventions [6], [4].

The dictionary consists in a file with two columns - the first column contains the word from the unique occurrences file and the second column contains the phonetic transcription of the word with a blank space between the phonemes. The two columns are separated by a blank space.

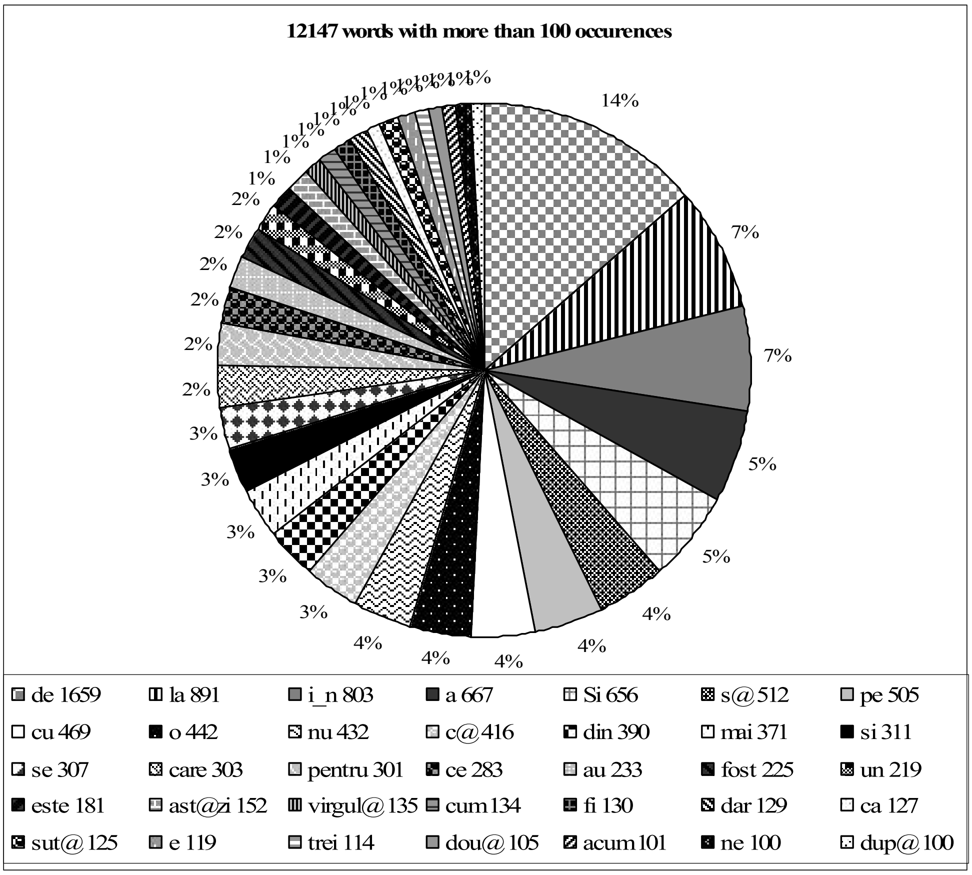

Fig. 3. The words with most occurrences in the new created database

Words labeling: the labeling was done manually using wavesurfer tool. The word labeled files were obtained using the following conventions: there is a short pause between every spoken word labeled with "sp", the long pauses are labeled with "sil"; there are long pauses considered between sentences or phrases or when the speaker is obviously deep breathing.

In order to considerably reduce the time for the labeling process which is done manually and is time consuming, maximum three labels for each *.wav file were used: the first and the last labels are "sil"s representing ambient noise, and the middle label contains all the spoken text with "sp"s and "sil"s.

Triphones labeling: from the phonetic word transcriptions the triphonetic transcriptions were created using HTK HLEd tool [5].

The histograms created for the triphonetic occurrences are illustrated in figures 4,5,6,7 and 8. From the word labeled files and the triphonetic transcriptions the triphonetic labeled files were generated.

3.4. Data Coding

For this phase [5] Mel Frequency Cepstral Coefficients (MFCCs) were used.

The HTK tool HCopy automatically converted the input data into MFCC vectors. The mfc files were obtained having as input the wave files.

The target parameters are to be Mel-Frequency Cepstral Coefficients (MFCCs).

The delta component is used and not the acceleration component (MFCC_E), the frame period is 10msec (HTK uses units of 100ns), the output is not saved in compressed format, and a crc checksum is not added. The FFT uses a Hamming window of 20 msec and the signal has first order preemphasis applied using a coefficient of 0.97. The filterbank has 26 channels and 12 MFCC coefficients are output. The variable ENORMALISE is by default true and performs energy normalization on recorded audio files. It cannot be used with live audio.

Creating these files reduces the amount of preprocessing required during training, which itself can be a time-consuming process.

4. Statistics for Romanian Language

The data base was created from scratch for future work and research in the spontaneous speech recognition domain. The results of the work, including the labeling, the dictionaries and all the convention took into consideration in creating the corpus, are illustrated in the histograms raised and the statistics regarding the number of the occurrences at word level and triphones level.

Instead of simple phonemes, set of three phonemes were used.

The reason of using triphones is that they permit the context analyzing, as they are entities that preserve the before and the after neighbors of a phonem, by including the left and right context phoneme in the triphonetic construction.

The most representative triphones in the newly created corpus are presented in the following figures, as they have the most occurrences in the used words.

Figure 4 illustrates the triphones that appear more than two thousand times in the corpus words. The triphones that have the most occurrences are “d+e” with 2305 occurrences and “d-e” with 1897 occurrences.

Figure 5 illustrates the next ten triphones with more than 729 and less than 984 occurrences among the words considered for the corpus. Triphones “l+a” and “l-e” are dominant in this histogram.

Figure 6 reveals the next ten different triphones which appear more than 618 times and less than 710 times among the words considered in the corpus.

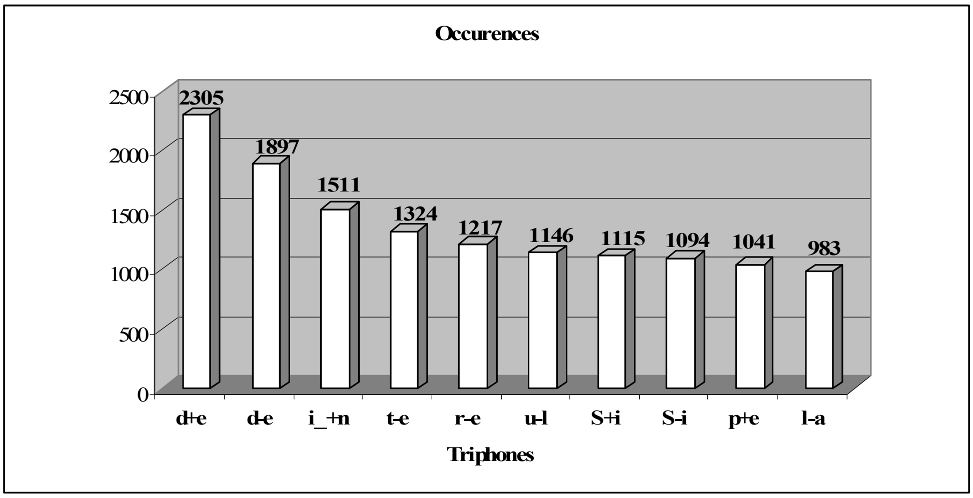

Fig. 4. The triphones with most occurrences in the corpus

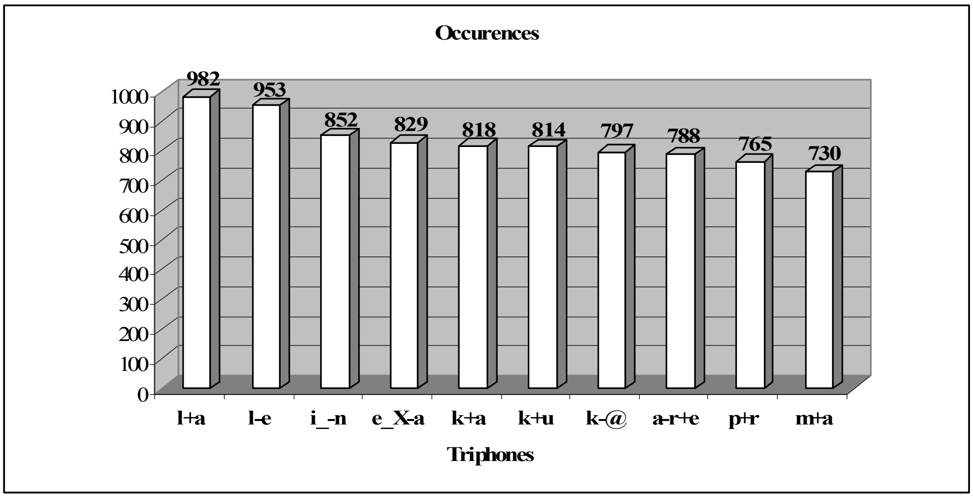

Fig. 5. The next ten triphones with most occurrences

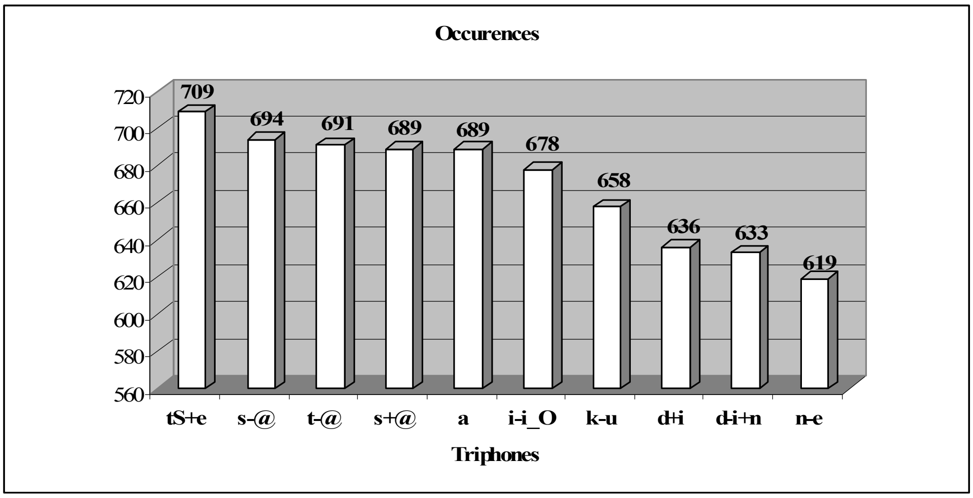

Fig. 6. Next ten triphones with most occurrences

The triphones percent representation according to the values in table 3 appears in figure 7.

TABLE 2. The first ten most representative triphones

Number of occurrences	Number of triphones
> 1000	9
500 - 1000	40
100 - 500	371
50 - 100	401
30 - 40	421
10 - 30	950
5 - 10	799
< 5	2104

From a total amount of 5095 triphones, which represent one hundred percent, the percentage distribution of the phonems represented in figures 4, 5, 6 is illustrated in figure 7.

In figure 7, one percent from the total amount of triphones has more than five hundred and less than one thousand appearances in the words considered for data base.

Seven percent from the total amount of triphones has more than one hundred and less than five hundred occurrences in the data base words.

Eight percent from the total amount of triphones has more than fifty and less than one hundred occurrences in the data base words. Another eight percent from the total amount of triphones has more than thirty and less than forty occurrences in the data base words.

Nineteen percent from the total amount of triphones has more than ten and less than one thirty occurrences in the data base words.

Sixteen percent from the total amount of triphones has more than five and less than ten occurrences in the data base words.

Forty-one percent from the total amount of triphones has more than one and less than five occurrences in the data base words. These triphones with fewer occurrences represent a little less than half of the total amount of triphones.

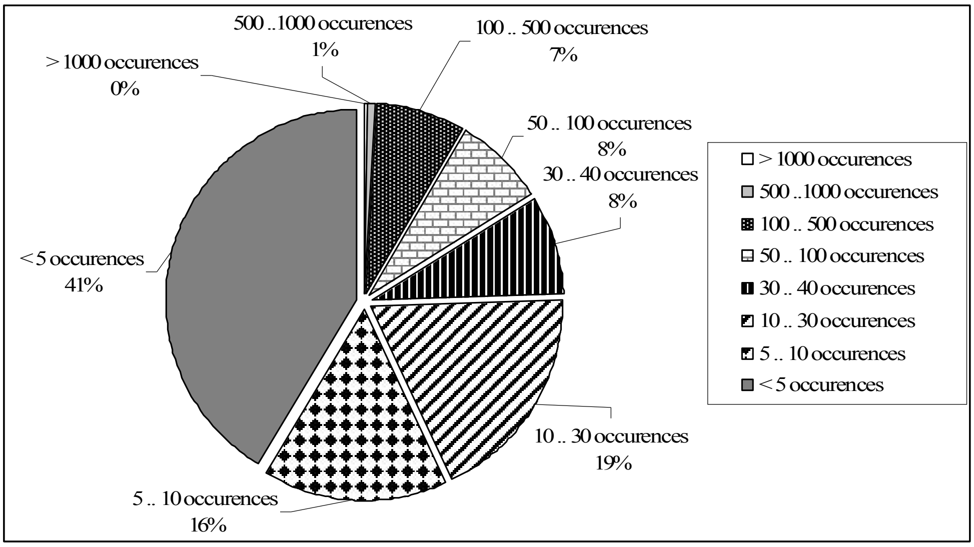

Fig. 7. Total number of triphones

5. Conclusions and Future Work

At the moment, a relatively medium size database for Romanian language has been created. However, to improve the analytic possibilities, on short term the goal is to increase the size of the database.

When creating database for Romanian language some issues have arisen.

A manual segmentation was performed when recording the audio files, as the TV programs are usually long (longer than 10 minutes or even hours). The manual segmentation of the files took place with durations between 60 and 180 seconds, for an easier processing.

The voice signal labeling at triphone level assumes many stages, so a manual labeling at audio file level was performed in the first place, followed by the automatic separation of every word in triphones by using the phonetic dictionary. The assumption that all triphones of a word have equal duration is not so true, but it was taken as a convention that has to be reconsidered in the next stage.

As a result of database processing information it has been observed that even though the database is relative medium size a big number of triphones has been obtained. This is specific to spontaneous speech.

Among triphones there are cases when some combinations have a bigger number of occurrences compared to others. For instance, triphones “d+e”, “d-e” and “i_+n” have more than 1500 occurrences.

For the time being, although there are triphones with lower number of occurrences these will not be ignored. There are 2903 triphones with less than ten occurrences and they represent fifty-seven percent from the total amount of triphones. When enlarging the size of the database it is expected that the triphones with less occurrences will have an increased number of entries.

Not ignoring triphones that have a small number of occurrences is assumed to be a characteristic of spontaneous speech and as such they have to be taken into consideration.

Based on the statistics represented above and database created so far the research will continue with the training of the HMM models generated for each triphone.

After the training, the project will include a dedicated phase to building testing block, testing activities and evaluation phase.

This part of the project will include several smaller steps such as: definition at a more detailed level of architecture to be used for testing purposes; definition of modules, scripts to be used during testing phase; training phase; converting the triphones to words; and as final activity, performance evaluation of the spontaneous speech recognition module.

Acknowledgments. The “IDEI” project has supported most part of the research for this paper.

References

[1]. S. Russell, P. Norvig, Artificial Intelligence – A Modern Approach, Prentice Hall, Second Edition (2003)

[2]. J. Martin, D. Jurafsky, Spoken Language Processing, Prentice Hall, Third Edition (2007)

[3]. X. Huang, A. Acero, H. Wuen-Hon, Spoken Language Processing – A Guide to Theory, Algorithm, and System Development, Prentice Hall (2001)

[4]. D. Burileanu, Basic research and implementation Decisions for a text-to-speech synthesis system in Romanian International Journal of Speech Technology (2002)

[5]. S.Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. (Andrew) Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, P. Woodland, The HTK Book (2005)

[6]. D. Munteanu, Contribuţii la realizarea sistemelor de recunoaştere a vorbirii continue pentru limba română, Teză de doctorat, Academia Tehnică Militară din Bucureşti (2006)

[7]. M. Drăgănescu, Gh. Ştefan, C. Burileanu, Electronica funcţională, vol. 1, Editura Tehnică, Bucureşti (1991)

[8]. V.B. Le, Reconnaissance automatique de la parole pour des langues peu dotees, these, Universite Joseph Fourier, Grenoble (2006)

[9]. T. Schultz, K. Kirchhoff, Multilingual Speech Processing, Academic Press (2006)