Health Evolution Monitoring for People with Speech Disorders by Means of Romanian Spontaneous Speech

Cristina Sorina Petrea, Diana Mirela Haneş, Andi Buzo, Vladimir Popescu and Corneliu Burileanu

Faculty of Electronics, Telecommunications and Information Technology,
University Politehnica of Bucharest, Romania
cburileanu@mesnet.pub.ro

Abstract

The conversational speech domain hasn't been exploited at full capacity. Research in the spontaneous speech recognition domain is still at incipient phase.

In the medical domain results are strongly required. A person with a speech disorder who's under treatment can be monitored by the means of the conversational speech.

The speech says very much about the speaker, especially when improvements are expected in his behavior. The before and after health state of a dyslexic or a person with dysarthria who is under treatment, can be obviously monitored and separated by the means of the speech. The improvements are easy to be observed in human speaking behavior, once a performant spontaneous speech recognition tool is implemented.

This paper presents the statistical results and the achievements obtained in the field of spontaneous speech recognition, beginning with the new Romanian corpus, built from scratch with words and triphones.

The research purpose is to identify the specificities of the spontaneous speech and to use them on full potential in order to implement a performant speech recognition tool meant to serve the contemporary needs.

1. Introduction

Spontaneous speech is an opened area in the large domain of speech recognition. Achievements are needed as they have high applicability in various domains, especially in medicine. The evolution of a dysarthric or a dyslexic and the phases he goes through from the beginning of the treatment until the end can be observed based on his speech.

Creating a speech recognition system that makes use of spontaneous speech, will also remove the need to type or hand write for a dyslectic, a person with dyspraxia or a person affected by speech disorders.

That's why the spontaneous speech is a gold mine that has to be properly exploited by researchers.

1.1. Spontaneous speech recognition characteristics

In spontaneous speech the speaker does not preserve rules; he talks freely, 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 conversational speech 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.

Some of the issues in processing 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.

1.2. State-of-the-art

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.

There are major differences between speech read from scripts and real time mind made and freely expressed spontaneously speech.

The most well known speech recognition software is "Dragon NaturallySpeaking" and is used to create documents, reports, emails and search the desktop and the Web. It makes use of continuous (no pauses), large vocabulary (300,000 words) [1].

Dyslectics, dyspractics, persons with speech disorders are not trained speakers for speech recognition purposes. They express themselves and ask for help in a natural manner, they are spontaneous speakers [2], [3].

Implementing the spontaneous speech recognition tool will bring novelty and improvement to the existing implementations, as it's suppose to use free speech of non-trained users, so everybody will be able to use it no matter the social condition, the context, the knowledge or other momentary impediments.

2. Proposed System Recognition Architecture

Building a speech recognizer from scratch involves in the first place a very good task and sub-task determination and separation [4], [5]. Figures 1 and 2 represent the proposed approach for the spontaneous speech recognition system with both training and testing phases. 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 [4], [5]. For small vocabularies (few tens of words),

the models can capture word properties, but sounds units are generally modeled (such as phonemes or triphones) [4], [5]. Nowadays, most successful approaches consider this model matching as a probabilistic process that has to account for the temporal variability and acoustic variability [6], [7], [8].

2.1. The training phase

As represented in figure 1, recorded voice files were used as input in the training phase.

The wave files labeling was most time consuming. The phonetic transcription using SAMPA conventions [11] was done in order to build the phonetic dictionary. After these steps, the labeling passed at triphone level.

The recognition system is based on Hidden Markov Models (HMMs) [4], [5]. One important step in the training phase was the definition of a HMM prototype for each triphone. 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 the HTK tools HInit and HCompV [10].

The voice signal MFCC parameters were needed [4], [5] to be used together with the labeled triphones for training the HMMs for each triphone. An iterative Viterbi alignment for the HMMs was used [4], [5], using HRest and HERest HTK tools [10].
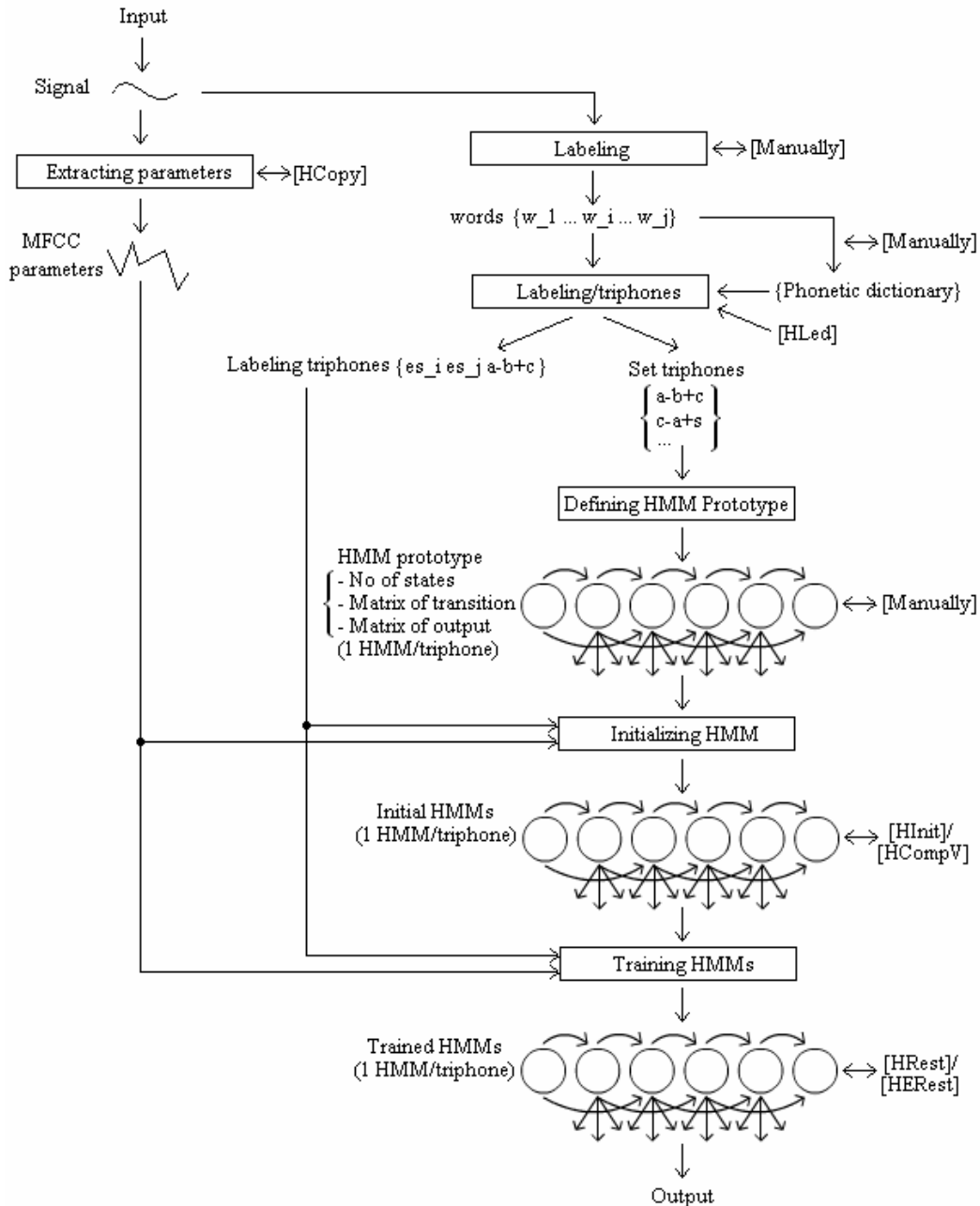

Fig. 1. The system architecture considered for the HMMs training phase

2.2. The testing phase

The testing phase uses all the output obtained in the training phase [4], [5], as represented in figure 2.

The testing sentences used as input were chosen from the audio files already existent in the database.

The wave files used as input were first parameterized and the MFCCs (Mel Frequency Cepstrum Coefficients) parameters were extracted using the HCopy HTK tool [10].

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

The next step involves the conversion of the triphones sequence into a word sequence.

The grammar defined and used at this phase was created using the HParse tool from HTK toolkit [10]. This grammar contains all the possible combinations of the words in the dictionary that can appear in the current speech. On a first approach, a convention has been established regarding the grammar construction. It mentions that given the fact that spontaneous speech is involved, a word in the dictionary can follow any other word. This convention has to be analyzed with maximum attention on the further approaches, depending on the characteristics of the spontaneous speech.

Having the reference text, the results can be evaluated. Reference text was 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.
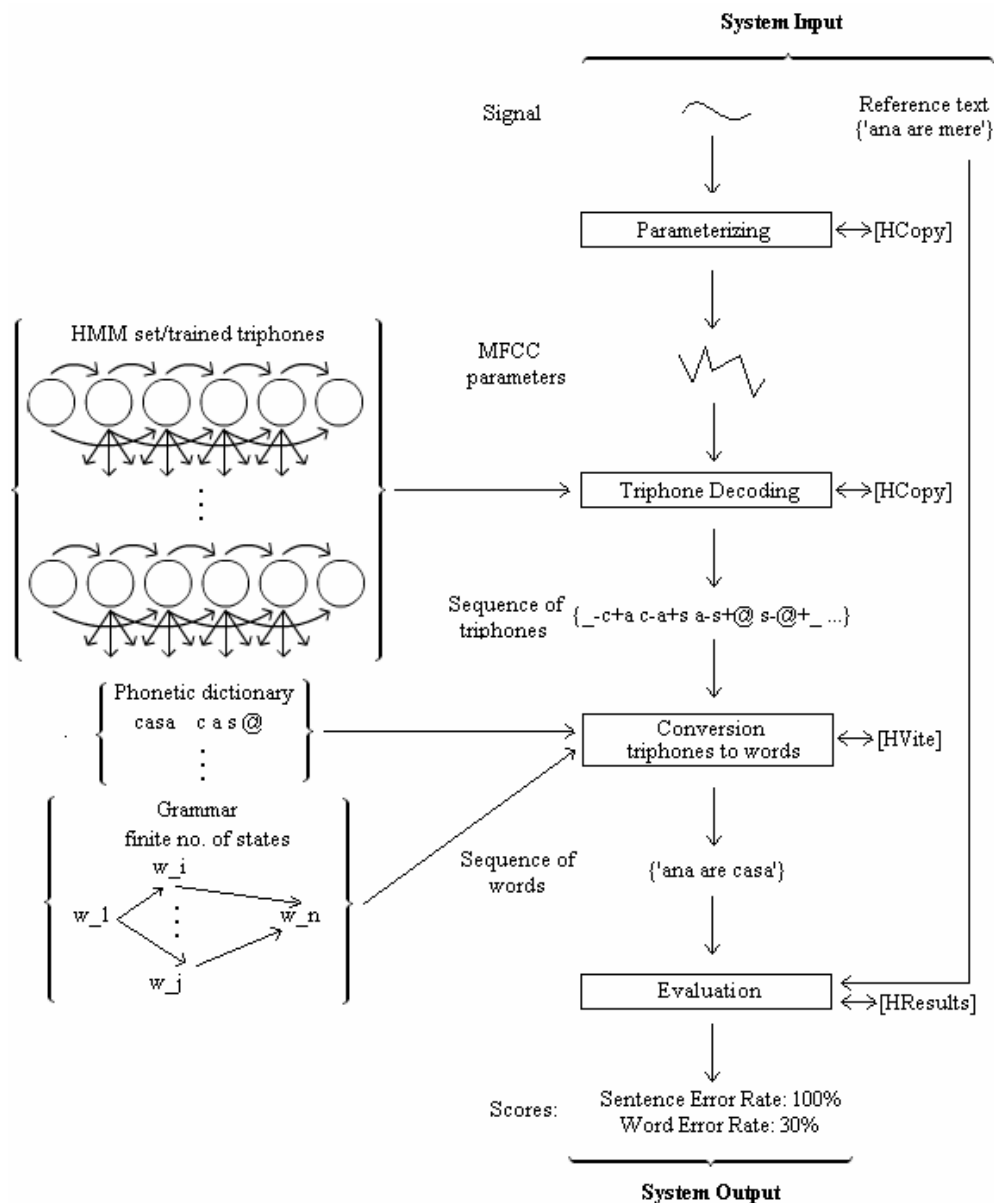

Fig. 2. The system architecture considered for the testing phase

3. Spontaneous Speech Database Building

The limited knowledge in the domain of spontaneous speech recognition requires to a lot of information to be collected. That is a strong reason in order to obtain a large amount of data for research. The corpora used in spontaneous speech recognition should have the property of being wide.

The new Romanian corpus that was obtained has medium to large size. Table 1 shows the main characteristics for the newly created database.

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

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 database contains twelve speakers taken from different domains, with different styles of life, knowledge, experience and speaking habits.

For each speaker, there are between five and thirty-eight wave files in the database.

Table 1. Database 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	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
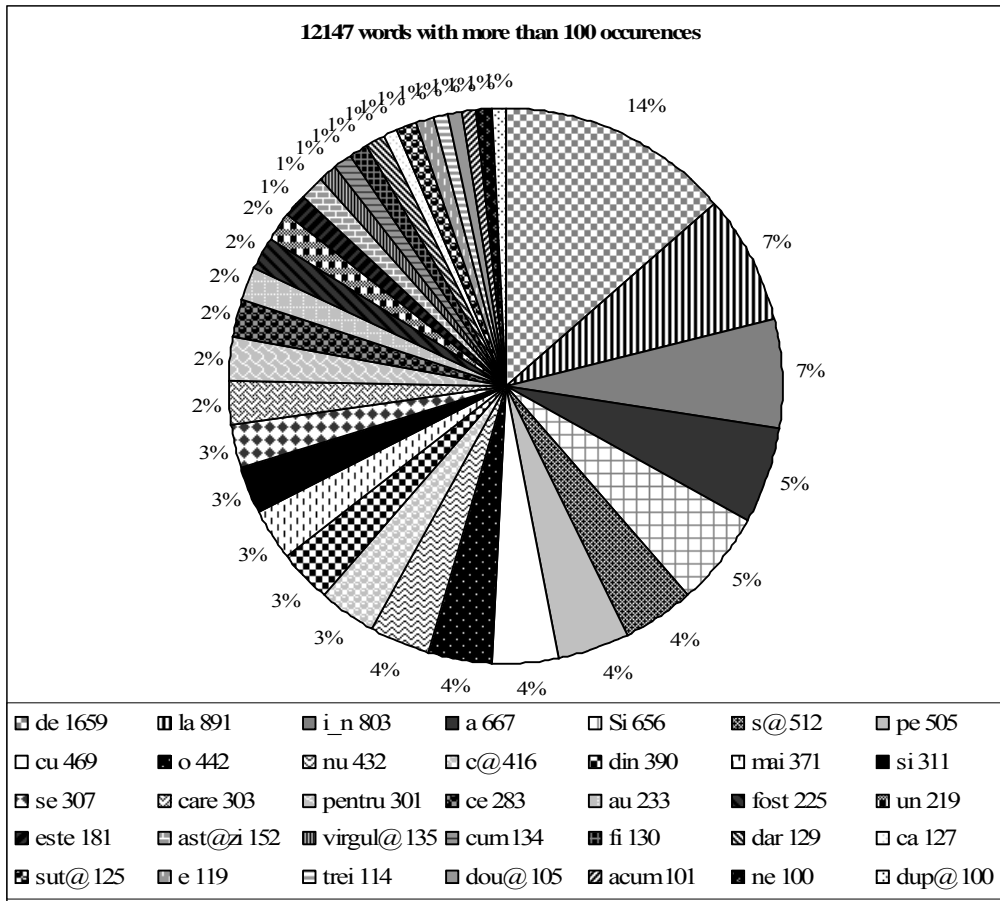
Fig. 3. The words with most occurrences in the new created database

3.1 Data Annotation Protocol

During the mentioned bellow phases spell check was performed several times.

Audio-to-text translation: the translation of the audio files into text files was done manually [4], [5]. 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 [11].

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 as double ones. For instance, "uuuuu" was translated just "uu".

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 [4], [5]: the phoneme translation was made using the SAMPA conventions [9], [11]. 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.

Words labeling: the labeling was done manually using wavesurfer tool. The word labeled files were obtained using 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 audio 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 [10]. The histogram created for the triphonetic occurrences is illustrated in figure 4 and figure 5.

From the word labeled files and the triphonetic transcriptions the triphonetic labeled files were generated.

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 on the next stage.

3.2 Data Coding

For this phase Mel Frequency Cepstral Coefficients were used [4], [5], having as input the wave files. The HTK tool HCopy [10] automatically converted the input data into MFCC vectors. The mfc files were obtained.

The delta component is used and not the acceleration component (MFCC_E_D), 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 ms 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.

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 database was created from scratch for future work and research in the spontaneous speech recognition domain. The results of the work are illustrated in the histograms raised and the statistics regarding the number of the occurrences at word level and triphone level.

Instead of simple phonemes, set of three phonemes was 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.
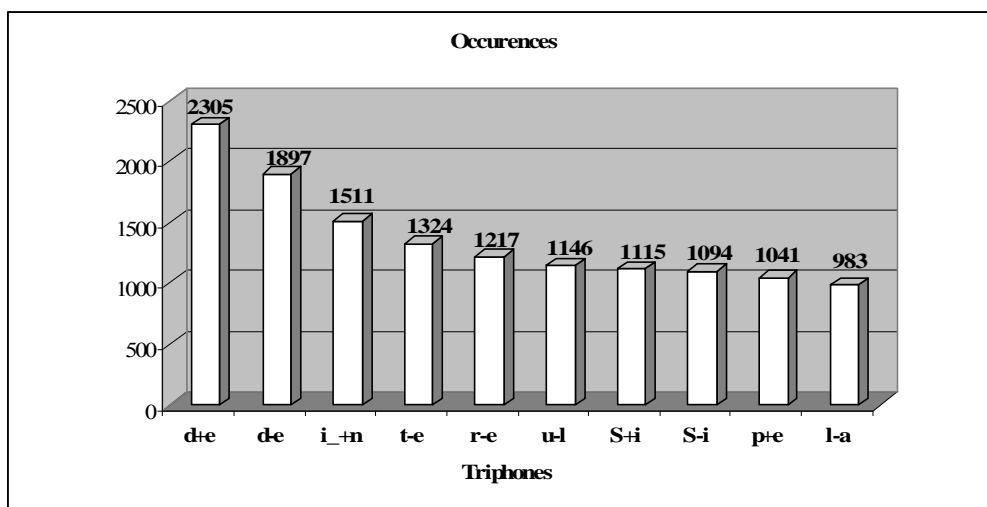

Fig. 4. The triphones with most occurrences in the corpus
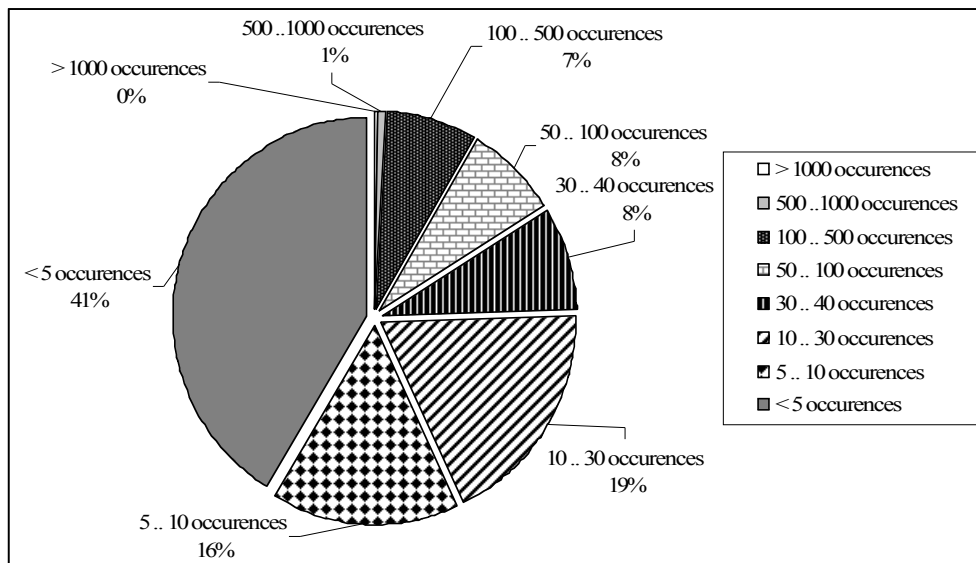
Fig. 5. Total number of triphones

Given the fact that the words are separated by the "sp" marker, which is not actually a phoneme, word beginnings and ends are realized as diphones. For instance, the word "c a s @" contains the phonetic constructions: "c-a", "c-a+s", "a-s+@" and "s+@". The most representative triphones in the newly created corpus are presented in the above figure, 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.

In figure 5, only one percent from the total amount of 5095 triphones has more than five hundred and less than one thousand appearances in the words considered in the database.

5. Conclusions and Future Work

At the moment, a medium to large size database for Romanian language has been created. The goal is to enlarge the database with specific spontaneous speech constructions.

The big number of triphones that has been obtained is specific to spontaneous speech. There are cases when some triphones 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. The explanation is that these are prefixes and suffixes and in the conversational speech, when stammering, the speaker tends to double the beginning or end of a word.

Triphones with lower number of occurrences will not be ignored. 2903 triphones have 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. Due to the nature of spontaneous speech, it is expected to obtain a completely different view over the triphones occurrences, on database size enlarging.

The corpus and the spontaneous speech recognition results will have applicability in medical monitoring and will help the persons with communication and writing difficulties.

As the already existent recognition tools which make use of continuous speech are frequently used, the spontaneous speech recognition tool intended to be implemented will be highly appreciated. It will ease the work of the user, taking off some of the existent constraints, like forcing to speak grammatically correct, stressing a dyslectic who is not able to correctly pick up his words, etc.

6. References

[1] www.dragontalk.com
[2] G. Reid, "An Overview of Current Research in Dyslexia". *University of Edinburgh, Scotland*
[3] www.minddisorders.com
[4] V. Rialle, J.B. Lamy, N. Noury, L. Bajolle, "Remote monitoring of patients at home: A Software Agent approach". *Computer Methods and Programs in Biomedicine, 2003.*
[5] M. Vacher, J.F. Serignat, S. Chaillol, D. Istrate, V. Popescu, "Speech and Sound Use in a Remote Monitoring System for Health Care"
[6] S. Russell, P. Norvig, "Artificial Intelligence – A Modern Approach", *Prentice Hall, Second Edition, 2003*
[7] J. Martin, D. Jurafsky, "Spoken Language Processing", *Prentice Hall, Third Edition, 2007*
[8] X. Huang, A. Acero, H. Wuen-Hon, "Spoken Language Processing – A Guide to Theory, Algorithm, and System Development", *Prentice Hall, 2001*
[9] D. Burileanu, "Basic research and implementation Decisions for a text-to-speech synthesis system in Romanian International Journal of Speech Technology", 2002
[10] 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
[11] D. Munteanu, "Contribuții la realizarea sistemelor de recunoaştere a vorbirii continue pentru limba română", Teză de doctorat, *Academia Tehnică Militară Bucureşti, 2006*