Spontaneous Speech Database for the Romanian Language with Medical Applicability

Cristina Sorina Petrea[1], Diana Mirela Haneş[1], Andi Buzo[1]
Vladimir Popescu[1, 2] and Corneliu Burileanu[1]

[1] Faculty of Electronics and Telecommunications, "Politehnica" University of Bucharest, Romania
[2] Laboratoire d'Informatique de Grenoble, Grenoble INP, France

Abstract. The research in the field of spontaneous speech for *Romanian language* has direct applicability in the medical process of patient remote monitoring and in helping the *dyslexics* and the *dyspraxic* people. It is an unexplored domain, with high potential for worthy results in speech recognition area. The goal is to achieve performance in helping persons with disabilities and common patients.
This paper describes 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 challenge is to create a spontaneous speech recognition tool for Romanian language with medical applicability for the benefit of persons with difficulties in mobility, communication and even common people.

1 Introduction

Spontaneous speech is an opened area in the large domain of speech recognition and has direct applicability in medical environment. It has applicability in patients' remote monitoring process and also in helping people with dyslexia and those with dyspraxia.

1.1 Spontaneous Speech Recognition Characteristics and Applicability

Nowadays, research work is done in medical remote monitoring domain, with the aim of detecting abnormal patient behavior at home in the context of residential health care [1], [2]. Speech is analyzed in order to extract informative key words like "Help me" and "Doctor" for distress identification.

Research is also intensively done in the dyslexia and dyspraxia domain. Dyslexia has been diagnosed in people of all levels of intelligence [10].

People suffering from dyslexia and dyspraxia have troubles with typing and hand writing and these persons monitored in their own homes or in the hospitals need special attention and help. Creating a speech recognition system that makes use of spontaneous speech, will remove the need to type or hand write for a dyslectic or a dyspractic and will provide the ability to help a supervised person on request by understanding his urgent needs.

Petrea C., Hanes D., Buzo A., Popescu V. and Burileanu C. (2009).
Spontaneous Speech Database for the Romanian Language with Medical Applicability.
In *Proceedings of the 1st International Workshop on Mobilizing Health Information to Support Healthcare-related Knowledge Work*, pages 78-86
DOI: 10.5220/0001818300780086

Stuttering and cluttering are warning signs of dyslexia [10]. Many dyslexics also can have problems with speaking clearly. That's why spontaneous speech is the right filed needing attention in helping these people.

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 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.

Some of the issues in processing spontaneous speech are: false starts, filled pauses, incoherence of the speaker with possible ungrammatical constructions. 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 the field of speech recognition reveals poor accuracy for the freely spoken spontaneous speech. As the spontaneous speech makes use of the acoustic and linguistic models that have been created 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 for desktop searching purposes. It makes use of continuous speech (no pauses), large vocabulary (300,000 words) [11].

Dyslectics, dyspractics, persons with mobility disabilities and also common patients are not trained speakers for speech recognition purposes. They express themselves and ask for help in a natural manner, they are spontaneous speakers.

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 Recognition System 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 (a few tens of words), these models can capture word properties, but sounds units are generally modeled (such as phonemes or triphones) [4], [5].

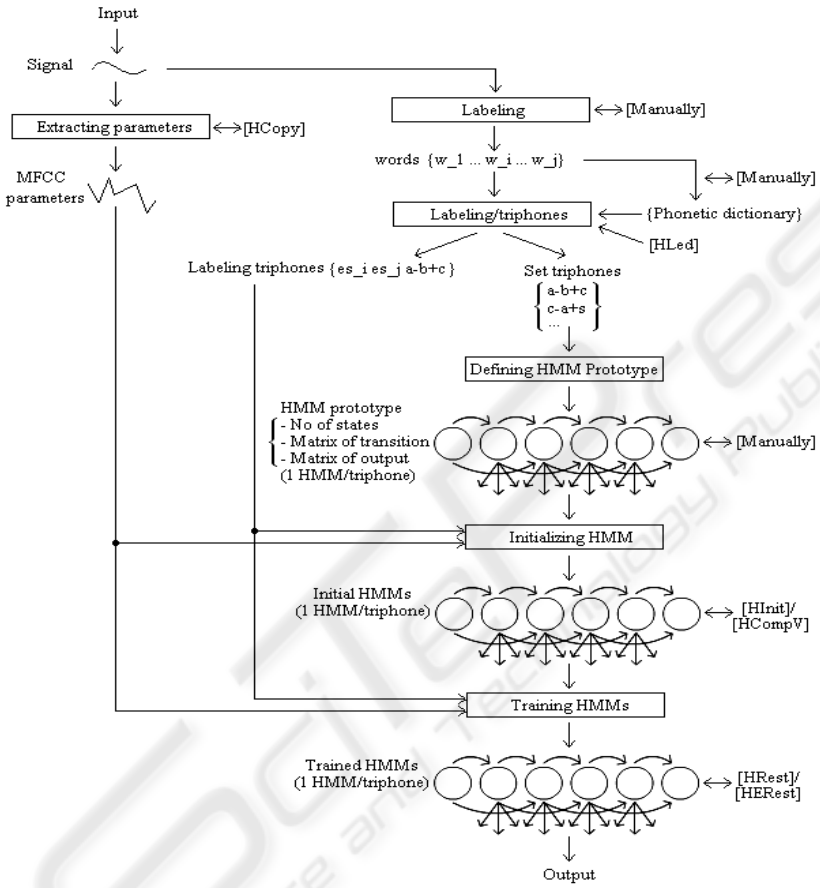

Fig. 1. The system architecture considered for the HMMs training phase.

The most successful approaches nowadays consider this model matching as a probabilistic process that has to account for the temporal variability and acoustic variability [3].

2.1 Training Phase

Recorded voice files have been used as input in the training phase. The wave files labeling was most time consuming.

The phonetic transcription using SAMPA conventions [8] 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 performed by using the HTK tools HInit and HCompV [7].

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 [7].

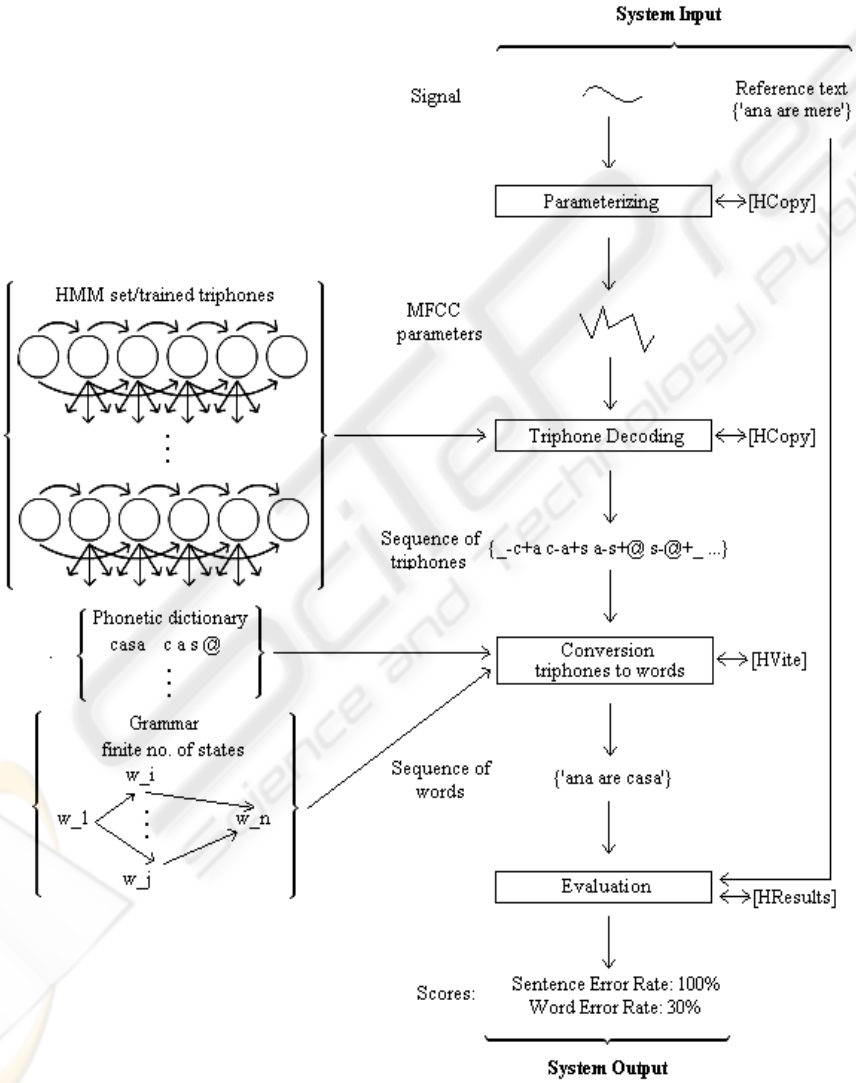

Fig. 2. The system architecture considered for the testing phase.

2.2 Testing Phase

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

Testing sentences were used as input and the voice signal was chosen from the already existent 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 [7].

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

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

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.

3 Spontaneous Speech Database Building

The corpora used in speech recognition should have the property of being wide. A series of problems and specific elements have been identified related to construction of databases for spontaneous speech recognition.

Table 1 shows the main characteristics for the newly created 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

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.

3.1 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 [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. 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.

All 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.

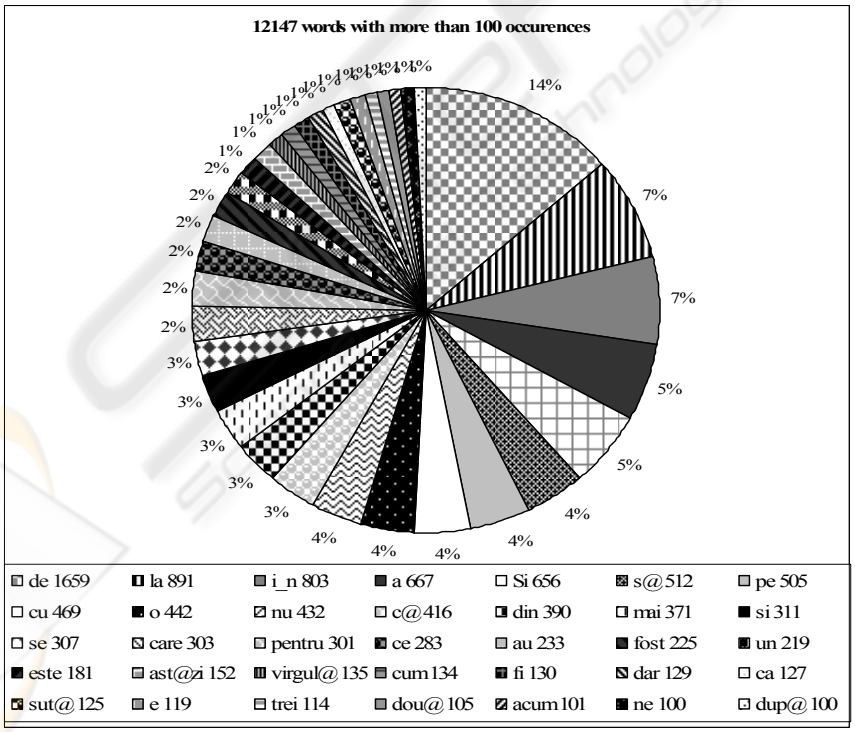

Fig. 3. The words with most occurrences in the new created database.

Creating the phonetic dictionary [4], [5]: the phoneme translation was made using the SAMPA conventions [8], [6]. 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 [7]. The histogram created for the triphonetic occurrences is illustrated in figure 4. 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 in 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 [7] 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), the frame period is 10ms (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 and the statistics regarding the number of the occurrences at word level and triphones level. Instead of simple phonemes, set of three phonemes was used.

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 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 figure, as they have the most occurrences in the used words.

Figure 4 illustrates the triphones that have more than 2000 occurrences in the corpus words. The triphones that have the most occurrences are "d+e" with 2305 occurrences and "d-e" with 1897 occurrences.

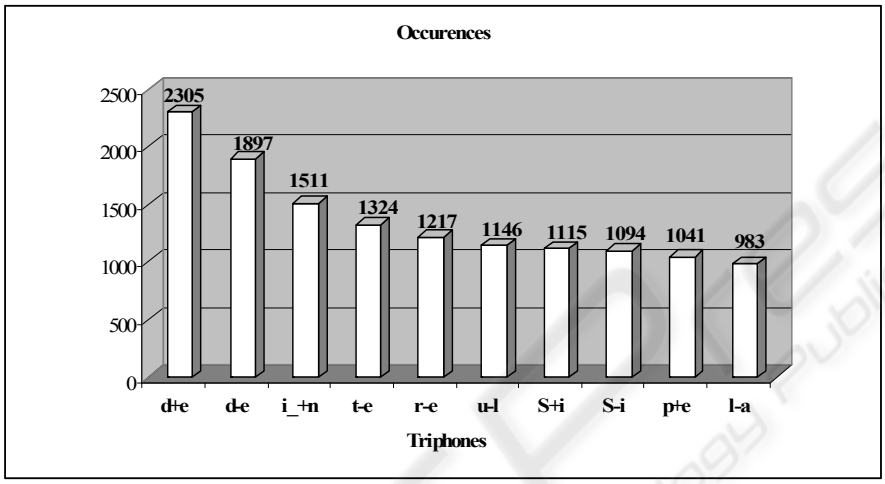

Fig. 4. The triphones with most occurrences in the corpus.

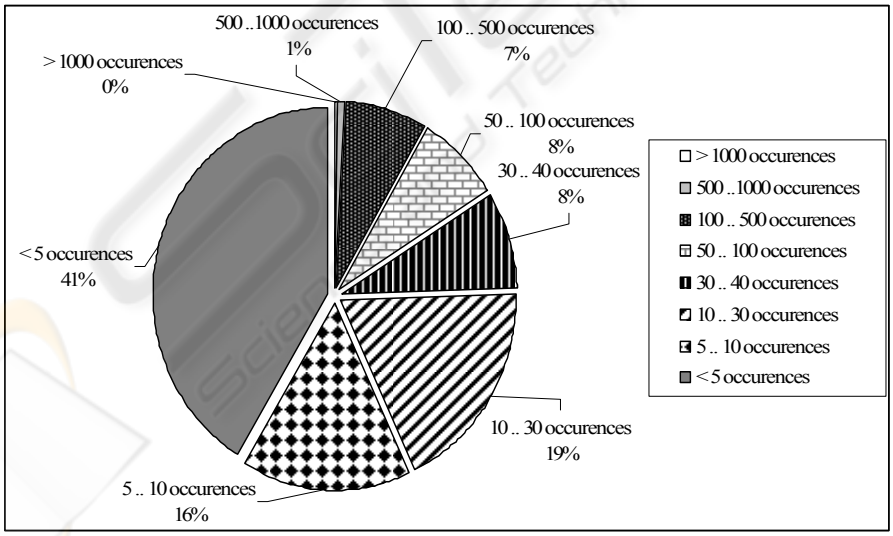

Fig. 5. Total number of triphones.

In figure 5, only one percent from the total amount of 5095 triphones has more than 500 and less than 1000 occurrences in the words considered in the database.

5 Conclusions and Future Work

At the moment, a medium size database for Romanian language has been created. On short term, the goal is to enlarge the database with common words and with specific words to be used in remote monitoring.

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. Triphones with lower number of occurrences will not be ignored. 2903 triphones have less than 10 occurrences and they represent 57% 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. 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.

The corpus and the spontaneous speech recognition results will have applicability in medical monitoring from remote locations and will help the persons with mobility, 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 pick up correctly his words, etc.

References

1. Rialle, V., Lamy, J.B., Noury, N., Bajolle, L.: Remote monitoring of patients at home: A Software Agent approach. Computer Methods and Programs in Biomedicine (2003)
2. Vacher, M., Serignat, J.F., Chaillol, S., Istrate, D., Popescu, V.: Speech and Sound Use in a Remote Monitoring System for Health Care
3. Russell, S., Norvig, P.: Artificial Intelligence – A Modern Approach, Prentice Hall, Second Edition (2003)
4. Martin, J., Jurafsky, D.: Spoken Language Processing, Prentice Hall, Third Edition (2007)
5. Huang, X., Acero, A., Wuen-Hon, H.: Spoken Language Processing – A Guide to Theory, Algorithm, and System Development, Prentice Hall (2001)
6. Burileanu, D.: Basic research and implementation Decisions for a text-to-speech synthesis system in Romanian International Journal of Speech Technology (2002)
7. Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X. (Andrew), Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (2005)
8. Munteanu, D.: Contribuţii la realizarea sistemelor de recunoaştere a vorbirii continue pentru limba română, Teză de doctorat, Academia Tehnică Militară din Bucureşti (2006)
9. Schultz, T., Kirchhoff, K.: Multilingual Speech Processing, Academic Press (2006)
10. http://www.dyslexic.com/
11. http://www.dragontalk.com/