THE PUBLISHING HOUSE
OF THE ROMANIAN ACADEMY

PROCEEDINGS OF THE ROMANIAN ACADEMY, Series A,
Volume 11, Number 1/2010, pp. 83–91

SPONTANEOUS SPEECH RECOGNITION FOR ROMANIAN IN SPOKEN DIALOGUE SYSTEMS

Corneliu BURILEANU[*], Vladimir POPESCU[*, **],
Andi BUZO[*], Cristina Sorina PETREA[*], Diana GHELMEZ-HANEŞ[*]

[*] Faculty of Electronics, Telecommunications and Information Technology, University "Politehnica" of Bucharest, Romania
[**] Laboratoire Informatique d'Avignon, University of Avignon, France
Corresponding author: Corneliu BURILEANU, E-mail: cburileanu@messnet.pub.ro

In this paper we present an attempt to develop a speech recognition module for the Romanian language in order to be used in a dialogue system. The main characteristics of such a dialogue system are first discussed. Further, we explain the design and acquisition of a spontaneous speech database for training the decoder: the design guidelines in developing the database, as well as several practical issues encountered, along with some triphones balancing statistics are pointed out. Then, the speech recognition architecture (based on components in the "Hidden Markov Modeling Toolkit" – HTK) is described in detail, emphasizing the two aspects, training and decoding. In the next section, a discussion of several preliminary recognition results is provided, emphasizing current limitations and the need to significantly increase the size of the database. A set of conclusions and perspectives are offered at the end of the paper.

Key words: Continuous speech recognition; Speech database; Hidden Markov Modeling.

1. INTRODUCTION

The personalized interaction between human subjects and computers represents a dominant challenge nowadays, as software services and products become more and more user-centered. Thus, spoken computer dialogue constitutes one of the most natural and convenient interaction means for the human.

This type of dialogue systems can be seen as advanced applications of spoken language technology. A dialogue system represents a voiced and relatively natural interface between the user and a software application. Thus, spoken dialogue systems subsume most of the fields in spoken language technology, including speech recognition and synthesis, natural language processing, and dialogue management (planning).

A dialogue system involves the integration of several components, which generally provide the following functions [1]:

● speech recognition: conversion of an utterance (represented as a sequence of acoustic parameters), into a word sequence; ● language understanding: analysis of a word sequence in order to obtain a meaning representation for this sequence, in the dialogue context; ● dialogue management: system-human interaction control, as well as the coordination of the other components of the dialogue system; ● task management: interfacing of the dialogue management and language understanding modules, with the application domain for the tasks performed by the system; ● answer generation: computation of the sequence of words constituting the answer generated by the system, situating it in the discourse context represented by the dialogue history, and in the pragmatic context, represented by the relationship between user and machine, as well as by their social roles; ● speech synthesis: conversion of the text representing the system's answers, into an acoustic waveform.

Among these components, some (dialogue and task management, partially language understanding and answer generation) are language independent and (in part) dependent on the application domain, whereas others (speech recognition and synthesis) depend on the language, being (in principle) independent of the application domain. Thus, for the components in the first category, reuse in new languages is, to a great extent, possible, if the application domains are kept, whereas the components in the second category have to be developed for each new language, in a manner that is independent of the application.

For the international languages (English, French) there are complete human-computer dialogue systems, in domains such as train or plane ticket reservation (the American system CMU Communicator, designed at Carnegie Mellon University, systems developed by France, in the ESPRIT European projects), resource-meeting room management in voice portal applications (the PVE – „Portail Vocal pour l'Entreprise" system, developed in cooperation with Grenoble University 1, CNRS and several companies, such as France Telecom and with government financing [10]). On the other hand, in other languages such as Romanian, considered "under-resourced, from a speech database point of view" (according to recent studies – [5], or [15]), the development of spoken dialogue systems is a long-term process.

The task of the speech recognition component in a spoken dialogue system consists in converting the utterance (in acoustic form) came from the user, into a sequence of discrete units, such as phonemes (sound units) or words. A major obstacle in accomplishing a reliable recognition resides in speech signal variability, which results from the *linguistic variability* (effect of several linguistic phenomena such as phonetic coarticulation), the *speaker variability* (effect of inter- and intra-speaker acoustic differences) and the *channel variability* (effect of the environmental noise and of the transmission channel noise).

The speech recognition component in a typical dialogue application has to take into account several additional issues:

– *speaker independence*: speech has to be collected from an acoustically representative set of speakers, and the system will use these data in order to recognize utterance came from (potential) users, whose voices were not used during training;

– *size of the vocabulary*: the number of words that are "intelligible" to the dialogue system depends on the application considered, as well as on the dialogue (management) complexity [8];

– *continuous speech*: the users are expected to be able to establish a conversation with the spoken dialogue system, using unconstrained speech and not commands uttered in isolation (isolated voice command systems for industrial robots have been developed also in Romania, in the ninth decade of the 20th century, at University "Politehnica" of Bucharest [2]). The issue of establishing the limits of the words is extremely difficult for continuous speech, since in the acoustic signal there is no physical border between them. Hence, linguistic or semantic information can be used in order to separate the words in users' utterances;

– *spontaneous speech*: since users' utterances are normally spontaneous and non-planned, there are generally characterized by disfluencies, false starts, stops in the middle and re-starts, or extra linguistic phenomena, such as cough. The speech recognition module must be able to extract, out of the speech signal, a word sequence allowing the semantic analyzer to deduce the meaning of the user's utterance.

In principle, 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). Such systems, based on statistical approaches, are available in the research community (the SPHINX system from Carnegie Mellon University, the HTK – "Hidden Markov Modeling Toolkit" toolkit from Cambridge University, the RAPHAEL system, from Laboratoire d'Informatique de Grenoble, etc.), as well as in the commercial domain (systems developed by Nuance, Dragon or Microsoft in the United States; those developed by France Telecom, Prosodie or As An Angel in France, etc.). At the same time, for the Romanian language, continuous speech recognition systems have been developed at the Military Technical Academy in Bucharest, and at the University "Politehnica" of Bucharest [11, 12].

A typical continuous speech recognition system works in two regimes: ***training***, which involves the creation of necessary knowledge, i.e., of the models being used (that are acoustic and linguistic), and ***recognition***, which involves the usage of the resources created during training, in order to convert an utterance (in acoustic form) came from the user, into a word sequence.

The training process takes as input a (large) set of acoustic data, labeled at word level and optionally phonetically transcribed, and involves three distinct steps:

1. acoustic processing of the input signal, involving, in turn, sampling, quantization, parameterization (linear prediction analysis or cepstral analysis on non-linear frequency scales – Mel scale cepstrum) ;

2. acoustic models estimation, at the level of the units considered (words, phonemes or triphones), involving the definition of their initial features (for hidden Markov models, widely used in this purpose, these characteristics mean: the number of states, the topology of the state transition matrix, the number of Gaussian mixtures that model emission probabilities for each state, weights and initial characteristics of these mixtures – mean vector and covariance matrix). This process involves, in general, the usage of some iterative algorithms for model parameter estimation, based on acoustic data.

3. computation of the language model for the language being used in the application, involving the usage of textual data reflecting typical utterances for the task concerned; these data drive the estimation of the succession probabilities between words, usually taken in groups of two to three (in that case, these models are called bigram models, and trigram models, respectively).

The recognition process considers in input an unlabeled segment (because unknown to the system) of speech signal and involves three distinct stages:

1. acoustic processing of the input speech signal, in a manner identical to the training process;

2. acoustic decoding of the parameter vectors sequence representing the input signal (template matching with the models estimated during training), obtaining a sequence of most likely acoustic units;

3. refining the results of the acoustic decoding process, using the language model, by confronting the word sequence obtained at the preceding step, to the language model.

At the end of these processing steps, the output of the recognition is represented by a number of alternative word sequences, for an utterance; sometimes, the differences between alternative word sequences are small and determined by semantically irrelevant words. The choice of the most relevant alternative, in a specified context, is the responsibility of other components in the dialogue system (namely, the semantic analysis component, or the dialogue manager).

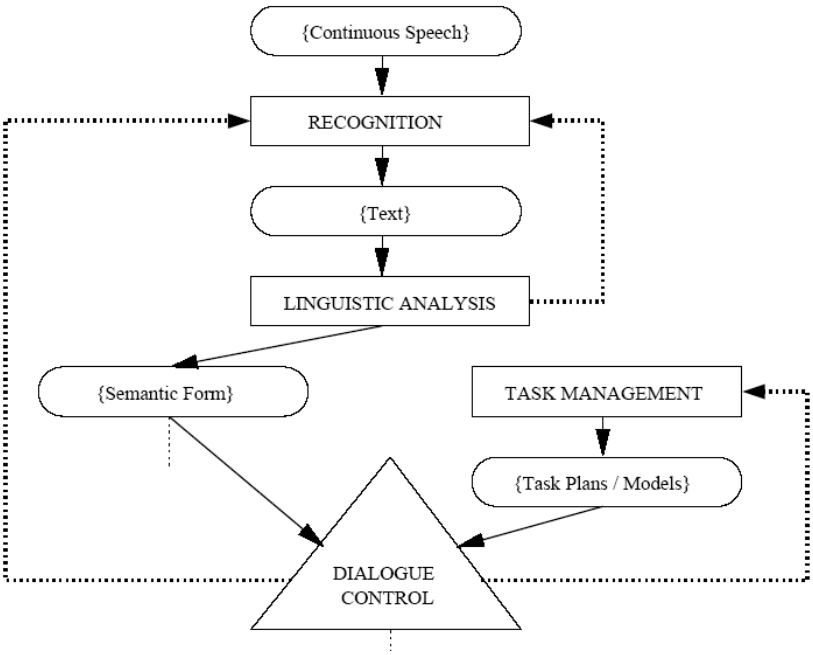

Fig. 1 – The speech recognition component in spoken dialogue architecture.

From the statements above, we observe that the concern of developing a speech recognition module suited for spoken dialogue systems is not trivial, although stand alone continuous speech recognition systems exist (e.g., for dictation applications) and begin to appear also in Romania. Thus, a continuous speech recognition module for spoken dialogue in the 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. The place of a speech recognition component in a typical dialogue system architecture (such as the framework proposed in [1]) is illustrated in Fig. 1, where it can be observed how the semantic analysis and dialogue control modules provide feedback to the speech decoder.

2. ACQUIRING A SPONTANEOUS SPEECH DATABASE IN THE ROMANIAN LANGUAGE

2.1. Basic Principles

Speech recognition can be seen as a pattern recognition process, which can be undertaken either via rule-based approaches, or through statistical methods [4]. This latter way is preferred nowadays, due to its good performances, while maintaining acceptable implementation costs [6]. A statistical approach however needs training data for automatically creating knowledge that the system can use. For speech recognition, the particularities of the training data depend on the nature of the application considered: speech recognition (*what* we say), or speaker recognition (*who* speaks). Thus, there are similarities and differences between the databases that are appropriate for each of these two application types.

The databases suited for speech recognition have to satisfy the following criteria: ensure a good coverage of the vocabulary and of the relevant acoustic units (phonemes, triphones), ensure a good inter-phoneme separation, be independent of the voice of a particular speaker (for speaker-independent speech recognition), or, on the contrary, fine-tuned to the voice of a particular speaker (for speaker-dependent speech recognition).

A database can be acquired via the following methods:

– direct recording; this yields a series of particular issues: choosing the recording place (studio, etc.), choosing the microphone (uni-directional, omni-directional, with or without active filter, etc.);

– acquisition of TV or radio broadcasts over the Internet; the particular issues in this situation are: homogeneity of the recording conditions (outdoor shows, studio recordings, movies, etc.), uniformity of the speech coding standards (A - PCM, μ- PCM, etc.), differences in the sampling frequencies (4, 8, 16 kHz, ...), control of the speaker set (so that a balanced speech signal is obtained from all the speakers);

– direct acquisition from radio or TV broadcast channels; in this case, the specific issues are: the analog-digital conversion of the signal, the homogeneity of the recording conditions;

A database for continuous speech recognition has the following components [13]: ● a set of speech signal samples; ● a set of correspondences between the speech signal samples and their features (duration of the signal, identities of the speakers, speech type – read, spontaneous, etc.); ● a set of labels, that state the words or phonemes that are uttered in each speech segment; ● a set of acoustic parameters, which "synthetically" represent the speech signal (Mel Frequency Cepstrum Coefficients – MFCC, Perceptual Linear Prediction Coefficients – PLP, etc.); a set of acoustic parameters is associated to each (suitably chosen) speech signal window.

Some essential problems have been determined, when speech databases are to be constructed:

a) *speech signal segmentation* – for a convenient and reliable treatment of the speech files by the human manipulator, speech signal lengths of 60 to 180 seconds are preferred;

b) *speech signal labeling* – this can be done at a word level (time-consuming process, accomplished manually), or a phoneme or triphone level (semiautomatic process, via bootstrapping, starting from an initial manual labeling [3]); this last process raises several reliability issues, because it is based on statistical algorithms (e.g. forced Viterbi alignment of hidden Markov models – HMMs [4]);

c) *speech signal parameterization* – here, several criteria have to be observed, such as maximizing the inter-phoneme variance while minimizing the intra-phoneme variance; these criteria can be observed either by the human expert (which is rather unreliable), or automatically (which is not robust when database extensions are envisaged).

2.2. Particular Features

The spontaneous speech database used for training the speech decoder has the following features: *source of the signal*: mainly acquisition of TV streaming in Romanian, broadcasted on the Internet; language: spoken Romanian; *duration of the recordings*: around 4 hours of speech; *number of speakers*: 12 (8 females, 4 men); *number of sessions per speaker*: from 3 to 20; *interval between sessions*: from one day to a couple of weeks; *number of word types*: around 3000 (medium vocabulary); *speech register*: spontaneous speech; *sampling frequency of the signal*: 8 kHz.

The following problems had to be addressed during acquisition:

– Segmentation of the speech signal files: very often, TV shows are relatively long (transmissions could last for tens of minutes or even hours); in order to improve the efficiency, manual segmentation of the audio files has been performed; this resulted in audio files with considerably shorter lengths: from 60 to 180 seconds.

– The sampling frequency of the voice signal acquired from the Internet varied from 8 to 44 kHz; for homogeneity reasons, the suppression of the speech samples at frequencies below 16 kHz has been performed, also operating a low-pass filtering at 16kHz for speech signals sampled at frequencies higher than 16 kHz.

– Choosing the optimum parameter configuration for the speech signal, so that a minimal intra-phoneme variation is ensured, while at the same time maintaining a maximal intra-phoneme variation of acoustic parameters.

– Voice signal labeling at the triphone level involves the following stages:

(i) Manual labeling at word level for all speech signal files.

(ii) Automatic separation of every word in triphones, by using the phonetic dictionary (manually built); the issue with this approach is that all triphones that made up a word are assumed to have the same duration; this assumption is obviously false and should be removed in the next step (iii).

(iii) Iterative forced Viterbi alignment of the HMMs (that had already been defined at a triphone level) of the acoustic feature vectors that correspond to the speech signal labeled at triphone level; every iteration has the purpose of maximizing the probability that the models represent the triphones considered.

2.3 Triphone Occurrence Statistics

The database described in Section 2.2 is meant to be used in an application for spontaneous speech recognition purposes. The results that characterize the database are represented through statistics that refer to the number of occurrences for both words and triphones. Given that between words “sp” has been used for differentiation purposes and that “sp” itself is not a phoneme, then the beginning and the end of the words may be actually considered as diphones.

For example, the Romanian word „c a s @” has the following phonetic transcription: “c-a”, “c-a+s”, “a-s+@”, “s+@”. The reason why it has been decided to use triphones is that they allow for a contextual analysis; the triphones are entities that maintain and transfer information about what is before and after the phoneme of interest.

In consequence, statistics about diphones (that occur at word boundaries; see the discussion above) and triphones for the Romanian database are represented in a graphic form. In Fig. 2 we represent the triphones with the highest number of occurrences in the Romanian database. They have occurred between 983 and 2305 times. The most significant are the (word boundary) triphones “d+e” and “d-e”.

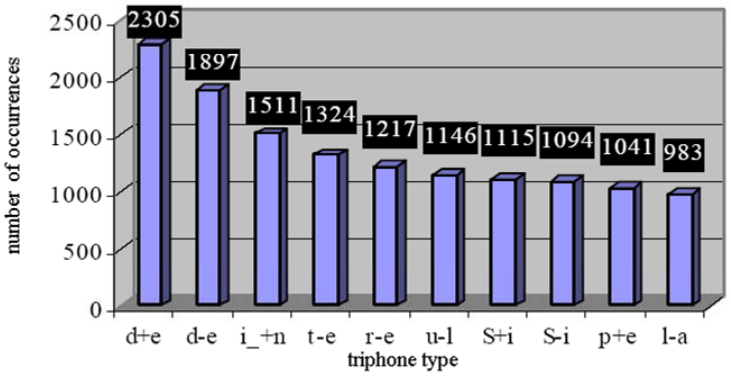

Fig. 2 – The most frequent triphones.

In Fig. 3 we represent the next ten triphones with the highest number of occurrences in the database. These triphones have between 982 and 730 occurrences in the database. Triphones with the highest number of occurrences are “l+a” and “l-e”.

The vocabulary for the Romanian database for spontaneous speech recognition has 5095 triphones. Based on them the following characteristics can be emphasized: 1% of the total number of triphones have

between 500 and 1000 occurrences; 7% of the total number of triphones have between 100 and 500 occurrences; 8% of the total number of triphones have between 50 and 100 occurrences; 8% of the total number of triphones have between 30 and 40 occurrences; 19% of the total number of triphones have between 10 and 30 occurrences; 16% of the total number of triphones have between 5 and 10 occurrences and the last 41% of the total number of triphones have between 1 and 5 occurrences.

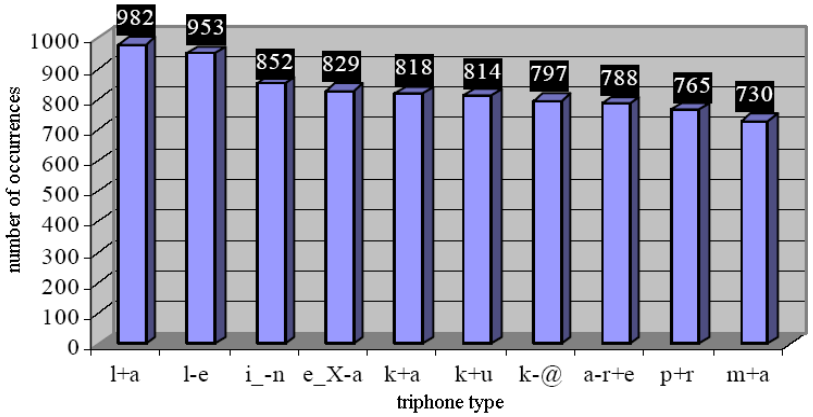

Fig. 3 – The next frequent triphones.

3. SPEECH RECOGNITION SYSTEM ARCHITECTURE

3.1. Outlook

Mainly for practical reasons, we have chosen to use the "Hidden Markov Modeling Toolkit" (HTK) for instantiating a general speech recognition front-end architecture in a spoken dialogue system. We used spontaneous speech, in the Romanian language, in closed laboratory room, with constant signal-to-noise ratio (> 25 dB), the microphone was headset-type, with constant distance between speaker's vocal cavity and the microphone, the vocabulary size was between 3000 and 10000 domain-independent words, speaker-independent, but tunable to the voice of a particular speaker, with less than 60 seconds of speech signal, acquired from this particular speaker and the voice detection was "on-line" speech decoding, with automatic "Voice activity detection" (VAD).

In this context, we have designed a sequential architecture, based on HTK, with the following specific features: ● MFCC coefficients were used for speech signal parameterization (for both training and decoding regimes); ● acoustic modeling was accomplished at the triphone-level; for each triphone an HMM is trained; ● HMM has, for each Romanian triphone 5 states (including one initial state and one final state, both non-emitting, and three emitting intermediary states), left-right topology (where transitions towards remote states and "backwards" transitions are not allowed) and a continuous output distribution with weighted linear combination of Gaussian mixtures for each emitting state.

Thus, the system is designed to work in two regimes, training and testing (recognition).

a) ***Training***: In this regime, the acoustic models (HMMs) are created, for each triphone considered in the instantiation of the architecture. The following steps are required:

1. speech signal parameterization;
2. manual word-level labeling of the speech signals in the database;
3. automatic extraction, from the labeling performed at step 2, of the set of words that represent the vocabulary of the system;
4. manual development of the phonetic dictionary, which contains the phonetic transcriptions of the words identified at step 3;
5. automatic derivation, according to the entries in the phonetic dictionary, of the set of triphones for the particular natural language considered (e.g. Romanian);

6. manual definition of prototype HMMs for the triphones identified at step 5; the prototype HMMs are identical for all the triphones and contain information regarding the number of states and the topology;

7. triphone-level labeling of the speech signal segments in the database; this process is automatic and is performed iteratively.

b) ***Recognition***: In this regime the resources created during training are used (the HMMs and the phonetic dictionary are used for recognizing the utterances). The process involves the following steps:

1. manual definition of a set of test sentences;

2. production, acquisition and parameterization of utterances of these sentences;

3. triphone-level decoding, which consists in determining the set of HMMs that have generated with maximal probability the sequence of acoustic parameters obtained at step 2;

4. converting the set of triphones obtained at step 3, in a word sequence, according to the phonetic dictionary;

5. performance evaluation: this process consists in iterating stages 1 to 4 on a sufficiently large set of test utterances, and then comparing, by dynamic alignment (the Levenshtein distance [3]), the reference sentences with the recognition output; thus, we compute both an utterance error rate and a word error rate.

3.2. Preliminary Results

In order to validate the speech recognition architecture, several experiments were made with a reduced set of words, like building a recognition system for a language with a very small vocabulary. The speech sequences have been chosen from the same speaker. The seven words in the Romanian language that were chosen for this experiment are listed in Table 1.

Table 1

Limited set of (frequent) Romanian words chosen in the recognition experiments

Words	Occurrences in the first phase	Occurrences in the second phase
în	27	34
şi	22	25
de	36	51
la	24	32
cu	12	21
din	14	22
un	10	20

It is important to mention that all the steps for recognition at triphone level described in Section 3.1 have been respected. It is also important to mention that these occurrences are isolated words, which means that short pauses denoted as “sp” are not trained. However, embedded training has been applied. Hence, no sub-word labeling but word-level labeling has been used instead.

The training of the HMMs was performed in two different phases. The second phase implies more observation data, about 40% more than the first phase. The goal here was to notice how a greater number of occurrences yield a higher word recognition rate. The recognition rate used as metric in this experiment is calculated as the number of words correctly recognized over the total amount of words present in the test sequences. In these conditions the recognition rate for the first phase is 52%, while the recognition rate for the second phase is 70%. It has been observed that the order in which the triphones are trained slightly affects the results, so these figures represent the best match. By exploiting the large database that we are building, we expect to obtain a recognition rate above 90%.

4. CONCLUSIONS AND PROSPECTS

A relatively medium-size database for the Romanian language has been created. In order to improve the analytic possibilities, our short-term goal is to increase the size of the database. When creating this database several 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 easier processing purposes. The voice signal labeling at a triphone level assumes many stages, hence a manual labeling at an utterance level has been performed in the first place, followed by the automatic separation of every word in triphones by using the phonetic dictionary. This was used for bootstrapping the training process, based on (forced) Viterbi alignment between the HMMs and this triphone-level transcription, followed, in a loop, by Baum-Welch HMM parameter estimation.

The great number of triphones that have been obtained is specific to spontaneous speech. There are cases when some triphone combinations have a greater number of occurrences, compared to others. For instance, triphones "d+e", "d-e" and "i_+n" have more than 1500 occurrences. This has a logical explanation consisting in the fact that the most frequent triphones are in fact prefixes and suffixes. In conversational speech, when stammering, the speaker tends to double the beginning or ending of a word. In a psychological way [14], it can be easily observed that a person who is speaking in front of an audience, or in front of an important person, or a nervous person, or a person in any other different situation that involves a higher level of stress or even attention, tends to hesitate. In this situation, in order to regain the fluency of the speech flow and try to transparently avoid the conversational blocking and embarrassing, the speaker tends to double some of the verbal constructions, in an unconscious manner. This way, the speaker tends to double, triple or multiply a whole word or phrase in order to gain time, precious time that offers the speaker the necessary interval for constructing his utterance.

The same psychological perspective shows that when the tension rises in a monologue or a conversation, this tension could influence a normal healthy person and make this person adopt a verbally erratic behavior. The physiology of the person changes, the face and body contract and tend to sweat, to become red or by contrary to become pale. The anxiety could cause different behaviors, but the most predominant one is the stuttering. It seems that the pattern found in all the behaviors of the persons that tend to speak freely is this repeated pronunciation of the first syllables of the words. Hence, spontaneous speech could be a way of observing common patterns on the subjects manifesting speaking anxiety.

The histograms determined for the word occurrences in the database show that: "de" (1659), "la" (891), "in" (803), "a" (667), "şi" (656) are the most frequent. The triphones with lower frequency counts will not be ignored. 2903 triphones have less than 10 occurrences and they represent 57% of the total amount of triphones. When enlarging the size of the database it is expected that the triphones with lower frequencies 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, after augmenting the database.

The corpus and the spontaneous speech recognition results will have applicability in recognition tools. 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.

For future prospects, the first problem is considered to be the database augmentation: nowadays, hundreds of hours are used at the acoustic level, in state-of-the-art systems. Hence, this is why we are currently working on increasing the size of the database. As soon as it becomes available the new statistics will be computed and observations extracted. Secondly, the tradeoff between recognition accuracy and computing (decoding) time has to be carefully considered: in dialogue applications, the former can be sacrificed for the latter, to a greater extent than in the case of related applications in broadcast news transcription. Hence, the computing time can be improved (reduced) via parallelization strategies, by extending previous work [11, 12], or by adopting particular search heuristics while decoding the signal [7]. Lastly, it might be interesting to improve the robustness at the acoustic level, by computing confidence measures on the phoneme or triphone-level HMM probability estimates (current work in progress).

ACKNOWLEDGEMENT

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.

REFERENCES

1. CAELEN, J., XUEREB, A., *Interaction et pragmatique*, Hermès, Paris, 2007.
2. DRĂGĂNESCU, M., ŞTEFAN, Gh., BURILEANU, C., *Electronica funcțională*, Vol. 1, Edit. Tehnică, Bucharest, 1991.

3. EVERMAN, G., et al., *The HTK Book*, Version 3.0, Cambridge University Engineering Department, 2005.
4. HUANG, X., ACERO, A., WUEN-HON, H., *Spoken Language Processing – A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.
5. LE, V. B., *Reconnaissance automatique de la parole pour des langues peu dotées*, Thèse de Doctorat, Université Joseph Fourier, Grenoble, 2006.
6. LEVINSON, S. E., *Mathematical Models for Speech Technology*, John Wiley & Sons, 2005.
7. LINARES, G., NOCERA, P., MASSONIE, D., MATROUF, D., *The LIA Speech Recognition System: From 10xRT to 1xRT, Text, Speech and Dialogue*, Lecture Notes in Computer Science Series, **4629**, pp. 302–308, Springer, Heidelberg, 2007.
8. MCTEAR, M. F., *Spoken Dialogue Technology: Enabling the Conversational User Interface*, ACM Computing Surveys, **34**, *1*, pp. 90–169, 2002.
9. MARTIN, J., JURAFSKY, D., *Spoken Language Processing* (Third Edition), Prentice Hall, 2007.
10. NGUYEN, H., *Dialogue homme–machine: Modélisation de multisession*, Thèse de Doctorat, Université Joseph Fourier, Grenoble, 2005.
11. POPESCU, V., BURILEANU, C., *Parallel Implementation of Acoustic Training Procedures for Continuous Speech Recognition*, in C. Burileanu (Ed.), *Trends in Speech Technology*, pp. 119–136, Romanian Academy Publishing House, Bucharest, 2005.
12. POPESCU, V., BURILEANU, C., RAFAILA, M., CALIMANESCU, R., *Parallel Training Algorithms for Continuous Speech Recognition, Implemented in a Message Passing Framework*, Proc. EUSIPCO, 2006 (CD-ROM).
13. ROBINSON, T., FRANSEN, J., PYE, D., FOOTE, J., RENALS, S., *WSJCAMO: a British English speech corpus for large vocabulary continuous speech recognition*, Proc. ICASSP, Vol. 1, pp. 81–84, 1995.
14. RUSSELL, S., NORVIG, P., *Artificial Intelligence – A Modern Approach* (Second Edition), Prentice Hall, 2003.
15. SCHULTZ, T., KIRCHHOFF, *Multilingual Speech Processing*, Academic Press, 2006.

Received August 24, 2009