Romanian Spoken Language Resources and Annotation for Speaker Independent Spontaneous Speech Recognition

Corneliu BURILEANU, Andi BUZO, Cristina Sorina PETRE, Diana GHELMEZ-HANES, Horia CUCU
Faculty of Electronics, Telecommunications and Information Technology
University “Politehnica” of Bucharest
Bucharest, Romania
e-mail: cburileanu@messnet.pub.ro, andione1@yahoo.com, cristina.sorina@yahoo.com,
dyanna1981@yahoo.com, horiacucu@gmail.com

Abstract — **This paper presents studies and early results with the scope to build a robust spontaneous speech recognition system in Romanian language. We have tried to give solutions to several issues that have arisen like building a large and accurate database within a reasonable time. A short description of the database is given and some statistics are collected in order to show its evolution in several stages of the project. Embedded training technique has been used for training triphones. As a consequence, the alignment problem has been studied and a solution is proposed for it. The final purpose of these attempts is to obtain substantial results in speech recognition for Romanian language that can be used as baseline for further results.**

Keywords - Spontaneous Speech Recognition, Romanian Triphonest Corpus, Embedded training.

I. Introduction

The resources in Romanian language needed by researchers in order to perform studies and obtain immediate results are very poor. There are no spoken databases developed for spontaneous Romanian language. There are no freely accessible databases for continuous Romanian language. The hardest and most time consuming activity in this domain is the creation of a new corpus, with all the related tasks like annotation or management, for instance.

The intention of this paper is to present the research work that has been done in the domain of Romanian language resource acquisition. A completely new Romanian language database has been created and it’s intended to be used for speaker independent continuous speech recognition. The database contains large spontaneous speech sections in order to bring a new approach on the research performed in Romania in the field of continuous speech recognition. Romanian language experts have directly contributed to the database creation.

Section II briefly describes the main characteristics of the database and the way it was built. Section III presents some practical issues that were faced during training, whereas the first results obtained are shown in Section IV.

II. Resources and Annotation

At the very first stage of database construction, the intentions were to keep the scalability property for the corpus in order to add as many words as the further investigations and research will need.

A series of problems and specific elements have been identified related to construction of the corpus. 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 (e.g., 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 several different methods [6]:

- 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, ...);
- 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 should have the following components [1]:

1. a set of speech signal samples;
2. 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.);
3. a set of labels, that state the words or phonemes that are uttered in each speech segment;
4. a set of acoustic parameters (Mel Frequency Cepstrum Coefficients – MFCC, Perceptual Linear Prediction Coefficients – PLP, etc.), which “synthetically” represent the speech signal and are derived from the speech signal files.

Some essential problems were met, when Romanian speech database was built[4]:

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 [2]); this last process raises several reliability issues, because it is based on statistical algorithms (e.g., forced Viterbi alignment of hidden Markov models – HMMs [3]);

c) speech signal parameterization – here, several criteria have to be observed, such as maximizing the inter-phoneme variance while minimizing the intra-phoneme variance (that is, the acoustic parameters have to exhibit a maximal variation from an acoustic realization of a phoneme to an acoustic realization of another phoneme, and minimal variation between several acoustic realizations of the same phoneme); 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).

The Romanian language database that was built, involved several enlargement stages.

First and second stage led to a medium size database with five hours of spontaneous spoken Romanian language, 44343 words with multiple occurrences and about 10160 words with single occurrences. Both stages involved mainly acquisition of TV streaming in Romanian, broadcasted on the Internet, gathering radio news, stories, TV shows, medical discussions, financial discussions, weather forecasts and other kind of information.

At a later time the database was enlarged, the next stages involved the acquisition of audio Romanian language stories, each story having a different narrator. Stages 3 brought to the original phonetic dictionary 19449 new words.

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;

TABLE I. Database evolution

Stage 1- database creation	Hours of aquisition		4	
	Words with multiple occurences		37.604	
	Words with single occurence		8,068 (21.5%)	
	Number of speakers		12	
	Male	Female	4	8
Stage 2 – database completion	Hours of aquisition		1	
	Words with multiple occurences		6.739	
	Words with single occurences		2,092 (31%)	
	Number of new speakers		5	
	Male	Female	2	3
Stage 3 – database completion	Hours of aquisition		4	
	Words with multiple occurences		40.0016	
	Words with single occurences		7770	
	Number of new speakers		7	
	Male	Female	4	3

- 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.

The next stage that is going to be performed for the database enlargement purpose is to use an exhaustive phonetic dictionary composed by Romanian linguistic experts. This dictionary will help in completing the corpus with 10000 representative words that are going to be pronounced by 5 to 8 different speakers. Each speaker is going to pronounce all the 10000 different words in order to obtain a robust speaker independent speech recognition.

At a later time after this stage, an automatic way to combine different words together in order to create sentences and phrases will be taken into consideration. [5]

III. Continuous speech recognition – Solutions to practical issues

For continuous speech recognition we have used the Hidden Markov Models (HMM) and the HTK tools for creating, estimating and testing them. We have chosen to model triphones because it adds more constraints to recognition.[7]

A. *Large database problem*

Training of the data for continuous speech recognition must be done by considering the specific issues that this task arises. Compared to isolated word recognition, in continuous speech words are spoken in too many different contexts and stressed in too many ways so a larger database is needed. If speaker independent recognition is targeted then even a larger database is needed. In this situation an annotation at phoneme level or even at word level does not scale. To overcome this issue annotation is done for several propositions and the label file contains only the order of the phonemes. Such files are trained in the *embedded* mode. Embedded training estimates the HMM parameters of all prototypes in the same time; it chains all the HMMs of phonemes/triphones of a label files by forming a long chain of HMM, which is trained using a Baum-Welch alignment. [3]

B. *Alignment problem*

Alignment is the main problem in the training process because of two issues, which makes it a fundamental factor in the success of recognition. First, Baum-Welch algorithm stops once a local maximum in likelihood probability. Second, HMM encounters difficulties to represent different durations of the same phoneme/triphone; by the way it is defined the probability to remain in the same frame for several frames decreases exponentially. To overcome this issue we have decided to model inter-word silence and because of their various duration, we have proposed 5 models of silence each having 50ms, 200ms, 500ms, 1s, 2s respectively. This way we expect to “mark” the start and the end of words. Another measure we take to help alignment is initializing HMMs of triphones with estimates obtained from phonemes trained with data labeled at a phoneme level. Few utterances are labeled at a phoneme label in order to obtain well trained HMMs. All triphones will *inherit* model parameters from the respective central phoneme’s HMM.

C. *Sensibility to human errors during annotation*

Embedded training is very sensible to an annotation error. If the error occurs at the beginning of label file it will propagate to the whole file. Moreover this will affect all models as erroneous data will be used in calculating means and variations of HMM outputs. To overcome this issue we intend to collect a database of isolated words. So we will firstly train some robust HMM of isolated words and based on them will train the continuous speech database. Then, the continuous speech database will be tested for recognition. Utterances with high word error rate will be checked for eventual errors that they may contain.

D. *Small number of occurrences*

A significant drawback in using triphones is the small numbers of occurrences for some of them and this makes them hard to train. To overcome this issue we will use the tied-states technique. Triphones, which have the same central phonemes have the same HMM parameters for central states. These states may be tied during training so they will have the same parameters at the end but benefitting from a higher number of occurrences.

E. *Context in continuous speech recognition*

In continuous speech phonemes and words are into a certain context. In a proposition not all the combinations of word sequences are possible. This hint gives us the possibility to consider some sequences to be more probable then the others. Triphones already add context to the phonemes. As for words, the N-gram technique can be used. It consists of analyzing large size of text corpora and calculating probabilities that N given words be consecutive in a sentence. N-grams will be applied later in our project and it is subject to further study as it arises several other issues that need solutions. [2]

IV. Early Results

So far, we have analyzed the evolution of database and have obtained a good recognition rate for isolated phonemes.

A. *Proposition-level annotated database evolution*

The proposition-level annotated database is a key factor in the success of the embedded training HMM system and its size, measured in number of hours of speech or number of basic elements (phonemes/triphones), is the most important aspect. Table II summarizes the key properties of our databases.

The first conclusion that can be extracted is that at this point the database extension introduced 25% more triphones (basic elements for which we had not got any occurrences in the first database). That is a first argument for another database extension. The extensions should continue until a new database extension will not introduce a significant number of new triphones.

A second statistic we’ve employed regards the “most common triphones” top. The idea behind this was that if a new database extension introduces new triphones, but this “most common triphones” top does not significantly change then the extension process could stop and the database could be considered representative for the Romanian language. The top is based on the occurrence ratio of the triphones (the number of occurrences for that particular triphone over the total number of triphones); a higher ratio means a more common triphone. We have given to the ratio the name of *weight*.

After our first database extension this top (first 100 most common triphones) contained 26 different triphones compared to the one created for the first database. For each of these triphones we have calculated the *weight*. This *weight* varied from 0.5% to 71.5% with an average of 33% and a standard deviation of 44%. This clearly means that the top has significantly changed after the extension, so the conclusion is: another database extension is absolutely necessary.

TABLE II. Database evolution in terms of basic recognition units (triphones)

	# Hours of speech	**# Triphones**	**# Distinct triphones**
First database	4	183007	5191
Extended database	9	352460	6525

B. Phoneme-level recognition rate

In order to properly initialize the HMM system a phoneme-level annotation database was created. The goal set was to have at least 20 occurrences of each of the 31 phonemes that we are currently recognizing. The most usual phonemes were of course annotated a lot more than 20 times.

Table III presents a subset of the phonemes, the number of training files and the recognition rate given the initialized and trained HMM system.

TABLE III. Phoneme-level recognition rate

Phoneme	**# Training files**	**Recognition rate**
k	26	96.1%
o1	22	95.4%
i3	22	95.4%
a	561	79%
n	48	75%
l	38	73.7%
i2	26	73.1%
@	30	70%

The recognition rate varies depending on the number of occurrences used for training, on phoneme's likeness and on the speaker. For example, the recognition rate for the "a" phoneme is relatively small, but stable even in the case of other speaker's files, because the HMM parameters have converged to some "general speaker" values. On the other hand, the huge recognition rate for the "k" phoneme varies a lot depending on the speaker, because the HMM parameters have not been trained enough. Despite the poor training for these phonemes, they have been used in the initialization step that aims to built suitable HMMs for the next steps, which are:

- Phoneme-level embedded training using the proposition-level annotated database. After this step we will be comparing whole words recognition rates.
- Triphone-level embedded training using the proposition-level annotated database. After this step we will also be comparing whole words recognition rates, but we expect the rates to be higher, as the training/recognition method takes into account more details.

V. Conclusion and Future Work

Spontaneous speech recognition requires a larger database than isolated word/continuous speech recognition.

The first database extension (from 4 to 9 hours of speech) introduced 25% more triphones, while the "most common triphones" top significantly changed. The extension process has to continue in parallel with the system training process.

Due to alignment issues a new step, consisting in isolated phonemes training, was introduced before the embedded phonemes training. This second step will be followed by an embedded triphones training step.

In the immediate future we intend to correct the existing database in order to assure an error-free one and building a language model using n-grams**.**

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] T. Robinson, J. Fransen, D. Pye, J. Foote, S. Renals, Wsjcamo, "A British English speech corpus for large vocabulary continuous speech recognition", Proc. ICASSP, Vol. 1, pp. 81-84, 1995.

[2] G. Everman, et al., "The HTK Book", Version 3.0, Cambridge University Engineering Department, 2005.

[3] D. Munteanu, "Contribuţii la realizarea sistemelor de recunoaştere a vorbirii continue pentru limba română", Teză de doctorat, p.34-35, p. 71-73, 2006.

[4] C. Burileanu, V. Popescu, A. Buzo, C. Petrea, D. Ghelmez-Hanes, "Spontaneous Speech Recognition for Romanian in Spoken Dialogue Systems", Romanian Academy Publishing House, Bucharest, 2009, in press.

[5] C. Petrea, D. Ghelmez-Hanes, V. Popescu, A. Buzo, C. Burileanu, "Spontaneous Speech Database for the Romanian Language with Medical Applicability", BIOSTEC, Porto, 2009.

[6] C. Petrea, D. Ghelmez-Hanes, V. Popescu, A. Buzo, C. Burileanu, "Spontaneus Speech Database for Romanian Language", ECIT, Iasi, 2008.

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