BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi" din Iaşi
Tomul LVII (LXI), Fasc. 6, 2011
Secția
ELECTROTEHNICĂ. ENERGETICĂ. ELECTRONICĂ

DEVELOPMENT OF A PRONUNCIATION DICTIONARY FOR THE ROMANIAN LANGUAGE

BY

J. DOMOKOŞ[*] and G. TODEREAN

Technical University of Cluj-Napoca

Received, May 31, 2011
Accepted for publication: July 16, 2011

Abstract. The scope of this paper is to present a novel automated grapheme-to-phoneme conversion system for Romanian, based on artificial neural networks, together with some transcription results based on a 1.004 words hand built pronunciation dictionary. Using this letter-to-sound transcription system we have also built a pronunciation dictionary for Romanian by transcribing the near 140.000 base form entries from the DEXOnline dictionary.

Key words: letter-to-sound conversion; grapheme-to-phoneme transcription; phonetic transcription; pronunciation dictionary.

1. Introduction

The recent achievements in spoken language technology make possible the realization of man–machine interfaces using spoken human–computer dialogue systems, one of the most natural and convenient way of interaction between users and software applications (Burileanu *et al*., 2010). This type of systems includes almost all major speech technology tasks from speech recognition and understanding to answer generation and speech synthesis.

Grapheme-to-phoneme conversion systems are very useful for speech recognition and speech production applications (Bisani & Ney, 2008; Davel &

[*] Corresponding author: *e-mail*: domi@ms.sapientia.ro

Barnard, 2008; Divay & Vitale, 1997; Damper *et al*., 1998) because they are at the base of automated segmentation of speech at phonetic level (Gómez & Castro, 2002), and predicting the pronunciation of a written word is an important sub-task of most speech production systems (Davel & Barnard, 2008).

This work is related to NaviRo (http://users.utcluj.ro/~jdomokos/) research project with the main objective to create a Romanian language voice driven navigation extension to the most popular web browsers. A subtask of this project is the development of a segmented and annotated Romanian language spontaneous speech database. At present we have recorded spontaneous conversations conducted by a mediator with 40 Romanian language speakers. The conversation topic is the use of personal computers for Internet browsing, revealing the frequently used commands for browsing. The mediator drives the conversation based on a set of previously formulated questions to ensure that it will not digress from the topic. The mediator was randomly chosen from the speaker list for each conversation.

Because the mediator knows the questions before starting the conversation, the recorded speech will be fully spontaneous just in the case of the interviewed person. The duration of each conversation is approximately 5…6 min. Therefore the recorded speech database contains 40 conversations and has a total length of over 200 min. The 40 speakers chosen were from Transylvania, mainly from Cluj-Napoca, Târgu Mureş and surroundings and are adult persons with ages between 20 and 40 years. There were selected 20 male speakers and 20 female speakers. The main criterion for choosing the speakers was the ability to work with computers. The recording was performed in a quiet office condition. The recording was made using 16 kHz sample rate and the recorded speech files are stored in 16 bit coded, stereo wav format, because this is the standard speech format for building acoustic models for wide-band microphone speech recognition.

Before starting automated speech segmentation we need an orthographic transcription of the recorded speech files. The orthographic transcription must be done manually by repeated listening of the conversations and noting what we are hearing according to some transcription rules. These orthographic transcriptions can be further phonetically transcribed. The process of transcribing orthographic texts to their spoken form is called *grapheme-to-phoneme conversion* or *letter-to-sound transcription*.

For grapheme-to-phoneme conversion there are several approaches in the literature (Bisani & Ney, 2008; Davel & Barnard, 2008; Divay & Vitale, 1997; Damper *et al*., 1998; Gómez & Castro, 2002; Braga & Coelho, 2006):

a) systems based on pronunciation dictionary;

b) phonetic rule-based transcription systems;

c) systems based on machine learning (using decision trees or artificial neural networks);

d) statistical systems based on hidden Markov models;

e) hybrid systems trying to use a combination of the above mentioned systems (like Default & Refine rule-based learning algorithm (Davel & Barnard, 2008).

The simplest systems use pronunciation dictionaries. Although these systems work very well for the words found in the dictionary, they have certain disadvantages (Bisani & Ney, 2008): dictionaries must be built by hand, which is a very slow and expensive process and the size of these dictionaries do not allow their use in the case of embedded systems or mobile devices. Moreover these systems are limited because they cannot handle new words that are not included in the dictionary. Development of a dictionary that contains all the words from a language together with their inflected forms is practically impossible because every day new words appear usually borrowed from other languages.

Linguistic transcription rule-based systems are very efficient but also have some disadvantages: number of rules is relatively high: about 1500 for English (Bisani & Ney, 2008), over 600 for French (Bisani & Ney, 2008), 112 for Romanian (Toma & Munteanu, 2009); establishing rules requires strong knowledge in the field of linguistics; the connections between the rules are usually very complex, and should therefore be analysed how to apply them; natural language often does not follow the rules and those exceptions must be treated (the most often used method is the use of an exception transcription dictionary); in some languages (as it is in the case of the Romanian language) transcription rules may present ambiguities (Bisani & Ney, 2008).

The third approach is based on training using hand built transcription dictionaries covering the most common words from a language. The most widely used systems are based on decision trees or neural networks (Bisani & Ney, 2008; Damper *et al*., 1998).

In the most important papers for Romanian language, grapheme-to-phoneme transcription is handled using rule-based systems (Toma & Munteanu, 2009), neural network based machine learning systems (Burileanu, 2002; Burileanu *et al*., 1999) and hybrid systems that use transcription rules and machine learning to solve the ambiguities of rules (Ordean *et al*., 2009; Jitcă *et al*., 2002, 2003).

To our best knowledge there is no pronunciation dictionary for Romanian language available in electronic form, as it is for example the CMU Pronouncing Dictionary (CMU, 2008) for English, that can be used for speech recognition and text-to-speech systems. The documentation studied (Burileanu, 2002; Burileanu *et al*., 1999; Toma & Munteanu, 2009; Ordean *et al*., 2009; Jitcă *et al*., 2002, 2003) shows that there exist such automatic grapheme-to-phoneme transcription systems for Romanian, and also some small hand built phonetically transcribed databases are reported which could be used for training such systems, but these applications and resources are not freely available.

2. Phonetic Transcription System Architecture

Our grapheme-to-phoneme conversion system is based on the system described by Sejnowski and Rosenberg (1987) and adapted for Romanian by Burileanu, Sima and Neagu (2002, 1999). Therefore the system is based on a parallel structure having 30 neural networks with 25 common inputs, each of them designed to detect the presence of an articulatory feature from the 30 features used to encode the Romanian language phonemes presented in Table 2, and to point out the presence or the absence of that feature at the network output.

Table 1
The Used Grapheme Set with their Binary Codification

Grapheme	Binary code	Grapheme	Binary code	Grapheme	Binary code
A	00001	D	01011	r	10101
Ă	00010	F	01100	s	10110
E	00011	G	01101	ş	10111
I	00100	gh (&)	01110	t	11000
î|â	00101	H	01111	ţ	11001
O	00110	J	10000	v	11010
U	00111	L	10001	z	11011
B	01000	M	10010	#	11100
C	01001	N	10011		
ch (%)	01010	P	10100		

Table 2
Error Percentage Values for the Used Articulatory Features

No	Articulatory feature	Error percentage %	No	Articulatory feature	Error percentage %
1	phonetic-zero unit	1.655	16	palatal	0.163
2	open	1.000	17	dental	1.964
3	medium	1.928	18	labiodental	0.036
4	closed	2.965	19	laryngeal	0.163
5	occlusive	1.400	20	lateral	1.655
6	semi-occlusive	0.218	21	type 2	2.856
7	fricative	0.909	22	type 3	1.255
8	liquid	0.527	23	type 4	0.127
9	vibrant	0.509	24	type 6	1.382
10	central	1.619	25	voiced	1.964
11	front	2.619	26	unvoiced	1.127
12	back	1.073	27	type 5	1.728
13	bilabial	0.600	28	oral	1.146
14	velar	0.236	29	nasal	0.309
15	prepalatal	0.145	30	type 1	1.891

The words that are intended to be transcribed are read from a text file edited with one word by line and are presented at the input of each neural network. The beginning and the end of each word is appended with two white space characters (#) and after this, the input words are split into five character long sequences and binary coded using the five bit codes presented in Table 1. Always the central grapheme from the sequence of five graphemes is analyzed; the other graphemes represent context information (two graphemes for left context and two graphemes for right context). Therefore at the input of each neural network we have $5 \times 5 = 25$ bit information; hence we can deduce the number of network inputs. The words at the input of the system are shifted character by character until all component graphemes are presented to the input in the same way as described by Burileanu *et al*. (1999).

The grapheme set contains 29 characters and is presented in Table 1 together with the used codification system. For the graphemes "î" and "â" we have used the same code because of their similar pronunciation. For the complex graphemes "ch" and "gh" we have choose to use 2 ASCII characters: % and & for having one character representation for all the graphemes. The symbol # marks white spaces between words.

Input words preprocessing, grapheme coding and phoneme coding according to a proceeding proposed by Burileanu *et al*. (1999) have been implemented through a Java application using regular expressions which also generates the training and testing sets for the neural networks. This application can also replace characters that are not part of the set of 29 graphemes used for input with their correspondences, *e.g*. "x" with "cs" or "gz", "w" with "v", "y" with "i" or "k" by "c".

The used phonemes set include a number of 33 phonemes plus the zero phonetic unit – 0 and the space between words marked with #, and along with their coding comes from (Burileanu, 2002). This set is supplemented with the short "i" phoneme /i_0/ coded as (1, 4, 11, 21, 27 – phonetic zero unit, closed, front, type 2, type 5) considered as a combination of phonemes /i/ (4, 11, 21), /j/ (4, 11, 27) and the phonetic zero unit /0/ (1). Therefore the 30 bits coding of the articulatory features for /i_0/ is (1, 4, 11, 21, 27) = = 100100000010000000001000001000, where the bit position represent the articulatory feature number. Finally our system has 36 output possibilities taking into account also the phonetic 0 unit and the space between words.

The 30 articulatory features are given by the outputs of the 30 neural networks; each network was trained to indicate the presence or the absence of one feature (value 0 if the feature was not detected, and value 1 for reporting the presence of that feature). The articulatory features were selected using previous proposals from linguistic specialists (Beldescu, 1984; DOOM, 2005) and extended with those proposed by Burileanu *et al*. (1999).

The used neural networks are totally connected multilayer perceptron type with two hidden layers (Fig. 1 *a*). The internal structure and the number of neurons in the hidden layers were determined based on some experimental

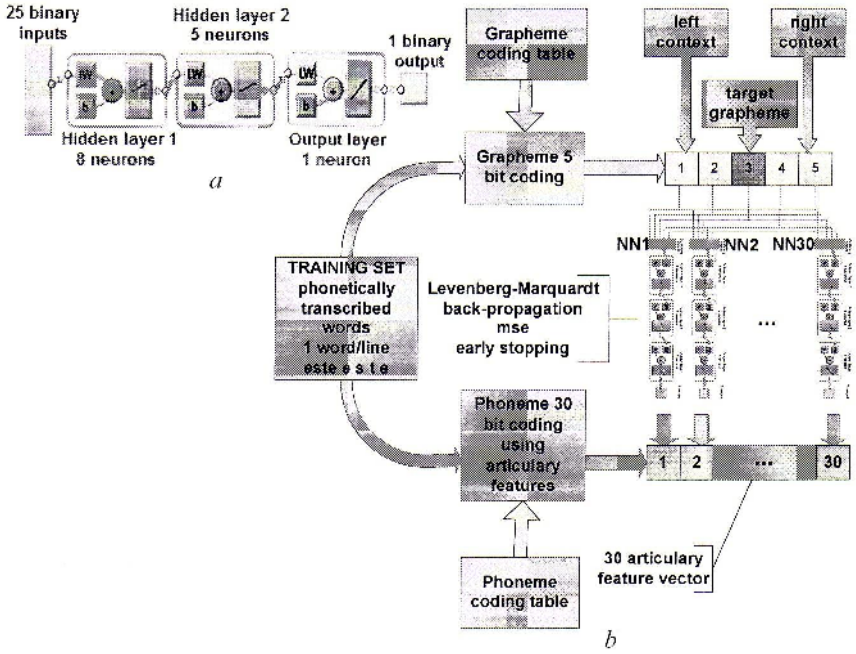

Fig. 1 – *a* – Feature detector neural network; *b* – system architecture when used for training.

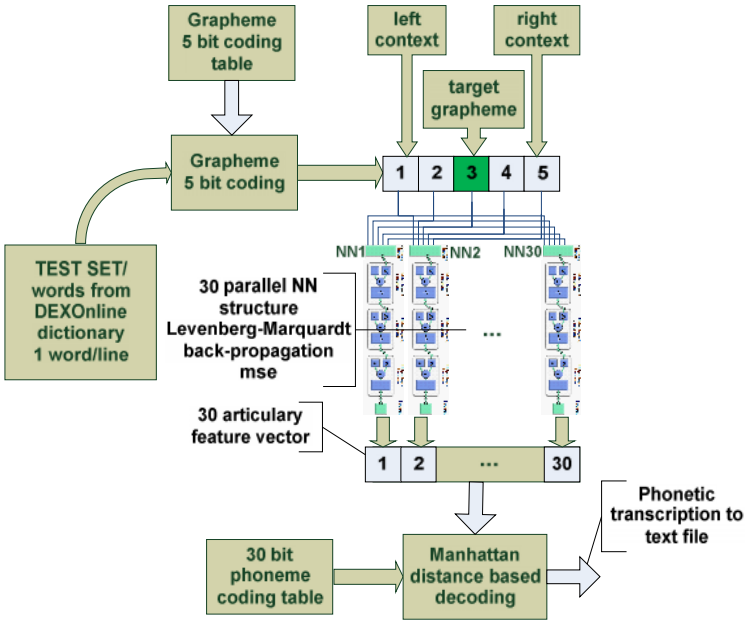

Fig. 2 – System architecture when used for grapheme-to-phoneme transcription.

testing. The structure with the best results has 25 inputs given by the number of bits used to encode the input graphemes, 8 neurons in the first hidden layer, 5 neurons in the second hidden layer and one output to indicate the corresponding articulatory feature presence or absence.

The neurons from the hidden layers have tansigmoidal transfer function and the output level has pure linear transfer function.

The structure of the 30 parallel networks system is presented in Fig. 1 *b* and Fig. 2 depending on which task is involved. Fig. 1 *b* shows the training phase of the system and Fig. 2 shows the system when used for grapheme-to-phoneme conversion. In comparison with the networks of different sizes presented by Burileanu (2002) (without giving the exact dimensions) our system brings a simplification on implementation. After several experiments, we have found that by altering the structure of hidden layers only minor improvements of the recognition results can be achieved.

3. Experimental Results

For training and preliminary testing the system we have manually built a database containing 1004 phonetically transcribed Romanian language words. The words were transcribed by phonetician experts and were collected from some linguistic resources available in published form (Beldescu, 1984; DOOM, 2005). The database contains a total number of 5497 phonemes. For training and testing, the phonetically transcribed word set was divided in three parts: training set, validation set and test set in proportion of 80%, 10% and 10%, respectively.

Network training was performed using Levenberg-Marquardt back-propagation training function, mean squared normalized error performance criterion and early stopping validation vectors are used to stop training early if the network performance on the validation vectors fails to improve or remains the same for 20 train epochs. Network training is quick, it takes about 1 min. for each network on a laptop computer, but training time can be further improved by perform parallel training process on multiple workstations or multiprocessor stations.

When testing the system, the vectors with the 30 articulatory features obtained from neural networks outputs are compared with the encoded vectors of the 36 phonemes used. The distance between the output vector and each coded vector is calculated using Manhattan distance function. System response is chosen as the vector with the smallest distance, replacing the correlation table method described by Burileanu *et al*. (1999).

The trained system performs grapheme-to-phoneme transcription measured on the test set with an accuracy of 92.83% at the phoneme level. Table 2 shows the error percentage for each articulatory feature used. The features with the biggest error percentage are the Closed, Front and type 2.

4. Phonetic Transcription Dictionary Development Based on DEXOnline Entries

Based on the input grapheme codes, the neural networks generate one by one the 30 bit code indicating the presence or the absence of the 30 articulatory features. The output feature vectors are compared with the encoded vectors of the phonemes used. The distance between the output vector and each coded vector is calculated using Manhattan distance function and the system response is chosen as the vector with the smallest distance.

To develop the pronouncing dictionary we have used the largest online dictionary for Romanian language, the DEXOnline dictionary (DEXOnline, 2010). DexOnline dictionary is freely available, can be downloaded from the Internet as a MySQLdump generated SQL file and used in accordance with the terms of GNU General Public License. The database can be easily restored on a MySQL database server. DEXOnline database is organized in multiple tables. The most important three tables for exporting dictionary words are: inflectedform, definition and lexem.

The inflectedform table contains all the inflected forms of the words recorded in the database. By selecting all the distinct wordforms from this table we get a total number of 992,979 records. This is the maximum size of pronouncing dictionary we can create based on DEXOnline. We have exported these words in distinct text files separated by the first grapheme of the words, one word per line, thus resulting input files with a reasonable number of records in the order of several tens of thousands per file.

The definition table is a smaller one containing the definitions recorded in database, 126,563 in number.

The lexem table contains just the base forms of words from the dictionary, totally 139,509.

The architecture of the grapheme-to-phoneme conversion system is presented in Fig. 2. Once the neural networks are trained to detect articulatory features the system can be used to perform grapheme-to-phoneme conversion. The words exported in text files from the DEXOnline database represent the input of the system. Each grapheme of the words are 5 bit coded according to the table of 5 bit binary grapheme codes and then sequences of 5 coded phonemes are presented at the neural networks input. The input graphemes are shifted until each grapheme is presented as target grapheme.

The generated transcription dictionary can be downloaded from the project website, and can be freely used. The dictionary is stored in text format with UTF-8 character encoding.

5. Conclusions

We appreciate that the results are very useful for new speech recognition system and text-to-speech system development. Although there are

reported results of over 98% accuracy of grapheme-to-phoneme transcriptions (Burileanu, 2002; Burileanu *et al*., 1999); Ordean *et al*. failed to repeat these experiments and together with other more recent works (Toma & Munteanu, 2009; Jitea *et al*., 2002, 2003) presents transcription results between 80…95% accuracy. In this situation our results can be considered good enough to build the pronunciation dictionary and can be further improved by increasing the number of manually transcribed words in the training set. Manually phonetically transcribed word database can be extended using the Pronunciation dictionary of Romanian language (Tătar, 1999) the extension is only a matter of time.

Compared to other existing systems we were able to demonstrate that reducing the number of neurons in the hidden layers of networks and using the same 8 × 5 grid for all the articulatory features, we can get the same good transcription results but with significantly lower training and setup times. The training time can be further improved by using C/C++ neural network implementations and performing parallel training on multiple workstations or multiprocessor systems. We have also investigated why the Closed, Front and type 2 articulatory features were so weekly recognized and the answer is that there were not enough examples in the training set.

We have created the first Romanian language pronouncing dictionary based on the words from the lexem table of DEXOnline and we conclude that for dealing with the biggest inflected form table an enlargement of the manually transcribed training set is needed. The pronunciation dictionary is freely available on the project web site which we think that will be a useful tool for all the Romanian speech technology researchers.

As future work we can mention that we are working to extend the manually transcribed training set to 5,000 words based on DEXOnline (1999). After finishing the development of the large training set we intend to retrain the system and regenerate the pronunciation dictionary. Our final goal is to generate an 1 million wordform pronouncing dictionary based on the inflected forms from the DEXOnline dictionary.

Acknowledgment. This paper was supported by the project "Development and Support of Multidisciplinary Postdoctoral Programmes in Major Technical Areas of National Strategy of Research – Development – Innovation" 4D-POSTDOC, contract no. POSDRU/89/1.5/S/52603, project co-funded by the European Social Fund through Sectoral Operational Programme Human Resources Development 2007-2013.

REFERENCES

Beldescu G., *Ortografia Actuală a Limbii Române*. Edit. Şt. şi Encicl., Bucureşti, 1984.

Bisani M., Ney H., *Joint-Sequence Models for Grapheme-to-Phoneme Conversion*. Speech Commun., **50**, 434–451 (2008).

Braga D., Coelho L., *Letter-to-Sound Conversion for Galician TTS Systems*. Proc. of the IV Jornadas en Tecnologia del Habla, Zaragoza, 2006.

Burileanu C., Popescu V., Buzo A., Petrea C. S., Ghelmez-Haneş D., *Spontaneous Speech Recognition for Romanian in Spoken Dialogue Systems*. Proc. of the Roman. Acad., **11**, *A*, 1/2010, 83–91 (2010).

Burileanu D., *Basic Research and Implementation Decisions for a Text-to-Speech Synthesis System in Romanian*. Internat. J. of Speech Technol., **5**, 211-225 (2002).

Burileanu D., Sima M., Neagu A., *A Phonetic Converter for Speech Synthesis in Romanian*. Proc. of the XIVth Congr. on Phonetic Sci. (ICPhS), San Francisco, **1**, 1999, 503-506.

Damper R.I., Marchand Y., Adamson M.J., Gustafson K., *Comparative Evaluation of Letter-to-Sound Conversion Techniques for English Text-to-Speech Synthesys*. Proc. of the 3rd ESCA/COCOSDA Workshop (ETRW) on Speech Synth., Blue Mountains, Australia, 1998.

Davel M., Barnard E., *Pronunciation Prediction with Default&Refine*. Comp. Speech a. Lang., **22**, 374-393 (2008).

Divay M., Vitale A. J., *Algorithms for Grapheme-Phoneme Translation for English and French: Applications for Database Searches and Speech Synthesis*. J. of Comput. Ling., **23**, *4*, 495-523 (1997).

Gómez J.A., Castro M.J., *Automatic Segmentation of Speech at the Phonetic Level*. Lecture Notes in Comp. Sci., *2396*, 2002.

Jitcă D., Apopei V., Grigoraş F., *An Ann-Based Method to Improve the Phonetic Transcription Module of a TTS System for the Romanian Language*. CD-ROM Proc. of the Europ. Conf. on Intell. Technol. (ECIT 2002), Jassy, 2002.

Jitcă D., Teodorescu H.-N. L., Apopei V., Grigoraş F., *An Ann-Based Method to Improve The Phonetic Transcription and Prosody Modules of a TTS System for the Romanian Language*. Proc. of the 2nd Speech Technol. a. Human-Comp. Dial. Conf. - SpeD, Bucharest, 2003, 43-50.

Ordean M. A., Şaupe A., Ordean M., Duma M., Silaghi G.C., *Enhanced Rule-Based Phonetic Transcription for the Romanian Language*. Proc. of the 11th Internat. Symp. on Symb. a. Numeric Algorithms for Sci. Comp. (SYNASC), Timişoara, 2009, 401-406.

Sejnowski T.J., Rosenberg C. R., *Parallel Networks that Learn to Pronounce English Text*. Complex Syst., **1**, 145-168 (1987).

Tătar A.L., *Dicţionarul de Pronunţare a Limbii Române*. Ed. a 2-a, Edit. Clusium, Cluj-Napoca, 1999.

Toma Ş.-A., Munteanu D., *Rule-Based Automatic Phonetic Transcription for the Romanian Language*. Proc. of the Computation World: Future Computing, Service Computation, Cognitive, Adaptive, Content, Patterns, Athens, 2009, 682-686.

* * * *DexOnline*. Transpunerea pe Internet a unor Dicţionare de Prestigiu ale Limbii Române. http://dexonline.ro/, 2010.

* * * *DOOM - Dicţionarul Ortografic, Ortoepic şi Morfologic al Limbii Române*. (Ediţia a II-a, revizuita şi adăugită), Institutul de Lingvistică „Iorgu Iordan - Alexandru Rosetti" al Academiei Române, Edit. Univers Enciclopedic, Bucureşti, 2005.

* * * *Pronouncing Dictionary*. CMU, http://www.speech.cs.cmu.edu/cgi-bin/cmudict, 2008.

DEZVOLTAREA UNUI DICȚIONAR DE PRONUNȚIE PENTRU LIMBA ROMÂNĂ

(Rezumat)

Scopul lucrării este de a prezenta un sistem automat de transcriere fonetică pentru limba română bazat pe rețele neuronale de tip perceptron multistrat. Sistemul a fost testat pe o bază de date construită manual, conținând 1004 cuvinte transcrise de experți lingviști. Folosind acest sistem s-a realizat un dicționar de pronunție pentru limba română efectuându-se transcrierea celor aproximativ 140 000 cuvinte din dicționarul DEXOnline.