Analysis of the Rule Based Phonetic Transcription Technique Applied to the Slovak Language

Jozef Ivanecký

Technical University of Košice
Faculty of Electrical Engineering and Informatics
Department of Cybernetics and Artificial Intelligence
ivanecky@gmail.com

Abstract. Correct phonetic transcription is a key requirement for any automatic speech recognition or text to speech system. In this paper we describe our effort toward automatic phonetic transcription for Slovak language. We give an overview on possible techniques suited for the phonetic transcription and we explore the ability to create a rule based transcription system. We focus also on the syllabical and morphological segmentation as necessary part of rule based transcription for Slovak language as well as the possibility to use it in the real application.

1 Introduction

Among mostly used techniques for computer speech recognition belongs today Hidden Markov Models (HMM) and Neural Networks (NS). In both cases a bigger amount of training data is required. Training input for such systems is recorded speech as well as transcription of recorded utterances. The quality of final system is very dependent exactly on the quality of the transcription.

Transcription itself is possible to create manually. In case of huge training set it is power consuming job and can result in bigger amount of errors. If such data are used for the training, a drop in the final accuracy can appear. On the other side not using such data in the recognition process can have also unexpected influence. The manual transcription can by done by two different ways:

- The input data are transcribed without listening what was really said (pronounced). In such case is necessary to generate all possible pronunciations. Advantage of such approach is, that we can obtain bigger amount of data useful in the future for a recognition process. Disadvantage is that transcriber does not have to cover also the pronunciation of the speaker.
- The input data are listened and transcription contains what was really said. Such approach is invaluable especially if we have data from different dialects regions. This approach is of course more time consuming but sometimes necessary.

The second approach is automatic transcription. In contrast with the first one it is possible to eliminate many human mistakes. On the other side the creation of such system can be very complicated. In some languages so complicated, that it is not used at all. Automatic transcription of Slovak but for speech synthesis is for example in [2]. Possibility of automatic transcription in Czech are discussed in [11], [1].

Rule based approach The easiest way to create phonetic transcription and start from the scratch are rule based systems. Such approach is also used when other techniques – required a training data set – is not possible to use. The requirements for creation of such system is existence of the rules – resp. existence of possibility to create such rules – for "correct" pronunciation in case, that an orthographic representation of the text is available. For the Slovak language very good source of such rules is [8].

One of the assumption for creation of rule based system is existence of relation between orthographic and orthoepic symbol sequence. It may be relation $1:1, 1:N$ or $N:1$. Relations $N:N$ has to be possible to split just to the rules containing $1:1, 1:N$ or $N:1$ only. If it is not possible, it is not possible to use rules to generate pronunciation for particular language. As example for Slovak language can be:

- '1 : 1' relation: '$a \rightarrow a$'; grapheme a always represents phone a.
- '1 : N' relation: '$i \rightarrow$ { I, $I\_\hat{}$ }'; grapheme i represents either phone I or phone $I\_\hat{}$. It is context dependent[1].
- 'N : 1' relation: '{ $\acute{i}$, $\acute{y}$ } $\rightarrow I$:'; graphemes $\acute{i}$, $\acute{y}$ always represent phone I:[2].

If it is possible for particular language to create the set of such rules then it is possible to create the system which will generate from input sequence of graphemes G the appropriate sequence of phones F. Quality of output sequence depends on the rules selection. For example if only orthoepic or also dialect pronunciation was considered etc. [4].

Statistical methods The methods discussed in this section need certain amount of training data. The training data are necessary for creation of certain statistics which are later used for generating the sequence of output phones from input orthographical representation. In the most of the cases decision trees are used. The basic idea is simple.

- The bigger training set is created. It consist of the W, Φ pairs where W is word in orthographic form and Φ is appropriate phonetic transcription. For each such pair the graphemes to phonemes assignment is created. This is the key task. First the assignment from Φ to W is created and then backward

[1] In this case if i is followed by short vowel, it can be either group of vowels or diphthong.

[2] Such relations is possible to simplify to set of '1 : 1' relations.

by simple rules assignment $W \rightarrow \Phi$. This is outcome from assumption, that number of graphemes is never smaller then number of phonemes. It is true for English. For Slovak just in case when we do not use stand-alone phone for the glottal stop. In Czech language – where glottal stop is more common – mentioned assumption is not correct. To generate assignment $\Phi \rightarrow W$ is possible to use for example HMM [7].
- From the pairs W, Φ created in previous step is for each symbol w created a decision tree based on the phone sequence ϕ generated for previous graphemes w and information about orthographical context of symbol w. So create tree is used later for conversion from orthographical to phonetical representation.

Advantage of such approach is fact that in case of big training set the final system will be able to deal with the foreign word better then rules based system. Such solutions are part TTS system where are very ofter a words, not included in the pronunciation dictionary. The hybrid system consisting of both discussed methods are also popular.

2 Rules based system for Slovak

The decision whether to create rules based transcriptions system for Slovak or use some statistical approach was determined by lack of sufficient amount of the training data required for the statistical approach. The rule based approach required the analyze of all phonetical phenomena in Slovak language. The analyzed problems can be devided in to following sections:

- `Vowels` – There is distinct relation between graphemes and phones in Slovak. This is true if vowel is between two consonants or between consonant and word boundary. The problem can be the transcription of vowel "ä" but it is discussed later.
- `Diphthongs` – Slovak has four diphthong: *ia, ie, iu, ô*. Not all combination of vowel *i* and *a, e, u* are diphthongs. Here was necessary to distinguish if sequence of vowels is a real diphthong or vowels sequence.
- `Vowels sequence` – The vowels sequence in Slovak are only in prefixed words, compounded words and foreign words. The pronunciation of vowels sequences in foreign words was adapted to pronunciation of Slovak words. The definition of the pronunciation rules was not problematic.
- `Hard vocal begin and glottal stop` – In Slovak language in contrast with for example Czech it appears very reary. Because there are no exact rules for pronunciation of this phenomena, we did not consider it in the rules definition.
- `Vowel` *ä* – The pronunciation of *ä* is considered as advanced pronunciation. In the standard pronunciation there are big regional differences. From this reason we used for transcription of *ä* phone *E*.
- `Vowels` *ö, ő, ü, ű* – The vowels *ö, ő, ü, ű* are in Slovak language only in loan–words. In Slovak pronunciation there are very often replaced by the

closest vowel. To stay within the Slovak phone set, we used such approach for the rules definition.

- `Voice assimilation` – In the case of consonants pronunciation the main problem is voice assimilation. The assimilation appear on the morphematic borders. While morphematic boundary on the words borders is clear, inside the word it seems to be as bottle neck of the proposed system. From the [9] is clear that to creation the morphematic boundary detector the morphematic dictionary is required. We solved this problem by simple morphematic dictionary.
- `Doubled consonants` – The doubled consonants in Slovak are pronounced on morphematic boundaries. In case of 3 or more consonants the doubled consonants appear very reary. Other simplification are applied here. Here we had to deal again with all problems related to morphematic boundaries detection.
- `Soft and hard consonants` – The graphemes *t,d, n, l* have two possible pronunciation in Slovak language. Either soft or hard one. To be able to deal with all pronunciation phenomena, the syllabic boundaries are required. We used statistical approach described in [5]. In case of pronunciation of graphemes *t,d, n, l* many exceptions exist. They were added in to the exception dictionary.
- `Consonants` [m, F, n, N, ň] – In case of the pronunciation of [m, F, n, N, ň] there are two main problems:
  - There is no exact definition of pronunciation of grapheme *n* in Slovak language.
  - Neither IPA or SAMPA has phonetic repertoire required for the resolution used in Slovak phonetic [6].

  From this reasons we did not fully follow the rules defined in [8]. With the simplification the rules definition was without special needs.
- `Other consonants sequences` – for the correct transcription of the consonant sequences is very often necessary to know morphematic boundaries. From this reason we have to deal here with the all morphematic problems described above. The rules without morphematic boundary needs were defined and exception were added in to the exception dictionary.
- `Other rules` – Because the pronunciation of graphemes *r* and *l* was not included in [8], to defined the rules we used [10]. The definition of the pronunciation rules was not problematic was not problematic in this case. The next set of rules here are the graphemes were pronunciation is unambiguous and therefore they were not mentioned in [8]. For the computer implementation they had to be exactly defined.

As we can see from above mentioned, there are several problematic domains in the rule based transcription of Slovak language:

- In the first case it is transcription of foreign and loan–words. We expected this problem and therefor we focused on domestic words. The rules for foreign words were defined only if it was not too complicated. Many of them are in the exception dictionary.

- The second problem is need of morphematic boundaries for some rules from the assimilation section. To create reliable morphematic analyze is necessary to use morphematic dictionary. This is the issue we were not currently able to deal with.
- The syllabic boundaries problem is necessary to solve in case of pronunciation of grapheme *j* and some others. The solution of syllabic boundaries is described in [5] and is possible to use it for transcription purposes. In this case the statistical approach was used. No external dictionaries are here required.
- The problem of non orthoepic pronunciation is specific problem which was solved just with the information from [8] and own experiences. From this reason many local pronunciation specifics were not included in the rules.

If it is possible to solve all the mentioned problems, it will be possible to create reliable rule based system for Slovak language.

3 Experiments

All the described rules were implemented in Perl language. From 255 pronunciation rules obtained from [8] and [10], 257 transcription rules were defined. Achieved results is possible to resume as follows:

- The vowels rules (vowels, diphthongs, vowels groups) were implemented in full range and testing showed the high reliability during the transcription generation.
- The consonants rules were also implemented in full range but quality of the transcription depend on the quality of the morphematic and syllabic segmentation. The next problem influencing the quality is amount of loan–words and exceptions. In the system is included also exceptions dictionary but language coverage is unknown.

During the implementation was very important to achieve the right order of the rules. The calling the rules in order as they are for example in [8] would lead to the incorrect results.

Before the rules are applied the input word is tested against pronunciation dictionary. In case the word does not belong to the exceptions, the next step is syllabic and morphematic segmentation. After that the transcription rules are applied. The rules are applied word by word considering context within the sentence. In case of transcription of large text the first step is splitting the text in to the single sentences. The sentences are then processed sequentially. Words within the sentence are processed as described above.

For the testing 100 randomly selected words were used. The words were selected from the test set for the syllabic segmentation and words with incorrect syllabic segmentation were removed. The reason for such solution was attempt to eliminate syllabic segmentation errors and get better picture about the quality of the transcription itself. For each word from the test the correct morphematic

segmentation was added in to the morphematic dictionary. Testing itself was done first without the morphematic dictionary and then with the morphematic dictionary. We wanted to know also influence of morphematic segmentation to overall quality. The generated transcription was compared with the manually created transcription.

The reason for transcription of entire sentences instead of isolated words was fact that the morphematic boundaries are also between two words and for correct transcription of assimilation phenomena the surrounding words have to be considered. The problem is detection of morphematic boundaries inside of words. The same we can say about syllabic boundaries and about the words belonging to exception dictionary.

In the following table is the transcription accuracy for the test set with and without morphematic dictionary:

	T. with morphematic dictionary	T. without morphematic dictionary
1 transcription	81 %	75 %
2 transcriptions	9 %	8 %
Overall	**90 %**	**83 %**

The analyze of the output showed that:

- Difference between the results achieved with and without morphematic dictionary is generated mostly by the words where voice assimilation rules were not applied. The voice assimilation is part of the transcription where morphematic boundaries are required. In one case it was pronunciation of doubled consonants. The problem were following 7 words: nepredpokladal (J E p r E **t** p O k l a d a l / J E p r E **d** p O k l a d a l), trikmi (t r I **g** m I / t r I **k** m I), podpíšu (p O **t** p I: S U / p O **d** p I: S U), odpálená (O **t** p a: l E n a:, O **t** p a: L E n a: / O **d** p a: l E n a:, O **d** p a: L E n a:), krúžkoch (k r U: **S** k O x / k r U: **Z** k O x), vládcom (v l a: **ts** ts O m / v l a: **d** ts O m), jesenná (j E s E **n** a: / j E s E **n n** a:).
- The second set of errors — 10 % — were mostly problems with the transcription of *t,d,n,l – ť, ď, ň, ľ*. Such kind of errors have to be fixed by exception dictionary because Slovak language has in case of pronunciation of *t,d,n,l – ť, ď, ň, ľ* too many exceptions. In this case it was group of following words: minifutbal (m I **J I** f U d b a l), kandidátska (k a n **JĪ** d a: ts k a), benevolenciu (b E **J E** v O l E n ts I_^ U, b E **J E** v O L E n ts I_^ U), kabinetu (k a b I **J E** t U), minimálnym (m I **J I** m a: l n I m), veterinárny (v E **c E** r I n a: r n I), Tibete (**c I** b E c E), teroristických (**c E** r O r I s c I ts k I: x), jednej (j E d **J E** I_^), pevného (p E U_^ **J E**: hÕ). From above mentioned word is clear that in the most cases problem are foreign words and loan–words. If we wanted to transcribe the pure Slovak words only, the accuracy would be above 95 % if we use also morphematic dictionary.

As an example of the system we show here the transcription of the sentence *Egyptská správa nehnuteľností zamestnáva tancmajstra s pedálikom.* In the fol-

lowing table is for each word from the input sentence its phonetic transcription in the SAMPA coding.

Graphemes	**Phones**
Egyptská	E g I p s k a:
správa	s p r a: v a
nehnuteľností	J E h\ n U c E L n O s c I:
zamestnáva	z a m E s t n a: v a
tancmajstra	t a n dz m a j s t r a
s	s
pedálikom	p E d a: l I k O m p E d a: L I k O m

4 Summary

In this paper we described the analyze of possibilities to generate phonetical transcription for Slovak language. The proposed method was also implemented and tested on the real data. The achieved result showed that the rules based transcription problem is not possible to solve without external data.

References

1. Černý, M. – Matoušek, V. – Mautner, P.: An automatic creation of the pronunciation dictionary. In: *Proceedings of 3rd Slovenian-German and 2nd SDRV Workshop.* Ljubljana, 1996.
2. Daržágín, S. – Franeková, Ľ. – Rusko, M.: *Konverzia a rečová syntéza slovenčiny.* Jazykovedný časopis, 45, 1994, 1, s. 31–43.
3. Ivanecký, J.: Automatická transkripcia slovenčiny v počítačovom rozpozávaní reči. In: *Slovenčina a čeština v počítačovom spracovaní.* Bratislava, Veda 2001, s. 109–116.
4. Ivanecký, J.: SAMPA v slovenčine a jej význam z pohľadu viacjazyčných systémov na rozpoznávanie reči. In: *Slovenčina a čeština v počítačovom spracovaní.* Bratislava, Veda 2001, s. 98–108.
5. Ivanecký, J.: Štatistický prístup pri určovaní slabičných hraníc. In: *Slovko 2003.* Bratislava, Veda 2003.
6. Ivanecký, J. – Nábělková, M.: *Fonetická transkripcia SAMPA a slovenčina.* Jazykovedný časopis, 53, 2002, s. 81–95.
7. Jelinek, F.: *Statistical Methods for Speech Recognition.* Cambridge – Massachusetts – London, The MIT Press 1998. 283 s.
8. Kráľ, Á.: *Pravidlá slovenskej výslovnosti.* Bratislava, Slovenské pedagogické nakladateľstvo 1983. 632 s.
9. Páleš, E.: *Sapfo – Parafrázovač slovenčiny*, Bratislava, VEDA 1994. 305 s.
10. Pauliny, E.: Slabičné [r][l] v slovenčine. *Slovo a slovesnosť*, 38, 1977, s 307–310.
11. Psutka, J.: *Komunikace s počítačem mluvenou řečí.* Praha, Nakladatelství Akademie věd české republiky 1995.