

# Towards Speech Technologies for Romani Language in Slovakia

Milan Rusko<sup>1</sup>, Sakhia Darjaa<sup>1</sup>, Marián Trnka<sup>1</sup>, Róbert Sabo<sup>1</sup>, Štefan Beňuš<sup>1,2</sup>

Institute of Informatics of the Slovak Academy of Sciences,  
Dúbravská cesta 9, 845 07 Bratislava, Slovakia

<sup>2</sup>University of Constantine, the Philosopher in Nitra,  
Štefánikova 64, 949 07 Nitra, Slovakia

{milan.rusko, sakhia.darjaa, marian.trnka, robert.sabo, stefan.benus}@sav.sk

## Abstract

This work summarizes activities at the Institute of Informatics of the Slovak Academy of Sciences (IISAS) in the research and development of speech technologies in the language of the Roma minority in Slovakia. Basic facts on orthography, phonetics, and prosody of the Romani language are given. The design of the text corpus, the speech database and speech synthesizers is described. The challenges that still have to be tackled in speech recognition are briefly mentioned. The current research in human-machine communication using robotic head is presented and its possible use in the education of Romani children is discussed.

**Keywords:** underresourced languages, Romani language, speech technologies, Romani speech synthesis, robotic head.

## Résumé

Táto práca sumarizuje doterajšie aktivity Ústavu informatiky Slovenskej akadémie vied vo výskume a vývoji rečových technológií v rómčine, popisuje súčasný výskum a plány do budúcnosti. Uvádžajú sa základné fakty o pravopise, fonetike a prozódii rómskeho jazyka. Popisuje návrh textového korpusu, rečovej databázy a rečových syntetizátorov. V krátkosti sa spomínajú výzvy, ktoré je ešte potrebné vyriešiť v rozpoznávaní reči. Prezentuje sa samotná práca na komunikácii človek-stroj pomocou robotického hlavy a diskutuje sa o jej možnom využití vo výučbe rómskych detí.

**Motto:** “Sa tu man šaj kames, kana dumakeri chib nadžanes? Sar saj prindžarav tiro, jilo kana tiro lav hin gadžikano?” (How could you want me if you do not know my language? How could I know your heart, if I do not understand the Gadžo (non-Roma) words?) – fragment of a poem by J. Berky-Luborecký.

## 1. Anglelav – Introduction

Digitally endangered languages are languages used by people who are too few in number or too poor to make them attractive to commercial software developers. This means that native speakers of these languages end up having two barriers to overcome to access computers – first, they have to learn English; then they have to learn IT skills. Their native language is marginalized, and becomes digitally endangered. ([www.mealldubh.org](http://www.mealldubh.org), 2008) From this point of view the language of Romanies in Slovakia (Romani chhib) without any doubt belongs to the digitally endangered languages. The total number of Roma in Europe amounts to 6.6 million people. (<http://romani.uni-graz.at>, 2008) Until recently Romani was only an oral language, without a written norm. During the last decades an attempt to create a written norm started in different countries and in several cases the written form was codified.

### 1.1 Serviko Romani čhib – The Language of Eastern Slovak Romanies

The language of European Romanies belongs to the group of Indian languages. Their language was always influenced by their habitat, where they stayed as a nomadic nation for certain time. Persian, Greek, Armenian, or even Slavonic words can still be found in this language. Slovak Romanies have in their language many words which were adopted from standard Slovak and even more expressions “borrowed” from local Slovak dialects. The language of Romanies living in the south

Slovakia uses also many expressions of the Hungarian origin.

During the population census in 2001, 89 920 people in Slovakia have claimed that their nationality is Roma. Nevertheless the real number of the members of Roma community is estimated to be as high as 380 000.

In Slovakia – as opposed to the Czech Republic the Romani language (as a group of varieties) does not seem to disappear, although many of the dialects and the local varieties (especially in Western, Central and Southern Slovakia) are endangered or close to extinction. But mainly in Eastern Slovakia in some socially isolated localities a gradual change from traditional bilingualism to monolingualism in Romani – or to a radical lowering of competence in Slovak (in generations which grew up to a productive age after the revolution in 1989) can be observed. This is in relation to the growth of unemployment and disintegration of social nets. (Elšík, V., 2007)

The Eastern Slovak (Serviko) Romani is one of the three main dialects that are spoken by Romanies in Slovakia. It is spoken by approximately 80–85% of Roma population in Slovakia, therefore Eastern Slovak Romani dialect was chosen as the basis for grammar, lexicon and phraseology of the codified language.

The term Servika comes from the words “Serbika”, “Serbos”, “Serbija”, reflecting the fact that they came from Serbia. Text.

## 1.2 Irišagos, Pheniben, Vakeribno Melodija – Pronunciation, Orthography, Intonation

In Slovakia the Romani orthography was codified in 1971 and recodified in the nineties. The codified form is based on the orthographical rules of Slovak. Basically the Slovak alphabet was adopted including diacritical marks, which are however used according to different rules. The most recent publication available on the topic (Hübschmannová, M., et al. 2006) gives an overview on the grammar and phonetic rules of the Romani language. In Romani a phoneme is always written with the same corresponding grapheme. This rule is consistently followed with the following phonemes: a, b, c, č, čh, d, e, f, g, h, i, j, k, kh, l, m, n, o, p, ph, r, s, š, t, th, u, v, z, ž. Unlike in the Slovak language, palatalized d,t,n,l are written with hacek-accents even before vowels i and e.

There is a general rule in Romani, that voicing and aspiration is neutralized at the end of words. In the written form however the graphemes corresponding to the aspirated and voiced phonemes are preserved (jakh [jak] – jakha [jakha] (eye-eyes)).

Combinations of vowels in the foreign words (in Slovak ia, ie, iu) are written as ija, ije, iju (geografija, gimnazijum). In contrast to Slovak, in which both /i/ and /y/ graphemes refer to the same [i] vowel, Romani does not use y grapheme.

The phoneme set used in codified Serviko Romani is very similar to that of standard Slovak. The only Romani specific phonemes that do not exist in Slovak are ?h, kh, ph and th, which are pronounced with a slight aspiration. Without the aspiration the words have essentially different meaning, e.g.: pherel (draw, pump) – perel (fall), khoro (jar) – koro (blind)(Hübschmannová, M., et al. 2006). Hence, aspirated sounds are separate phonemes in Serviko Romani.

A problematic issue, significant also for prosody is the problem of vowel quantity. Long vowels are not marked by acute accent.

The rule that voicing is lost at the end of words holds also for Slovak and therefore does not cause any difference in pronunciation between the languages.

Word stress is generally placed on the pre-final syllable in Serviko Romani, which also holds for Eastern Slovak dialects. However, too many exceptions exist from this basic rule, so we had to use a lexicon of word stress exceptions. (As our rule-based intonation model was built for Standard Slovak, which has accents always on the first syllable, we had to revise the whole model and change the rules of accentuation. The rules for phoneme lengths prediction, which are based mainly on the mean value of the phoneme length remained unchanged.)

## 2. Chibakro modulatoris – Speech synthesizer

Since the beginning of the millennium several versions of Romani synthesizers have been developed at the Institute of Informatics. Technically they can be considered as four generations of synthesizers: Diphone synthesizer, Unit-selection synthesizer, HMM (Hidden Markov Model)

synthesizer and DNN (Deep Neural Network) synthesizer. We will briefly describe each version.

### 2.1 Diphone Synthesizer

The first Romani speech synthesizer was based on concatenation of recorded realizations of diphones, and it was a slightly modified version of our Slovak synthesizer. To prepare a diphone database we had to define a set of words that contain the Romani specific diphones that were not present in the diphone set of the Slovak speech synthesizer. For the baseline version adding a set of diphones containing aspirated phonemes – ph, kh, ch, th was sufficient. Table 1 presents examples of Romani words containing various vowel - aspirated consonant combinations.

čh / tʃ <sup>h</sup>	lačharel ačhavel čhamenger	e čhercheña bilače prečhinel	vičhinel dičhiben	očohano fočhipena	čhuvalskro odučharel
kh / k <sup>h</sup>	khamoro naarakheha	te khelel jekhetane	dikhipena lokhiben	o khosno polokhe	te khuvél mukhavkerekel
ph / p <sup>h</sup>	phabaj zaphenel	phenel barephikeskero	phirel priphandel	phosavel dophenel	phuv phurikano
th / t <sup>h</sup>	tharel sathemeskro	themeskero prethovel	ithiskero prithovibe	te thovel odothar	thudeskero thuvalel

Table 1: Examples of Romani words containing aspirated consonants.

Grapheme to phoneme conversion was based on a sophisticated set of rules supplemented by a pronunciation vocabulary and a list of exceptions. The Romani version of this block was created by changing Slovak rules according to Romani pronunciation rules mentioned in section 1.2.

With a relatively small intervention in the text preprocessing, grapheme to phoneme rules, phoneme inventory and the intonation model of the Slovak speech synthesizer we managed to create a basic level diphone speech synthesizer in Serviko Romani. The quality of the speech was adequate to the type of the synthesizer. The segmental quality was not perfect, but was acceptable. In spite of the shortcomings – robotic and buzzy speech quality and very simple prosody model – the produced speech was reasonably intelligible.

### 2.2 Unit-Selection Synthesizer

The Unit-selection Romani synthesizer was using our own synthesis engine presented in Rusko, Trnka and Darjaa (2006). The algorithm did not calculate the joint and cost functions, but was merely relying on phonetical-phonological pre-selection of elements (mainly syllables). The main features determining the selection were phonological context, pitch and phoneme length. The unit-selection synthesizer with a CART trees (Breiman et.al., 1984) based prosody model brought much better naturalness and allowed for more advanced experiments in prosody modeling.

### 2.3 Statistical parametric HMM synthesizer

The synthesizer using Hidden Markov Models was developed using the HTS toolkit (Zen et.al., 2007). It was designed in order to be able to generate not only emotionally neutral speech, but also warning messages in Romani. This feature was meant to be used for automatic generation of voice messages in the warning information system in case of fire, flood, state security threats, or other crisis situations.

#### 2.3.1 Text Resources

To get a better idea of the structure of the language and to prepare a set of sentences for the speech database recording, a bigger volume of texts was needed. For the Romani language, the quantity and quality of the texts available is highly insufficient. Moreover, many of the texts that have been published are written in the particular local dialect of their author and not in the standardized form of the language.

To have at least some amount of texts for initial efforts, we used the archive of the Slovak Romani newspaper “Romano nevo fil” from the years 2003 to 2010, unpublished Romani fairy tales by Vladimír Zeman and several tenths of pages of texts that we were provided by Stanislav Cina, who is our Romani language expert and experienced bi-lingual voice talent. We obtained only about 600 kB of texts in total from these sources, which formed our corpus of Romani texts. These were analyzed and a basic set of 1574 phonetically rich sentences was selected for the recording of the emotionally neutral part of the Romani speech database. The same amount of Slovak texts was prepared and recorded for training of the Slovak baseline neutral voice.

#### 2.3.2 The Expressive Speech Synthesizer

In 2012 we designed an expressive speech database CRISIS. It consists of 90 prompted short warning messages (160 sentences) per level of urgency and per language. An original three-step method of recording this expressive speech database was proposed and successfully employed. The sentences were uttered by one male bilingual speaker in three levels of urgency. The first one represents neutral speech and served mainly as a reference level to the higher two levels. The second level represents assertive warnings or commands, and in the third level the speaker uttered the messages in extremely intense and urgent way – “as if human lives were directly endangered and the speaker had to try to save them” (Rusko et.al., 2012).

As it was mentioned in the previous paragraph, a larger neutral speech databases were recorded by the same speaker in both languages to create higher quality neutral baseline voices, that can be later adapted to the three final voices with different levels of expressivity.

The HTS system (Zen et.al., 2007) was used for creating the speech synthesizer. The baseline HMMTTS voice was trained from the emotionally neutral bigger speech database in the corresponding language. This voice was then adapted to three levels of expressivity using the recordings of the emotional speech database CRISIS and

applying the Constrained Structural Maximum A-Posteriori Linear Regression (CSMAPLR) technique (Nakano et.al. 2006).

According to the informal listening tests the synthesized speech kept the voice quality, rhythm, intonation, and the resulting expressive load from the source recordings very well. Different levels of urgency of the messages were reliably distinguishable across the three adapted synthesizers (level 1 – normal/neutral, 2 - urgent, and 3 – extremely urgent) in both Slovak and Romani languages. The results suggest that the used “three step method” of expressive speech database development is suitable for gathering a good quality expressive and hyper-expressive speech database for the design of speech synthesizers for emergency situations.

### 2.4 Statistical parametric DNN synthesizer

The Slovak Deep Neural Network (DNN) synthesizer was developed using the Merlin toolkit for building DNN models for statistical parametric speech synthesis (Zhizheng, Watts, and King, 2016). We combined it with our own front-end text processor and the WORLD vocoder (Morise, Yokomori, and Ozawa, 2016). WORLD vocoder decomposes the input speech into three parameters: fundamental frequency (f0), spectral envelope and aperiodicity (representation of excitation via the band-aperiodicity function). The used DNN has six feed-forward hidden layers having 1024 hyperbolic tangent units each.

The DNN synthesis based on the WORLD vocoder is considerably more natural than the HMM-based synthesis we used before.

## 3. Towards speech recognition in Romani

The issues making the development of automatic speech recognition in Romani are the same as for the other under-resourced languages. The general issues are lack of written texts, lack of speech recordings, lack of speakers suitable for studio recording, lack of annotators knowing the language, and lack of funding. The specific problem is, that the official codified language is spoken only by several tenths to hundreds of Roma people in Slovakia. All other speak their local dialect that can be significantly different from the codified one.

The experiments with Romani speech recognition are ongoing. They are trying to take the advantage of the fact that high quality Slovak acoustical models have already been developed and are trying to overcome the problems with missing data on the Romani-specific triphones and other phenomena. Due to the very limited amount of text data, building a more general state-of-the-art language model is practically impossible. Therefore, the designers can only work on applications that would do with grammars or simple language models.

For the initial experiments we use the ASR system based on Kaldi Speech Recognition Toolkit (Povey et al., 2011). We hope our first results will be ready for publication soon.

#### 4. Teaching L2 and L1 communication skills with a robotic head

A Furhat is a physical 3D humanoid head that employs the optical projection of an animated facial model (Al Moubayed et. al, 2012). The face projections has functionality allowing for eye brow movement, blinking, and various emotional expressions that are easy to adjust or scale.

In our current work (Beňuš, Sabo, and Trnka, 2019) we investigate how the social robotic head Furhat might be used in human-machine communication research.

We designed a novel communicative game “Guess the animal” to study various parameters of human speech, dialogue phenomena, as well as the effectiveness and convenience of the communication. About 100 healthy and 10 handicapped subjects played the game; they were recorded and filled a questionnaire.



Figure 1: A handicapped person playing the Guess the animal with Furhat.

Both healthy subjects and handicapped people would certainly not prefer Furhat over a human. Nevertheless, they expressed a strong positive evaluation on the usefulness of the robot in training and teaching communicative skills. So the social robotic head can probably also be used to assist human teachers in improving skills in second language acquisition with healthy students and with people of various communication handicaps. Moreover, we think this method could be successfully used to raise interest in learning and training the communicative skills also in the codified version of the language of Roma minority. This could help the children speaking different dialects of Romani to acquire and accept the codified form of their language.

#### 5. Conclusion

However the aim of the work was not to create a perfect speech synthesizer, recognizer, or social assistant but to study the under-resourced language, find its main peculiarities, and prepare the basic speech processing background needed for further development of much more

comprehensive speech technology applications like pedagogical tools and information systems in Romani.

#### 6. Acknowledgements

This work was supported by VEGA grant nr. 2/0161/18. The authors have included in this work some parts of texts of their publications (Rusko et.al., 2006, 2008, 2012), and (Beňuš, Š., Sabo, R., and Trnka, M., 2019). This was necessary for giving a consistent picture of various phases of the research and providing basic facts on the Romani language that have already been published earlier.

#### 7. Bibliographical References

- Al Moubayed, S., et. al, (2012), Furhat: A Back-Projected Human-Like Robot Head for Multiparty Human-Machine Interaction, in *Esposito A., et.al. (eds) Cognitive Behavioural Systems. LNCS, vol 7403*. Springer, Berlin, Heidelberg.
- Beňuš, Š., Sabo, R., and Trnka, M., 2019, Word guessing game with a social robotic head, (19th Conference Information Technologies: Applications and Theory, ITAT 2019), in: *CEUR Workshop Proceedings: Information technologies - application and theory 2019, 2019, vol. 2473*, pp. 1-5.
- Breiman, L. et.al., (1984) *Classification and Regression Trees*. Chapman Hall, New York.
- Elšík, V. (2007), *personal communication*.  
<http://www.mealldubh.org/index.php/2006/02/05/strength-inconfederation/> (2008)
- <http://romani.uni-graz.at/rombase/index.html> (2008)
- Hübschmannová, M., et al. (2006). *Rules of Romani Orthography* (in Slovak). State Paedagogical Institute, Bratislava.
- Morise, M., Yokomori, F., and Ozawa, K., (2016), WORLD: a vocoder-based high-quality speech synthesis system for real-time applications, *IEICE transactions on information and systems*, vol. E99-D, no. 7, pp. 1877-1884.
- Nakano, Y., et.al. (2006), Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis, *Proc. of ICSLP 2006*.
- Povey et al., D., (2011), The Kaldi Speech Recognition Toolkit, in: *Proceedings of ASRU 2011*.
- Rusko, M., Trnka, M., and Darjaa, S. (2006). Three Generations of Speech Synthesis Systems in Slovakia. In: *Proceedings of XI International Conference Speech and Computer, SPECOM 2006*. Sankt Peterburg.
- Rusko, M., et.al., (2008) Making Speech Technologies Available in (Serviko) Romani Language. In: *Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) TSD 2008. LNCS (LNAI), vol. 5246*, Springer, Heidelberg, pp. 501–508.
- Rusko, M., et.al., (2012), Expressive speech synthesis database for emergent messages and warnings generation in critical situations. In: *Language Resources for Public Security Workshop (LRPS 2012), LREC 2012 Proceedings.*, Istambul, pp. 50–53.
- Zen, H., et.al., (2007) The HMM-based speech synthesis system version 2.0. In: *Proc. of ISCA SSW6, Bonn*.
- Zhizheng, W., Watts, O., and King, S., (2016), Merlin: An Open Source Neural Network Speech Synthesis System, *Proceedings of 9th ISCA Speech Synthesis Workshop (SSW9)*, September 2016, Sunnyvale, CA.