Application of Russian Language Phonemics to Generate Macedonian Speech Recognition Model Using Sphinx

Conference Paper · September 2016

CITATION
1

READS
716

3 authors:

Riste Mingov
Ss. Cyril and Methodius University in Skopje
4 PUBLICATIONS **96** CITATIONS
SEE PROFILE

Eftim Zdravevski
Ss. Cyril and Methodius University in Skopje
157 PUBLICATIONS **1,433** CITATIONS
SEE PROFILE

Petre Lameski
Ss. Cyril and Methodius University in Skopje
102 PUBLICATIONS **929** CITATIONS
SEE PROFILE

Application of Russian Language Phonemics to Generate Macedonian Speech Recognition Model Using Sphinx

Riste Mingov *, Eftim Zdravevski, and Petre Lameski

Faculty of Computer Science and Engineering
Ss.Cyril and Methodius University, Skopje, Macedonia
{riste.mingov}@gmail.com
{eftim.zdravesvki,petre.lameski}@finki.ukim.mk

Abstract. Speech-recognition provides possibilities for improved user experience and new level of features in various applications. Although there are widely available open-source and proprietary systems for speech recognition and synthesis for the more widely used languages, there are no available and robust enough systems for the Macedonian language. Building a speech recognition system requires a lot of high-quality recordings, which is expensive operation. To overcome this issue we propose applying some background knowledge and using existing speech synthesis engines to train speech recognition system. Since the Macedonian language belongs to the group of Slavic languages, there are many similarities between them. In this paper we apply this fact to generate a speech synthesis module for the Macedonian language based on the Russian language model. Furthermore, we use the speech synthesis module to build a speech recognition module using the CMUSphinx Toolkit. Finally the results are presented and they confirm that a system with substantial quality can be built without the need of manual recordings in the specific language.

Keywords: speech recognition, acoustic model, CMUSphinx, supervised adaptation

1 Introduction

Speech synthesizing and speech recognition performed by a computer program is already becoming omnipresent in the every-days life. All major operating systems manufacturers already have a speech synthesizer and speech recognition capabilities implemented in their products. To build a custom speech recognition module one could use one of the many available speech recognition tool-kits.

CMU Sphinx is an open source speech recognition toolkit consisted of multiple tools ranging from speech recognisers to acoustic model trainer. It comes

* This work was partially financed by the Faculty of Computer Science and Engineering at the Ss.Cyril and Methodius University, Skopje, Macedonia.

with a variety of libraries and a pre-trained acoustic models for some of the most widely used languages in the world. One of the main limitation of CMU Sphinx and other similar libraries is that most of the world languages are not covered. For such languages a custom acoustic model must be developed in order to use CMU Sphinx or other libraries alike. Training an acoustic model in Sphinx requires more than 70 hours of transcript recordings from more than 200 speakers. From practical point of view, this is a difficult and expensive task to accomplish. In order to obtain a speech recognition module without the costly process of recording and transcribing speeches, we could use an approach similar to the one presented in [1].

Instead of training a new acoustic model, we ca adapt an existing acoustic model to perform the task of speech recognition for a different language than the one it is intended for. Since our goal is to create a speech recognition module for the Macedonian language, we have chosen the closest available language in the CMU Sphinx toolkit, the Russian language because it is from the same group of languages as the Macedonian. The CMU Sphinx approach of training an acoustic model is similar to the way a person would learn a second language.

2 Related work

Although CMU Sphinx [2] suggests a different method of creating acoustic models [3], many authors have tried to use similar techniques of adapting acoustic models and most adaptations use more popular languages [4, 5]. These adaptations use only the phones of a parent language and do not divide the languages in language groups. We selected Russian to be the basic acoustic model because Macedonian and Russian both belong to the Slavic groups of languages. Using this method we aim to achieve better by leveraging the fact that pronunciation of the phones is more similar within the same language group. Some authors, such as [6], have tried to create a generalized set of phonemes to be used in more languages. In contrast, we are using more specialized set of phonemes designed only for one language or language group. The state-of-the-art approaches for speech recognition use deep learning [7] and long-tern short-term memory neural networks [8]. These approaches, however, need a lot of available and labeled data to train a good speech recognition model. The focus of this paper is to use an existing model from a language that belongs to the same language group and adapt it to the target language, without introducing any additional data to the model.

3 Design and Implementation

3.1 CMUSphinx acoustic models

The first step of our approach is to gather all possible phones in an acoustic model and decide which acoustic model is best suited as template for the target language. CMU Sphinx has support for several languages such as: US English,

UK English, French, Mandarin, German, Dutch and Russian. Additionally, it supports building models for other languages. From the available general models the closest to the Macedonian is the Russian model. The acoustic model is described in a DICT file. An example content of the Russian model DICT file is given in Table 3.1.

Table 1. Example of acoustic model definitions in a DICT file from CMU Sphinx

Russian	Phones
...	
вистует	vv i s t uu i t
виступ	vv ii s t u p
виступа	vv i s t u p aa
виступав	vv i s t u p aa f
виступала	vv i s t u p aa l ay
виступали	vv i s t u p aa ll i
...	

3.2 Collecting Macedonian words

In order to mark the phones of the words, next we need to collect all available words from the Macedonian language. In order to accomplish this, we crawled the Macedonian dictionary from the digital Macedonian dictionary web site [9]. We managed to obtain a set of 52993 different words. In the context of the topic of the paper, a very convenient characteristic of the Macedonian language is that each phone is a letter, so it is fairly simple to map them. Most of the letters in the Russian language are same as the letters in the Macedonian language, so the mapping of these phones is a trivial task. Some letters (phones) that are encountered in Macedonian language do not exist in the Russian language, but they can be represented by combining multiple Russian phones. The phone mapping used to represent the Macedonian language using the Russian phones is presented in Table 2.

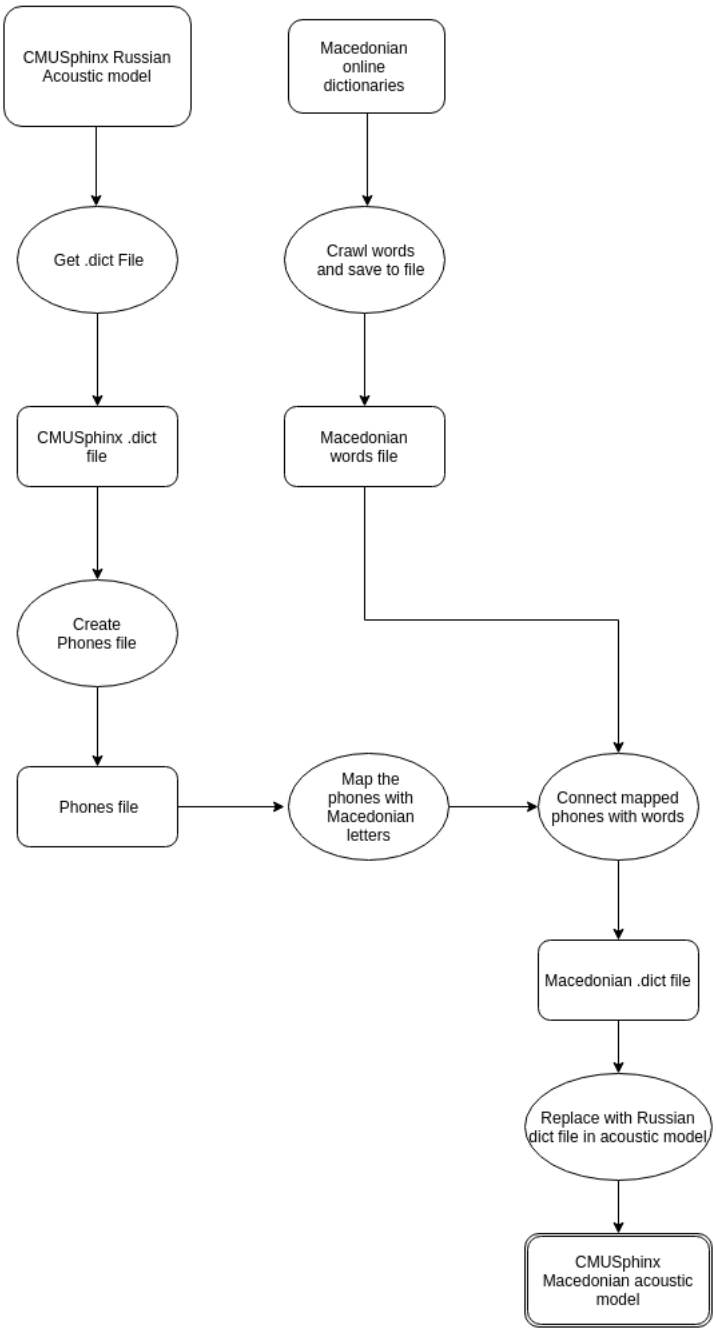
3.3 Creating the Macedonian acoustic model

After the mapping of each phone, the full Macedonian dictionary is processed and as a result is generated the phone representation of each Macedonian word in a new DICT file in a similar way as presented in Table 3.1. Next we replace the DICT file of the Russian acoustic model with the adapted Macedonian DICT file and we can use the new model as a Macedonian CMUSphinx acoustic model. The full process of generating the Macedonian recognition model is depicted in Fig. 3.2.

Table 2. Phone mapping of the Macedonian phones with the Russian phones

Macedonian	Russian	CMUSphinx Phone
А	А	A
Б	Б	B
В	В	V
Г	Г	G
Д	Д	D
Ѓ	Г Ј	G J
Е	Е	EE
Ж	Ж	ZH
З	З	Z
Ѕ	Д З	D Z
И	И	I
Ј	Й	J
К	К	K
Л	Л	L
Љ	Л Й	L J
М	М	M
Н	Н	N
Њ	Н Й	N J
О	О	OO
П	П	P
Р	Р	R
С	С	S
Т	Т	T
Ќ	К Й	K J
У	У	U
Ф	Ф	F
Х	Џ	H
Ц	Ц	C
Ч	Ч	CH
Џ	Д Ж	D ZH
Ш	Ш	SH

Fig. 1. Process of generating the Macedonian speech recognition model from the Russian model



3.4 Experimental setup

To test our adapted acoustic model we created an Android application using the Pocketsphinx library from the CMU Sphinx toolkit. The application starts by randomly selecting 100 words from the Macedonian dictionary, and displays them one by one in a single view. When a word is displayed the user should pronounce it, while the application attempts to recognize what the user is saying. If the application recognises the word, the user can confirm if this is correct or not. The same process is repeated by until 100 words are processed. During testing some false positives may occur, so we added the option the user to mark falsely recognized words as false positives. This means that the user did not say anything but the application recognized a word. After the experiments all results were collected and analysed, as described in the following section

4 Results

The application was used by male and female participants. For each language the approach was tested on random subset of approximately 2000 different words. The obtained results show an average precision of 71.75% with 7.05% rate of false positives and 21.2% rate of not recognized words. The minimum length for the miss-recognized words is 4 the maximum is 14 and the average is 9. The most frequently miss-recognized phones were: ц, г, ч, j, б, ш, њ, ф, ж, ќ, х, џ, s. Also, the generated model performs better for male voices than for female voices which could be a problem of the underlying Russian speech recognition model. There are other works that report gender bias in the speech recognition of other applications and modules too [10, 11]. The average results obtained by our experiments can be seen in Table 3.

Table 3. Testing results by groups

Group	Guessed	Missed	False positive
All	71.75%	21.2%	7.05%
Male	83.6%	9.9%	6.5%
Female	59.9%	32.5%	7.6%

5 Conclusion

We can conclude that the proposed approach works good for most of the common phones between two languages, but performs poorly for the phones that are not common. Further research is needed to find a way to represent the phones that are significantly differently pronounced by the two languages. The results obtained show that it is possible to generate a new recognition model with reasonable accuracy based on an existing model from the same group of languages.

Table 4. Most missed phones

Phones	Times missed	Percentage miss
ц	76	16.13%
г	73	15.5%
ч	67	14.22%
j	67	14.22%
б	56	11.9%
ш	36	7.64%
нь	24	5.1%
ф	20	4.24%
ж	14	3%
ќ	13	2.76%
х	11	2.33%
џ	6	1.27%
s	4	0.85%
љ	3	0.63%
ѓ	1	0.21%

The phone mapping for the Russian and Macedonian language and the usage of an existing model for recognizing words spoken in a similar language are the main contributions of this paper. Further research is needed to verify the proposed approach for other groups of languages, to verify that the approach is indeed applicable for all language from the same group. The paper shows promising results for the application of the Russian language model for generating a Macedonian speech recognition model. Our goal for this paper was to find a simple and effective solution to the speech recognition for languages that do not have enough digital media to train an acoustic model [12] by choosing a parent language from a same language group in order to achieve maximum efficiency. Currently we are only using the CMUSphinx library, but the approach should be applicable to other libraries that support phone mapping to words.

Bibliography

[1] Pascale, F., Yuen, M.C., Kat, L.W.: Map-based cross-language adaptation augmented by linguistic knowledge: from english to chinese. In: Proc. Eurospeech, Citeseer (1999) 871–874

[2] El Amrani, M.Y., Rahman, M.H., Wahiddin, M.R., Shah, A.: Building cmu sphinx language model for the holy quran using simplified arabic phonemes. Egyptian Informatics Journal (2016)

[3] Varela, A., Cuayáhuitl, H., Nolazco-Flores, J.A.: Creating a mexican spanish version of the cmu sphinx-iii speech recognition system. In: Progress in Pattern Recognition, Speech and Image Analysis. Springer (2003) 251–258

[4] Köhler, J.: Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks. In: Acoustics, Speech and Signal Processing, 1998. Proceedings of the 1998 IEEE International Conference on. Volume 1., IEEE (1998) 417–420

[5] Schultz, T., Waibel, A.: Polyphone decision tree specialization for language adaptation. In: Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. Volume 3., IEEE (2000) 1707–1710

[6] Köhler, J.: Comparing three methods to create multilingual phone models for vocabulary independent speech recognition tasks. In: Multi-Lingual Interoperability in Speech Technology. (1999)

[7] Tan, S., Sim, K.C.: Towards implicit complexity control using variable-depth deep neural networks for automatic speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2016) 5965–5969

[8] Hori, T., Hori, C., Watanabe, S., Hershey, J.R.: Minimum word error training of long short-term memory recurrent neural network language models for speech recognition. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE (2016) 5990–5994

[9] : Дигитален речник на македонскиот јазик. `http://makedonski.info` Accessed: 2016-05-26.

[10] Rodger, J.A., Pendharkar, P.C.: A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. Int. J. Hum.-Comput. Stud. **60**(5-6) (2004) 529–544

[11] Tatman, R.: Google's speech recognition has a gender bias. `https://makingnoiseandhearingthings.com/2016/07/12/googles-speech-recognition-has-a-gender-bias/` Accessed: 2016-08-11.

[12] Biondi, M.S., Catania, V., Di Natale, R., Cilano, Y., Intilisano, A.R., Monteleone, G., Panno, D.: G2pil: Agrapheme-to-phoneme conversion tool for the italian