INTERSPEECH 2016
September 8–12, 2016, San Francisco, USA

ASR for South Slavic Languages Developed in Almost Automated Way

Jan Nouza, Radek Safarik, Petr Cerva

SpeechLab, Technical University of Liberec, Czech Republic
{jan.nouza, radek.safarik, petr.cerva}@tul.cz

Abstract

Slavic languages pose several specific challenges that need to be addressed in an ASR system design. Since we have already built an engine suited for highly-inflected languages, we focus on adopting it for new languages, now. In this case, we present an efficient way to adapt the system to all (seven) South Slavic languages, using methods and tools that benefit from language similarities, easily adjustable G2P rules or common phonetic subsets. We show that it is possible to build accurate language and acoustic models in an almost automated way, entirely from resources found on the web. The AMs are trained via cross-lingual bootstrapping followed by lightly supervised re-training from public data, like broadcast and parliament archives. Tests done on a set of main broadcast news in each language show WER values in range 16.8 to 21.5 %, which includes also errors caused by OOL (out-of-language) utterances often occurring in this type of spoken programs.

Index Terms: speech recognition, inflected language, South Slavic languages, multi-lingual system, cross-lingual training

1. Introduction

Slavic languages are spoken by some 320 million people, mainly in the central, eastern and southern parts of Europe. The two biggest are Russian (~160 million speakers) and Polish (~50 million). Other 8 ones have at least 1 million native speakers, each. For many years, these languages stood outside the main focus of speech research community. It had several reasons, one of them being their linguistic complexity. The Slavic languages are known for their rich morphology where nouns, pronouns, adjectives, numbers and verbs are inflected in accord with grammatical context. Inflected word-forms are created from lemmas by prefixes, suffixes and/or changes in stems. This results in very large vocabularies, usually with hundreds of thousands of items. Moreover, some suffixes differ only in one phoneme, which makes many word-forms sound very similar and confusing. A side effect of the rich morphology is a relatively free word order in sentence, which diminishes the role of N-grams in ASR.

In late 2000s, we accomplished the development of a robust LVCSR system for Czech language. It runs with 500K+ vocabularies and has been used in applications, like dictation, broadcast monitoring or audio archive processing [1]. Later, we ported it to other two West Slavic languages: Slovak [2] and Polish [3]. Recently, we are working on a large project, whose goal is to make the system run with most other Slavic languages. At the moment, the highest priority is given to the South Slavic ones, because they are spoken in countries that either already are in EU (Slovenia, Croatia, Bulgaria), or are close to become full members (Serbia, Macedonia, Bosnia and Montenegro). The system will be used mainly for broadcast transcription and monitoring in these countries.

2. Related work and our approach

In some of the mentioned states, there exist (or existed) research teams working in the ASR field. An LVCSR system for Slovene has been built at the University of Maribor [4]. Several papers on Croatian speech recognition (e.g. [5]) have been published by Ipsic et al. from the University of Rijeka. In Serbia, there is a team in Novi Sad that deals with applications of TTS and ASR tools [6]. An attempt to build a Bulgarian LVCSR system based on Microsoft API was described in [7]. Most of these works represent academic research whose goal was to investigate possibilities of existing technologies for the national languages. The authors of [4] tested a less traditional approach to LVCSR when words were built from stems and suffixes. The idea promised an alternative to very large ASR vocabularies for inflective and agglutinative languages. It was tested also by other teams, e.g. for Czech [8], Finish [9] or Turkish [10]. Yet, as computers have become more powerful, the sub-word based approaches lost some of their appeal.

Croatian and Bulgarian are two South Slavic languages covered in GlobalPhone database [11], which is often used in experiments with multi- and cross-lingual approaches in ASR. They were utilized, e.g. by Vu et al. for testing a method for rapid development of language models [12] and for an initial training of acoustic models from a multi-lingual data-pool by using so called A-stabil confidence score [13].

Our approach is similar to the last mentioned ones as it also automates most works needed for AM and LM training. It does not require any annotated speech data in the target languages. Instead, it employs audio files publicly available on web pages of radio-television (RTV) companies or national parliaments (NP), and matches them to related texts using an existing LVCSR system. The segments with closer match are extracted and their transcriptions are used for iterative AM retraining. At the initial phase, the LVCSR system operates with an AM borrowed from a donor language (or languages), later it is trained on a mix of donor and target speech data, and, eventually, only the latter is kept to get genuine AMs for the target languages. In this way, we were able to create AMs well suited for broadcast applications in seven languages.

3. South Slavic languages

In Table 1, there are some basic facts about all the languages. Being official in 7 European states, they are spoken by some 35 million people. Croatian, Slovene and Bosnian use Latin script, Bulgarian and Macedonian Cyrillic, while Serbia and the two remaining countries allow to use both, with a straightforward mapping between them.

http://dx.doi.org/10.21437/Interspeech.2016-747

Table 1. *South Slavic languages split into 3 groups*

Language	Abbrev.	Script	Speakers
Croatian	HR	Latin	7 million
Serbian	SR	Lat./Cyr.	9.5 million
Bosnian	BS	Latin	3 million
Montenegrin	MN	Lat./Cyr.	0.2 million
Slovene	SL	Latin	2.5 million
Macedonian	MK	Cyrillic	2.5 million
Bulgarian	BG	Cyrillic	9 million

The languages can be divided into 3 groups, the largest being the Serbo-Croatian one, which includes also Bosnian and Montenegrin. Another group is made by Bulgarian and Macedonian, which are rather specific as they do not use declension of nouns but, at the same time, they attach several types of definite articles at the end of many words. Slovene is unique and some of its features make it closer to West Slavic languages. Within each group, the languages are mutually intelligible, in spite of many differences in lexicons, spelling, pronunciation or script. From the ASR point of view, each language requires its own vocabulary and an LM, while AMs may be shared within each group, if necessary.

4. Corpora, lexicons and language models

4.1. Text corpora

The best source of multi-domain texts are web-pages of major newspapers and broadcasters. We have designed a web parser that can be adjusted to any web source type and that transfers HTML files to an XML structure, from which we distill the content we are interested in. This helps us to avoid, e.g. discussions attached to web articles which contain many typos, colloquial words, or plain-ASCII text (without diacritic marks) and which would contaminate the lexicon. By employing a language classifier (based on letter N-grams) we try to remove text parts written in non-target languages, e.g. those spoken by minorities (say Italian in Slovenia, Albanian in Montenegro).

Before any further processing, all texts have been converted into Latin script. For Serbian, Bosnian and Montenegrin we used the official Cyrillic-to-Latin conversion table [20]. For Macedonian and Bulgarian, we created a 1-to-1 mapping of Cyrillic letters to the Latin ones with same or close pronunciation. In this way we avoided the troubles that would otherwise arise when reading, typing or editing non-Latin texts. Moreover, it allowed us to use the same string manipulation routines (needed, e.g., for digit transcription or grapheme-to-phoneme conversion) for all seven languages.

One of the critical issues in text pre-processing are digits. In Slavic languages, they can get many different inflected forms. We have created a versatile tool that translates dates, years, and basic forms of cardinal, ordinal and decimal-point numbers. For each language, it requires just a small set of elementary terms (words used to express units, decades, hundreds, thousands, decimal point, etc) and several language specific patterns to convert a digit string to a text form[14].

4.2. Vocabularies

In case of inflected languages, we must expect a typical ASR vocabulary size in range 200K to 500K words if we include those seen at least 5 times. This lower limit usually assures an OOV (out-of-vocabulary) rate below 2 %. We applied this value to get vocabularies for most languages, except Bosnian and Montenegrin. Their corpora were too small compared to the other ones and hence we had to merge them with that of the closest language: Bosnian with Croatian and Montenegrin with Serbian. This means, that the vocabularies (and LMs) for Bosnian and Montenegin are in fact supersets of the Croatian and Serbian ones. The sizes of the vocabularies together with the volumes of the cleaned corpora are summarized in Table 2.

Table 2. *Resources automatically extracted from Internet*

Language	Text data	Vocabulary	Speech data
Croatian	1.10 GB	304K	45 hours
Serbian	1.23 GB	307K	40 hours
Bosnian	0.78 GB	312K	1.2 hours
Montenegrin	0.11 GB	309K	2.4 hours
Slovene	0.91 GB	300K	42 hours
Macedonian	0.83 GB	265K	40 hours
Bulgarian	0.98 GB	283K	41 hours

4.3. Pronunciation

In Slavic languages, a grapheme-to-phoneme (G2P) relation is rather straight. It is true namely for those with Cyrillic script where even foreign names/terms are transcribed phonetically. This can be utilized also for predicting the pronunciation of these words in the remaining languages. We have built a versatile G2P converter that implements the rules common for all Slavic tongues, such as typical pronunciation of most letters, voiced-unvoiced assimilation, consonant cluster reduction. For an individual language, some of these rules can be easily modified, e.g. by omitting devoicing of phoneme 'v' in Croatian, or the same for final consonants in Slovene.

For all the languages, we suffice with 6 vowels: 'a', 'e', 'i', 'o', 'u' and schwa. The last one is used also for Bulgarian letter 'ъ'. There are 28 consonants in the common inventory (incl. the palatalized phonemes 'nj' and 'lj' that are typical for all the languages). If necessary (e.g. during the bootstrapping), this common inventory is very helpful. One exception is Slovene, where pronunciation of all vowels and some consonants (namely 'v' and 'l') depends on syllable stress and some other rather specific rules. In this case, we still use the common inventory, but let the system train and absorb all the variants.

The pronunciation lexicons are built automatically and we allow multiple variants for selected items, namely for foreign words, abbreviations and numbers where either different rules may be applicable, or where speakers often vary. For example, many abbreviations (like, e.g. 'USA') are pronounced either by using local letter names, or English ones, or by adding schwa after consonants, and sometimes as regular syllabic words. As these alternatives compete during the training and recognition phases, they do not harm ASR even if some may be wrong.

4.3. Language models

We employ N-gram LM which is slightly modified by including frequently collocated word strings in the vocabulary. Our experiments showed, that for languages with very large lexicons and free word order, a decoder running with bigrams (and multi-words) achieves almost the same results (with less computation demands) as in case of standard trigrams.

5. Speech data and AM training

Acquiring enough speech data and preparing it for AM training is the key issue in porting an ASR system to a new language. We have developed a scheme that allows us to do it efficiently, with minimum human work and with data that are

freely available on the Internet. Its algorithmic and technical details are described in [15].

5.1. Speech data harvested from web

The first step is to find web pages containing text and audio that are supposed to be somehow related. A typical example are web pages of RTV stations that regularly publish news and (occasionally) complement them by video or audio shots. Another, and more suitable, sources are TV programs with subtitles, or archives of national parliament sessions with video recordings and their transcripts. The provided texts may (but may not) contain phrases, sentences, or quotations that appear in the audio signal. To identify them, we employ the existing LVCSR (see section 6) equipped with the lexicon and LM for the target language, together with any available AM.

The segments where the ASR output closely matches parts of the text are extracted and stored with their (ASR produced) phonetic annotations. To quantify a match score, we use the standard word recog. rate formula, although it has a different meaning here. The segments with 100% match are moved to the AM training set. Those with the score higher than a certain level (e.g. 70 %) are stored in a buffer that will be used in next iterations when a new (supposedly better) AM is available. The data in the buffer are ordered according to their score, which allows us to optimize their effective reuse. Optionally, those with the score above 90 % are checked by a human listener who uses a special tool that highlights the differences and allows for fast correction either in the ASR output or in the reference text. This human intervention is not necessary, but especially at earlier stages it helps to detect some types of errors introduced within the previous automated steps (e.g. in pronunciation). The scheme can process large amounts of data and effectively eliminate those with no or very little match.

We have found the above specified type of web data for all the languages we have worked on. Especially from RTV and NP archives one can get tens or even hundreds of hours of speech that is suitable for our purpose.

5.2. Cross-lingual bootstrapping

Usually, when starting a work on a new language, no annotated speech (and no AM) for that language is available. We have to utilize an existing AM from another (donor) language. Since the two may differ in their phonetic sets, we need to do a temporal mapping between them. The inventory of 41 Czech phonemes and 7 noises is used as the common platform. It allows us to cover any sound in a Slavic language by the same or acoustically close Czech phoneme, or by their combination. For example, palatalized consonants occurring in some South Slavic languages are well modeled by standard consonants followed by phoneme 'j'. This common platform is used also when we combine speech data from several languages to train a multi-lingual AM.

The bootstrapping phase runs in iteration steps. In each, we process all available data from the given source. At the beginning, the train set is made of a small amount (~10 hours) of data from the donor language(s). At the end of each iteration, the newly annotated data (i.e. those with 100% match) are added to the train set and a new AM is trained. The bootstrapping phase finishes when the amount of annotated target data exceeds that of the donor ones. After that, the latter is removed, the annotations are re-mapped back to the original phonetic set, and the first genuine AM for the target language is trained.

Table 3. *Figures illustrating main phases in Bulgarian data mining and AM training. (WER was measured on a small development set of read speech - see text.)*

AM version	Target [hours	Donor [hours]	WER [%]
Czech (CZ)	0	10.1	28.7
Slovak (SK)	0	10.2	27.1
Polish (PL)	0	10.0	31.6
Croatian (HR)	0	9.6	25.7
Multi-lingual (CZ+SK+PL+HR)	0	39.9	26.1
Mixed (BG+HR), 1st iter. BNP	2.3	9.6	23.4
Mixed (BG+HR), 8th iter. BNP	10.3	9.6	19.7
Bulgarian after all BNP data	15.2	0	18.9
Bulgarian after adding BNT data	36.8	0	11.9
Bulgarian after adding BNR data	41.2	0	11.2

5.3. Adding more data and AM iterative training

In general, the data mining and speech annotation process gets more efficient when the target AM is available. We are able to collect more data per iteration and we can try to mine other sources to get robust models trained on more voices and various acoustic conditions. In this project, we finished the data collection process, when we got over 40 hours, or when no other data was available (for smaller languages). The final amounts of annotated speech are summarized in Table 2. We have made separate AMs for Croatian, Serbian, Macedonian, Slovene and Bulgarian. The AM for Bosnian and Montenegrin was trained on data from all Serbo-Croatian languages.

5.4. Bulgarian as case example

To illustrate the complete process, let us take Bulgarian as an example. Before we started, we had already had data and models from our previous work (Czech, Slovak and Polish) and Croatian from this project. We found suitable speech sources at the web pages of the Bulgarian National Parliament (BNP) [16], and two national broadcasters: BNT [17] and BNR [18]. The main phases of the data mining process are illustrated in Table 3. We show the amounts of data in target and donor languages and provide WER figures achieved on a small development set (33 minutes of speech read by 4 native speakers). We tested AMs from all 4 available languages, and since the Croatian yielded the best results, it was chosen as the donor.

We started the bootstrapping phase using the data from 20 randomly selected parliament sessions (from 2013-2015). The first iteration yielded 2.3 hours of matched annotated speech. We got over 10 hours after the 8th iteration and created the first Bulgarian AM. It was used to mine the remaining data in the BNP buffer. We stopped it when the amount of newly added annotated speech dropped below 15 minutes. After that we launched the mining scheme from the BNT source. Its TV archive offers thousands of short news with videos. Texts attached to them occasionally contain quotations of speakers occurring in the shots. Within 16 iterations we were able to mine out 21.6 hours. The same was repeated with data from Bulgarian Radio (BNR) archive. We finished the process when the amount of annotated BG speech reached 41.2 hours. It would be possible to get much more data namely from the huge BNP archive, however it would mean that some speakers (parliament members) could have too much data in the training set. The whole process of data mining and iterative AM training took about 3 weeks during which approx. 250 hours of Bulgarian (freely available) audio data have been examined.

Table 4. *Broadcast news test sets and their parameters*

Language	Duration [min]	# Words	OOV [%]	OOL [min]
Croatian	104	15319	0.99	0.7
Serbian	89	12791	0.41	0.3
Slovene	109	14943	0.68	4.0
Macedonian	94	12916	0.52	1.6
Bulgarian	100	15197	0.61	0.1

Table 5. *Performance on broadcast test sets from Table 4*

Language	WER [%]	WER [%] after excluding OOL passages
Croatian	20.69	20.07
Serbian	19.17	18.90
Slovene	21.49	16.16
Macedonian	16.83	14.54
Bulgarian	20.94	20.86

6. Practical evaluation on broadcast data

For evaluation, we employed the same own LVCSR system that performs broadcast transcription tasks in Czech, Slovak and Polish. It uses log-filter banks for signal parameterization and a DNN-HMM framework for speech decoding. The DNNs have 5 hidden layers (1024-1024-768-768-512 neurons) and have been trained via triphone GMMs using Torch toolkit [21]. Within the data mining process we used the same system, but with the GMMs only, as it saved time needed for repetitive retraining on continuously increasing amount of data.

The tests were run on real broadcast records downloaded from archives of major TV and radio stations in Croatia, Serbia, Slovenia, Macedonia and Bulgaria. For each language we took 3 main news shows from different stations, each about 30 min. long. They were complete; from the opening jingles to the closing ones. They included all types of speech occurring in news programs: clean speech read in studio, speech with background music or noise, spontaneous utterances recorded in streets, or a dubbed speech with a talk in a foreign language in background). Reference transcriptions have been made by native speakers. (We did not find them for Montenegrin and Bosnian, and this is why these two languages were excluded from testing.) All the shows were broadcast in February 2016, while the lexicons and LMs had been created during 2015. The main parameters of the test sets are summarized in Table 4.

The results are presented in Table 5. We can observe that WER values are in range 17 to 22 %, which is comparable to other languages we have been working on so far. In case of clean speech, most errors are due to omitted or inserted short (one-phoneme) words and also due to confused word-forms with acoustically similar suffixes. More serious errors occurred namely in parts with spontaneous and noisy speech. This is obvious but we admit it could be partly influenced also by the fact that the automatic data mining process - due to its nature - extracts mainly clean and clear speech passages.

Another aspect that we have to take into account is that typical news programs may contain parts where a foreign language is spoken (e.g. by world politicians, interviewed foreigners, etc). This happened in large scale namely in Slovenian news and that is why we marked these out-of-language (OOL) passages in the whole test set and optionally excluded them from the WER computation. For Slovene, it represented an absolute WER reduction of 5 %. In other cases, the effect was smaller. The modified results are shown in the last column of Table 5.

We have made the complete test set available so that anybody can use it for own experiments [22].

7. Discussion

The presented approach relies on the availability of audio and text files that are (or are expected to be) mutually related. If both contain common sentences or phrases, the described method has a fair chance to discover them, extract them and prepare them for AM training. If not, the method simply skips such files. The scheme is reliable since it allows only correctly annotated data to become a part of the AM training set. It gives the final AM high probability to be as accurate as possible. In a recent study [19], it is shown that phonetic annotations used for AM training in which 2 % of all phonemes are wrongly annotated can increase WER of a state-of-the-art LVCSR system by 2 to 5 % relatively.

Moreover, the method produces training data where each audio file is complemented by both orthographic and phonetic transcriptions. Hence, we get a training material in a format similar to standard speech databases. This also means that the two transcriptions can be checked, analyzed and edited (manually or automatically) if necessary, e.g. when some pronunciation or transcription rules need to be modified.

8. Conclusions

In the paper, we present a highly automated procedure that allowed us to develop ASR systems for seven South Slavic languages within a relatively short period. For all modules of the systems we used only publically (and freely) available data from the Internet.

To make the development efficient, we tried to benefit from language relations and similarities as much as possible. We have built a common platform that included: a) Latin alphabet coding for all the languages (i.e. also for those with the Cyrillic one), b) a common phonetic inventory that is helpful mainly during the initial bootstrapping phase, c) a versatile G2P tool applicable (and easily modifiable) to most Slavic languages, d) a versatile digit-to-text translator that works with most number generating patterns occurring in these languages.

The experiments performed on real data prove that the ASR systems achieve results that are applicable for automatic monitoring of broadcast stations in this European region. After employing them in daily use, we intend to get more data and utilize it for further improvements, namely in the acoustic modeling part.

Within the current research project we plan to adopt the same scheme for the remaining Slavic languages. It will be the East Slavic ones that also pose several specific challenges. We have already discovered web sources with speech and text data required by the methods described in this paper.

9. Acknowledgements

The research was supported by the Technology Agency of the Czech Republic in project no. TA04010199 called MultiLinMedia and partly also by the Student Grant Scheme (SGS 2016) at the Technical University of Liberec.

10. References

[1] Nouza, J, et al. "Speech-to-text technology to transcribe and disclose 100,000+ hours of bilingual documents from historical Czech and Czechoslovak radio archive". *Proc. of Interspeech* 2014, pp. 964-968.

[2] Nouza, J, et al. "Czech-to-Slovak adapted broadcast news transcription system." *Proc. of Interspeech*, 2008, pp. 2683–2686.

[3] Nouza, J, Cerva, P, Safarik, R. "Cross-Lingual Adaptation of Broadcast Transcription System to Polish Language Using Public Data Sources", *Proc of LTC*, Poznan, 2015, pp. 181-185

[4] Rotovnik, T., Maučec, M., Kačič, Z. "Large vocabulary continuous speech recognition of an inflected language using stems and endings". *Speech communication*, 49(6), 2007, pp. 437-452.

[5] Martinčić-Ipšić, S., Pobar, M., & Ipšić, I. "Croatian large vocabulary automatic speech recognition". *AUTOMATIKA*, 52(2), 2011, pp. 147-157.

[6] Delić, V., Sečujski, M., Jakovljević, N., Janev, M., Obradović, R., Pekar, D. "Speech technologies for Serbian and kindred South Slavic languages". *Advances in Speech Recognition*, 2010, pp.141-164.

[7] Mitankin, P., Mihov, S., Tinchev, T. Large "Vocabulary Continuous Speech Recognition for Bulgarian". *Proc. of RANLP* 2009, pp. 246-250.

[8] Ircing, P., Psutka, J. V., Psutka, J. "Using morphological information for robust language modeling in Czech ASR system". *Audio, Speech, and Language Processing*, IEEE Transactions on, 17(4), 209, pp. 840-847.

[9] Kurimo, M., et al. "Unlimited vocabulary speech recognition for agglutinative languages". In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. 2006, pp. 487-494

[10] Sak, H., Saraclar, M., & Gungör, T. "Morphology-based and sub-word language modeling for Turkish speech recognition *Proc. of ICASSP*, 2010, pp. 5402-5405

[11] Schultz, T. "Globalphone: a multilingual speech and text database developed at Karlsruhe university", In *Proc. of Interspeech,* 2002.

[12] Vu, N. T. et al. "Rapid bootstrapping of five eastern European languages using the rapid language adaptation toolkit". In *Proc. of Interspeech*, 2010, pp. 865-868.

[13] Vu, N. T., Kraus, F., Schultz, T. "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil." In *Proc. of ICASSP*, 2011, pp. 5000-5003

[14] Chaloupka, J. "Automatic Symbol Processing for Language Model Building in Slavic Languages". To appear in *SloNLP (16th Conference on Information Technologies - Applications and Theory)*, Slovak Republic, 2016

[15] Nouza, J., Cerva, P., Kucharova, M. "Cost-Efficient Development of Acoustic Models for Speech Recognition of Related Languages", *Radioengineering*, vol. 22, no. 3, pp. 866-873, 2013

[16] http://www.parliament.bg/tv/ (Bulgarian national parliament)

[17] http://bnt.bg (Bulgarian national TV)

[18] http://bnr.bg (Bulgarian national radio)

[19] Safarik, R, Mateju L. "Impact of Phonetic Annotation Precision on Automatic Speech Recognition Systems." *Proc. of TSP conference*, Vienna, 2016

[20] https://en.wikipedia.org/wiki/Serbian_Cyrillic_alphabet

[21] http://torch.ch

[22] https://gitlab.ite.tul.cz/SpeechLab/SouthSlavicTestData