RAPID BOOTSTRAPPING OF A UKRAINIAN LARGE VOCABULARY CONTINUOUS SPEECH RECOGNITION SYSTEM

Tim Schlippe, Mykola Volovyk, Kateryna Yurchenko, Tanja Schultz

Cognitive Systems Lab, Karlsruhe Institute of Technology (KIT), Germany

ABSTRACT

We report on our efforts toward an LVCSR system for the Slavic language Ukrainian. We describe the Ukrainian text and speech database recently collected as a part of our *GlobalPhone* corpus [1] with our Rapid Language Adaptation Toolkit [2]. The data was complemented by a large collection of text data crawled from various Ukrainian websites. For the production of the pronunciation dictionary, we investigate strategies using grapheme-to-phoneme (g2p) models derived from existing dictionaries of other languages, thereby reducing severely the necessary manual effort. Russian and Bulgarian g2p models even decrease the number of pronunciation rules to one fifth. We achieve significant improvement by applying state-of-the art techniques for acoustic modeling and our day-wise text collection and language model interpolation strategy [3]. Our best system achieves a word error rate of 11.21% on the test set on read newspaper speech.

Index Terms— speech recognition, rapid language adaptation, Ukrainian, Slavic language, pronunciation dictionary

1. INTRODUCTION

Our goal was to rapidly bootstrap and improve an automatic speech recognition (ASR) system for Ukrainian with low human effort and at reasonable cost. We used our Rapid Language Adaptation Toolkit (RLAT) [2] for collecting a large Ukrainian speech and text corpus. RLAT aims to significantly reduce the amount of time and effort involved in building speech processing systems for new languages and domains. It is envisioned to be achieved by providing innovative methods and tools that enable users to develop speech processing models, collect appropriate speech and text data to build these models, as well as evaluate the results allowing for iterative improvements. For this study we further advance the language-dependent modules in RLAT. To face the challenge of the rich morphology and high out-of-vocabulary (OOV) rate thereby improving language model (LM) quality, we use our "snapshot" function which gives informative feedback about the quality of text data crawled from the Web. This funtion enables a day-wise text collection and LM interpolation strategy which we have already successfully applied to Bulgarian, Croatian, Czech, Polish, and Russian ASR [3]. Another challenge was the rapid and economic creation of a qualified Ukrainian pronunciation dictionary. Dictionaries provide the mapping from the orthographic form of a word to its pronunciation, which is useful in both text-to-speech and ASR systems. They are used to train the systems by describing the pronunciation of words according to manageable units, typically phonemes [4]. Dictionaries can also be used to build generalized grapheme-to-phoneme (g2p) models, for the purpose of providing pronunciations for words that do not appear in the dictionary [5]. The production of dictionaries can be time-consuming and expensive if they are manually written by language experts. Therefore several approaches to automatic dictionary generation from word-pronunciation pairs of the target language have been introduced in the past [6][7][8]. [9] and we [10][5] describe automatic methods to produce dictionaries using word-pronunciation pairs found in the Web. However, we did neither possess Ukrainian word-pronunciation pairs nor find those in sufficiant amount in the Web. Therefore we investigated strategies using g2p models derived from existing dictionaries of other languages, thereby reducing severely the necessary manual effort.

In the next section, we give a brief introduction to the structure of the Ukrainian language. In Section 3, we present work that is related to Ukrainian ASR. Section 4 describes our speech and text data collection. In Section 5 we present our baseline recognizer resulting from the rapid initialization based on RLAT. We investigate the dictionary creation using g2p models derived from existing dictionaries of other languages in Section 6. Section 7 describes our optimization steps including a data-driven acoustic modeling of semi-palatalized phonemes and our day-wise text collection and LM interpolation strategy. We conclude in Section 8 with a summary of current results and an outlook to future work.

2. THE UKRAINIAN LANGUAGE

Ukrainian is the official language of Ukraine. In the state census in 2001, 67.5% (or 32.5 million) of the population in Ukraine declared Ukrainian to be their native language [11]. However, 42.8% of Ukraine's habitants use Ukrainian at home, 38.7% speak Russian and 17.1% speak both languages [12]. Ukrainian speakers who use Russian at home may have a slight Russian accent when speaking Ukrainian. With over 37 million speakers all over the world, there are

in particular big Ukrainian speaking communities in Russia, Canada, Moldova, USA, Kazakhstan, Belarus, Romania, Poland, and Brazil [13]. Together with Russian and Belarusian, the Ukrainian language forms the subgroup of East Slavic languages. The Cyrillic alphabets for Russian and Ukrainian are different. Both have 33 letters [14]. However, the Ukrainian alphabet does not have the graphemes э, ё, ы, ъ, but some other letters such as є, і, ї, ґ plus the apostrophe (’). Some graphemes belonging to both languages correspond to different phonemes. For example, г is pronounced as the consonant /g/ in Russian and as the voiced glottal fricative /ɦ/ in Ukrainian. Like other Slavic languages Ukrainian has a rich morphology. Further peculiarities are the occurrence of palatalized consonants (e.g. ряд - /rʲɐd/) [15], the existence of long geminates as in Polish (e.g. знання - /znɑˈɲːɑ/), the use of the apostrophe similar to the Russian hard sign, and the affricates /d͡z/ and /d͡ʒ/ that are not represented by separate letters but by the digraphs дз and дж. [16] define rules for the g2p relation and investigate the properties of the Ukrainian version of the Cyrillic alphabet. The IPA transcription they use is based on the tables given by [15].

3. RELATED WORK

[17] developed an LVCSR system for the experimental system of a computerized stenographer for the proceedings of the Ukrainian Parliament. They report an ASR accuracy of 71.5% with a bigram LM and a context-independent acoustic model with 56 acoustic model units. The dictionary was created automatically using context-dependent Ukrainian g2p convertion rules. Due to the different speaking and pronunciation style of the speakers, they analyzed the use of personal dictionaries for the decoding and report an improvement of 1% absolute on average. At present, there are no speech and language databases for Ukrainian in the ELRA catalogue or in other multilingual corpora like SpeechDat, Speecon, and Speech Ocean. Research on Ukrainian ASR has been carried out in Ukraine [14]. A corpus of continuous and spontaneous Ukrainian speech has been collected there [18]. Using this corpus for training, [19] report 59.61% accuracy for spontaneous speech, [20] on average 10% word error rate in a dictation system. [20] use a g2p converter which is described in [21] to generate Ukrainian pronunciations. Usually linguists define 32 Ukrainian consonants and 6 vowels [15][16]. [17], [19] and [22] use those phonemes plus additionally 13 semi-palatalized consonants for ASR. [22] and [23] also investigate the discrimination of stressed and unstressed vowels in Ukrainian and Russian ASR but this leads to comparable results.

Our contribution is the collection of Ukrainian speech and text data as a part of our *GlobalPhone* [1] corpus. *GlobalPhone* is a multilingual speech and text data collection in 20 languages available from ELRA[1]. We create a dictionary automatically using context-dependent g2p rules and then check and revise it manually. For a cheaper and faster creation, we additionally demonstrate that we can reach comparable quality using g2p models derived from existing dictionaries of related languages. Finally, we apply state-of-the art techniques for acoustic modeling such as context-dependent modeling and data-driven modeling of the Ukrainian semi-palatalized phonemes. Using the day-wise LM interpolation and a vocabulary adaptation, we obtain a 3-gram LM with high n-gram coverages, low perplexity and low OOV rate on our development and test sets.

umoloda.kiev.ua	day.kiev.ua	ukurier.com.ua
pravda.com.ua	chornomorka.com	tsn.ua
champion.com.ua	ukrslovo.org.ua	epravda.com.ua

Table 1. List of crawled Ukrainian Websites.

4. UKRAINIAN RESOURCES

4.1. Text Corpus

To build a large corpus of Ukrainian text, we used RLAT [2] to crawl text from 9 websites as listed in Tab. 1, covering Ukrainian online newspaper sources. RLAT enables the user to crawl text from a given webpage with different link depths. The websites were crawled with a link depth of 10, i.e. we captured the content of the given webpage, then followed all links of that page to crawl the content of the successor pages (link depth 2) and so forth until we reached the specified link depth. After collecting the text content of all pages, the text was cleaned and normalized in the following three steps: (1) Remove all HTML tags and codes, (2) remove special characters and empty lines, and (3) identify and remove pages and lines from other languages than Ukrainian based on large lists of frequent Ukrainian words and on the Ukrainian character set. We complemented the text with fragments from the Ukrainian literature by P. Myrny, I. Nechuy-Levytsky, and O. Honchar and lyrics. The websites and the literature works were used to extract text for the LM and to select prompts for recording speech data for the training (*train*), development (*dev*), and evaluation (*test*) set.

4.2. Speech Corpus

To develop and evaluate our Ukrainian recognizer, we collected speech data in *GlobalPhone* style [1], i.e. we asked speakers of Ukrainian in Ukraine and Germany to read prompted sentences of newspaper articles. The corpus contains 13k utterances spoken by 46 male and 73 female speakers in the age range of 15 to 68 years. All speech data was recorded with a headset microphone in clean environmental conditions. The data is sampled at 16 kHz with a resolution of 16 bits and stored in PCM encoding. The Ukrainian *GlobalPhone* database is presented in Tab. 2. We recorded 39 Ukrainian speakers with Ukrainian as their first language

[1] http://catalog.elra.info

and 80 with Russian as their first language. Information about native language, age, gender, etc. is preserved for each speaker to allow for experiments based on the speakers' characteristics. The dev set was used to determine the optimal parameters for our ASR system.

Set	Male	Female	#utterances	#tokens	Duration
train	38	61	11k	69k	11 h 45 mins
dev	4	6	1k	7k	1 h 14 mins
test	4	6	1k	7k	1 h 08 mins
Total	46	73	13k	83k	14 h 07 mins

Table 2. Ukrainian *GlobalPhone* Speech Corpus.

5. BASELINE SPEECH RECOGNITION SYSTEM

According to [15] and [16], we use 38 basic phonemes consisting of 6 vowels and 32 consonants. As described in [17], [19], and [22], we additionally use 13 semi-palatalized consonants which leads to our final 51 Ukrainian phonemes as acoustic model units. Based on [16], [22] and [23], we abstain from distinguishing stressed and unstressed vowels. Our goal in this work was to build an ASR system that works for all collected speakers. Therefore all the 11.75 hours of the training set were used to train the acoustic models (AMs) of the Ukrainian speech recognizer. Our corpus, however, allows future experiments with individual systems for speakers with and without Russian accent or to investigate adaptation techniques. As in [24], we used our multilingual phone inventory to bootstrap the system which is included in RLAT [2], the preprocessing with Melscale Frequency Ceptral Coefficients (MFCCs) and state-of-the-art techniques for acoustic modeling to rapidly build a baseline recognizer for Ukrainian. For our context-dependent AMs with different context sizes, we stopped the decision tree splitting process at 2k quintphones. With the training transcriptions, we built a statistical 3-gram LM (*TrainTRL*) which contains their whole vocabulary (7.4k). It has a perplexity (PPL) of 594 and an OOV rate of 3.6% on the dev set. The pronunciations for the 7.4k words were created in a rule-based fashion and were manually revised and cross-checked by native speakers. The word error rate (WER) of the baseline system trained with all the 11.75 hours is 22.36% on the dev set and 18.64% on the test set. We also simulated scenarios where less training data were available. Fig. 1 shows the WER of the proposed techniques with smaller amounts of training data.

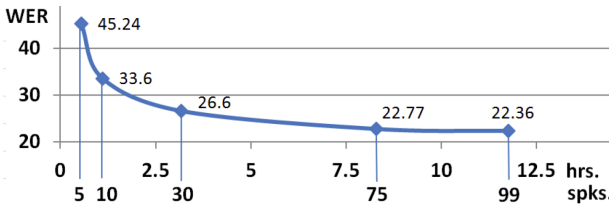

Fig. 1. WER over size of audio data for training (in hours)

6. CROSS-LINGUAL DICTIONARY PRODUCTION

The production of dictionaries can be costly in terms of time and money if no word-pronunciation pairs in the target language for a data-driven automatic dictionary generation are available. Often native speakers or linguists have to define rules and computer experts have to implement and apply them; e.g. for the creation of the Ukrainian dictionary, 882 search-and-replace rules based on [16] were elaborated and applied to produce phoneme sequences corresponding to our Ukrainian words. For the fast and cost-saving creation of a dictionary, we investigated generic strategies using g2p models derived from existing dictionaries of other languages, thereby reducing severely the necessary manual effort. We tested the support of Russian (*ru*), Bulgarian (*bg*), and German (*de*) g2p models that have been generated from our existing *GlobalPhone* dictionaries plus English (*en*) created from a dictionary that is based on the CMUdict[2]. Tab. 3 lists their phoneme and grapheme coverages on Ukrainian. For *en* and *de* we used the existing official standardized Ukrainian transliterations on grapheme level (*) [25]. As Ukrainian, all tested languages are of the Indo-European language family. *ru* and *bg* also belong to the Slavic languages.

Language	Grapheme coverage	Phoneme coverage
Russian (ru)	88%	57%
Bulgarian (bg)	88%	67%
German* (de)	0%	39%
English* (en)	0%	37%

Table 3. Language relationship to Ukrainian.

6.1. Cross-lingual Dictionary Generation Strategy

To cross-lingually generate pronunciations for the Ukrainian words, we propose the following strategy:

1. *Grapheme Mapping*: Mapping Ukrainian graphemes to the graphemes of the related language (*Rules before g2p*)
2. Applying g2p model of the related language to the mapped Ukrainian words
3. *Phoneme Mapping*: Mapping resulting phonemes of the related language to the Ukrainian phonemes (*Rules after g2p*)
4. *Optional*: Post-processing rules to revise shortcomings (*Post-rules*)

Step	ru	bg	de	en
1	биг	биг	bih	bih
2	ru_b ru_i ru_g	bg_b bg_i bg_g	de_b de_i	en_b en_ih
3	ua_b ua_i ua_h	ua_b ua_i ua_h	ua_b ua_i	ua_b ua_y
4	ua_bj ua_i ua_h	ua_bj ua_i ua_h	ua_bj ua_i	ua_b ua_y

Table 4. Cross-lingual pronunciation production for биг.

As *GlobalPhone* dictionaries contain phonemes based on the International Phonetic Alphabet (IPA) scheme [26], we mapped the phonemes of the related language to the Ukrainian phonemes based on the closest distance in the IPA

[2]http://www.speech.cs.cmu.edu/cgi-bin/cmudict

chart in *step 3*. We stopped to include *Post-rules* once obviously no further improvement was possible due to the quality of the underlying g2p model of the related language. Tab. 4 shows the output for the Ukrainian word бігт (running) after each step of our cross-lingual dictionary generation strategy. The correct pronunciation in the handcrafted dictionary is *ua_bj ua_i ua_h*. As European languages are written in segmental phonographic scripts which display a somewhat close g2p relationship, with one grapheme roughly corresponding to one phoneme, we also trained and decoded a system with a pure graphemic dictionary (*grapheme-based*) for comparison. This approach gave encouraging results in former studies [27][28][29] and even outperforms manually cross-checked phoneme-based dictionaries for some languages.

6.2. Performance

Tab. 5 indicates that we can generate qualified dictionaries using *ru* and *bg* g2p models. Comparing the new pronunciations derived from the two languages to those of the handcrafted Ukrainian dictionary in terms of phoneme edit distance results in small phoneme error rates (PERs). Furthermore, using the new dictionaries for training and decoding leads to WERs on the dev set that outperform *grapheme-based* (23.82% WER) and even the performance of the handcrafted dictionary (22.36% WER). We need only 18% of the number of the 882 search-and-replace rules to generate a qualified Ukrainian dictionary using *ru* g2p models and 21% using *bg* g2p models. *de* and *en* g2p models did not outperform *grapheme-based*. We assume that the dictionaries generated with *bg* and *ru* g2p models outperform our handcrafted dictionary since due to the properties of *bg* and *ru* some semi-palatalized phonemes get lost which may be less important for Ukrainian ASR. Thus we apply a special technique to model those phonemes for further experiments.

	# *Rules before g2p*	# *Rules after g2p*	PER (%)	WER (%)	# *Post-rules*	PER (%)	WER (%)
ru	43	56	12.4	22.80	57	1.7	**21.63**
bg	40	79	10.3	23.70	65	2.8	**22.09**
de	(68)*	66	32.7	27.10	39	28.6	26.36
en	(68)*	63	46.8	34.86	21	36.6	34.02

Table 5. Effort (# rules) and quality using cross-lingual rules.

7. SYSTEM OPTIMIZATION

7.1. Acoustic Modeling of Semi-Palatalized Phonemes

In addition to the fact that skipping some semi-palatalized phonemes in our cross-lingual dictionary generation experiments leads to ASR improvement, the auditory discrimination between semi-palatalized and non-palatalized phonemes is very small. To enhance the modeling of the 13 semi-palatalized phonemes, we therefore apply a data-driven phone modeling technique which had been successfully applied to the tonal vowels in Vietnamese and Hausa [24][30]. In this method the semi-palatalized and the non-palatalized variant of a phoneme share one base model. However, the information about the semi-palatalized articulation is added to the dictionary in form of a tag. Our Janus Recognition Toolkit [31] allows to use these tags as questions to be asked in the context decision tree when building context-dependent AMs. This way, the data decide during model clustering if the semi-palatalized articulation and the non-palatalized articulation have a similar impact on the basic phoneme. If so, the semi-palatalized and the non-palatalized variant of that basic phoneme would share one common model. In case the semi-palatalized articulation information is distinctive (of that phoneme and/or its context), the question about the semi-palatalized articulation information may result in a decision tree split, such that different variants of the same basic phonemes would end up being represented by different models. Tab. 6 shows that better performance can be obtained with our data-driven semi-palatalized phone modeling compared to modeling all semi-palatalized phonemes (*With semi-palatalized*) and excluding semi-palatalized articulation information (*Without semi-palatalized*).

Acoustic Modeling	WER (%) on dev
With semi-palatalized (baseline)	22.36
Without semi-palatalized	21.73
Data-driven Semi-Palatalized Phone Modeling	**21.65**
Grapheme-based	23.82

Table 6. Results with Semi-Palatalized Phonemes.

7.2. Language Model Improvement

By interpolating the individual LMs built from only 5 day long "snapshop" crawls of 3 further Ukrainian online newspapers (texts with 94M running words) and the TrainTRL, we created a new LM as in [3]. The interpolation weights were tuned on the dev set transcriptions by minimizing the PPL of the model. We increased the vocabulary of the LM by selecting frequent words from the additional text material which are not in the transcriptions. A 3-gram LM with a total of 40k words with a PPL of 373 and 0.53% OOV rate on the dev set performed best. It resulted in the lowest WER of 13.03% on the dev set and 11.21% on the test set with the system that also contains the data-driven semi-palatalized phone modeling.

8. CONCLUSION

We have described the rapid development of a Ukrainian LVCSR system. We collected 14 hours of speech from 119 Ukrainian speakers reading newspaper articles. After a rapid bootstrapping, based on a multilingual phone inventory, using RLAT, we improved the performance by investigating the peculiarities of Ukrainian. The initial recognition performance of 18.64% WER was improved to 11.21% on the test set. For the fast and cost-saving creation of the dictionary, we investigated strategies using g2p models derived from existing dictionaries of other languages, thereby reducing severely the necessary manual effort. We plan to investigate these strategies with other source and target languages.

9. REFERENCES

[1] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A Multilingual Text Speech Database in 20 Languages," in *ICASSP*, 2013.

[2] A. W. Black and T. Schultz, "Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing," in *ICASSP*, 2008.

[3] N. T. Vu, T. Schlippe, F. Kraus, and T. Schultz, "Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit," in *Interspeech*, 2010.

[4] O. Martirosian and M. Davel, "Error Analysis of a Public Domain Pronunciation Dictionary," in *PRASA*, 2007.

[5] T. Schlippe, S. Ochs, and T. Schultz, "Grapheme-to-Phoneme Model Generation for Indo-European Languages," in *ICASSP*, 2012.

[6] S. Besling, "Heuristical and Statistical Methods for Grapheme-to-Phoneme Conversion," in *Konvens*, 1994.

[7] A. W. Black, K. Lenzo, and V. Pagel, "Issues in Building General Letter to Sound Rules," in *ESCA Workshop on Speech Synthesis*, 1998.

[8] M. Bisani and H. Ney, "Joint-Sequence Models for Grapheme-to-Phoneme Conversion," *Speech Communication*, 2008.

[9] A. Ghoshal, M. Jansche, S. Khudanpurv, M. Riley, and M. Ulinski, "Web-derived Pronunciations," in *ICASSP*, 2009.

[10] T. Schlippe, S. Ochs, and T. Schultz, "Wiktionary as a Source for Automatic Pronunciation Extraction," in *Interspeech*, 2010.

[11] "Ukrainian Population Census 2001: Historical, Methodological, Social, Economic and Ethnic Aspects," 2001, `http://2001.ukrcensus.gov.ua`.

[12] Oleksandr Kramar, "Russification Via Bilingualism," *The Ukrainian Week*, 2012, `http://ukrainianweek.com/Society/47497`.

[13] "Ethnologue," `http://www.ethnologue.com`.

[14] A. Karpov, I. Kipyatkova, and A. Ronzhin, "Speech Recognition for East Slavic Languages: The Case of Russian," in *SLT-U*, 2012.

[15] T. Bilous, *IPA for Ukrainian*, 2005.

[16] S. N. Buk, J. Macutek, and A. A. Rovenchak, "Some Properties of the Ukrainian Writing System," *CoRR*, 2008.

[17] V. Pylypenko and V. Robeyko, "Experimental System of Computerized Stenographer for Ukrainian Speech," in *SPECOM*, 2009.

[18] V. Pylypenko, V. Robeiko, M. Sazhok, N. Vasylieva, and O. Radoutsky, "Ukrainian Broadcast Speech Corpus Development," in *SPECOM*, 2011.

[19] T. Lyudovyk, V. Robeiko, and V. Pylypenko, "Automatic Recognition of Spontaneous Ukrainian Speech based on the Ukrainian Broadcast Speech Corpus," in *Dialog'11 Conference*, 2011.

[20] V. Robeiko and M. Sazhok, "Real-time Spontaneous Ukrainian Speech Recognition System based on Word Acoustic Composite Models," in *UkrObraz*, 2012.

[21] M. Sazhok and V. Robeiko, "Bidirectional Text-To-Pronunciation Conversion with Word Stress Prediction for Ukrainian," in *UkrObraz*, 2012.

[22] S. Lytvynov and A. Prodeus, "Modeling of Ukrainian Speech Recognition System using HTK Tools," *Electronics and Communications*, vol. 1.

[23] D. Vazhenina and K. Markov, "Phoneme Set Selection for Russian Speech Recognition," in *NLPKE'11*, 2011.

[24] T. Schlippe, E. G. Komgang Djomgang, N. T. Vu, S. Ochs, and T. Schultz, "Hausa Large Vocabulary Continuous Speech Recognition," in *SLT-U*, 2012.

[25] East Central and South-East Europe Division of the United Nations Group of Experts on Geographical Names, "Romanization System in Ukraine," 2011.

[26] International Phonetic Association, *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*, Cambridge University Press, 1999.

[27] S. Kanthak and H. Ney, "Context-dependent Acoustic Modeling using Graphemes for large Vocabulary Speech Recognition," in *ICASSP*, 2002.

[28] M. Killer, S. Stueker, and T. Schultz, "Grapheme based Speech Recognition," in *Eurospeech*, 2003.

[29] S. Stueker and T. Schultz, "A Grapheme based Speech Recognition System for Russian," in *SPECOM*, 2004.

[30] N. T. Vu and T. Schultz, "Vietnamese Large Vocabulary Continuous Speech Recognition," in *ASRU*, 2009.

[31] H. Soltau, F. Metze, C. Fuegen, and A. Waibel, "A One Pass-Decoder Based On Polymorphic Linguistic Context Assignment," in *ASRU*, 2001.