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 function 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-phoneme 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-phoneme pairs found in the Web. However, we did neither possess Ukrainian word-phoneme pairs nor find those in sufficient 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 ы, ь, ы, ь, but some other letters such as е, і, ı, r 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 /h/ in Ukrainian. Like other Slavic languages Ukrainian has a rich morphology. Further peculiarities are the occurrence of palatalized consonants (e.g. ɲàd - ɲ/ɜd/) [15], the existence of long geminates as in Polish (e.g. ężań - /ɛɲaɯ/), the use of the apostrophe similar to the Russian hard sign, and the affricates /dz/ and /dʒ/ that are not represented by separate letters but by the digraphs ǳ and ʤk. [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 conversion 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% 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 ELRA1. We create a dictionary au-

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. Mymy, 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.

Table 1. List of crawled Ukrainian Websites.

<table>
<thead>
<tr>
<th>umoloda.kiev.ua</th>
<th>day.kiev.ua</th>
<th>ukurier.com.ua</th>
</tr>
</thead>
<tbody>
<tr>
<td>pravda.com.ua</td>
<td>chornomorka.com</td>
<td>tsn.ua</td>
</tr>
<tr>
<td>champion.com.ua</td>
<td>ukrsovo.org.ua</td>
<td>epravda.com.ua</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Set</th>
<th>Male</th>
<th>Female</th>
<th>#utterances</th>
<th>#tokens</th>
<th>Duration</th>
</tr>
</thead>
<tbody>
<tr>
<td>train</td>
<td>38</td>
<td>61</td>
<td>11k</td>
<td>69k</td>
<td>11 h 45 mins</td>
</tr>
<tr>
<td>dev</td>
<td>4</td>
<td>6</td>
<td>1k</td>
<td>7k</td>
<td>1 h 14 mins</td>
</tr>
<tr>
<td>test</td>
<td>4</td>
<td>6</td>
<td>1k</td>
<td>7k</td>
<td>1 h 08 mins</td>
</tr>
<tr>
<td>Total</td>
<td>46</td>
<td>73</td>
<td>13k</td>
<td>83k</td>
<td>14 h 07 mins</td>
</tr>
</tbody>
</table>

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 Cepstral Coefficients (MFCCs) and state-of-the-art techniques for acoustic modeling to rapidly build a baseline recognizer for Ukrainians. For our context-dependent AMs with different context sizes, we stopped the decision tree splitting process at 2k quint-phones. 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.

6. CROSS-LINGUAL DICTIONARY PRODUCTION

The production of dictionaries can be costly in terms of time and money if no word-phonunciation 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 CMUdict2. 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.

<table>
<thead>
<tr>
<th>Language</th>
<th>Grapheme coverage</th>
<th>Phoneme coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russian (ru)</td>
<td>88%</td>
<td>57%</td>
</tr>
<tr>
<td>Bulgarian (bg)</td>
<td>88%</td>
<td>67%</td>
</tr>
<tr>
<td>German* (de)</td>
<td>0%</td>
<td>39%</td>
</tr>
<tr>
<td>English* (en)</td>
<td>0%</td>
<td>37%</td>
</tr>
</tbody>
</table>

Table 3. Language relationship to Ukrainian.


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)

<table>
<thead>
<tr>
<th>Step</th>
<th>ru</th>
<th>bg</th>
<th>de</th>
<th>en</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>6ir</td>
<td>6ir</td>
<td>bih</td>
<td>bih</td>
</tr>
<tr>
<td>2</td>
<td>u_b u_i u_a_g</td>
<td>b_g b_i g</td>
<td>d_e_d_e_i</td>
<td>a_n_e_n_i</td>
</tr>
<tr>
<td>3</td>
<td>u_b u_a u_i u_h</td>
<td>u_b u_a u_i u_h</td>
<td>u_a u_a_y</td>
<td>u_b u_a_y</td>
</tr>
<tr>
<td>4</td>
<td>u_a_b_j u_i u_h</td>
<td>u_b_i u_i u_h</td>
<td>u_b_i u_i</td>
<td>u_b u_a_y</td>
</tr>
</tbody>
</table>

Table 4. Cross-lingual pronunciation production for 6ir.

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

2http://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 įaj (running) after each step of our cross-lingual dictionary generation strategy. The correct pronunciation in the handcrafted dictionary is _ua_by _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.

![Table 5. Effort (# rules) and quality using cross-lingual rules.](image)

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).

![Table 6. Results with Semi-Palatalized Phonemes.](image)

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


