Rapid Bootstrapping of a Ukrainian Large Vocabulary Continuous Speech Recognition System

Tim Schlippe, Mykola Volovyk, Kateryna Yurchenko, Tanja Schultz
tim.schlippe@kit.edu

Cognitive Systems Lab

Karlsruhe Institute of Technology

1. Overview

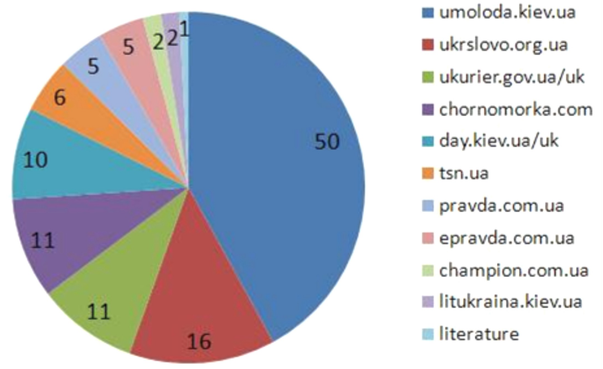
Goal of this work

- Collect speech and text data in the Ukraine for the East Slavic language Ukrainian as a part of the *GlobalPhone* corpus (*Schultz et al., 2013*) with our Rapid Language Adaptation Toolkit (RLAT) (*Vu et al., 2010*)
- Develop a Ukrainian LVCSR system rapidly
- Use grapheme-to-phoneme models derived from existing dictionaries of other languages to reduce necessary manual effort for dictionary generation
- Apply state-of-the-art techniques for acoustic modeling and our day-wise text collection and language model interpolation strategy (*Vu et al., 2010*)

2. Ukrainian Resources

Text Corpus

- Text crawled from online newspapers using RLAT
- Complemented fragments from Ukrainian literature and lyrics
- Applied language-dependent text normalization
- Selected prompts to record speech data for the training, development, and evaluation set
- Training set text used for the language model



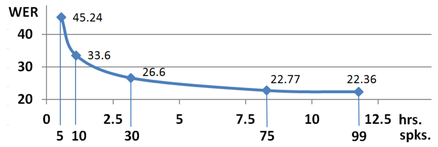
Text sources and the numbers of speakers reading prompts from them

Speech Corpus

- Speech data collection in *GlobalPhone* style (*Schultz, 2013*), i.e. we asked Ukrainian speakers to read prompted sentences of newspaper articles

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

3. Baseline ASR System



- 51 phonemes
  - 38 basic phonemes set for Ukrainian (*Buk, 2008*) (*Bilous, 2005*)
  - 13 semi-palatalized consonants *(Pylypenko, 2008) (Lytvynov, 2009) (Liudovyk, 2011)*
- Pronunciation dictionary – manually generated with 882 simple "search and replace rules" from *(Buk, 2008)*
- Bootstrapping with RLAT using multilingual phone inventory *MM7 (Schultz and Waibel, 2001)*
- Preprocessing: 143 MFCC (adjacent frames) → 42 (LDA)
- 11.75 hours from 99 speakers to train acoustic models
- Fully-continuous 3-state left-to-right HMM, Context-dependent AM: decision tree splitting stopped at 500 quinphones
- 3-gram language model: Vocabulary size=7.4k, PPL=594, OOV rate=3.6%
- Word error rate on development set: **22.36%**, on evaluation set: **18.64%**

4. Cross-lingual Dictionary Production

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 grapheme-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

Cross-lingual pronunciation production for биг

	# *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

Effort (# rules) and quality using cross-lingual rules

5. System Optimization

Acoustic Modeling of Semi-Palatalized Phonemes

- Information about semi-palatalized articulation is added as tag to pronunciation dictionary
- Tag added as question in clustering procedure
- Data decide during model clustering on the impact on the basic phoneme
- If no impact → share 1 common model; Otherwise → split models

Language Model Improvement

- 5 day long "snapshop" crawls of 3 further Ukrainian online newspapers (texts with 94M running words)
- Interpolation of the individual LMs based on minimizing the PPL of the model on the dev set transcriptions
- 3-gram LM with a total of 40k words (PPL=373, OOV Rate=0.53%) performed best
- Word error rate on development set: **13.03%**, on evaluation set: **11.21%**

	WER (%) on dev	WER (%) on test
Acoustic Modeling of Semi-Palatalized Phonemes		
With semi-palatalized (baseline)	22.36	18.64
Without semi-palatalized	21.73	
Data-driven Semi-Palatalized Phone Modeling	**21.65**	
Grapheme-based	23.82	
Language Model Improvement		
5 day "snapshop" and vocabulary increase	**13.03**	**11.21**