Grapheme-to-Phoneme Model Generation for Indo-European Languages

Tim Schlippe, Sebastian Ochs, Tanja Schultz
tim.schlippe@kit.edu

Cognitive Systems Lab

Karlsruhe Institute of Technology

1. Overview

Motivation

- Quality of pronunciation dictionary is important for Speech Recognition
- g2p models might be of different quality depending on training data

Goal of Work

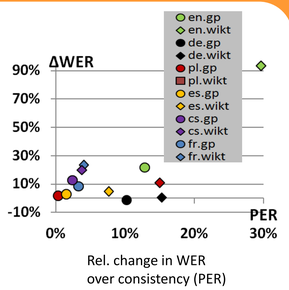
- Creation of pronunciation dictionaries for new languages and domains rapidly and economically based on statistical grapheme-to-phoneme (g2p) models

Goals of this particular study

- Comparison of g2p models [Bisani and Ney, 2008] between:
  - Languages: English (en), German (de), Polish (pl), Spanish (es), Czech (cs), French (fr)
  - Different training data quality:
    1. *GlobalPhone* word-pronunciation pairs (successfully applied to LVCSR): ***GP***
    2. *Wiktionary* word-pronunciation pairs (provided by Internet community): ***wikt***
- Evaluation criteria:
  - **Consistency check** (with Phoneme Error Rate (PER))
    Generalization ability of the g2p models
    - Consistency within each pronunciation dictionary
    - Comparison to validated *GlobalPhone* pronunciation dictionary
  - **Complexity check**
    g2p model sizes (number of non-pruned 6-grams plus their backoff scores)
  - **Automatic Speech Recognition (ASR) performance**
    Word error rate using pronunciations generated with the g2p models

3. Evaluation of g2p Models: ASR Performance

	GlobalPhone (base form)	GlobalPhone g2p (1-best)	Wiktionary g2p (1-best)	GlobalPhone (*GP*) Consistency (PER)	Wikt. (*wiktOnGP* / (*wikt*)) Consistency (PER)	
cs	15.59	17.58	18.72	2.41	3.75	(4.47)
de	16.71	16.50	16.81	10.21	15.27	(7.74)
en	14.92	18.15	28.86	12.83	29.65	(34.44)
es	12.25	12.59	12.82	1.99	7.63	(10.78)
fr	20.91	22.68	25.79	3.28	4.02	(3.77)
pl	15.51	15.78	17.21	0.36	15.02	(4.86)



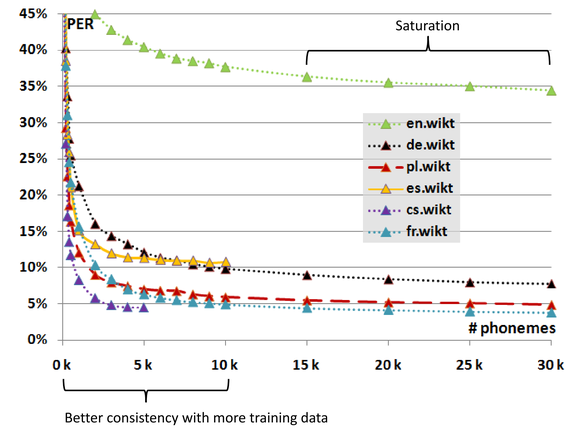
- Use *GP* and wikt *g2p* models trained with 30k phoneme tokens and corresp. graphemes to reflect saturated g2p model consistency (5k and 10k for cs and es)
- Replace pronunciations in dictionaries of *GlobalPhone* ASR systems with pronunciations generated with g2p models
- Reasonable performance degradations given the cost and time efficient generation process

2. Evaluation of g2p Models: Consistency and Complexity

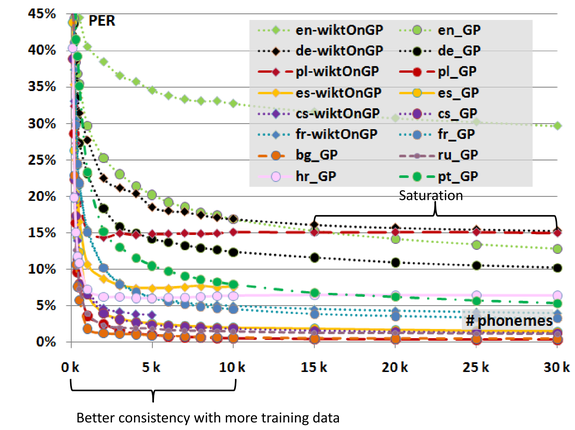
	Train	Test
GP	GlobalPhone	GlobalPhone
wikt	Wiktionary	Wiktionary
wiktOnGP	Wiktionary	GlobalPhone

Consistency check setup.

Consistency of *Wikt*



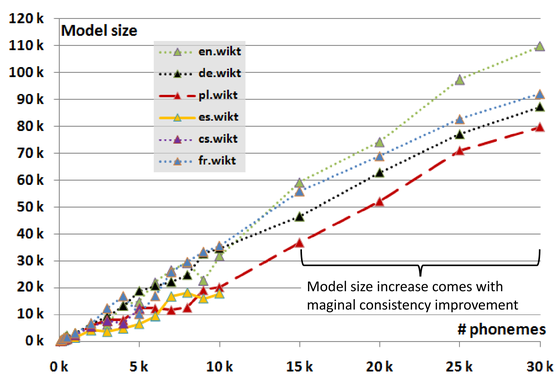
Consistency of *GP*



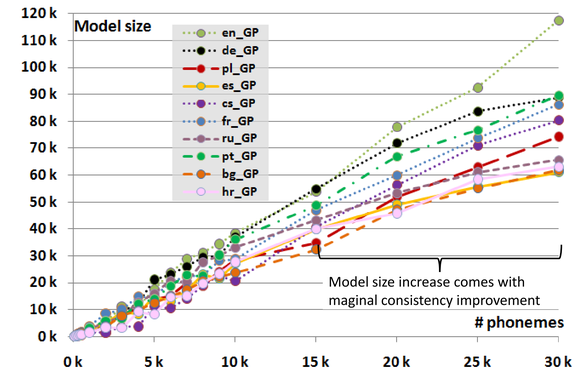
- 6-fold cross validation for consistency and complexity check, evaluation on 30% of respective dictionary
- Standard deviation in consistency less than 1% PER with only 1k phoneme tokens (with corresp. graphemes) (Trend to smaller deviations with more training material)
- *GP* consistency: Large range of PER (pl, bg, cs, es, ru < fr, hr, pt, de < en)
- PER varies with amount of training data betw. 100 and 10k phoneme tokens (with corresponding graphemes)
- 15k phoneme tokens necessary for reasonable results per language,
- When automatically creating pronunciations based on *Wiktionary* (trained with only 5k phoneme tokens)
  - Czech (PER 3.7%): each 27th phoneme
  - French (PER 6.4%): each 16th phoneme
  - Spanish (PER 7.6%): each 13th phoneme

  to be changed to meet the validated quality of *GlobalPhone*

***Wikt* g2p model complexity**



***GP* g2p model complexity**



- Model complexity keeps increasing for larger amounts of data but this has minor impact on quality