GLOBALPHONE: A MULTILINGUAL TEXT & SPEECH DATABASE IN 20 LANGUAGES

Tanja Schultz, Ngoc Thang Vu, Tim Schlippe
tanja.schultz@kit.edu

Cognitive Systems Lab

Karlsruhe Institute of Technology

1. Overview

- GlobalPhone: multilingual database with high quality read speech with transcriptions and pronunciation dictionaries in 20 languages
- More than 400 hours transcribed audio data, 2000 native speakers
- Excellent basis for multilingual speech processing (Speech Recognition, Speech Synthesis, Speaker ID, Language ID)
- Benchmark numbers and language models available

Links: http://csl.ira.uka.de/GlobalPhone

3. Rapid Language Adaptation Toolkit (RLAT)

Provides **innovative methods** and **interactive web-based tools** to:

- Develop speech processing components for new languages at low cost
- Continuously harvest, normalize, and process text data from web
- Create prompts for recordings, create vocabulary lists
- Select appropriate phone sets for new languages efficiently
- Automatically generate pronunciation dictionaries
- Iteratively build and evaluate speech processing components

Links: http://csl.ira.uka.de/rlat-dev

2. The GlobalPhone Corpus

Language Coverage

- Covers Europe, Africa, America, Asia
- Large variety of language peculiarities relevant for Speech & Language processing
  - Phonetic characteristics, phonotactics
  - Writing systems, word segmentation
  - Morphological variations

Data Acquisition

- Recorded in countries where the language is officially spoken, native speakers
- Two batches: 1996 – 1997, 2003 – 2013
- Read speech, National news articles (web)
- Quiet environmental conditions
- High quality, Close-speaking microphone
- 16bit PCM encoding, 16kHz, mono quality
- Validated transcripts
- Information on speakers, sessions, etc.

Language	Training [hrs:min]	Development [hrs:min]	Evaluation [hrs:min]
Arabic	12:00	TBA	TBA
Bulgarian	16:47	2:16	1:56
Croatian	11:48	2:02	1:45
Czech	26:49	2:22	2:41
French	24:55	TBA	2:01
German	14:54	1:57	1:28
Hausa	6:36	1:02	1:06
Japanese	21:51	1:26	1:40
Korean	16:34	2:09	2:04
Mandarin	26:38	1:59	2:25
Portuguese	22:45	1:38	1:47
Polish	18:39	2:47	2:16
Russian	21:08	2:41	2:36
Shanghai	9:50	TBA	TBA
Spanish	17:35	1:40	2:03
Swedish	17:39	2:03	1:58
Tamil	15:50	1:04	1:00
Thai	19:05	2:03	1:58
Turkish	13:04	1:57	1:53
Ukrainian	11:32	1:13	1:07
Vietnamese	22:15	1:40	1:30
Total	368:14	33:59	35:14

4. GlobalPhone Dictionaries

Languages	#Phones	#Words	#Dict entries
Bulgarian	44	275k	275k
Croatian	32	21k	23k
Czech	41	277k	277k
French	39	122k	195k
German	43	39k	41k
Hausa	33	43k	48k
Japanese	31	9k	13k
Korean	39	1.3k	3k
Mandarin	49	73k	73k
Portuguese	45	59k	59k
Polish	36	34k	34k
Russian	47	39k	40k
Spanish	42	31k	39k
Swedish	48	25k	25k
Tamil	41	288k	292k
Thai	44	23k	25k
Turkish	31	34k	34k
Ukrainian	51	40k	40k
Vietnamese	38	30k	39k

5. GlobalPhone Language Models

Language *	3-gram PPL LM-BM	3-gram PPL LM	OOV [%]	#Vocab	#Tokens
Bulgarian (w)	454	351	1.0	274k	405M
Croatian (w)	721	647	3.6	362k	331M
Czech (w)	1421	1361	4.0	267k	508M
French (w)	324	284	2.4	65k	-
German (w)	672	555	0.3	38k	20M
Hausa (w)	97	77	0.5	41k	15M
Japanese (s)	89	76	1.0	67k	1600M
Korean (c)	25	18	0	1.3k	500M
Mandarin (c)	262	163	0.8	13k	900M
Portuguese (w)	58	49	9.8	62k	11M
Polish (w)	951	904	0.8	243k	224M
Russian (w)	1310	1150	3.9	293k	334M
Spanish (w)	154	108	0.1	19k	12M
Swedish (w)	423	387	5.3	73k	211M
Tamil (s)	730	624	1.0	288k	91M
Thai (s)	70	65	0.1	22k	15M
Turkish (w)	53	45	13.2	29k	7M
Ukrainian (w)	594	373	0.5	40k	94M
Vietnamese (s)	218	176	0	30k	39M

* Word-based units (w), syllable-based (s), character-based (c)

6. Speech Recognition Systems

- State alignments using multilingual phone inventories
- Bottle Neck features trained with multilingual (12) MLP
- Traditional 3-state HMMs
- 500 – 3,000 quintphones
- Data-driven tone modeling (Hausa, Mandarin, Thai, and Vietnamese)
- Vowel length modeling (Hausa)
- Error Rates (word/syllable) on 19 languages