DNN-Based Speech Recognition for GlobalPhone Languages

By
Martha Yifiru Tachbelie, Ayimunishagu Abulimiti, Solomon Teferra Abate, Tanja Schultz

@
Cognitive Systems Lab, University of Bremen, Germany

ICASSP2020 Barcelona

Outline

- Motivation
- The Corpus
  - Languages Covered
  - Data Acquisition
- Pronunciation Dictionaries
- Language Models
- Acoustic Models
- Speech Recognition Results
- Conclusions

Motivation

- There are more than 7000 languages in the world
- However, large-scale data resources for research are available for about 100 languages only due to various challenges:
  - costs for collection
  - lack of language conventions
  - lack of a standardized writing system
  - difficulty to find experts with both technical background and native language expertise
- In 2002, we released a multilingual text and speech corpus, GlobalPhone (GP), for 15 languages to address the lack of databases
- Later, GP is extended to more than 20 languages

Motivation

- The status of GP and GMM based reference benchmark Automatic Speech Recognition (ASR) system performances of 20 languages was provided (Schultz et al., 2013)
  - does not reflect current state-of-the-art performances based on recent developments in Deep Neural Networks (DNN)
- The current paper intends:
  - to provide new reference benchmarks for GP based on hybrid HMM-DNN
  - In addition, four Ethiopian languages (Amharic, Tigrigna, Oromo and Wolaytta) and Uyghur are considered
    - very similar to GP in terms of speaking style (read), number of speakers and size of speech

The Corpus

- GP is a multilingual database of high-quality read speech with corresponding transcriptions and pronunciation dictionaries

Languages Covered in GP

- Area – languages from most continents of the world
  - Europe
  - North and South America
  - Asia
  - African
- Phonetic characteristics
  - tonal sounds, consonantal clusters
  - nasals, plosive sounds,
  - uvular and palatalized sounds
- Writing systems
  - logographic, phonographic segmental,
  - phonographic consonantal, phonographic syllabic,
  - abugida/ethiopic, linear nonfeatural, and phonographic featural scripts
- Morphological variations
  - agglutinative, compounding, and non-concatinative root-pattern morphology

Data Acquisition in GP

- Texts were selected from national newspaper articles
  - report national and international political news, as well as economic news
- The recording was performed in countries where the language is officially spoken
- About 100 adult native speakers were asked to read about 100 sentences, for each language
- Data were recorded with a close-speaking microphone and is available in identical characteristics for all languages:
  - PCM encoding, mono quality, 16bit quantization, and 16kHz sampling rate
- Recordings were done in ordinary rooms or offices, in the majority without background noise
- GP contains over 400 hours of speech spoken by more than 2000 native adult speakers
  - divided into speaker disjoint sets for training (80%), development (10%), and evaluation (10%)

Data Acquisition in GP

- Training Speech size in hh:mm

Languages	Training Speech		Languages	Training Speech
Amharic	20:00		Portuguese	18:06
Bulgarian	16:48		Russian	21:00
Croatian	11:48		Spanish	17:30
Czech	26:42		Swedish	17:42
French	21:54		Thai	11:36
German	14:54		Tigrigna	22:06
Hausa	6:36		Turkish	13:12
Japanese	30:46		Ukrainian	10:42
Mandarin	26:42		Uyghur	12:24
Oromo	22:48		Vietnamese	20:48
Polish	19:18		Wollayta	29:42

Pronunciation Dictionaries (PDs)

- Phone-based pronunciation dictionaries are available for each GP language
  - cover the words which appear in the training transcriptions
  - constructed in a rule-based manner using language specific phone sets

Language Models (LMs)

- We used the GP language models available at
  - https://www.csl.uni-bremen.de/GlobalPhone/
- For the five newly added languages:
  - different sizes of text corpus obtained from the web, except for Wolaytta
    - the training transcription has been used
  - we developed trigram language models using SRILM

PDs and LMs

Languages	#Phones	#PD Vocab	OOV	LMToken	PPL
Amharic	40	310k	3.06	4M	41.2
Bulgarian	44	275k	1.07	405M	341.62
Croatian	32	23k	2.09	331M	934.75
Czech	41	277k	4.04	508M	1223.5
French	38	122k	6.028	220M	356.87
German	43	39k	0.059	20M	675.86
Hausa	33	43k	0.32	15M	76.63
Japanese	31	58k	0.18	1600M	89.41
Mandarin	49	73k	0	900M	268.06
Oromo	59	21k	11.73	1.2M	266.17
Portuguese	45	59k	1.09	11M	45.8
Polish	36	49k	0.1	224M	880.83
Russian	47	40k	2.09	334M	1070.74
Spanish	42	43k	4.65	12M	113.44
Swedish	48	25k	0	211M	325.91
Thai	44	23k	0.22	15M	16.64
Tigrigna	44	299k	4.9	4M	211.41
Turkish	31	34k	1.25	7M	55.04
Ukrainian	49	40k	0.0002	94M	105.76
Uyghur	37	40k	13.9	250k	260.59
Vietnamese	59	39k	3.17	39M	1227.01
Wolaytta	57	25k	9.34	226k	254.9

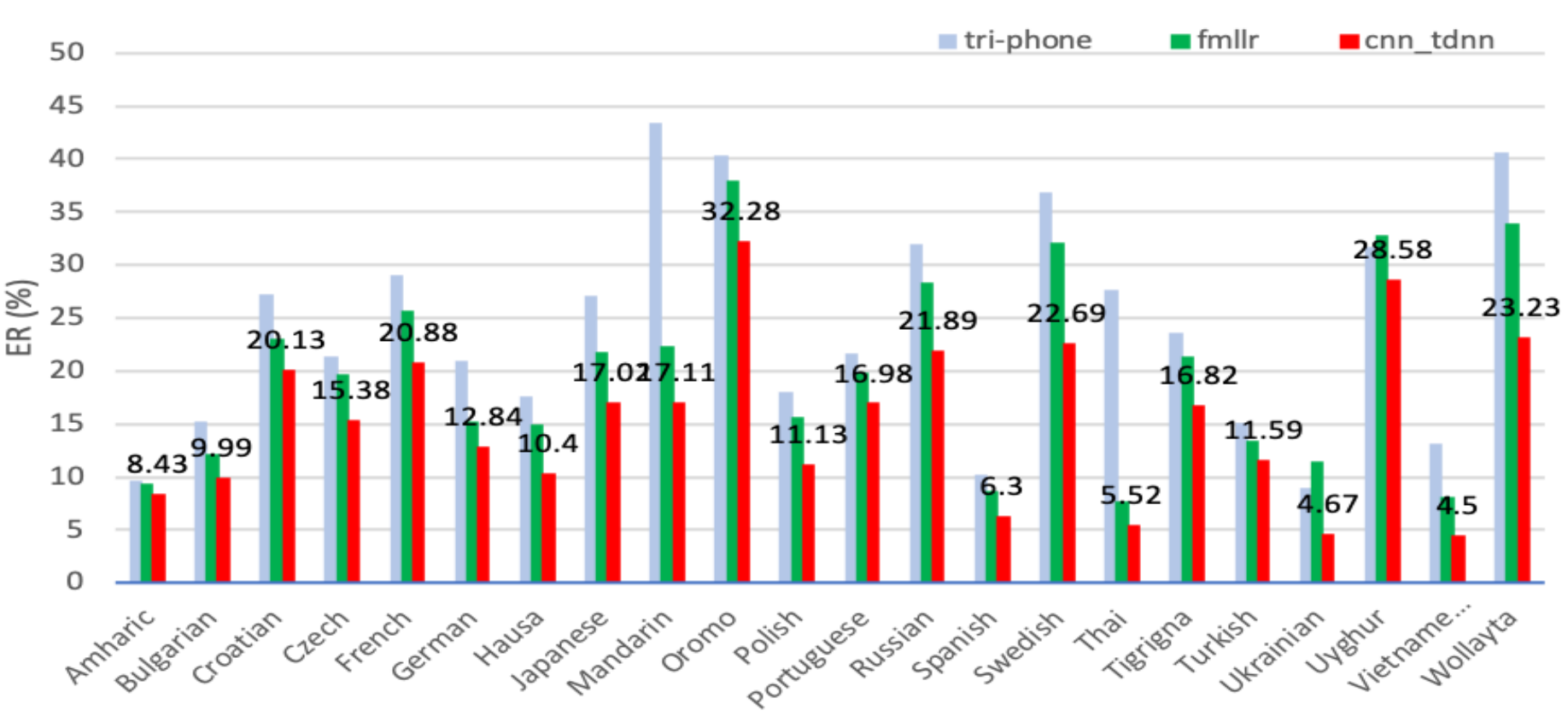
Acoustic Models

- HMM-GMM based context dependent acoustic model is developed for each language
  - The acoustic model uses a fully-continuous 3-state left-to-right HMM
  - Speaker Adaptive Training (SAT) has been done using an affine transform, fMLLR
- The best model of each language, which is mostly the fMMLR, is used to obtain alignments for DNN training
- DNN architecture and hyper-parameters used:
  - Factored Time Delay Neural Networks with additional Convolutional layers (CNN-TDNNf)
  - 15 hidden layers (6 CNN followed by 9 TDNNf) and a rank reduction layer
  - Trained for 7 epochs
- Results are reported in word, syllable and character error rate
- Kaldi ASR toolkit is used to built ASR systems

Speech Recognition Results

Cognitive Systems Lab

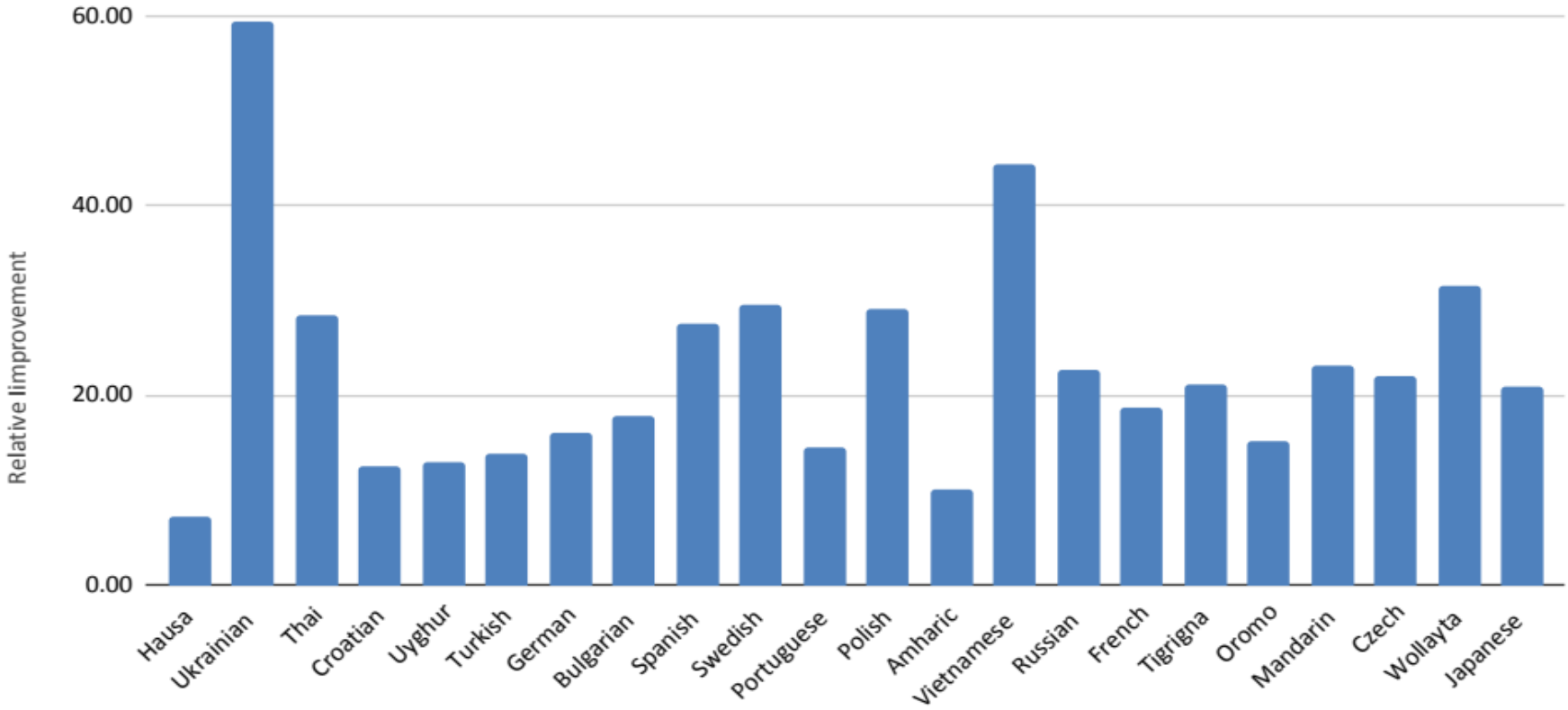
- Error rates of ASR systems



- The numbers are for DNN-based ASR systems

Results

- Relative Improvement DNN systems over GMM



- Languages sorted based on training speech size

Conclusions

- We presented reference benchmark ASR system performances based on hybrid HMM-DNN for 22 languages
- Regardless of the training speech size, error rate reduction has been obtained using DNN
- Average relative error rate reduction of 22.69% has been achieved
    - Relative error rate reduction are between 7.14% and 59.43%