The GoVivace Multilingual Automatic Speech Recognition System For Low-Resource Indian Languages

Systems Submitted to MUCS 2021, Subtask - 1

Nagendra Kumar Goel, Mousmita Sarma, S Moris, Zikra Iqbal, Supreet

Go-Vivace Inc.
McLean, VA, USA

12 August, 2021

Outline

Background

Notable Highlights

- The core acoustic models are developed using lattice-free MMI based DNN approach.
- We have primarily experimented two set ups with :
  1. A combined phone set and
  2. A pooled phone set
- We have experimented two DNN architectures:
  1. Factorized TDNN
  2. Multi-stream CNN-TDNN with spectral augmentation
- Language Model is trained using all training text and external text downloaded from internet through bootcat.
- The best submitted system achieved a Word Error Rate (WER) of 23.39% on the test set and 25.53% on the blind test set.

GoVivace at the pace of technology

Background

Significant Systems Submitted for blind test scoring

1. TDNN-F with pooled phone set and 4 gram LM trained on the train text provided by the organizer.
2. TDNN-F with pooled phone set and 4 gram LM trained on the train text provided by the organizer and external text downloaded from the internet using BootCaT[1].

[1] https://bootcat.dipintra.it/

System Descriptions

Reference Lexicon and OOV

- Collected a set of most frequent words for each of the languages in the range of 2000 to 5000 words.
- Created individual lexicon for each languages using the Indic TTS-unified parser [2].
- Phonetisaurus[3] Grapheme to Phoneme (G2P) has been trained for each language and derive pronunciations for out of vocabulary (OOV) words extracted from the training text.
- Combined phone set approach: A total of 66 non silence phones are used.
- Pooled phone set approach : 300 non silence phones are used in this training to build the tree.
- 4 silence phones (silence, laughter, non-speech noise and spoken noise).

[2] https://www.iitm.ac.in/donlab/tts/unified.php

[3] Novak et al., Phonetisaurus: Exploring grapheme-to-phoneme conversion with joint n-gram models in the WFST framework, Natural Language Engineering , Volume 22 , Issue 6 , November 2016 , pp. 907-938

System Descriptions

LF-MMI Based DNN Training

- Phone alignment lattices for DNN training are generated from triphone based speaker-adaptive GMM acoustic models.
- A Factorized Time delay deep neural network (TDNN-F) configuration based Deep Neural Network (DNN) acoustic model trained with lattice-free MMI objective function [4].
- In-domain data provided by the challenge organizer are used for training.
- 40 dimensional high resolution MFCCs and speaker discriminative I-vector features.
- 5-way speed perturbation and volume perturbation is performed on the training data.

[4] D. Povey et al., "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI", Proceedings of INTERSPEECH, 2016.

System Descriptions

Decoding

- A 4-gram language model is trained using the train text for six languages provided by the organizers as well as external text downloaded from internet using BootCaT.
- The final decoding graph has been created using this LM and Minimum Bayes Risk (MBR) Decoding has been performed to derive the hypothesis transcript
- All our experiments are performed using kaldi toolkit.

Results

Table: Word Error Rate (WER) % on test set

System	Hindi	Marathi	Oriya	Tamil	Telugu	Gujarati	Average
BASELINE	40.41	22.44	39.06	33.35	30.62	19.27	30.85
Combined phone + 4gram LM	37.17	21.77	38.92	32.81	29.62	19.72	30.00
Pooled phone tdnn-f + 4gram LM	28.8	18.17	36.33	31.42	27.96	17.58	26.71
+ external text	22.88	17.97	31.22	27.89	24.95	15.48	23.39
Pooled phone Multistream tdnn-f + 4gram LM + external text	27.29	21.08	33.71	31.33	28.29	17.91	26.60

Results

Table: Word Error Rate (WER) % on blind test set

System	Hindi	Marathi	Oriya	Tamil	Telugu	Gujarati	Average
BASELINE	37.2	29.04	38.46	34.09	31.44	26.15	32.73
Pooled phone							
+ 4gram LM	25.93	28.45	33.73	31.99	28.69	23.97	28.79
+ external text	21.77	25.73	29.05	28.92	26.5	21.22	25.53

Thank you