Baseline Acoustic Models for Brazilian Portuguese Using Kaldi Tools

Cassio T. Batista, Ana Larissa da S. Dias, Nelson C. Sampaio Neto
Federal University of Pará, Institute of Exact and Natural Sciences – Belém, Brazil
{cassiotb,nelsonneto}@ufpa.br, ana.dias@itec.ufpa.br

Introduction

- Kaldi's new state-of-the-art hybrid approach (HMM-DNN) for ASR acoustic modeling has been successfully applied to many languages
  - English, German, Spanish, Italian, Arabic, ...
- Brazilian Portuguese (BP): the FalaBrasil group stands out in LVCSR using CMU Sphinx and HTK toolkits, however with HMM-GMM only

Problem

- Related works lack of solutions for BP using deep-learning techniques for LVCSR
- Apparently, no previous work has attempted to develop an HMM-DNN hybrid recipe for Brazilian Portuguese using Kaldi tools [1]

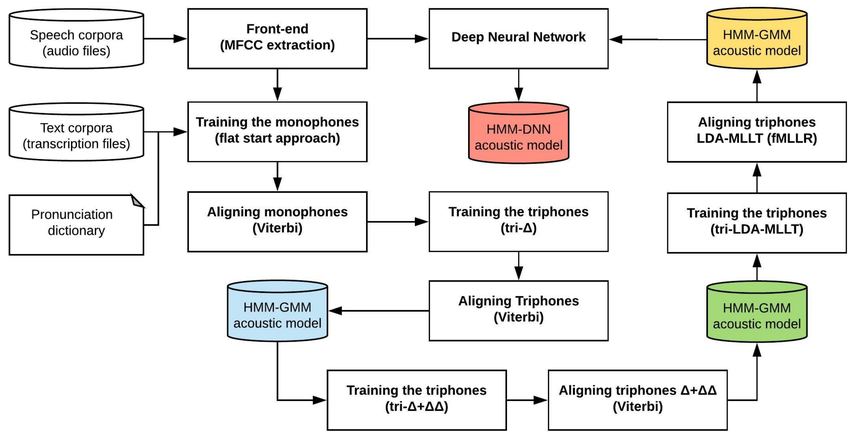
Solution

- A recipe adapted from the WSJ scripts is now publicly available at FalaBrasil group's GitLab repo (`https://gitlab.com/falabrasil`)

Building ASR Systems for BP

- Comparison between Kaldi and CMU Sphinx
- Audio corpora: ≈ 171h, 16 kHz mono WAV
- Language model: 3-gram trained with SRILM over CETENFolha text corpora (pp. ≈ 170)
- Phonetic dictionary: FalaBrasil's G2P software

Kaldi's Hybrid HMM-DNN Acoustic Model Training Pipeline



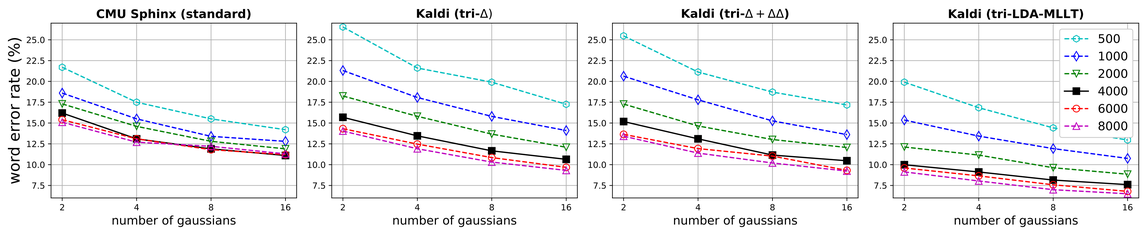
Experimental Tests Setup

- Tied-states (leaves or senones): 500 up to 8,000
- Gaussians per mixture (densities): 2 up to 16
- HMM-GMM triphone-based AMs were trained with both Kaldi and CMU Sphinx toolkits
- Re-estimation: CMU Sphinx uses Baum-Welch algorithm while Kaldi performs Viterbi training (which includes Viterbi alignment at each step)
- HMM-DNN hybrid model was trained on the top of Kaldi's best HMM-GMM model
- Hardware: HP EliteDesk 800 with Intel® Core™ i5-4570 CPU, 8 GB RAM and 1 TB disk storage

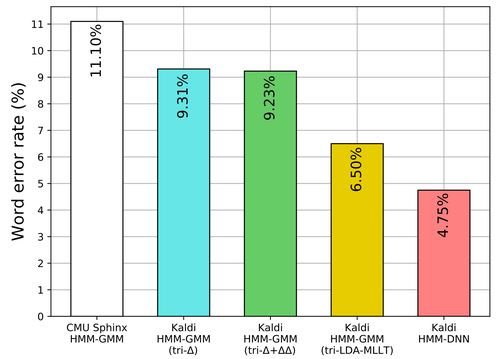
DNN Tools and Parameters

Tool or Param.	Value
DNN codebase	`nnet2` ("Dan's DNN")
Script	`train_pnorm_fast.sh`
Hidden layers	2
Activation function	p_norm
`pnorm_output_dim`	3,000
`pnorm_input_dim`	300
`num_epochs`	8
`num_epochs_extra`	5
Minibatch size	512
Learning rate	0.02 down to 0.004

Word Error Rate (WER) Results: Tied-states vs. Gaussians



Best WERs Comparison



Conclusions

- CMU Sphinx loses to Kaldi in WER even when comparing HMM-GMM models only, probably because of different re-estimation procedures (Baum-Welch vs. Viterbi)
- 57.21% improvement showed by Kaldi's HMM-DNN acoustic model over the best CMU Sphinx's HMM-GMM model

Acknowledgements

CAPES
FALA BRASIL

Reference

[1] D. Povey, A. Ghoshal, G. Boulianne, N. Goel, M. Hannemann, Y. Qian, P. Schwarz, and G. Stemmer, "The kaldi speech recognition toolkit," in *In IEEE 2011 workshop*, 2011.