Uso de Redes Neurais Profundas e Recorrentes Para Reconhecimento de Fala em Português Brasileiro

João Victor da Silva Dias Canavarro

Instituto de Ciências Exatas e Naturais
Faculdade de Computação
Laboratório de Visualização, Interação e Sistemas Inteligentes
Orientador: Prof. Dr. Nelson Cruz Sampaio Neto

10 de Outubro de 2020

Introdução

- Interações convencionais:
  - Teclado, *mouse*, *touchscreen*, *joystick*, controle remoto, etc.
- Interações não convencionais:
  - acionadores externos, **reconhecimento automático de voz**, etc.

Reconhecimento de Voz

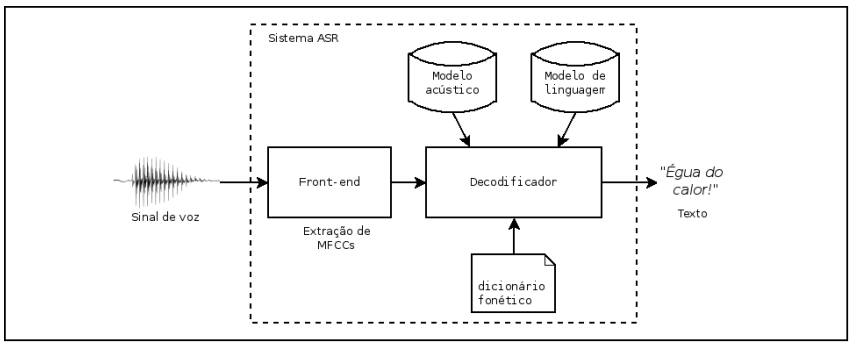


Figura: Esquema tradicional de um sistema automático de reconhecimento de voz

- Aplicações:
  - Tecnologias assistivas, *eye-trackers*, linguística (**alinhamento fonético**), síntese de voz

Ferramentas Utilizados

- Kaldi: um pacote *open-source* de reconhecimento de voz
  - Possui suporte para ambas HMM-GMMs (mistura de gaussianas) e HMM-DNNS (redes neurais profundas)
  - Suporte para PT-BR

- Praat: *software* utilizado por linguistas na análise da fala

Objetivos

- Desenvolver um sistema de reconhecimento de voz para PT_BR utilizando o pacote Kaldi para treinamento dos AMs e LMs
- Disponibilizar recursos à comunidade científica

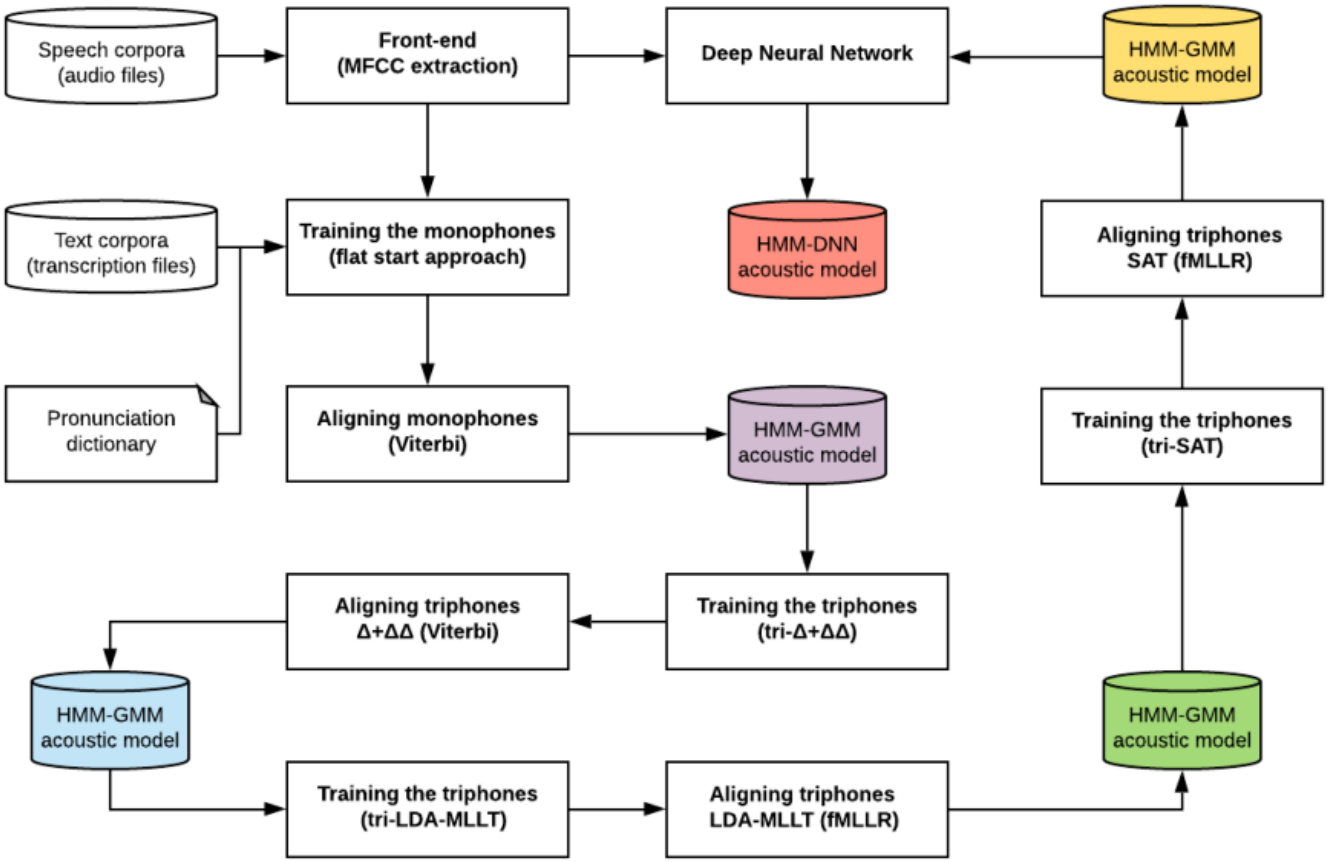
Metodologia

Base de Áudio Transcrita

Dataset	Ref.	Horas	Palavras	Oradores
LapsStory	[7]	5h:18m	8,257	5
LapsBenchmark	[7]	0h:54m	2,731	35
Constituição	[8]	8h:58m	5,330	1
Defesa do Consumidor	[8]	1h:25m	2,003	1
Spoltech LDC	[9]	4h:19m	1,145	475
West Point LDC	[10]	5h:22m	484	70
CETUC	[11]	144h:39m	3,528	101
Total		170h:51m	14,518	687

Metodologia

Treinamento dos AMs



Testes

- *Dataset* de Avaliação
  - 200 enunciados falados por um orador masculino
  - Total de 7min58s de áudio alinhado manualmente
- Característica comparada: limite fonético
  - Diferença entre o tempo final da ocorrÃªncia do fonema em ambos alinhamentos, pelo alinhador e alinhado manualmente

Resultados

	Tolerância (ms)			
Ferramenta / Modelo	< 10 (%)	< 25 (%)	< 50 (%)	< 100 (%)
HTK [20]	33.95	65.73	86.40	96.54
Kaldi monophones	45.57	83.89	**96.71**	99.39
Kaldi triphones	**48.36**	**85.35**	**96.71**	**99.71**
Kaldi triphones LDA-MLLT	47.66	83.82	96.53	**99.71**
Kaldi triphones SAT	46.62	83.03	96.08	99.55
Kaldi DNN	46.49	82.65	96.15	99.66

Figura: Distribuição cumulativa dos limites fonéticos

Resultados

Ferramenta / Modelo	μ (ms)	mediana (ms)	σ
HTK [20]	26.043	15.961	32.378
Kaldi monophones	15.233	11.196	16.327
Kaldi triphones	**14.438**	**10.357**	15.178
Kaldi triphones LDA-MLLT	14.726	10.577	**15.095**
Kaldi triphones SAT	15.359	10.834	16.314
Kaldi DNN	15.306	10.904	15.864

Figura: Média, mediana e desvio padrão dos alinhadores avaliados, em comparação ao alinhamento manual

Conclusão

- Os modelos acústicos treinados utilizando o Kaldi obtiveram resultados superiores à outros com suporte à língua portuguesa, e tão satisfatórios quanto modelos para outras línguas
- Desenvolvimento de uma interface para utilização do alinhador
- Fatores positivos:
  - Avanços na área de reconhecimento de voz para PT-BR
  - Disponibilização dos recursos desenvolvidos: `https://ufpafalabrasil.gitlab.io/`

Agradecimentos

UNIVERSIDADE FEDERAL DO PARÁ

PROPESP
Pró-Reitoria de Pesquisa
e Pós-Graduação | UFPA

Uso de Redes Neurais Profundas e Recorrentes Para Reconhecimento de Fala em Português Brasileiro

Obrigado!

João Canavarro (jvcanavarro@ufpa.br)
Universidade Federal do Pará (UFPA)
Belém – Pará – Brasil