PyTASRAF
FrameWork em *Python* para Treinamento e Avaliação de Sistemas de Reconhecimento Automático de Fala

Autor: Érico Cretton Andrade[a] Orientadores: Prof. Dr. Fábio Violaro[b] e Dr. Edmilson Morais[c]

(a) FEEC/UNICAMP, (b) FEEC/UNICAMP, (c) FEEC/UNICAMP e VOCALIZE – Soluções em Tecnologias da Fala e da Linguagem

Introdução

As primeiras pesquisas na área de Reconhecimento Automático de Fala (RAF) datam da década de 50. Entretanto, somente nos últimos anos essa tecnologia atingiu a qualidade necessária para o desenvolvimento dos primeiros sistemas comerciais. A elevada disponibilidade de *corpora* de texto e fala, o desenvolvimento de novos algoritmos e o grande poder de processamento dos computadores atuais, explicam porque esses avanços se tornaram possíveis.

Objetivos

Desenvolvimento de um ambiente integrado para treinamento e avaliação de sistemas de reconhecimento automático de fala (RAF). O software desenvolvido, denominado pyTASRAF, foi desenvolvido em linguagem *Python* e faz uso das ferramentas HTK (*Hidden Markov Model ToolKit*) da Universidade de Cambridge, UK e SRILM (*the* SRI *Language Modeling ToolKit*) pelo STAR (*Speech Technology Research Lab.*) da Universidade de Stanford, US.

Fundamentos

O objetivo de um sistema RAF consiste em estimar, durante a etapa de treinamento, e utilizar durante a etapa de reconhecimento, a função probabilística $P(M \mid \mathbf{X}, \Theta)$

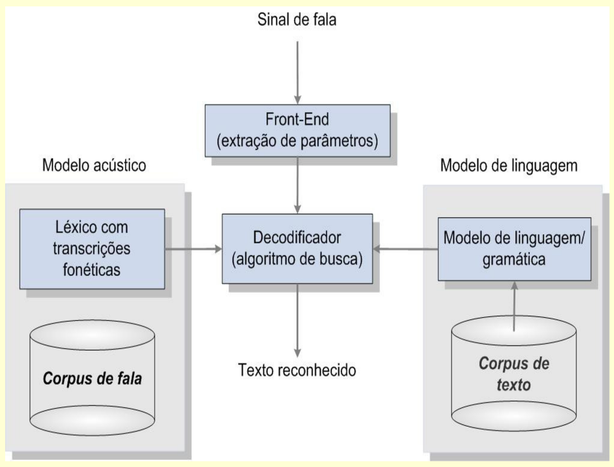
$$\hat{M} = \underset{M}{\operatorname{argmax}} P(M \mid \mathbf{X}, \Theta) = \underset{M}{\operatorname{argmax}} \frac{p(\mathbf{X} \mid M, \Theta) \cdot P(M \mid \Theta)}{p(\mathbf{X} \mid \Theta)}$$

sendo $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$ uma seqüência de vetores acústicos derivados do sinal a ser reconhecido, $s(n)$, através de um procedimento de **Extração de características**. Uma vez que a sentença pode ser construída a partir da concatenação de palavras $M = \{W_1, W_2, ..., W_N\}$, a tarefa de um sistema RAF também pode ser interpretada como a determinação da seqüência de palavras mais prováveis $\hat{M}$, dada a seqüência de vetores acústicos $\mathbf{X}$ e o conjunto de parâmetros Θ.

O termo $p(\mathbf{X} \mid \Theta)$ independe de M e, portanto, não necessita ser calculado. O termo $p(\mathbf{X} \mid M, \Theta)$ é denominado Modelo Acústico e o termo $p(M \mid \Theta)$ é denominado Modelo de Linguagem.

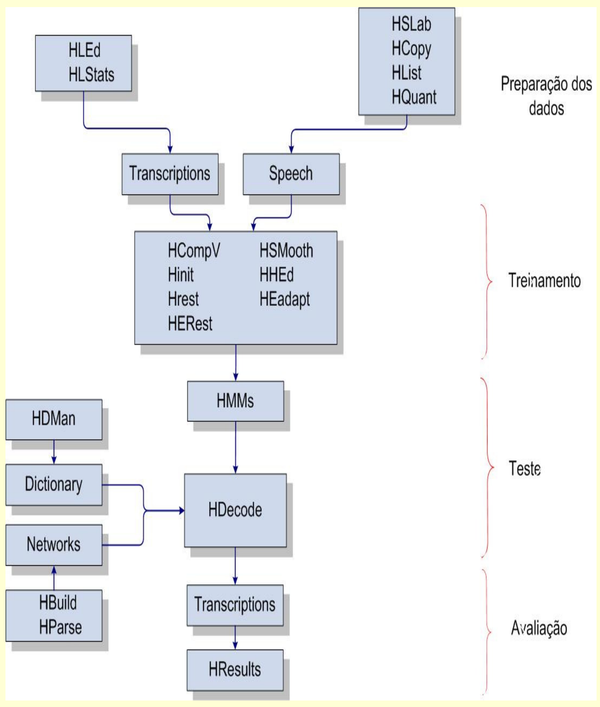
A determinação de $\hat{M}$ é realizada através de um algoritmo de busca e esse processo é denominado Decodificação.

Portanto, um sistema RAF pode ser dividido em quatro módulos principais: (1) Extração de características, (2) Modelagem Acústica, (3) Modelo de Linguagem e (4) Decodificação. A figura a seguir ilustra estes quatro módulos.



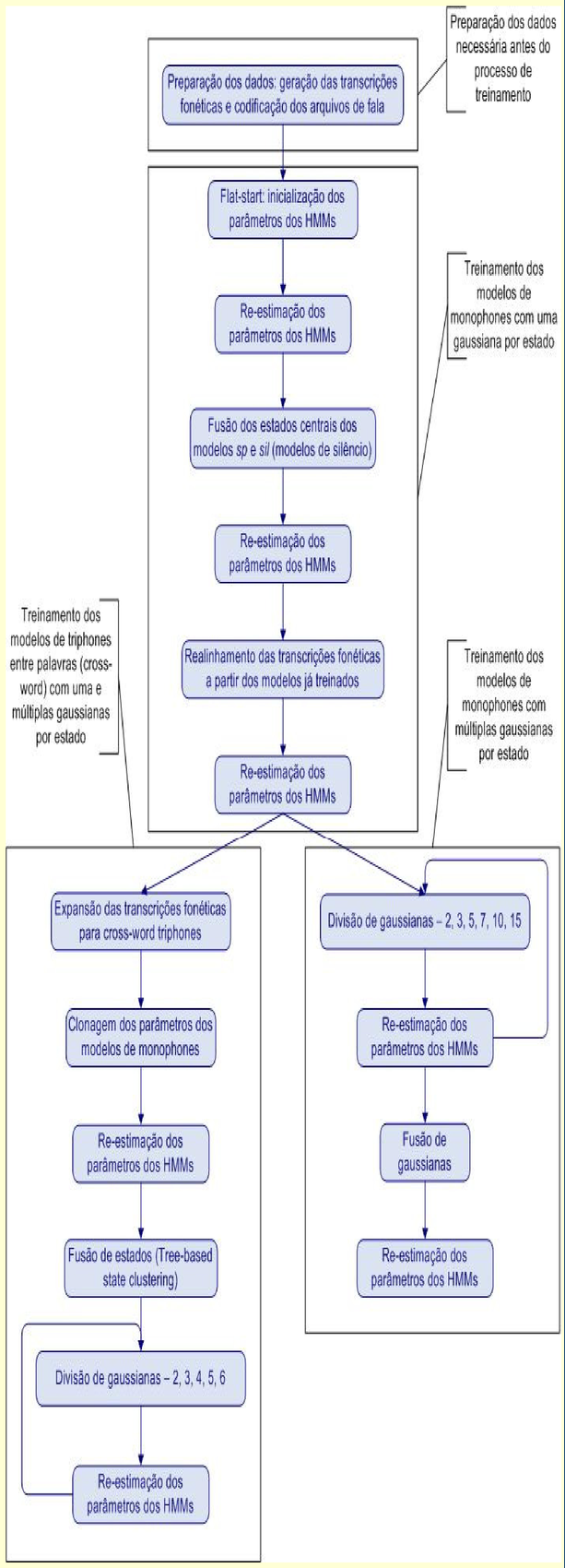
HTK

Ferramentas disponíveis no *Hidden Markov Model ToolKit* para a construção dos quatro módulos principais de um sistema RAF.



pyTASRAF

Framework desenvolvido em *Python* para treinamento e avaliação de sistemas de reconhecimento automático de fala.



Experimentos

Treinamento do Modelo Acústico

Corpus de fala com 2700 sentenças e nove locutores (300 sentenças por locutor)

Treinamento do modelo de linguagem

Corpus com 390.000 sentenças

Decodificação

Léxico com 64000 palavras

Resultados

Taxa de acerto de palavras: 91% utilizando modelos *cross-word triphones*.

Tempo médio de reconhecimento por palavra: 5 segundos.

Considerações Finais

Com o HTK, foi possível executar vários experimentos com diferentes propostas de treinamento dos modelos para um sistema RAF.

A utilização do *Python* facilitou a integração dos diversos módulos do HTK, por se tratar de uma linguagem de alto nível e de fácil entendimento.

Agradecimentos

À VOCALIZE – Soluções em Tecnologia da Fala e da Linguagem, pelos fundamentos teórico/práticos, pela infraestrutura e pelo material necessário para a realização desse trabalho.

Ao Professor Fábio Violaro pela orientação e pela atenção dispensada em todas as fases da pesquisa que culminou no desenvolvimento deste *framework*.