Reconhecimento de Fala em Português Brasileiro

Fabiano Weimar dos Santos
xiru@xiru.org

Eu sou...

- “Apenas um rapaz latino americano...”
- Mestre em Computação pela UFRGS (2009), Bacharel em Computação pela UCS (2004)
- Envolvido com IA desde 1999
- Consultor e Desenvolvedor Python, Zope e Plone desde 2000
  - Core-developer do Plone
  - Contribuidor em diversos Plone Products
- Sysadmin do provedor PyTown.com

Roteiro

- Introdução
- O que é Reconhecimento de Fala
- Como funciona?
  - E o idioma Português Brasileiro...
- Dicas: como implantar
- Jabá...

Introdução

- Nossa interface natural com o mundo é a fala (não um teclado e mouse)
- A área de reconhecimento de fala é pesquisada desde a década de 80
- Não é uma tarefa simples!
- Reconhecimento de Fala = Processamento de Sinais + Fonética + Linguística Computacional + Inteligência Artificial

O que é Reconhecimento de Fala

- Não é "reconhecimento de voz"
- Preocupada em reconhecer o que está sendo dito
- Reconhecer quem está falando é uma outra área: reconhecimento de locutor
- Reconhecimento de **fala contínua sem restrições** é ainda um problema em aberto

Síntese x Reconhecimento

- Todo leigo confunde:
  - Síntese: Conversão de Texto para "Fala"
  - Reconhecimento: Conversão de Fala para "Texto"
- Se comparada ao reconhecimento, a síntese é trivial

Aplicações

- Para Câmaras: Apoio na Taquigrafia
- Indexação e busca de conteúdo no que é dito em comissões em plenário, TV e rádio
- URA

Como Funciona?

Respire fundo...

Fundamentação Teórica

Consideremos A como a representação de um evento acústico; W como uma string de n palavras. Se P(W|A) denota a probabilidade que as palavras W foram faladas, dado os eventos acústicos A observados, então o reconhecedor deve decidir em favor de uma palavra W que satisfaça

$$W_{max} = \text{argmax}_{w} P(W|A)$$

Fundamentação Teórica

Podemos reescrever a equação como P(W|A) = (P(W)P(A|W)) / P(A), onde P(W) é a probabilidade da palavra W ser dita, P(A|W) é a probabilidade que quando a palavra W é dita o evento acústico A será observado e P(A) é a probabilidade de A ser observado

Fundamentação Teórica

Como a maximização é feita com a variável
A fixa (pois o evento acústico observado é determinado), temos:

$$W_{max} = \text{argmax}_w \, P(W)P(A|W)$$

Essa fórmula define que processos um reconhecedor de fala deve solucionar...

Fundamentação Teórica

- Um Modelo Acústico para calcular P(A|W)
  - **HMM**, Rede Neural, DTW
- Um Modelo de Linguagem para calcular P(W)
  - **n-gram**, CFG
- Busca de hipótese para W_{max}
  - Não trivial, pois o espaço de busca é muito grande e tempo real pode ser "desejável"

Fundamentação Teórica

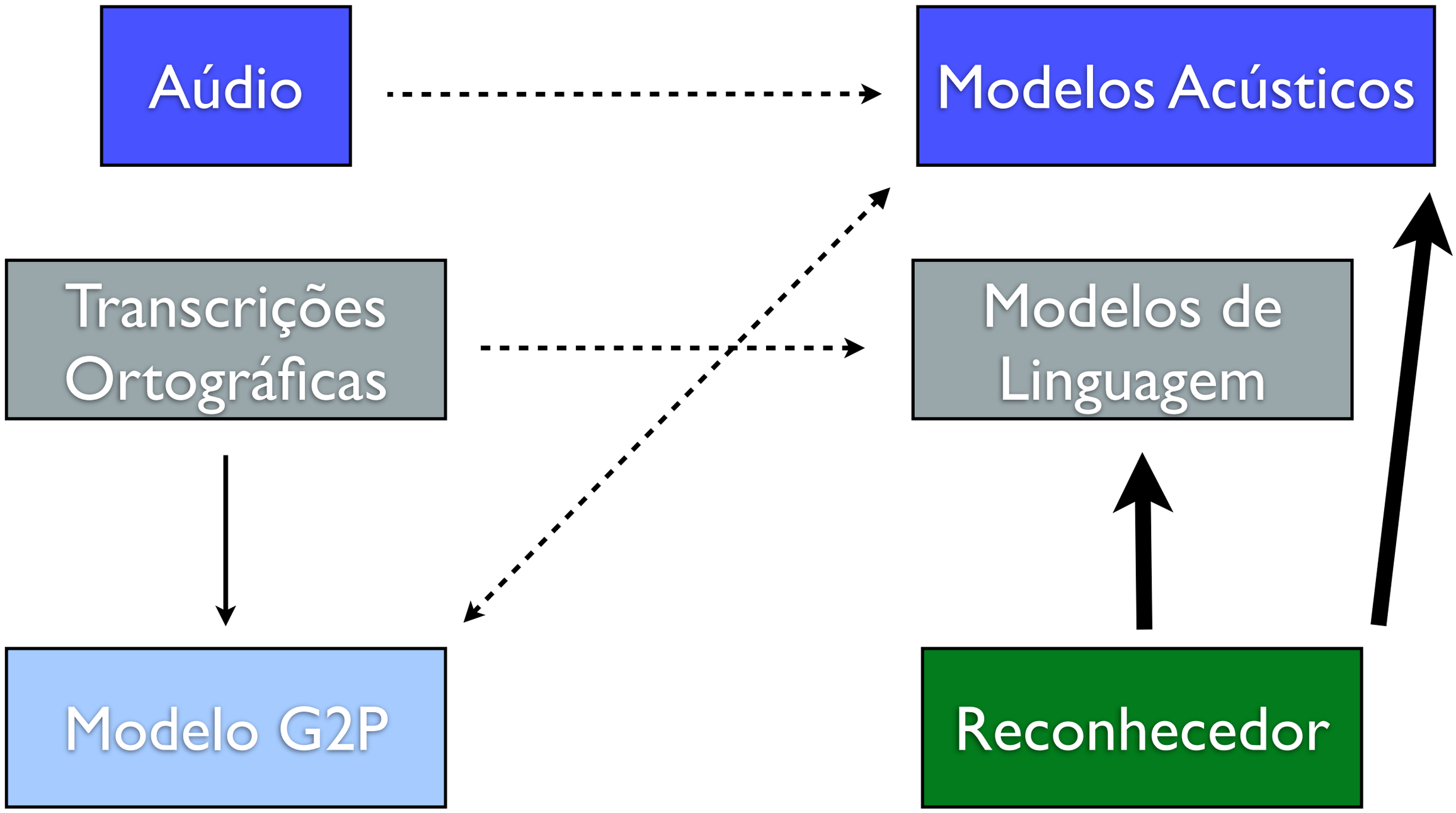
- Um Modelo Acústico para calcular P(A|W)
  - **HMM**, Rede Neural, DTW
- Um Modelo de Linguagem para calcular P(W)
  - **n-gram**, CFG
- Busca de hipótese para W_{max}
  - Não trivial, pois o espaço
    grande e tempo real pode

Trifones "tied state" com múltiplas Gaussianas

Fundamentação Teórica

- Um Modelo Acústico para calcular P(A|W)
  - **HMM**, Rede Neural, DTW
- Um Modelo de Linguagem para calcular P(W)
  - **n-gram**, CFG
- Busca de hipótese para W_{max}
  - Não trivial, pois o espa
    grande e tempo real po

Modelos estatísticos baseados em 3-gram suavizados por algorítmos de desconto e interpolação


Aúdio
Modelos Acústicos
Transcrições Ortográficas
Modelos de Linguagem
Modelo G2P
Reconhecedor

Reconhecedores

- HTK (Hidden Markov Model Toolkit)
- Sphinx (CMU Sphinx Open Source Speech Recognition Engine)
- Julius (Open-Source Large Vocabulary CSR Engine Julius)

Linguística Computacional

- Autenticidade
- Adequação
- Representatividade
- Extensão

Modelos de Linguagem

- Geralmente adota-se modelos estatísticos de linguagem, baseados em n-grams
- Quanto maior a ordem do n-gram, melhor é a representação do contexto, mas mais esparso torna-se o espaço de busca
- Necessidade de grandes quantidades de dados para a criação de modelos representativos

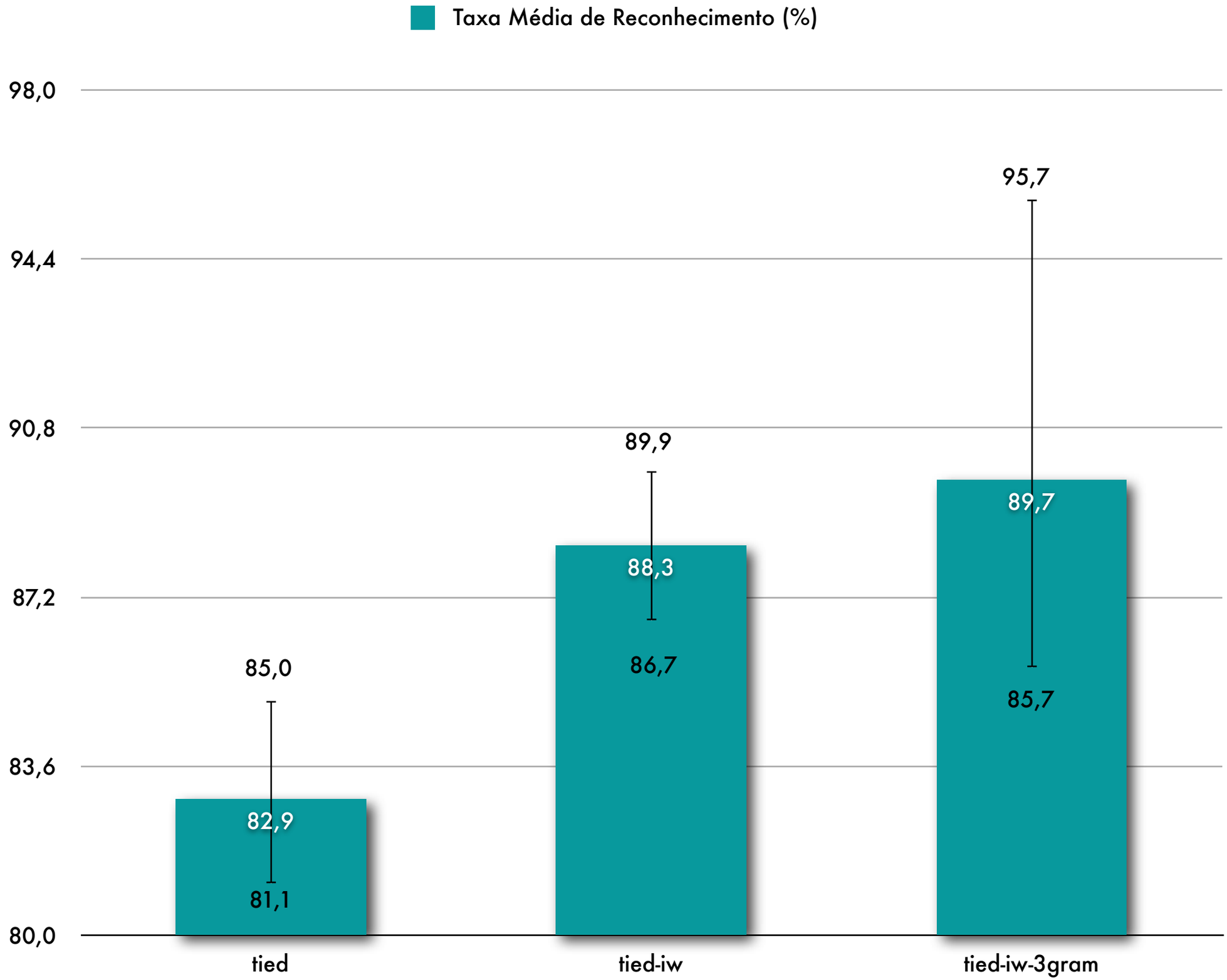
Modelos de Linguagem

- Representatividade limitada pela extensão
  - Modelos de linguagem pequenos são eficientes, mas tem aplicabilidade restrita
  - Modelos de linguagem com grande vocabulário tem maior aplicabilidade, mas são menos eficientes (alta perplexidade)
  - Modelos de linguagem realmente grandes (teóricos) são eficientes e "irrestritos", mas são difíceis de manter (limitações computacionais)

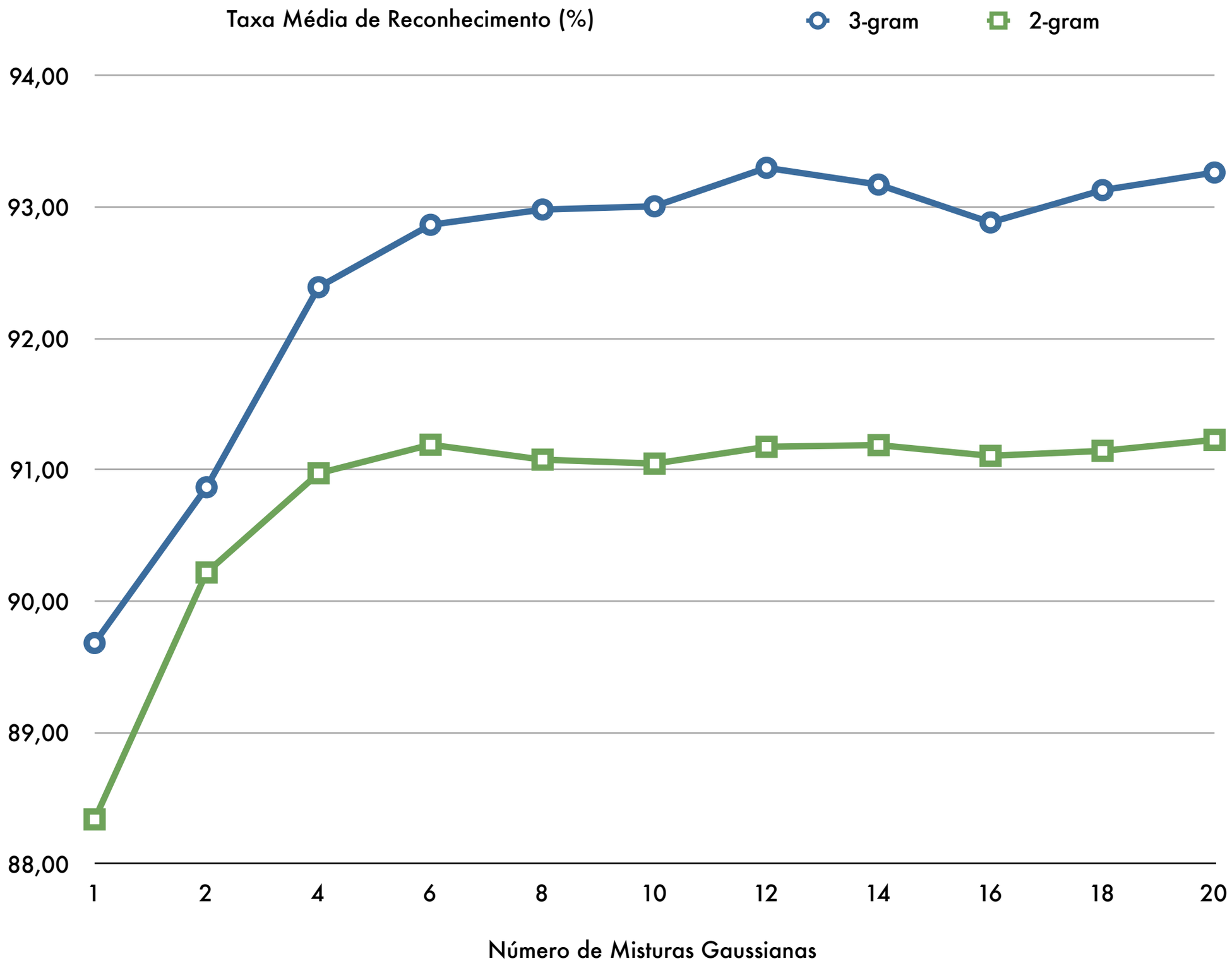
Modelos Acústicos

- Em condições ideais, usa-se todo o texto disponível para criar os modelos de linguagem e os respectivos áudios para criar os modelos acústicos
- Geralmente isso não é tão simples pois:
  - A quantidade de áudio costuma ser limitada
  - A aquisição de corpus textuais costuma ser mais "simples"

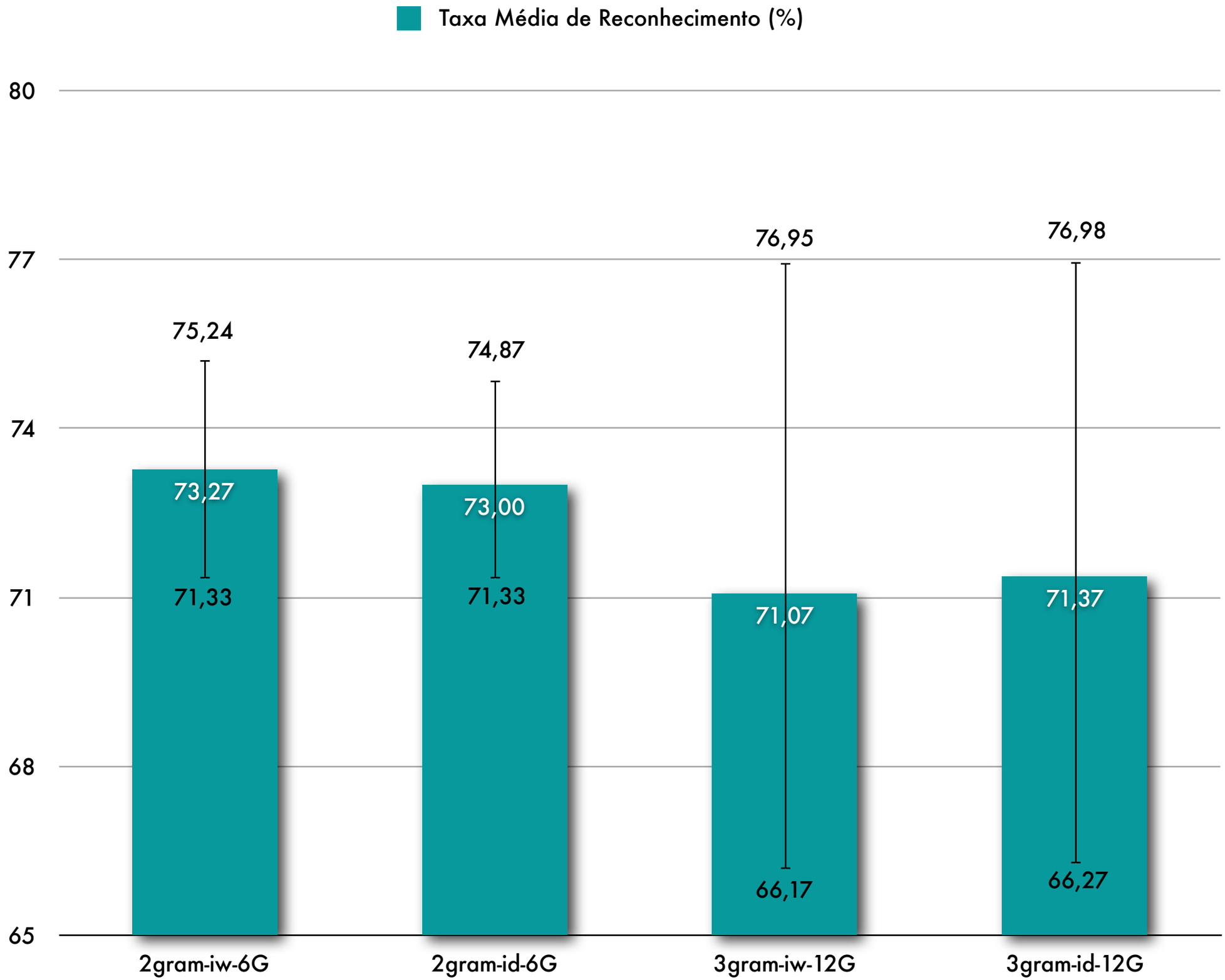
Modelo Acústico - Trifones “tied state”



Múltiplas Misturas Gaussianas



Grande Vocabulário - CETEN-Folha



Dicas: como implantar

- Organize seus dados
  - Armazene o áudio de tudo que é dito
  - Armazene as transcrições ortográficas
- Cuidado com soluções prontas
- Sistema de Reconhecimento deve aprender com erros (e não apenas confiar na adaptação de locutor)

Dicas: como implantar

Não seja otimista demais.
Bons resultados requerem muito trabalho (e tempo).

Jabá...

Obrigado!

Fabiano Weimar dos Santos
xiru@xiru.org