Síntese de Voz em Tempo Real

Eduardo Tenório
embat@cin.ufpe.br
embatbr@gmail.com

Introdução

- Produz voz humana **artificialmente**

Introdução

- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)

Introdução

- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)
- Onde usar?

Introdução

- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)
- Onde usar?
  - Leitura de tela para cegos

Introdução

- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)
- Onde usar?
  - Leitura de tela para cegos
  - Pessoas com problemas de fala (Hawking)

Introdução

- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)
- Onde usar?
  - Leitura de tela para cegos
  - Pessoas com problemas de fala (Hawking)
  - Interface humano-máquina (Google Glass, Siri...)

Introdução

- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)
- Onde usar?
  - Leitura de tela para cegos
  - Pessoas com problemas de fala (Hawking)
  - Interface humano-máquina (Google Glass, Siri...)
  - Atores virtuais

Introdução

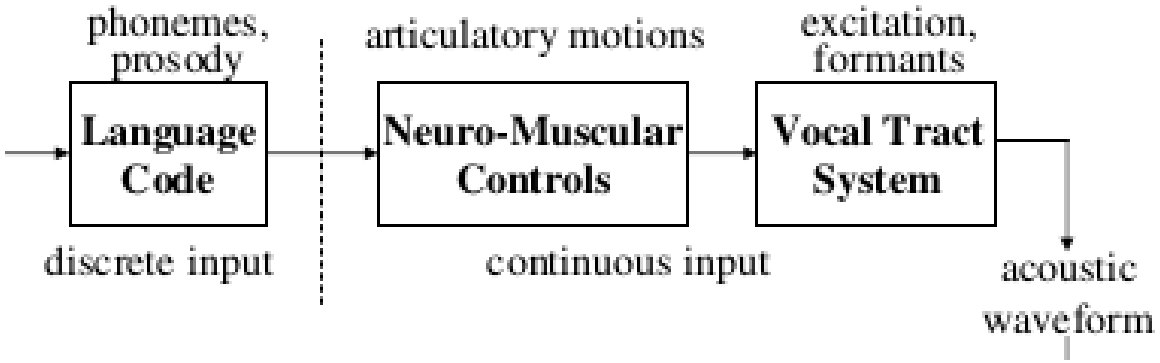
- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)
- Onde usar?
  - Leitura de tela para cegos
  - Pessoas com problemas de fala (Hawking)
  - Interface humano-máquina (Google Glass, Siri...)
  - Atores virtuais
  - *[more options]*

Introdução

- Produz voz humana **artificialmente**
- Síntese de Voz → Síntese de Voz via Texto (**TTS**)
- Onde usar?
  - Leitura de tela para cegos
  - Pessoas com problemas de fala (Hawking)
  - Interface humano-máquina (Google Glass, Siri...)
  - Atores virtuais
  - *[more options]*
- Síntese de Voz via **Interface Cerebral** (recente)

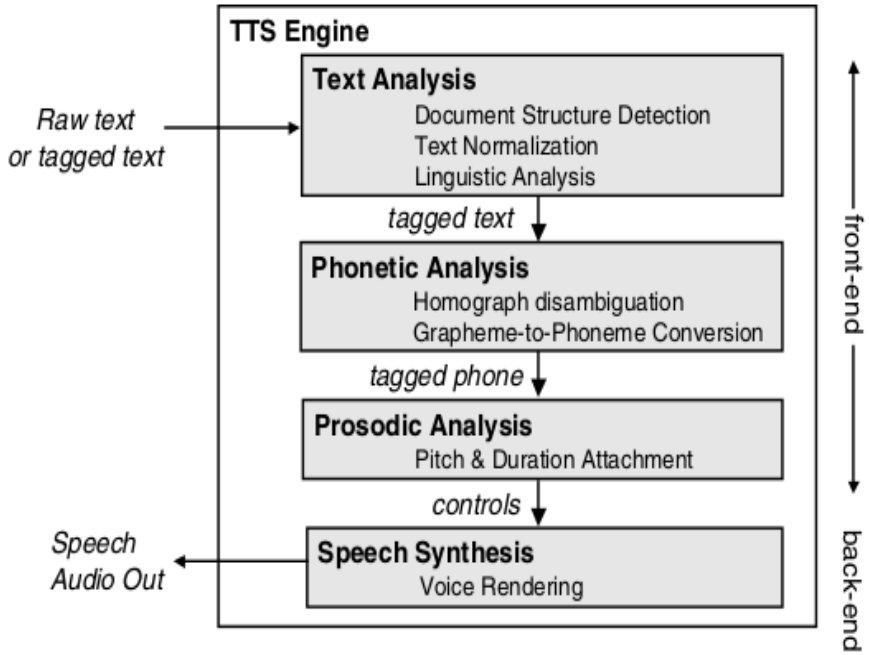
Text-To-Speech Synthesis

- Um sistema TTS **simula** parte da *speech chain:*
  - Codificação da linguagem
  - Controles neuro-musculares
  - Trato vocal



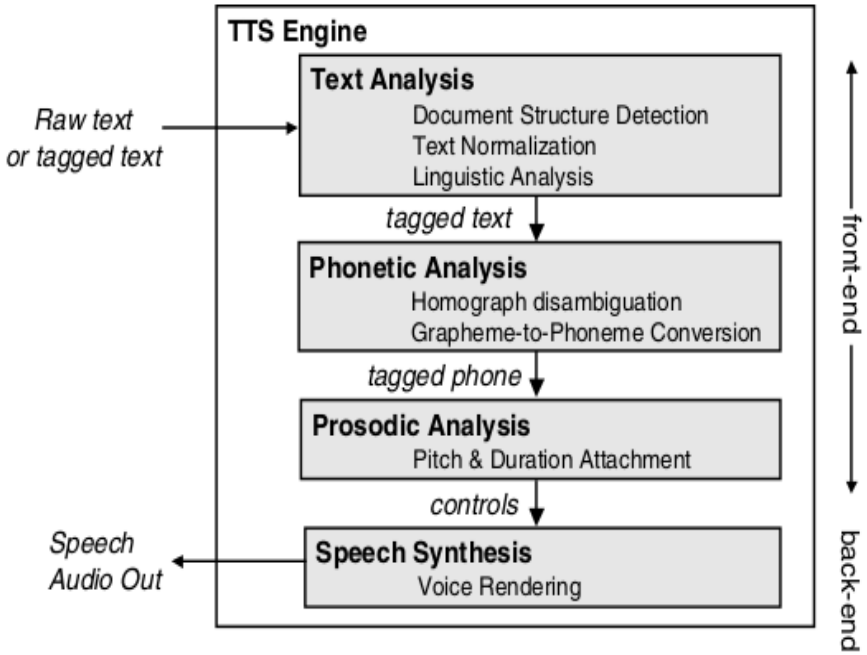
Text-To-Speech Synthesis

- Motor TTS:
  - Entrada: Texto
  - Saída: Voz



Text-To-Speech Synthesis

- Motor TTS:
  - Entrada: Texto
  - Saída: Voz
- Trabalha com:
  - Pronúncia do texto
  - Estrutura sintática
  - Semântica e ambiguidade



Text-To-Speech Synthesis

- *Document Structure Detection*:
  - Listas vs. texto corrente
  - Fim de frase
  - Fim de parágrafo
  - Pontuação
  - “This is Dr. Frankenstein.”

Text-To-Speech Synthesis

- *Text Normalization*:
  - “I live on Bourbon St. in St. Louis”
  - “$10”
  - “4:20”
  - “06/06/2014”
  - “She worked for DEC”
  - “I read Foucault”

Text-To-Speech Synthesis

- *Linguistic Analysis*:
  - *Part of speech* (POS): substantivo, verbo, etc.
  - Pausa entre frases
  - Ênfase nas palavras certas
  - Tipo da fala: raivoso, emotivo, relaxado, etc.
  - Um *parser* linguístico é muito lento

Text-To-Speech Synthesis

- *Homograph Disambiguation*:
  - Pronúncia correta de homografos
  - Checar o contexto
  - “an **ab**sent boy” vs. “do you choose to ab**sent** yourself?”
  - Isso já é **Processamento de Linguagem Natural**!

Text-To-Speech Synthesis

- *Grapheme-to-Phoneme Conversion*:
  - Converte o texto para *tagged phone*.
  - Uso de dicionário de pronúncia
  - Cada palavra é procurada independentemente
  - Regras de conversão para as exceções

Text-To-Speech Synthesis

- *Pitch & Duration Attachment*:
  - Provê ao sintetizador um conjunto de sinais de controle (sequência de sons, durações, *pitch*)
  - Sequência de sons deriva da ordem das palavras
  - Durações e *pitch* podem ser gerados baseados em regras próprias
  - Estresse, pausas e etc. tornam a voz mais natural

Text-To-Speech Synthesis

- *Voice Rendering*:

Text-To-Speech Synthesis

- *Voice Rendering*:
  - ***Rule-based*** *systems*
  - ***Data-driven*** *systems*

Text-To-Speech Synthesis

- *Voice Rendering*:
  - ***Rule-based*** *systems*
    - Baseiam-se em modelos físicos
  - ***Data-driven*** *systems*

Text-To-Speech Synthesis

- *Voice Rendering*:
  - ***Rule-based*** *systems*
    - Baseiam-se em modelos físicos
    - Voz gerada *from scratch*
  - ***Data-driven*** *systems*

Text-To-Speech Synthesis

- *Voice Rendering*:
  - ***Rule-based*** *systems*
    - Baseiam-se em modelos físicos
    - Voz gerada *from scratch*
    - Útil para sistemas simples
  - ***Data-driven*** *systems*

Text-To-Speech Synthesis

- *Voice Rendering*:
  - ***Rule-based*** *systems*
    - Baseiam-se em modelos físicos
    - Voz gerada *from scratch*
    - Útil para sistemas simples
  - ***Data-driven*** *systems*
    - Necessita de uma base de dados

Text-To-Speech Synthesis

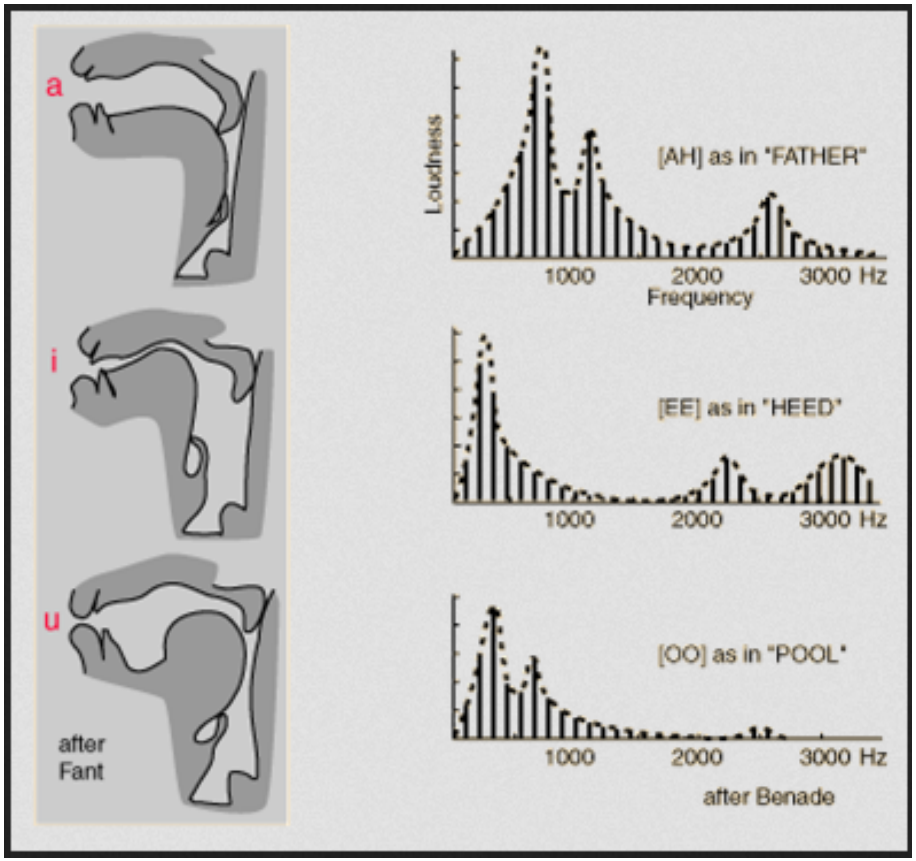
- *Voice Rendering*:
  - ***Rule-based*** *systems*
    - Baseiam-se em modelos físicos
    - Voz gerada *from scratch*
    - Útil para sistemas simples
  - ***Data-driven*** *systems*
    - Necessita de uma base de dados
    - Abordagem dominante

Text-To-Speech Synthesis

- Três grandes grupos:
  - *Formants Synthesis*
  - *Articulatory Synthesis*
  - *Contatenative Synthesis*

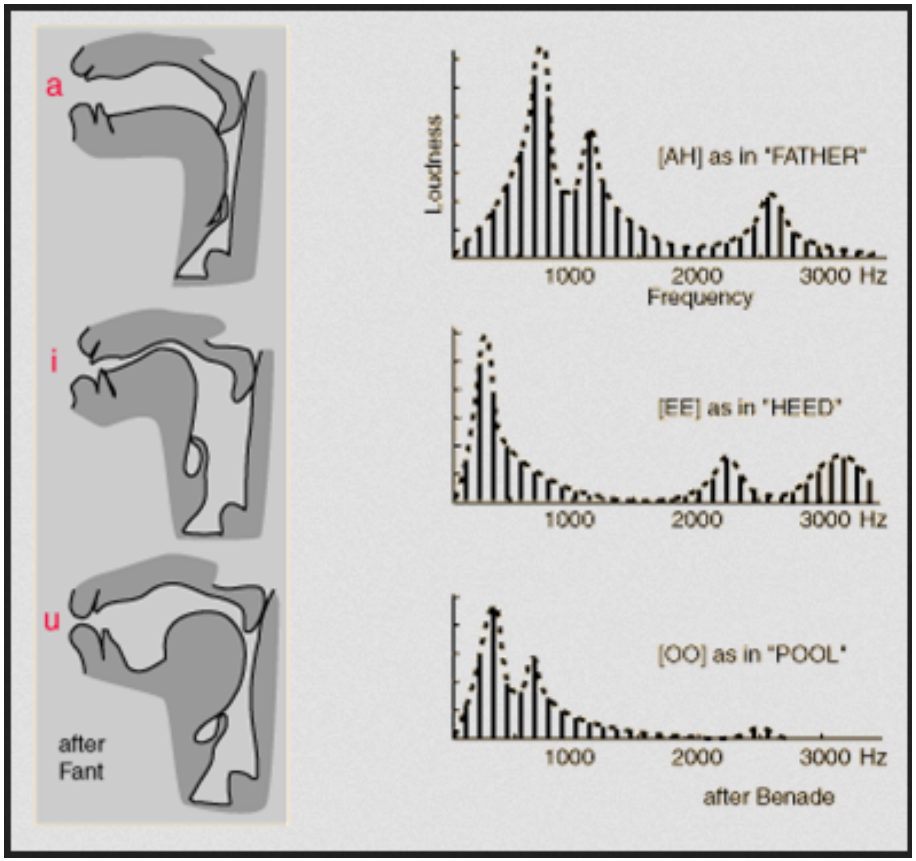
Text-To-Speech Synthesis

- *Formants*:



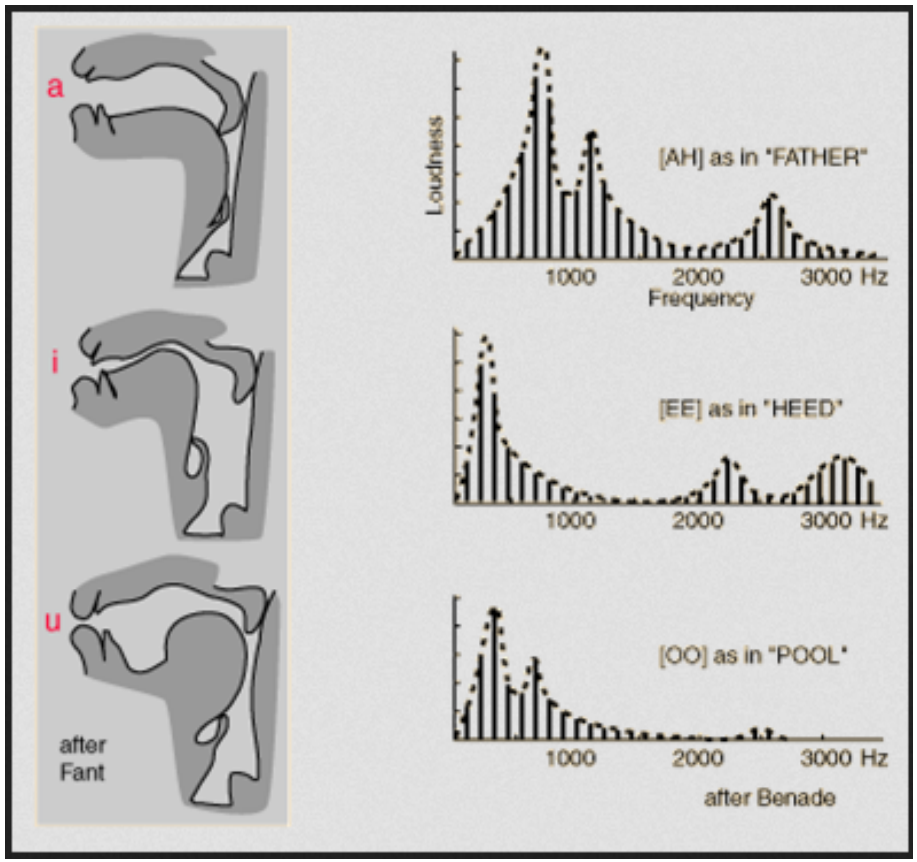
Text-To-Speech Synthesis

- *Formants*:
  - Picos no espectro da frequência



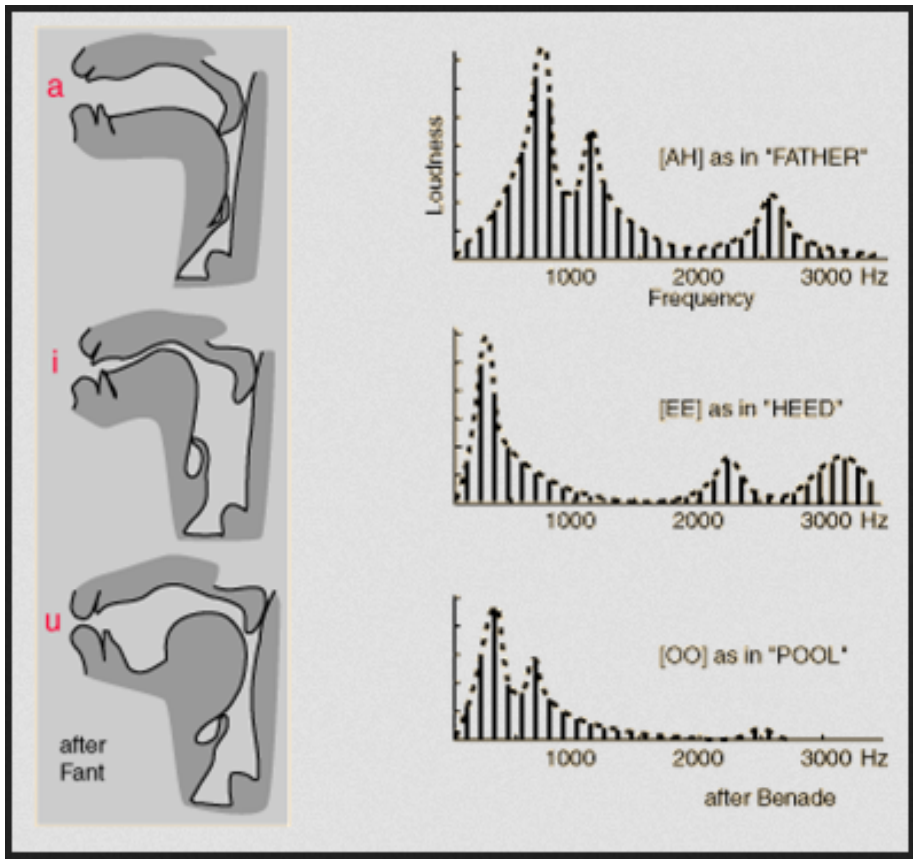
Text-To-Speech Synthesis

- *Formants*:
  - Picos no espectro da frequência
  - Função do trato vocal simulada satisfatoriamente



Text-To-Speech Synthesis

- *Formants*:
  - Picos no espectro da frequência
  - Função do trato vocal simulada satisfatoriamente
  - Frequências de ressonância do sistema



Text-To-Speech Synthesis

- *Articulatory*:

Text-To-Speech Synthesis

- *Articulatory*:
  - Modelagem direta de todo o sistema vocal

Text-To-Speech Synthesis

- *Articulatory*:
  - Modelagem direta de todo o sistema vocal
  - Voz de alta qualidade

Text-To-Speech Synthesis

- *Articulatory*:
  - Modelagem direta de todo o sistema vocal
  - Voz de alta qualidade
  - Um dos métodos mais difíceis de implementar

Text-To-Speech Synthesis

- *Articulatory*:
  - Modelagem direta de todo o sistema vocal
  - Voz de alta qualidade
  - Um dos métodos mais difíceis de implementar
  - Difícil adquirir dados para criar o modelo

Text-To-Speech Synthesis

- *Articulatory*:
  - Modelagem direta de todo o sistema vocal
  - Voz de alta qualidade
  - Um dos métodos mais difíceis de implementar
  - Difícil adquirir dados para criar o modelo
  - *Trade-off* acurácia/implementação

Text-To-Speech Synthesis

- *Articulatory*:
  - Modelagem direta de todo o sistema vocal
  - Voz de alta qualidade
  - Um dos métodos mais difíceis de implementar
  - Difícil adquirir dados para criar o modelo
  - *Trade-off* acurácia/implementação
  - Piores resultados

Text-To-Speech Synthesis

- *Articulatory*:
  - Modelagem direta de todo o sistema vocal
  - Voz de alta qualidade
  - Um dos métodos mais difíceis de implementar
  - Difícil adquirir dados para criar o modelo
  - *Trade-off* acurácia/implementação
  - Piores resultados
  - *Silver bullet* da síntese de voz (se for criado um modelo satisfatório)

Text-To-Speech Synthesis

- *Concatenative*:

Text-To-Speech Synthesis

- *Concatenative*:
  - Concatena unidades de voz pré-gravadas

Text-To-Speech Synthesis

- *Concatenative*:
  - Concatena unidades de voz pré-gravadas
  - Unidades podem ser **palavras**, **sílabas**, **semi-silábas**, **fonemas**, **difonemas**...

Text-**T**o-**S**peech Synthesis

- *Concatenative*:
  - Concatena unidades de voz pré-gravadas
  - Unidades podem ser **palavras**, **sílabas**, **semi-silábas**, **fonemas**, **difonemas**...
  - Unidades mais longas:

Text-To-Speech Synthesis

- *Concatenative*:
  - Concatena unidades de voz pré-gravadas
  - Unidades podem ser **palavras**, **sílabas**, **semi-silábas**, **fonemas**, **difonemas**...
  - Unidades mais longas:
    - Mais natural

Text-To-Speech Synthesis

- *Concatenative*:
  - Concatena unidades de voz pré-gravadas
  - Unidades podem ser **palavras**, **sílabas**, **semi-silábas**, **fonemas**, **difonemas**...
  - Unidades mais longas:
    - Mais natural
    - Menos pontos de concatenação

Text-To-Speech Synthesis

- *Concatenative*:
  - Concatena unidades de voz pré-gravadas
  - Unidades podem ser **palavras**, **sílabas**, **semi-silábas**, **fonemas**, **difonemas**...
  - Unidades mais longas:
    - Mais natural
    - Menos pontos de concatenação
    - Necessita de mais memória

Text-To-Speech Synthesis

- *Concatenative*:
  - Concatena unidades de voz pré-gravadas
  - Unidades podem ser **palavras**, **sílabas**, **semi-silábas**, **fonemas**, **difonemas**...
  - Unidades mais longas:
    - Mais natural
    - Menos pontos de concatenação
    - Necessita de mais memória
    - Tende a ser impraticável

Text-To-Speech Synthesis

- *Concatenative*:
  - Unidades mais curtas:

Text-To-Speech Synthesis

- *Concatenative*:
  - Unidades mais curtas:
    - Menos natural

Text-To-Speech Synthesis

- *Concatenative*:
  - Unidades mais curtas:
    - Menos natural
    - Mais pontos de concatenação

Text-To-Speech Synthesis

- *Concatenative*:
  - Unidades mais curtas:
    - Menos natural
    - Mais pontos de concatenação
    - Necessita de menos memória

Text-To-Speech Synthesis

- *Concatenative*:
  - Unidades mais curtas:
    - Menos natural
    - Mais pontos de concatenação
    - Necessita de menos memória
    - Coleta de amostras e técnicas de rotulação mais complexas

Text-To-Speech Synthesis

- *Concatenative*:
  - Unidades mais curtas:
    - Menos natural
    - Mais pontos de concatenação
    - Necessita de menos memória
    - Coleta de amostras e técnicas de rotulação mais complexas
    - Difonemas são as unidades mais usadas:

Text-To-Speech Synthesis

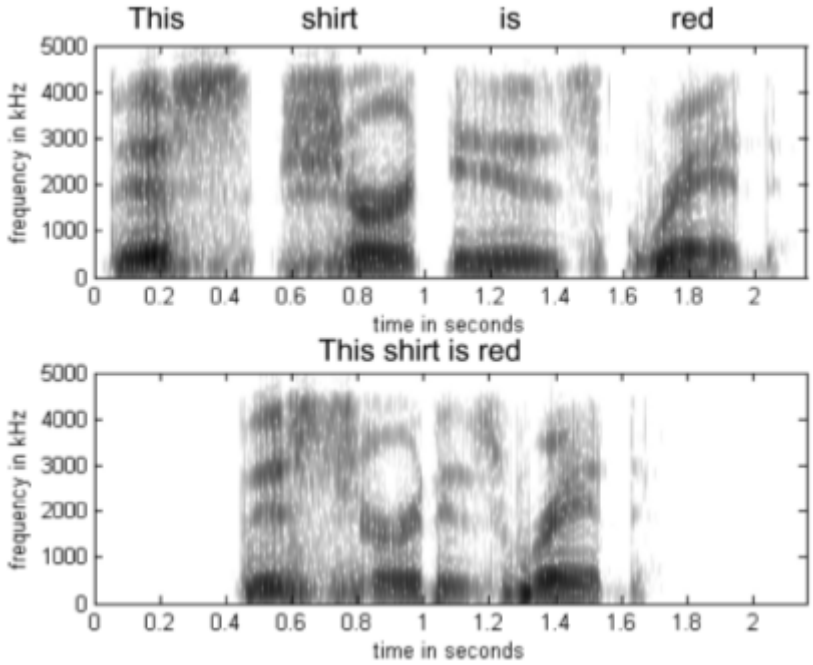
- *Concatenative*:
  - Unidades mais curtas:
    - Menos natural
    - Mais pontos de concatenação
    - Necessita de menos memória
    - Coleta de amostras e técnicas de rotulação mais complexas
    - Difonemas são as unidades mais usadas:
      - Transição mais suave entre fonemas

Text-To-Speech Synthesis

- *Concatenative*:
  - Unidades mais curtas:
    - Menos natural
    - Mais pontos de concatenação
    - Necessita de menos memória
    - Coleta de amostras e técnicas de rotulação mais complexas
    - Difonemas são as unidades mais usadas:
      - Transição mais suave entre fonemas
      - Da metade do 1º à metade do 2º fonema

Text-To-Speech Synthesis

- *Concatenative*:
  - O espectrograma de baixo **não** é uma superposição do de cima

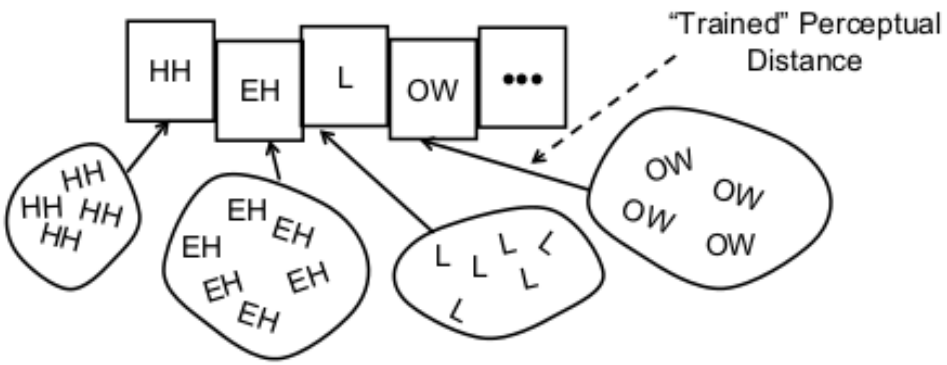


Text-To-Speech Synthesis

- *Concatenative*:
  - Unit Selection é o *estado-da-arte:*

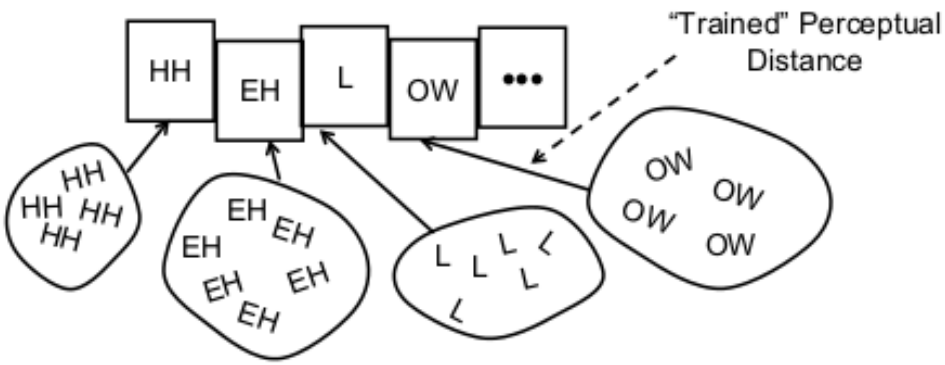
Text-To-Speech Synthesis

- *Concatenative*:
  - Unit Selection é o *estado-da-arte:*
    - Escolhe entre múltiplas **instâncias**



Text-To-Speech Synthesis

- *Concatenative*:
  - Unit Selection é o *estado-da-arte:*
    - Escolhe entre múltiplas **instâncias**
    - A instância que melhor casa com o *target* é escolhida (menos modificações)



Text-To-Speech Synthesis

- *Unit Selection:*

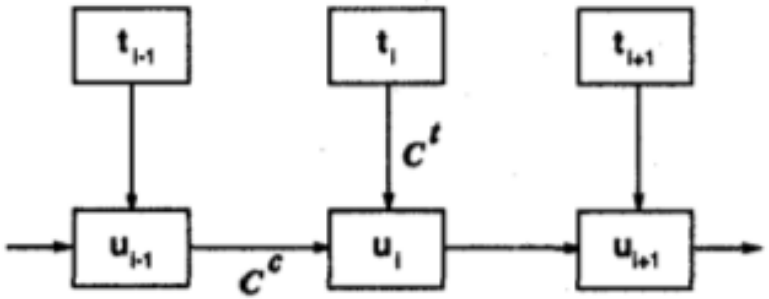


Figure 1. Unit Selection Costs

Text-To-Speech Synthesis

- *Unit Selection:*
  - Minimizar ***target cost***: estimativa da incompatibilidade entre uma unidade e o *target*

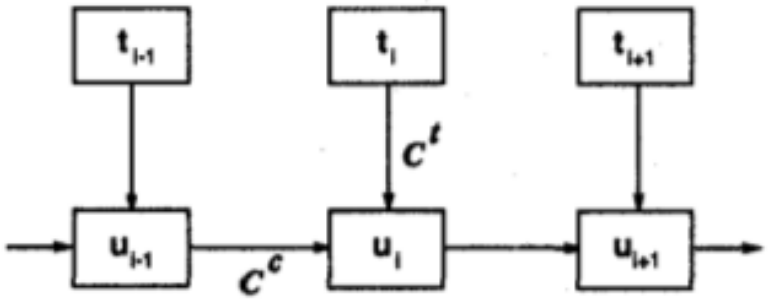


Figure 1. Unit Selection Costs

Text-To-Speech Synthesis

- *Unit Selection:*
  - Minimizar ***target cost***: estimativa da incompatibilidade entre uma unidade e o *target*
  - Minimizar ***join cost***: estimativa da incompatibilidade acústica com o fonema anterior

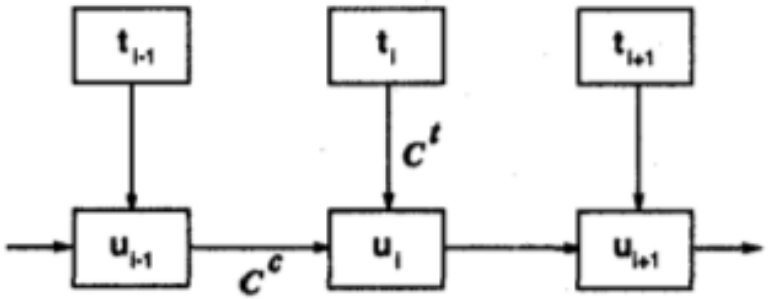


Figure 1. Unit Selection Costs

Text-To-Speech Synthesis

- *Unit Selection*:
  - Unidades consecutivas possuem *join cost* **zero** (concatenação natural)
  - O *unit selection* é a tarefa de determinar a sequência cujo custo total é o menor

$$C^t(t_i, u_i) = \sum_{j=1}^{p} w_j^t C_j^t(t_i, u_i) \qquad [1]$$

$$C^c(u_{i-1}, u_i) = \sum_{j=1}^{q} w_j^c C_j^c(u_{i-1}, u_i) \qquad [2]$$

$$C(t_1^n, u_1^n) = \sum_{i=1}^{n} C^t(t_i, u_i) + \sum_{i=2}^{n} C^c(u_{i-1}, u_i) + C^c(S, u_1) + C^c(u_n, S) \qquad [3]$$

Text-To-Speech Synthesis

- *Unit Selection*:
  - Ou, selecionar o melhor caminho (*Viterbi search*)

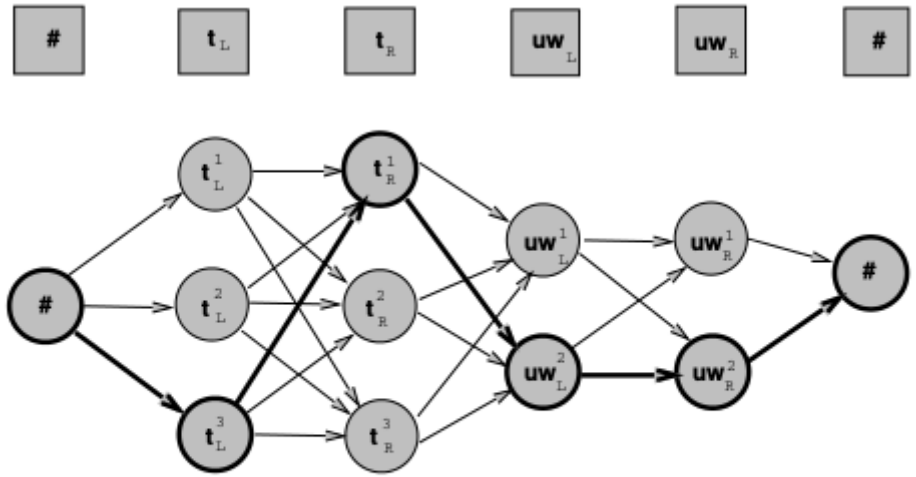


Figure 1: For half-phones in the word "two", a search finds the lowest cost path, selecting one candidate unit from each column for synthesis.

Brain-Machine Interface

- *Back to the formants...*

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)
- Spikes inseridos na região responsável pelos articuladores da fala

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)
- Spikes inseridos na região responsável pelos articuladores da fala
- Estima o **primeiro** e o **segundo** formantes

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)
- Spikes inseridos na região responsável pelos articuladores da fala
- Estima o **primeiro** e o **segundo** formantes
- Delay de 50 ms do neurônio até a fala

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)
- Spikes inseridos na região responsável pelos articuladores da fala
- Estima o **primeiro** e o **segundo** formantes
- Delay de 50 ms do neurônio até a fala
  - Respeita o deadline (200 ms)

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)
- Spikes inseridos na região responsável pelos articuladores da fala
- Estima o **primeiro** e o **segundo** formantes
- Delay de 50 ms do neurônio até a fala
  - Respeita o deadline (200 ms)
- Eficácia de 70%

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)
- Spikes inseridos na região responsável pelos articuladores da fala
- Estima o **primeiro** e o **segundo** formantes
- Delay de 50 ms do neurônio até a fala
  - Respeita o deadline (200 ms)
- Eficácia de 70%
- Difícil gerar fonemas mais complexos

Brain-Machine Interface

- *Back to the formants...*
- Técnica recente (2004)
- Spikes inseridos na região responsável pelos articuladores da fala
- Estima o **primeiro** e o **segundo** formantes
- Delay de 50 ms do neurônio até a fala
  - Respeita o deadline (200 ms)
- Eficácia de 70%
- Difícil gerar fonemas mais complexos (*s**t*)

Exemplos

- Klatt Synthesizer (1980)
  - Formant
- VocaliD
  - Concatenative
- CW Speak (codewelt.com/proj/speak)
- IBM ViaVoice
- **Cepstral** (www.cepstral.com/en/demos)
- **AT&T Natural Voices** (www2.research.att.com/~ttsweb/tts/demo.php)
- Além dos produtos Apple...

É um Sistema de Tempo Real?

É um Sistema de Tempo Real?

- Claro!

É um Sistema de Tempo Real?

- Claro!
- Deve responder sem muito atraso

É um Sistema de **Tempo Real**?

- Claro!
- Deve responder sem muito atraso
- **Ordem** e **Velocidade** da fala importam:

É um Sistema de **Tempo Real**?

- Claro!
- Deve responder sem muito atraso
- **Ordem** e **Velocidade** da fala importam:
  - Fonemas fora de ordem → fala sem sentido

É um Sistema de **Tempo Real**?

- Claro!
- Deve responder sem muito atraso
- **Ordem** e **Velocidade** da fala importam:
  - Fonemas fora de ordem → fala sem sentido
  - Velocidade inconstante → prejuízo ao ouvinte

É um Sistema de **Tempo Real**?

- Claro!
- Deve responder sem muito atraso
- **Ordem** e **Velocidade** da fala importam:
  - Fonemas fora de ordem → fala sem sentido
  - Velocidade inconstante → prejuízo ao ouvinte
  - Deadline **hard** ou **firm**

É um Sistema de **Tempo Real**?

- Claro!
- Deve responder sem muito atraso
- **Ordem** e **Velocidade** da fala importam:
  - Fonemas fora de ordem → fala sem sentido
  - Velocidade inconstante → prejuízo ao ouvinte
  - Deadline **hard** ou **firm**
- Mais por causa da evolução dos computadores:

É um Sistema de **Tempo Real**?

- Claro!
- Deve responder sem muito atraso
- **Ordem** e **Velocidade** da fala importam:
  - Fonemas fora de ordem → fala sem sentido
  - Velocidade inconstante → prejuízo ao ouvinte
  - Deadline **hard** ou **firm**
- Mais por causa da evolução dos computadores:
  - O sinal da fala é lento

É um Sistema de **Tempo Real**?

- Claro!
- Deve responder sem muito atraso
- **Ordem** e **Velocidade** da fala importam:
  - Fonemas fora de ordem → fala sem sentido
  - Velocidade inconstante → prejuízo ao ouvinte
  - Deadline **hard** ou **firm**
- Mais por causa da evolução dos computadores:
  - O sinal da fala é lento
  - Não precisa de tantas técnicas sofisticadas...

É um Sistema de Tempo Real?

- Também pode ser embarcado (not so good):

É um Sistema de Tempo Real?

- Também pode ser embarcado (not so good):
  - DSPs são usados para sistemas específicos

É um Sistema de Tempo Real?

- Também pode ser embarcado (not so good):
  - DSPs são usados para sistemas específicos
  - Baixa potência

É um Sistema de Tempo Real?

- Também pode ser embarcado (not so good):
  - DSPs são usados para sistemas específicos
  - Baixa potência
  - *Concatenative* é possível (sem *Unit Selection*)

É um Sistema de Tempo Real?

- Também pode ser embarcado (not so good):
  - DSPs são usados para sistemas específicos
  - Baixa potência
  - *Concatenative* é possível (sem *Unit Selection*)
  - TD-PSOLA

É um Sistema de Tempo Real?

- Também pode ser embarcado (not so good):
  - DSPs são usados para sistemas específicos
  - Baixa potência
  - *Concatenative* é possível (sem *Unit Selection*)
  - TD-PSOLA
  - Database comprimido

Thanks