Síntese de Voz

Text-to-Speech (TTS) Synthesis

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

Síntese de Voz

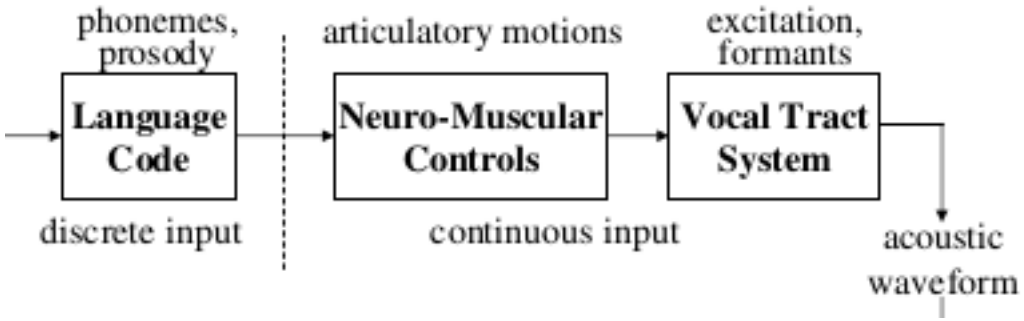
- Introdução
- Métodos
    1. *Formants*
    2. *Articulatory*
    3. *Concatenative*
    4. *Unit Selection*
    5. *HNM (**H**armonic plus **N**oise **M**odel)*
    6. *HMM (**H**idden **M**arkov **M**odel)*
- Conclusão

Introdução

- Produz voz humana **artificialmente**
- Um sistema TTS converte **texto** em **voz**
- Onde usar?
  - Leitura de tela para cegos
  - Voz para pessoas com dificuldades de fala
  - Interface humano-máquina (Google Glass, Siri...)
  - Atores virtuais
  - *[preencha com algo interessante]*

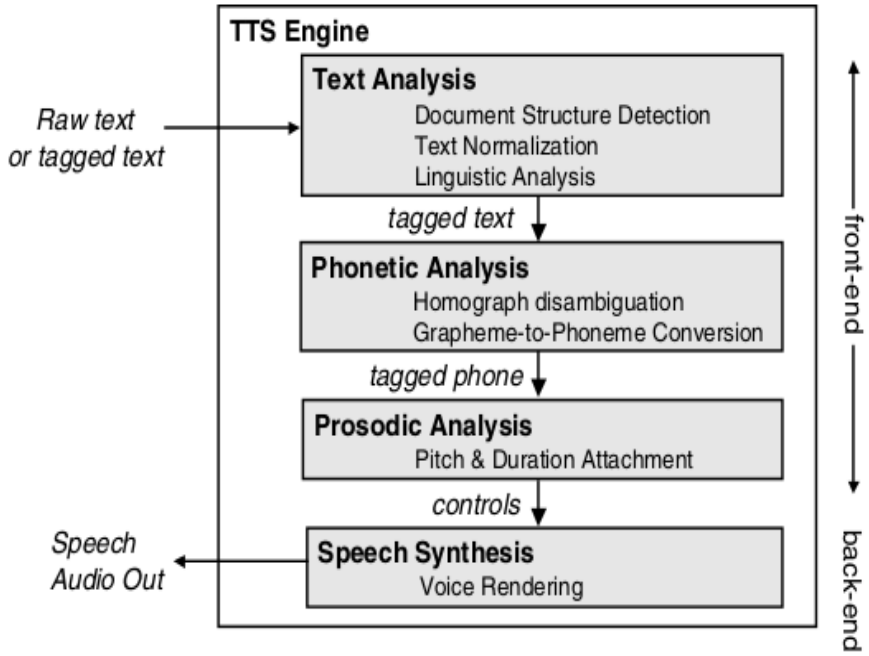
Introdução

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



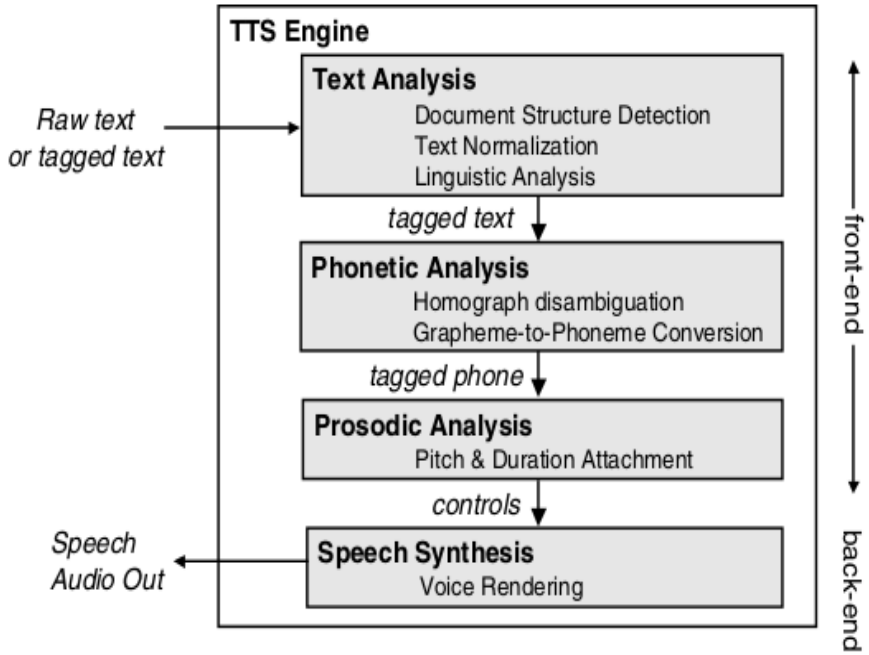
Introdução

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



Introdução

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



Introdução

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

Introdução

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

Introdução

- *Linguistic Analysis*:
  - *Part of speech* (POS): substantivo, verbo, etc.
  - Pausa entre frases
  - Presença de anáfora (“Latino é cantor. **Ele** é horrível!”)
  - Ênfase nas palavras certas
  - Tipo da fala: raivoso, emotivo, relaxado, etc.
  - Um *parser* linguístico é muito lento

Introdução

- *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**!

Introdução

- *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

Introdução

- *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

Introdução

- *Speech Synthesis*:
  - Aqui o bicho pega!
  - ***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

Métodos

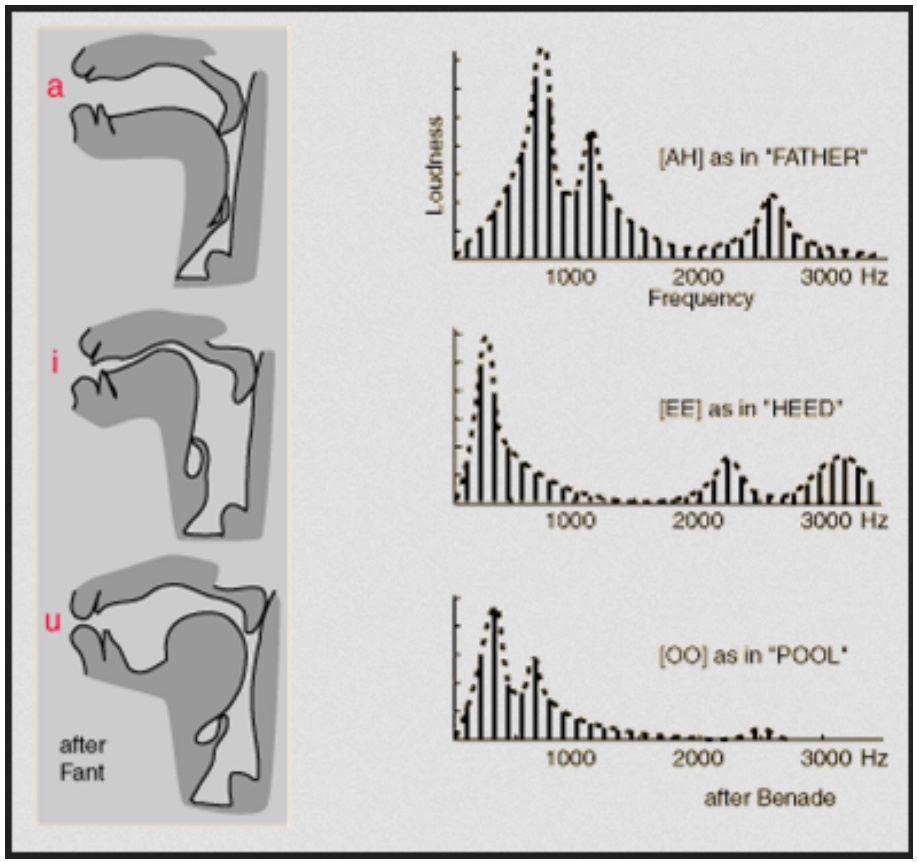
- Os métodos são basicamente três:
  - *Formants Synthesis*
  - *Articulatory Synthesis*
  - *Contatenative Synthesis*

Métodos

- Os métodos são basicamente três:
  - *Formants Synthesis*
  - *Articulatory Synthesis*
  - *Contatenative Synthesis*
    - *Unit Selection* é um **aprimoramento**
    - *HNM* é usado em **conjunto** com o *Unit Selection*
    - *HMM* parte de outra premissa, mas mantém a abordagem *data-driven*

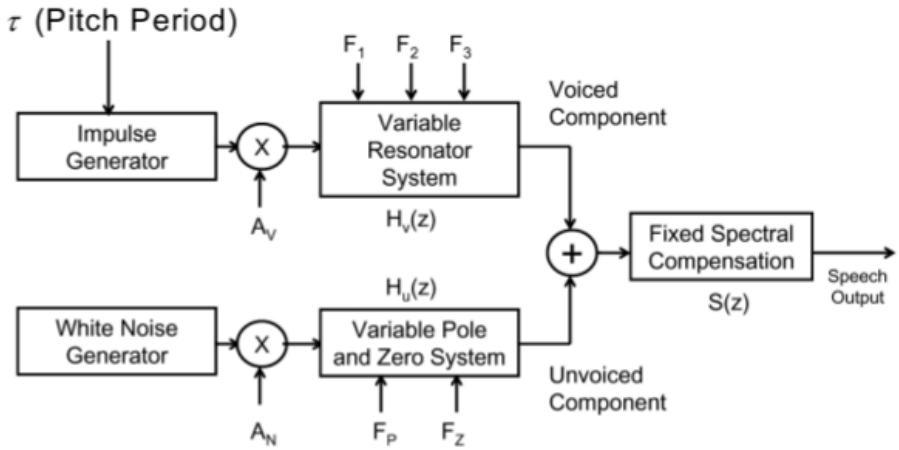
Métodos

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



Métodos

- *Formants*:
  - Compensador permite simular efeitos anasalados, fricativos e plosivos
  - A especificação de 20 ou mais parâmetros pode gerar uma reconstrução satisfatória do sinal



Métodos

- *Formants*:
  - Vantages:
    - Parâmetros altamente correlacionados com a produção e propagação do som no trato vocal
    - Inteligibilidade alta e computacionalmente leve
  - Desvantagens:
    - Difícil especificar os parâmetros automaticamente
    - Naturalidade extremamente deficiente

Métodos

- *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)

Métodos

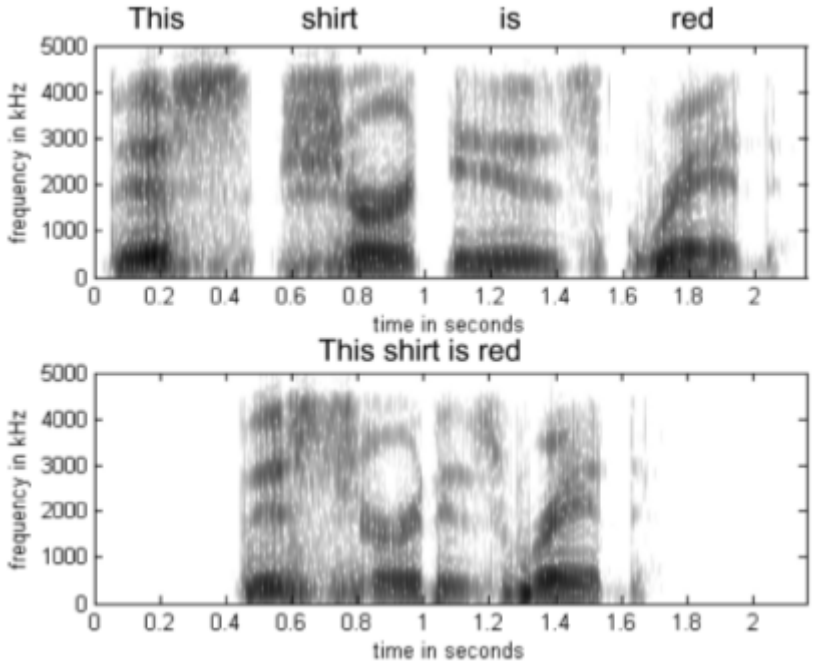
- *Concatenative*:
  - Concatena unidades de voz pré-gravadas
  - Unidades podem ser **palavras**, **sílabas**, **semi-silábas**, **fonemas**, **difonemas** ou **trifonemas**.
  - Unidades mais longas:
    - Mais natural
    - Menos pontos de concatenação
    - Necessita de mais memória
    - Tende a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

Métodos

- *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

Métodos

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



Métodos

- *Concatenative*:
  - Difonemas:

    Entrada: *“I want”*.

    Fonemas: /#/ AY/ /W/ /AA/ /N/ /T/ /#/

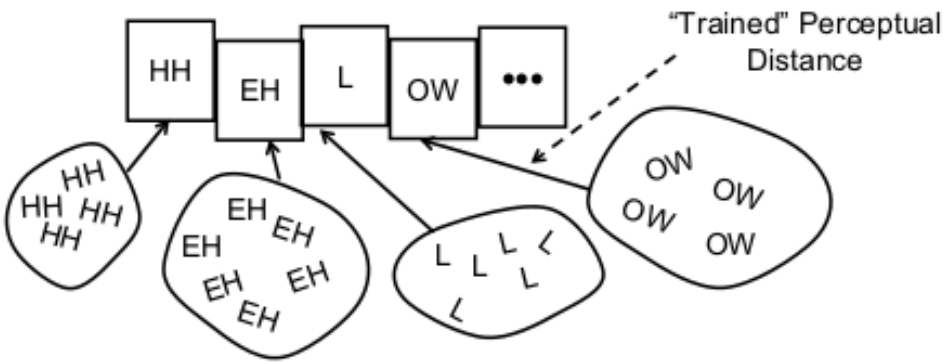
    Difonemas: /# AY/ /AY-W/ /W-AA/ /AA-N/ /N-T/ /T- #/
  - Os pontos de concatenação ainda soam pouco naturais (transição não suave)
  - Nem sempre o difonema representa o som correto

Métodos

- *Unit Selection*:
  - Em *Concatenative Synthesis*, difonemas precisam ser modificados através de processamento de sinais para resultar na prosódia desejada
  - Pontos de concatenação tornam a fala pouco natural
  - Efeitos de co-articulação não são limitados apenas ao fonema anterior

Métodos

- *Unit Selection*:
  - Unidade com várias **instâncias** (variações na prosódia)
  - A instância que melhor casa com o *target* é escolhida (menos modificações)



Métodos

- *Unit Selection*:
  - As unidades ainda sofrem transformações (*PSOLA*, *TD-PSOLA*, *HNM...*)
  - Contudo, quanto maior a base, maior a chance de encontrar um *match*
  - Reduz a necessidade de aplicar as modificações de prosódia

Métodos

- *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

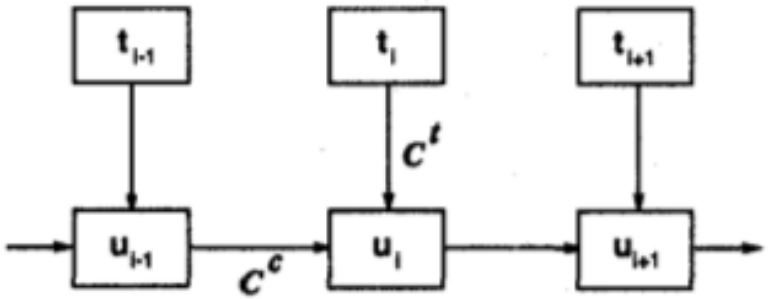


Figure 1. Unit Selection Costs

Métodos

- *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

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

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

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

Métodos

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

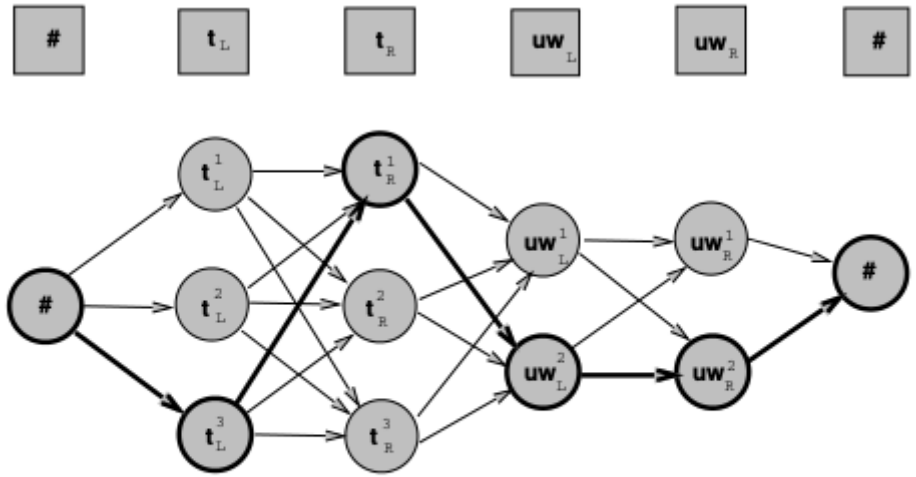


Figure 1: For half-phones in the word "two", a search finds the lowest cost path, selecting one candidate unit from each column for synthesis.

Métodos

- *Unit Selection*:
  - Base mínima com 30 minutos de gravação já torna o método realizável (na prática, 1-10 horas)
  - Desvantagens:
    - Base de dados grande
    - Rotulação incorreta e não detecção de contexto leva a fragmentos de voz com qualidade péssima
    - Margem de erro considerável

Métodos

- *Unit Selection* + *HNM*:
  - Modificações prosódicas são necessárias para sintetizar voz com alta qualidade
  - Inteligibilidade, naturalidade e agradabilidade superiores ao TD-PSOLA

Métodos

- *Unit Selection* + *HNM*:
  - Modificações prosódicas são necessárias para sintetizar voz com alta qualidade
  - Inteligibilidade, naturalidade e agradabilidade superiores ao TD-PSOLA
  - Fala composta pelas partes harmonica (quase periódica) e ruidosa (não periódica)
    - Separadas pela *maximum voiced frequency*, *Fm*, que varia no tempo
    - F < Fm → harmônicos
    - F > Fm → componente ruidosa

Métodos

- *Unit Selection* + *HNM*:
  - Espectro dividido em duas bandas
  - A banda baixa é modelada como uma soma de harmônicos
  - Onde arg{ak(ti)} = arg{ck(ti)} = arg{dk(ti)}

$$s_h(t) = \sum_{k=-L(t)}^{L(t)} A_k(t) e^{j\, k\omega_0(t)\, t} \qquad (1)$$

$$A_k(t) = a_k(t_i) \qquad (2)$$

$$A_k(t) = a_k(t_i) + t\ b_k(t_i) \qquad (3)$$

$$A_k(t) = a_k(t_i) + t\ c_k(t_i) + t^2\ d_k(t_i) \qquad (4)$$

Métodos

- *Unit Selection* + *HNM*:
  - Ao usar Ak(t) = ak(ti), modelo HNM1, a fala produzida já possui uma ótima qualidade
  - A banda alta é formada pela convolução de um modelo autorregressivo (envelopado) com o ruído branco

$$s_n(t) = e(t)[h(\tau, t) \star b(t)] \qquad (5)$$

Métodos

- *Unit Selection* + *HNM*:
  - A parte não períodica é o resultado da parte harmônica subtraída da fala completa
  - Descontinuidades nos pontos de concatenação são considerados apenas para a parte harmônica
  - Incompatibilidades nos *pitches* são removidas efetuando uma interpolação linear num ponto de concatenação, $t = ti$

Métodos

- *HMM*:
  - *Unit Selection* possui como desvantagem:
    - Um grande e crescente banco de dados
    - Pouca flexibilidade (recria o que foi gravado)

Métodos

- *HMM*:
  - *Unit Selection* possui como desvantagem:
    - Um grande e crescente banco de dados
    - Pouca flexibilidade (recria o que foi gravado)
  - *HMM* utiliza uma abordagem estatística para inferir parâmetro a partir dos dados

Métodos

- *HMM*:
  - *Unit Selection* possui como desvantagem:
    - Um grande e crescente banco de dados
    - Pouca flexibilidade (recria o que foi gravado)
  - *HMM* utiliza uma abordagem estatística para inferir parâmetro a partir dos dados
  - Vantagens:
    - Necessita de menos memória
    - Mais variações nos exemplos (converter uma voz em outra)

Métodos

- *HMM*:
  - Consiste de duas fases principais: **treinamento** e **síntese**
  - Na fase de treinamento geralmente utiliza como característica o MFCC e suas derivadas primeira e segunda
  - As características são extraídas por frame, colocadas em um vetor e usadas pelo algoritmo *Baum-Welch* para criar um modelo de cada fonema
  - Síntese em dois estágios:
    - Estimação das características
    - Transformação em sinais de áudio

Métodos

- *HMM*:
  - Ainda assim não é melhor que o *Unit Selection* (estado da arte)
  - Modelagem pode ser melhorada utilizando:
    - *Hidden semi-Markov Models*
    - *Trajectory HMMs*
    - *Stochastic Markov Graphs*

Métodos

- *HMM*:
  - Integração com HNM leva a um sistema mais rápido e barato de desenvolver
  - Produz fala de qualidade superior ao *Unit Selection*
  - Contudo, mais difícil de entender...
  - … mas uma vez entendido, mais fácil de implementar

Conclusão

- Métodos de síntese de voz já são suficientemente bons
- O estado da arte ainda requer uma quantidade de dados grande demais
- Deve-se tentar entender melhor o funcionamento do sistema vocal humano para modelá-lo
- E aprimorar métodos estocásticos