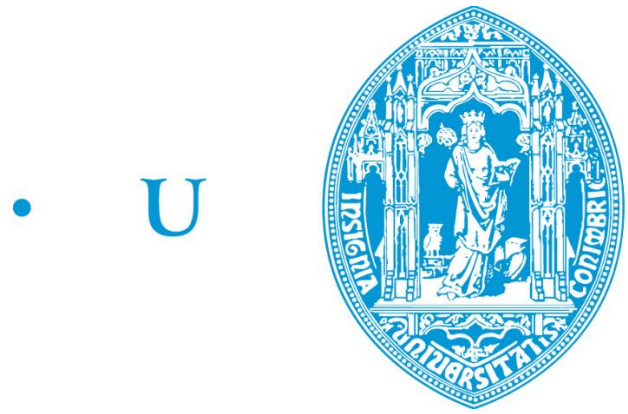

Tiago António Coroado da Silva Ferreira

# SISTEMA ONLINE DE SÍNTESE DE FALA EM PORTUGUÊS

Dissertação de Mestrado em Engenharia Eletrotécnica e de Computadores

Setembro de 2014

UNIVERSIDADE DE COIMBRA



Universidade de Coimbra

Faculdade de Ciências e Tecnologia

Departamento de Engenharia Eletrotécnica e de Computadores

Mestrado Integrado em Engenharia Eletrotécnica e de Computadores

# Sistema Online de Síntese de Fala em Português

Tiago António Coroado da Silva Ferreira

Orientadores:

Professor Doutor Fernando Santos Perdigão

Doutora Sara Maria Fernandes Rato e Costa Marques Candeias

Júri:

Professor Doutor Vitor Manuel Mendes da Silva (Presidente)

Professor Doutor Fernando Santos Perdigão (Orientador)

Professor Doutor Luís Alberto da Silva Cruz (Vogal)

Setembro de 2014

## Agradecimentos

*Quero começar por agradecer ao meu orientador Professor Doutor Fernando Perdigão pela sua grande ajuda no desenvolvimento desta dissertação, pela sua constante disponibilidade e pelo grande apoio que me deu para o meu sucesso neste trabalho.*

*Agradeço também à minha co-orientadora Doutora Sara Maria Fernandes Rato e Costa Marques Candeias pela sua contribuição e disponibilidade nesta dissertação.*

*Quero também agradecer à minha família, em especial aos meus pais e à minha irmã pelo constante apoio e motivação durante todos estes anos. Sem eles não teria conseguido chegar aqui e ter sucesso na minha vida académica.*

*Agradeço aos meus colegas de laboratório Doutor Arlindo Veiga e Jorge Proença pela preciosa ajuda que me deram neste trabalho e pela sua disponibilidade.*

*Agradeço aos meus amigos e colegas de curso pelo apoio e companhia que tornaram estes cinco anos numa das melhores fases da minha vida proporcionando momentos que ficarão para o resto da minha vida.*

## Resumo

A área de síntese de fala tem vindo a conhecer grandes desenvolvimentos e a ter uma maior adoção nos sistemas que dependem de uma ligação entre uma máquina e um utilizador humano por forma a amenizar e facilitar esta interação. Estas ferramentas tornam os sistemas mais *user-friendly* e facilitam a adaptação do utilizador à aplicação.

As principais preocupações no desenvolvimento destes sistemas prendem-se, sobretudo, no aumento da qualidade do áudio produzido e na diminuição de recursos utilizados, especialmente na redução da base de dados utilizada pelo sistema.

As duas soluções mais usadas para este propósito são os sistemas de concatenação e os sistemas de modelos estatísticos. O objetivo desta dissertação é desenvolver um sistema de síntese de fala HTS (*HMM-based Text to Speech Synthesis System*) baseado em modelos de Markov não-observáveis (*Hidden Markov Models* ou "HMM") que possa ser implementado numa página web, ou seja, um sistema online de síntese de fala. Está provado que este tipo de sistema consegue melhores resultados face aos sistemas concatenativos, não só ao nível da qualidade da fala produzida, mas também reduz o tamanho da base de dados usada na síntese.

Esta solução consegue modelar com sucesso a fala humana, usando modelos estatísticos para descrever a variação da fala na frequência usando a representação espectral na escala de Mel (MGC), a excitação da fala (vozeada ou não vozeada) e a duração de cada segmento de fala.

Para criar modelos de fala humana é necessário gravar locuções de fala que contenham pouca entoação e que cujos locutores tenham boa dicção para se obter uma base de dados de treino com boa qualidade. Esta é uma boa base para a síntese de uma fala natural e inteligível.

Na fase do treino dos modelos são calculados os parâmetros que descrevem cada uma das frases de treino a partir da análise das locuções gravadas.

Finalmente, usa-se o resultado do treino, uma "voz", que consiste nos modelos de fala, para fazer síntese de fala.



Palavras-Chave: HTS, HMM, MGC, TTS, Síntese, Markov

## Abstract

Speech synthesis systems have been improved and implemented in systems that rely on a connection between a machine and its user. They turn technological devices more user-friendly and facilitate the human-machine adapting process for the interaction.

The main concerns about such systems are mostly related to the need of increasing audio quality and reducing the usage of hardware resources by minimizing the size of the database used by the system.

The two most used solutions for this purpose are the concatenative and the statistic modeling approaches. The main goal of this dissertation is to develop a HTS speech synthesis system (Text-to-Speech HMM-based Synthesis System) based on the hidden-Markov model (HMM). It is proven that this type of systems achieve better results than the concatenative ones not only on the quality of the synthesized speech, but mostly on the reduction of the database used in the synthesis procedure

This solution can successfully model human speech using statistic models in order to describe speech frequency variation by using spectral representation in Mel scale (MGC), speech excitation (voiced or unvoiced) and the duration of each speech segment.

To create statistic models of human speech it is necessary to record several speech utterances with neutral intonation from speakers with good diction in order to obtain a database with enough quality that leads to a natural and intelligible synthesized speech.

In the training phase, the parameters that describe the utterances are calculated by analyzing the recorded utterances.

Finally, the result of this training is a "voice" that basically comes up by using speech models and will be used to do speech synthesis.

Keywords: HTS, HMM, MGC, TTS, Synthesis, Markov

# Índice

# Lista de Figuras

# Lista de Tabelas

## Lista de Acrónimos

CB – Carlos Braz

DG – Diana Guardado

G2P – Grapheme-To-Phoneme

GV – Global VarianceHMM – Hidden-Markov Model

HTK – Hidden Markov Model Toolkit

HTS – HMM-based Speech Synthesis System

IPA – International Phonetic Alphabet

MDL – Minimum Description Length

MGC – Mel Generalized Cepstrum

MLF – Master Label File

MSD-HMM – Multi-Space Probability DistributionPDF – Probability Density Function

POS – Part-of-speech

RM – Ricardo Mariano

SAMPA – Speech Assessment Methods Phonetic Alphabet

SoX – Sound eXchange

SPTK – Speech Signal Processing Toolkit

TTS – Text-To-Speech

# Capítulo 1 - Introdução

## 1.1 - Motivação

Com o desenvolvimento dos equipamentos portáteis e de aplicações que necessitem de grande interação com o utilizador, torna-se prático ter a possibilidade de comunicação entre a máquina e o utilizador através da fala.

É neste seguimento que surgem os sistemas de síntese de voz, correntemente chamados de sistemas *text-to-speech* [1] ou simplesmente TTS, em que a partir de um texto arbitrário seja possível criar um excerto de áudio que seja o mais realista possível e fiel ao texto introduzido. Além destes sistemas existem também os sistemas de reconhecimento de voz, conhecidos como STT (*speech-to-text*) ou ASR (*automatic speech recognition*), que possibilitam a comunicação no sentido inverso, traduzindo um excerto de fala sob a forma escrita. Pode-se dizer que estes dois sistemas complementam-se no processo de comunicação entre um utilizador e uma aplicação.

Espera-se que um sistema TTS sirva uma grande variedade de dispositivos e, por isso, deve ser um sistema rápido, eficiente e que exija poucos recursos, ou seja, tenha um *footprint* reduzido.

Na implementação de um sistema TTS surgem como paradigmas os sistemas baseados em concatenação de excertos de fala e os sistemas baseados em modelos estatísticos. Os sistemas baseados em concatenação usam segmentos de fala que podem ser reduzidos a fones simples ou fones com contexto, dependendo do sistema, e posteriormente são concatenados, com o devido processamento, para formar a fala completa. Os sistemas baseados em modelos estatísticos modelam a fala usando modelos de fones, sua duração e modelos de prosódia.

Um fonema é a unidade fonológica mais básica de um idioma. Um fone é a realização acústica de um fonema, ou seja, é um segmento de som correspondente a um fonema.

Apesar da adoção cada vez maior dos sistemas de síntese de voz a partir de modelos HMM [19], os sistemas concatenativos ainda são usados na atualidade. Nestes sistemas de síntese por concatenação é possível obter uma boa qualidade de fala, mas à custa de uma base de dados enorme, ou seja, um grande *footprint*. Sendo assim os sistemas baseados em HMM tornam-se mais eficientes em termos do uso dos recursos disponíveis.

Devido a este fator, a escolha na elaboração desta dissertação recaiu sobre o sistema baseado em modelos de Markov não-observáveis (HMMs, *Hidden Markov Models*) para modelar os fones da fala, a sua duração bem como a prosódia das frases. Esta abordagem é muito vantajosa em relação aos sistemas concatenativos, pois além de apresentar um *footprint* muito mais reduzido,

esta solução oferece, ainda, suavização do sinal de fala, evitando interrupções audíveis do sinal de fala, como é comum em sistemas concatenativos.

### 1.2 - Objetivo

O objetivo principal desta dissertação consiste em criar uma aplicação que consiga fazer síntese de voz de forma automática a partir da introdução de um texto arbitrário. Pretende-se ainda que a aplicação possa ser evocada a partir de uma página web por forma a se criar um sistema de síntese online.

Para criar uma aplicação de síntese de fala é necessário dispor de um grande conjunto de locuções de frases, onde as combinações de contexto de fonemas da língua esteja presente, com uma dada prosódia de locução, de forma a treinar os modelos HMM do sistema e criar uma "voz" para a síntese. No contexto desta dissertação o termo "voz" consiste nos modelos treinados a partir das locuções de um único locutor.

O treino de uma "voz" consiste em analisar as locuções gravadas de forma a criar ficheiros de etiquetas de fonemas, de tom da fala, de conteúdo espectral, etc. O sistema avalia, então, o contexto de cada fone (contexto fonológico) nas locuções de forma a "aprender" como é que certo fonema em certa situação é pronunciado, em termos de excitação (F0 ou frequência fundamental), coeficientes cepstrais e duração dos estados dos HMM.

Usualmente estes sistemas são treinados com um único locutor, para reduzir a variabilidade da fala entre locutores. As locuções sintetizadas ficam assim com o timbre do locutor usado no treino.

No âmbito deste trabalho foram usadas as ferramentas HTS [3], que desde há vários anos constituem uma referência da síntese com HMMs.

Este trabalho tem como base uma dissertação feita previamente [2] e visa melhorar a qualidade da fala obtida anteriormente e criar uma aplicação simples, rápida e eficiente de síntese de fala em português. Com isto pretende-se dizer que a aplicação deve ser capaz de sintetizar áudio a partir de um texto ou ficheiro de texto num espaço curto de tempo sem ser preciso uma capacidade tremenda da máquina em que é usada. São ainda usados modelos mais completos para alinhamento do áudio das locuções com as transcrições das frases treinadas e rotinas melhoradas para as mais diversas funções. Nesta dissertação foram usadas as locuções gravadas nessa dissertação anterior que são de excelente qualidade, não sendo necessário fazer a recolha de locuções de novos locutores.

No decorrer da dissertação foram adicionadas algumas funcionalidades à aplicação que inicialmente não estavam previstas e que efetuam a geração dos ficheiros de etiquetas de fones com contexto para a realização do treino dos modelos HMM.

### 1.3 - O sistema HTS

O sistema de treino de voz para síntese denominado de HTS (*HMM-based Speech Synthesis System*) [3] foi desenvolvido no Instituto de Tecnologia de Nagoya, no Japão, e é uma modificação para o sistema HTK (*Hidden Markov Model Toolkit*) [4] que é usado na área de reconhecimento de fala. Modificando o HTK para o seu uso correto na síntese de fala é possível usar as mesmas ferramentas base do HTK. O HTS está disponível com uma licença *Simplified BSD* que permite a sua utilização sem grandes restrições.

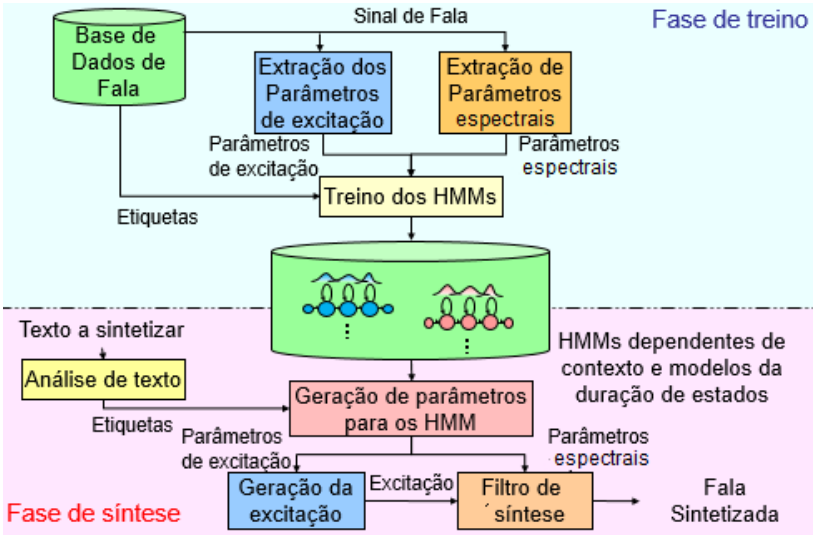



*Figura 1.1 - Diagrama do sistema completo do HTS. Editado de [12]*

O sistema de síntese baseado em HMM está esquematizado na figura 1.1.

Podemos dividir o sistema HTS em duas partes: a fase de treino dos modelos HMM e a fase de síntese.

Na fase de treino o sistema recebe a base de dados de fala, ou seja, as locuções gravadas de um locutor, e a partir de ficheiros de etiquetas de fones e dos parâmetros de excitação (tom) e espectrais criam-se os modelos HMM a usar na fase de síntese.

Na fase de síntese o sistema recebe o texto a sintetizar e, a partir dos modelos HMM resultantes da fase de treino, gera um sinal de fala.

A representação matemática de um sinal de fala neste sistema é feita usando um modelo de tempo discreto da síntese do sinal de fala em termos de excitação e filtro. Este modelo está representado na figura 1.2.

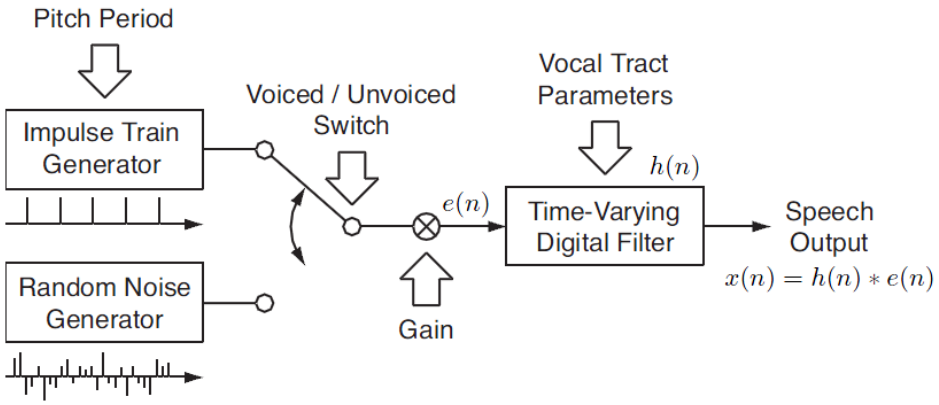



*Figura 1.2 - Modelo de tempo discreto para representação de fala. Retirado de [13]*

Neste modelo de produção de fala temos dois tipos de excitação: um trem de impulsos de período F0 que corresponde à frequência fundamental da fala (*pitch*) para o caso de a fala ser vozeada e ruído branco para o caso da fala ser não vozeada. A função de transferência do filtro $H(z)$ (com resposta a impulso $h[n]$) modela a estrutura do trato vocal a partir dos coeficientes espectrais. Este modelo é uma boa aproximação considerando intervalos de tempo curtos para a síntese da fala. Ao utilizar coeficientes espectrais na escala de mel, a resposta em frequência do filtro adequa-se à sensibilidade do ouvido humano e à sua variação não linear ao longo da gama de frequências.

O filtro de síntese $H(z)$ implementado no HTS é dado pela seguinte fórmula:

$$H(z) = \exp\left[\sum_{m=0}^{M} c(m) z^{-m}\right] \tag{1.1}$$

onde $c(m)$ é o vetor de coeficientes espectrais de tamanho $M + 1$ [12].

Para ajustar o filtro de acordo com a sensibilidade do ouvido humano faz-se uma distorção da sua transformada de z para $z_\alpha$, definindo o sistema como um filtro passa tudo de primeira ordem:

$$z_\alpha^{-1} = \frac{z^{-1} - \alpha}{1 + \alpha\, z^{-1}} \tag{1.2}$$

em que $\alpha$ é o fator de distorção de frequência que para este caso (frequência de amostragem de 48kHz) é de 0.55. Este fator é usado para calcular a frequência distorcida segundo a expressão:

$$\tilde{\omega} = \frac{(1-\alpha^2)\sin\omega}{(1-\alpha^2)\cos\omega - 2\,\alpha} \tag{1.3}$$

Fundamentalmente passa-se de um sistema de predição linear para um sistema de predição usando a escala de Mel.

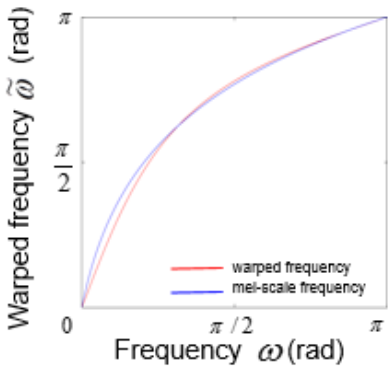



*Figura 1.3 - Exemplo da escala de Mel em frequência. Retirado de [12]*

Para treino dos modelos é necessário que se usem as ferramentas do HTS numa ordem lógica e apropriada. Para isso é usada uma das versões do HTS-demo que contém scripts TCL e Perl, disponibilizados como exemplo de treino do sistema. Estes scripts foram alterados em conformidade com os dados para português europeu. Trata-se dos ficheiros de áudio em formato RAW, os ficheiros de etiquetas dos fones relativos às locuções e os ficheiros de questões para a construção das árvores de decisão. Nesta dissertação usou-se como base a demo *HTS-demo_NIT_BP_F001*, pois esta versão da demo do HTS foi feita para o português brasileiro e surge, por isso, como uma boa base para o português europeu. Para o treino de modelos de português europeu esta versão teve de ser alterada nos seguintes parâmetros:

- Conjunto de fones
- Frequência de amostragem e os parâmetros derivados desta
- Intervalo de frequências para determinação do tom (F0) de cada locutor
- Questões para as árvores de decisão relativas aos fonemas e prosódia

Como a pronúncia do português brasileiro difere do português europeu é necessário alterar o conjunto de fones usados no sistema para um que se adeque ao português europeu.

Nesta dissertação foram usadas locuções gravadas a uma frequência de amostragem de 48kHz e além deste parâmetro é necessário alterar alguns parâmetros como o número de amostras por trama, o período das tramas, o tamanho da FFT e o número de coeficientes cepstrais.

O intervalo de frequências para determinação do tom de cada locutor deve ter em consideração se o locutor é masculino ou feminino. Um locutor feminino terá valores de tom muito maiores que um locutor masculino.

Foi também necessário alterar as questões usadas nas árvores de decisão de acordo com o conjunto fonético utilizado. Aqui utilizaram-se como base as questões utilizadas num projeto anterior [2] e foram revistas para estarem de acordo com as definições utilizadas nesta dissertação.

A demo do HTS necessita, ainda, de ferramentas do SPTK (*Speech Signal Processing Toolkit*) [5] para manipulação dos dados, extração do tom e cálculo dos coeficientes espectrais, e de ferramentas do SoX (*Sound eXchange*) [6] para manipulação de ficheiros de áudio. Nesta dissertação foi ainda usada a ferramenta Praat [7] para extração do tom como alternativa à ferramenta pré-definida do SPTK, o *Snack*.

No processo de criação dos ficheiros de etiquetas de monofones, ao se proceder ao alinhamento das locuções gravadas com as transcrições das frases foi ainda necessário usar algumas ferramentas do HTK. Neste passo foi necessário utilizar um reconhecedor de fonemas, modelos de fones que foram desenvolvidos no laboratório, um dicionário de transcrições fonéticas e um sistema de alinhamento para transcrição automática das locuções.

Os modelos de fones usados nesta etapa foram criados no laboratório e é um produto de muito tempo de desenvolvimento que resulta num recurso muito valioso para este trabalho.

O dicionário de transcrições fonéticas deve ter uma grande variedade de palavras para ser possível fazer uma transcrição o mais correta possível. Esta é uma etapa crítica para todo o processo de treino pois trata-se da atribuição de excertos de fala a cada fonema e que caracterizarão este fonema durante todo o processo de treino e servirão para correspondência com os fonemas da frase de síntese. É claro que é impossível ter todas as palavras e todas as variações de uma palavra num dicionário destes. No entanto, este problema pode ser resolvido com a criação de um dicionário estendido com as transcrições adicionais necessárias para cada situação, ainda que dependendo da lista de palavras em falta possa ser uma tarefa demorada e morosa. Além desta dificuldade acrescenta-se a necessidade de transcrever os grafemas de acordo com a pronunciação do locutor que, na maior parte dos casos, não difere da convenção normal, mas que pode causar alguns problemas de etiquetagem. Tome-se o exemplo da palavra vermelho que pode ter duas pronunciações “vermelho” ou “vermâlho”.

O contexto dos fones é ainda estendido para além dos trifones criando um contexto com dois fonemas anteriores e posteriores ao central (os dois anteriores à esquerda e os dois posteriores à direita) originando pentafones.

A síntese propriamente dita é feita usando uma ferramenta do sistema HTS denominada por *hts_engine* [8], que a partir da "voz" modelada pelo HTS consegue gerar fala seguindo uma frase que lhe é dada como parâmetro. Este sistema de síntese foi desenvolvido pelos mesmos investigadores que criaram o HTS e serve, então, como complemento final do sistema de treino. O sistema de treino e o sistema de síntese são as duas partes fulcrais do processo realizado nesta dissertação.

Como entrada do sistema de treino temos, portanto, um texto que pode ser composto por várias palavras e frases. Para se obter uma voz, o texto de entrada tem de ser transcrito para um alfabeto fonético que serve de base às unidades fonéticas do sintetizador (adotando aqui a entidade fonema na representação da classe de cada som-fone). Para isso, o texto é reduzido a uma unidade ortográfica deste, os grafemas, e convertido, para a sequência de fonemas correspondente. O alfabeto fonético usado nesta dissertação é o SAMPA *(Speech Assessment Methods Phonetic Alphabet*) [9], que usa carateres do código ASCII que podem ser introduzidos por um teclado normal de computador, para representar cada fonema. Esta solução torna-se prática pois é possível representar cada fonema apenas com um símbolo do teclado e tem por base o alfabeto fonético internacional (IPA) [10]. Escolheu-se usar o SAMPA com um só caractere por fonema (quando existem diacríticos nasais ou de tónica) para representar um fonema para aumentar a simplicidade e legibilidade dos fonemas. A nível de alfabeto foram ainda distinguidas as vogais em posição tónica das restantes para dar ênfase no processo de treino a estes fones e melhorar a qualidade da fala na síntese.

| Vogais | | | | | Consoantes | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Tipo | Fonema | Grafema | Exemplo | Transcrição Fonética | Tipo | Fonema | Grafema | Exemplo | Transcrição Fonética |
| Orais | ª | a | abadia | ª b ª d í ª | Nasais | v | v | abusiva | ª b u z í v ª |
| | a | a, á | abala | ª b á l ª | | m | m | acabamento | ª k ª b ª m Ë t u |
| | e | e | àquele | á k e l @ | | n | n | cineasta | s i n i á S t ª |
| | i | i | àquilo | á k i l u | Fricativas | s | s, ç, c | observa | O b s É r v ª |
| | o | o, ô | boiar | b o i á r | | S | ch, s, z, x | prancha | p r Ã S ª |
| | u | u, o | tudo | t u d u | | f | f | ferreira | f @ R â i r ª |
| | O | o, ó | boné | b O n É | | z | z, s, x | presença | p r @ z Ë s ª |
| | E | e, é | caráter | k ª r á t E r | | Z | j, g, s, x, z | colagem | k u l á Z ã ï |
| | @ | e | de | d @ | Líquidas | J | nh | colarinho | k u l ª r í J u |
| Nasais | ã | ã, am, an | andorra | ã d ô R ª | | l | l | explorou | @ S p l u r ô |
| | ë | en | credenciação | k r @ d ë s i ª s Ã ü | | r | r | monteiro | m õ t â i r u |
| | ï | im, in | impossível | ï p u s í v E l | | R | rr | salvaterra | s a l v ª t É R ª |
| | õ | om, in | compostela | k õ p u S t É l ª | | L | lh | velho | v É L u |
| | ü | um, un | hungria | ü g r í ª | Plusivas | p | p | campeão | k ã p i Ã ü |
| | | | | | | t | t | tiago | t i á g u |
| | | | | | | k | k, c | fonseca | f õ s ê k ª |
| | | | | | | b | b | gabão | g ª b Ã ü |
| | | | | | | d | d | holanda | O l Ã d ª |
| | | | | | | g | g | inglaterra | ï g l ª t É R ª |

*Tabela 1.1 - Tabela do alfabeto fonético utilizado*

O alfabeto fonético utilizado encontra-se especificado na tabela 1.1. As vogais presentes na tabela podem ser diferenciadas pela sua inclusão numa sílaba tónica.

Para a conversão de grafemas em fonemas é utilizado o sistema G2P (grapheme-to-phoneme) desenvolvido no laboratório [11], que além de dispor de um dicionário de transcrições fonéticas de palavras bastante grande, usa também modelos para transcrever palavras que não existam no dicionário de pronunciação. Juntando estes dois métodos é possível obter-se uma transcrição fonética bastante fiável de qualquer palavra.

Para o conversor funcionar de forma eficiente e fiável é necessário que o texto que chega ao conversor fonético esteja normalizado, ou seja, os números cardinais sejam substituídos pela sua representação ortográfica, as datas e as horas escritas por extenso, a normalização de abreviaturas como Dr (doutor), Sr (senhor), Engº (engenheiro), ºC (graus centígrados) ou cm (centímetro), entre outras, a pronunciação dos endereços de email ou de websites, como a leitura de um "." como "ponto" e www como três palavras em vez de três letras juntas sem pronunciação e a leitura de números romanos corretamente.

# Capítulo 2 - Modelos de Markov Não Observáveis (HMM)

## 2.1 - Introdução aos HMM

O sistema HTS sobre o qual se baseia esta dissertação, usa HMM ou modelos de Markov não observáveis para representar cada fonema sob a forma de um modelo estatístico.

Os HMM são usados atualmente nas áreas de reconhecimento de fala e de síntese de fala. Além destas áreas de maior interesse desta dissertação, os HMM são também usados em outras áreas de reconhecimento de padrões temporais não relacionadas com a fala como o reconhecimento de gestos, de escrita ou até *part-of-speech* (POS) *tagging*, onde se classifica a palavra em relação à sua função na frase.

Os HMM baseiam-se na cadeia de Markov em que cada estado da cadeia é representado por uma função densidade de probabilidade (PDF). No entanto, a sequência de estados usados nos HMM não é conhecida e, por isso, denominam-se como modelos de Markov não observáveis (*hidden*). No caso do HTS a sequência de estados de HMM é composta por 5 estados para o caso do tom e do espectro e por 3 estados para a duração de cada fone. O modelo HMM diz-se duplamente estocástico, pois além dos estados da cadeia de Markov, a cada estado está associada uma PDF relativa à distribuição dos dados nesse estado.

Um HMM usa uma máquina de estados finita que gera uma sequência de estados discreta (a cadeia de Markov). Só é possível usar esta aproximação para os sinais de fala considerando uma análise em tempo curto, pois sabendo que os sinais de fala não são estacionários no tempo, ou seja, a suas propriedades variam constantemente no tempo, apenas é possível definir um estado para um segmento extremamente curto de um sinal de fala considerando que neste pequeno espaço de tempo o sinal de fala é estacionário.

Sendo assim, em cada intervalo de tempo existe uma observação que é representada por um estado na cadeia de Markov e a cada intervalo de tempo há uma transição entre estados da cadeia de Markov de acordo com a matriz da probabilidade de transição de estados $\mathbf{A} = \{a_{ij}\}$ em que $a_{ij}$ é a probabilidade de transição do estado *i* para o estado *j*. Esta matriz é um dos parâmetros que define um HMM. Os outros parâmetros são a PDF de cada estado que é definida como a distribuição de saída de cada estado e pela probabilidade de ocupação inicial do estado.

A PDF de cada estado é composta por uma soma ponderada de *M* Gaussianas, dada pela expressão

$$b_i(\boldsymbol{o}) = \sum_{m=1}^{M} w_{im}\, \mathcal{N}(\boldsymbol{o}; \mu_{im}, \boldsymbol{\sigma_{im}}) \tag{2.1}$$

em que para uma observação **o** no estado *i* temos que a densidade de probabilidade é dada pela soma das *M* Gaussianas definidas pela densidade de probabilidade Gaussiana com média $\boldsymbol{\mu_{im}}$ e matriz de covariância **σ**. Cada uma das Gaussianas tem um peso próprio na soma dado pelo vetor **w**. Uma PDF Gaussiana é definida como:

$$\mathcal{N}(o; \mu_{im}, \mathrm{U}_{im}) = \frac{1}{\sqrt{(2\pi)^L |\sigma|}} \exp\left[-\frac{1}{2}(o-\mu_{im})^T \sigma_{im}^{-1}(o-\mu_{im})\right] \tag{2.2}$$

onde *L* é o número de componentes do vetor de observações do HMM.

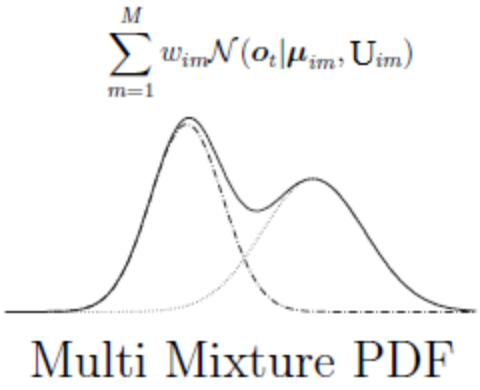



*Figura 2.1 - Mistura de PDF Gaussianas. Editada de [13]*

Na figura acima temos um exemplo de uma mistura de duas PDFs Gaussianas representadas pelas linhas a tracejado.

Uma cadeia de Markov pode ter dois tipos de estrutura: um modelo ergódico em que de um estado se pode transitar para outro qualquer ou um modelo esquerda direita em que só se pode transitar do estado atual para o estado à direita. No contexto do HTS faz mais sentido usar um modelo esquerda-direita, pois representa-se em cada cadeia uma sequência de eventos fonéticos, ou seja, uma sequência de eventos para cada fonema.

Como já foi referido, a cadeia de Markov usada no HTS tem cinco estados emissores no caso dos modelos de tom e espectro. No entanto, surgem ainda dois estados não emissores, ou seja, que

apenas servem para concatenação entre HMMs. Na figura temos uma ilustração de uma cadeia de Markov neste formato.

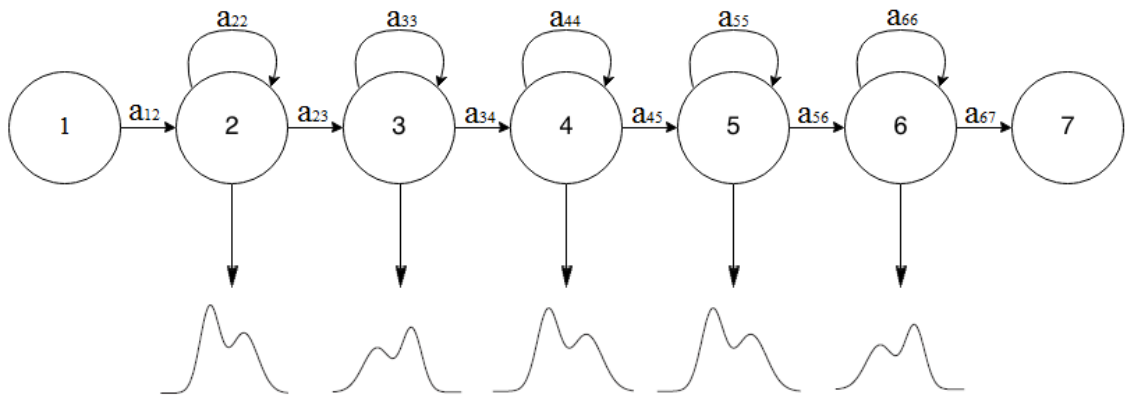



*Figura 2.2 - Cadeia de Markov com 5 estados emissores (estado 2 a 6).*

Cada linha da matriz de transição entre estados é formada pelas probabilidades de permanecer no mesmo estado, $a_{22}$ por exemplo, e pelas probabilidades de transitar para o estado seguinte, ou seja, à direita, $a_{23} = 1 - a_{22}$, por exemplo, sendo estas a duas únicas possibilidades em cada estado. A cada estado é associada uma mistura de PDF Gaussianas, conforme sugerido pela figura.

Cada HMM caracteriza um monofone (um fone independente do contexto à esquerda ou à direita na sequência de fones) ou um pentafone, conforme a fase do treino do HTS. Concatenando vários HMMs obtêm-se conjuntos fonéticos maiores formando palavras e consequentemente frases.

Tendo como base os dados apresentados até agora, podemos concluir que para se definir um modelo HMM basta uma matriz de transição de estados **A**, e ainda, por estado, de *M* vetores de média e *M* matrizes de covariância (formada à custa de *M* vetores de variância; isto é, as matrizes são diagonais), além dos *M* pesos de cada componente da mistura Gaussiana.

Os modelos HMM são usados para descrever as observações acústicas da fala em termos da representação espetral dos seus fones. Mas para a síntese de fala, temos ainda de associar um dado valor do tom da fala a uma observação (que será zero no caso de fala não vozeada como acontece em consoantes não vozeadas). Uma observação está então dividida em duas camadas: parâmetros relativos ao espetro e um parâmetro do tom da fala (a frequência fundamental do sinal de fala, usualmente referido como F0). No HTS o parâmetro de tom corresponde ao logaritmo de F0. Como foram usadas locuções amostradas a uma frequência de 48kHz, por apresentarem melhor qualidade

de áudio, o número de coeficientes para definição do cepstrum aumenta em relação ao número de coeficientes usados para amostragem a 16kHz, como na dissertação anterior deste tema [2]. O número de parâmetros cepstrais tem um total de 105 coeficientes: 35 para os coeficientes estáticos, 35 para os delta, correspondendo a uma aproximação da primeira derivada temporal dos coeficientes, e 35 para os delta-delta, correspondendo à segunda derivada. As fórmulas de cálculo que se seguem [12] representam cada coeficiente como $c_t$:

$$\Delta c_t = \frac{\partial c_t}{\partial t} \approx 0.5(c_{t+1} - c_{t-1}) \tag{2.3}$$

$$\Delta^2 c_t = \frac{\partial^2 c_t}{\partial t^2} \approx c_{t+1} - 2c_t + c_{t-1} \tag{2.4}$$

A ilustração do método de cálculo dos coeficientes delta e delta-delta (também denominados coeficientes dinâmicos) pode ser vista na imagem seguinte:

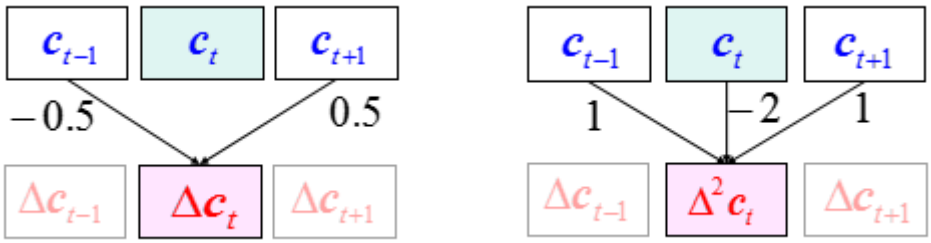



*Figura 2.3 - Cálculo dos coeficientes dinâmicos. Retirado de [12]*

Os restantes parâmetros consistem na definição do logaritmo de F0, ou seja, indicam se o estado atual retrata um segmento de fala vozeado ou não vozeado e a sua frequência fundamental. De notar que no caso do não vozeado o F0 é nulo e o seu logaritmo seria menos infinito. Por forma a contornar este pormenor considera-se que o valor do logaritmo, neste caso, é um número negativo muito grande, $-10^{10}$.

Para a representação da variação do tom é necessário ir mais além que os tradicionais HMM, pois os segmentos não vozeados não têm definição em frequência e surge uma lacuna entre a representação em frequência dos segmentos vozeados e não vozeados, pois se temos uma representação contínua dos valores da frequência nos segmentos vozeados, nos segmentos não vozeados temos apenas um valor discreto.

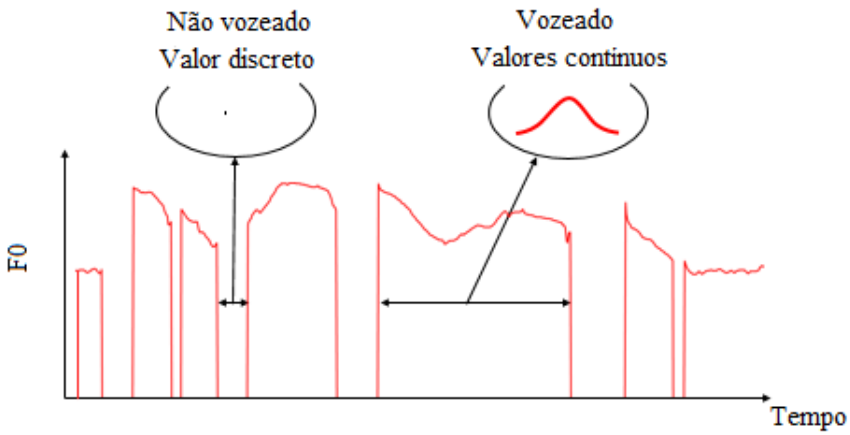



*Figura 2.4 - Exemplo da variação do tom (F0). Editado de [12]*

Na figura acima vemos que nas regiões vozeadas temos a definição do tom no conjunto de valores contínuos e nas regiões não vozeadas o tom tem apenas um valor discreto.

Por isso, é necessário representar a variação em frequência do segmento e indicar se é vozeado ou não vozeado. Surge, então, o conceito de MSD-HMM (*multi space probability distribution HMM*) [14]. Neste tipo de HMM temos dois níveis de representação para o tom: um de dimensão unitária para representação contínua do valor de log(F0) no atual segmento e outro para a representação do valor discreto que define o segmento como não vozeado. Temos então dois espaços de probabilidade, cada um com o seu peso que determina a probabilidade de ser vozeado ou não, para dois tipos de representação do tom. A esta representação juntam-se mais dois parâmetros MSD-HMM: os delta e delta-delta do tom.

## 2.2 - Síntese a partir dos HMM

Tendo os HMMs definidos no treino, o problema passa a centrar-se em como gerar parâmetros para a fala a partir dos HMM treinados, ou seja, parâmetros de tom e de espectro de potência do segmento de áudio. Isto consiste em reverter o processo de treino e a partir das definições de um HMM (médias e variâncias de tom e de cepstrum) gerar um sinal de fala.

A solução passa por obter nos HMM dados uma sequência de vetores de parâmetros que maximize a probabilidade $P(O|\lambda,T)$ [13], ou seja, cuja sequência de vetores de observação **O** seja a mais aproximada possível à pretendida em relação aos HMM $\lambda$ e com uma sequência de $T$ observações. Para isso utiliza-se o algoritmo de Viterbi [20]. Este algoritmo encontra os estados mais prováveis para um HMM que satisfaça as condições apresentadas em cima.

Nesta fase foram obtidos os estados dos HMM para os fonemas das frases a serem sintetizados. Após concatenação dos HMM de vários fonemas podemos ter discrepância entre as

médias e variâncias entre estados que podem resultar em transições abruptas na produção do som e que eliminem grande parte da naturalidade do som.

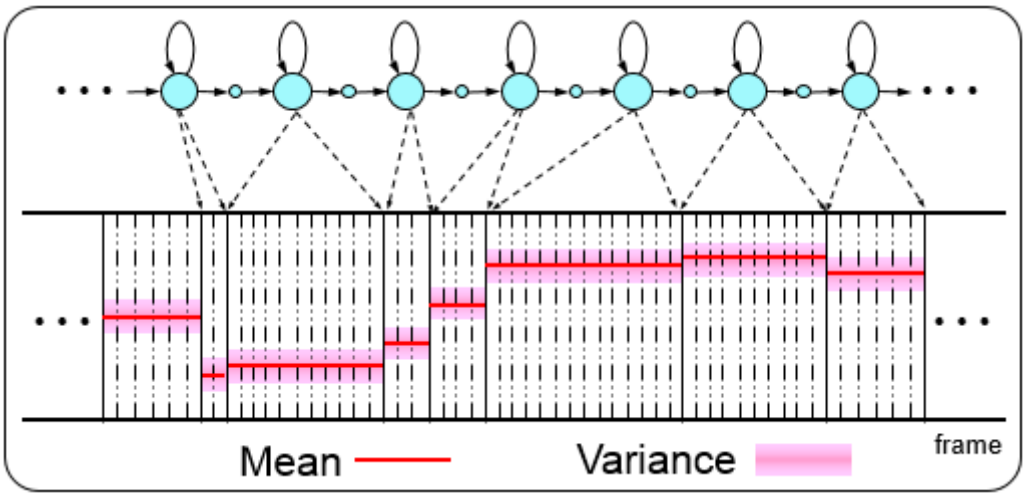



*Figura 2.5 - Parâmetros de HMM concatenados. Retirada de [12]*

Por isso, é utilizada a informação dos coeficientes dinâmicos (delta e delta-delta) para suavizar as transições entre estados e melhorar a qualidade do áudio, dentro das restrições impostas pela variância global. A utilização dos coeficientes dinâmicos permite suavizar as transições entre HMM, pois os 3 coeficientes (estático, delta e delta-delta) limitam-se mutuamente e como resultado as suas trajetórias tornam-se mais realistas e as transições mais suaves, como exemplificado na figura seguinte.

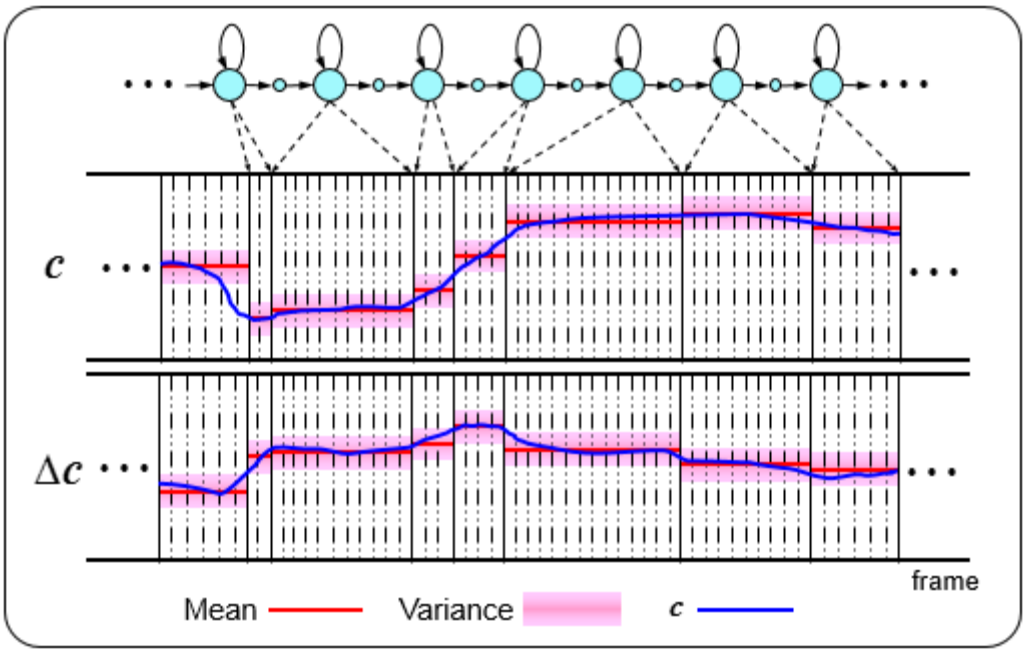



*Figura 2.6 - Parâmetros dos HMM concatenados usando coeficientes dinâmicos. Retirada de [12]*

O resultado deste processo é mais evidente ao analisarmos a síntese de segmentos de áudio que contenham um fone de silêncio não vozeado (sil) e outro vozeado.

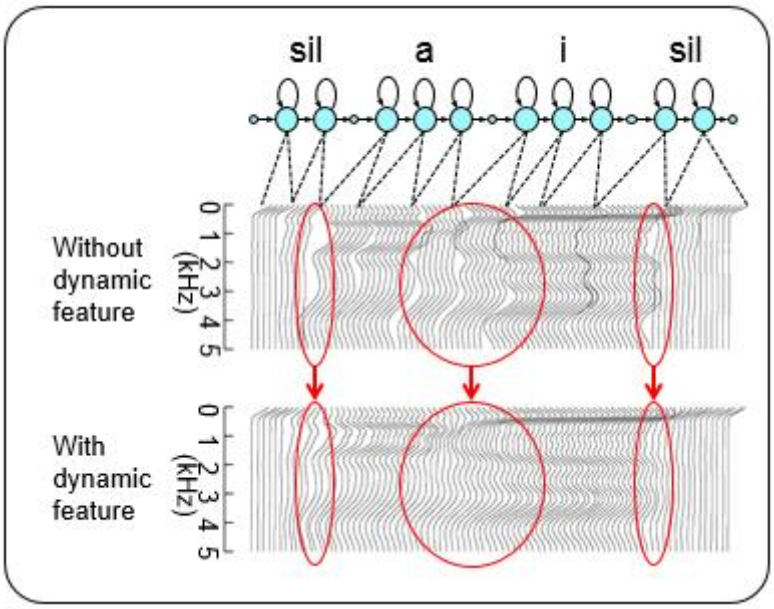



*Figura 2.7 - Análise em frequência da síntese a partir de HMM com e sem coeficientes dinâmicos. Retirado de [12]*

A partir desta imagem pode-se observar que as transições abruptas entre os silêncios (sil) e os outros fones são amenizadas usando os coeficientes dinâmicos (*dynamic feature*) nas zonas assinaladas a vermelho.

# Capítulo 3 - Treino da voz

## 3.1 - Ficheiros necessários para o treino

Para sintetizar fala é necessário ter uma "voz" que consiste nos modelos HMM que são criados a partir das locuções gravadas e que são a base do sistema de síntese. Por isso, os ficheiros de entrada do sistema de treino serão a personalização desta voz em conformidade com o locutor e o idioma em que as locuções são gravadas.

O sistema de treino do HTS tem os seguintes ficheiros de entrada:

- Ficheiros de áudio das locuções
- Ficheiros de etiquetas de monofones alinhados temporalmente com as locuções
- Ficheiro de etiquetas de pentafones alinhados temporalmente com as locuções em conformidade com os ficheiros de etiquetas de monofones
- Ficheiros de questões para a criação das árvores de decisão

Os ficheiros de áudio derivam das gravações efetuadas no âmbito da dissertação anterior [2] a partir da leitura das frases ricas foneticamente pelos locutores.

Os ficheiros de etiquetas de monofones servem para ter uma base de aprendizagem da forma como o locutor pronuncia as palavras, seja em termos de linguagem, entoação ou duração (tempo de cada fonema).

Os ficheiros de etiquetas de pentafones são criados com base nos ficheiros dos monofones e acrescentam contexto a cada fonema, dando informação de coarticulação entre fonemas. Forma-se, então, um conjunto de cinco fonemas para contextualização de cada fonema individual à qual se junta toda a informação a nível de sílabas, palavras, frases, pontuação e a classificação gramatical da palavra de contexto (*part-of-speech* ou *POS*). Mais à frente estes parâmetros são discutidos.

## 3.2 - Criação das árvores de decisão

Na análise da fala é necessário levar em conta o contexto de cada fonema, ou seja, o mesmo fone pode ter diferentes tipos de parâmetros, por exemplo a entoação e a duração, consoante a sua posição na sílaba, na frase, no texto e, principalmente, consoante os fonemas que o antecedem ou sucedem. Neste passo é introduzida a nova informação de prosódia como a tonicidade das sílabas e dos fonemas, a pontuação, entre outros.

Existe um número muito grande de possibilidades de sequências fonéticas pelo que é impossível ter numa base de dados de fala todas as ocorrências possíveis. Ao agrupar em *clusters* os fones que partilhem o mesmo contexto é possível usar estes *clusters* para a geração de modelos HMM para fones com contexto que não estejam presentes no treino do sistema. Por isso é que o HTS usa o agrupamento de modelos com o mesmo contexto (*tree-based clustering*).

O contexto de um fone é definido ao identificar conjuntos de fones vozeados, fones de vogais, fones cuja posição da língua na sua pronúncia seja igual, seja a nível de altura (alta, central ou baixa) de profundidade (anterior, posterior ou central) e a abertura da boca na sua pronúncia. Pegando nestes conjuntos iniciais e definindo outros resultantes da interceção destes, geram-se conjuntos mais restritos e mais próximos de definir um *cluster*. Seguindo esta lógica para o número de palavras, sílabas, fones e frases aumenta-se a variedade de perguntas e a sua área de incidência no contexto.

Para efetuar esta divisão são usadas árvores de decisão com base em perguntas definidas nos ficheiros em questão e que são binárias pois apenas têm duas respostas possíveis: ou o fone a ser classificado está num conjunto de fones que têm o parâmetro designado na pergunta e a resposta é afirmativa ou caso contrário o fone não está neste conjunto e a resposta é negativa.

No âmbito da síntese, as árvores de decisão criadas servirão para definir os estados de HMM que melhor descrevem um fone num dado contexto.

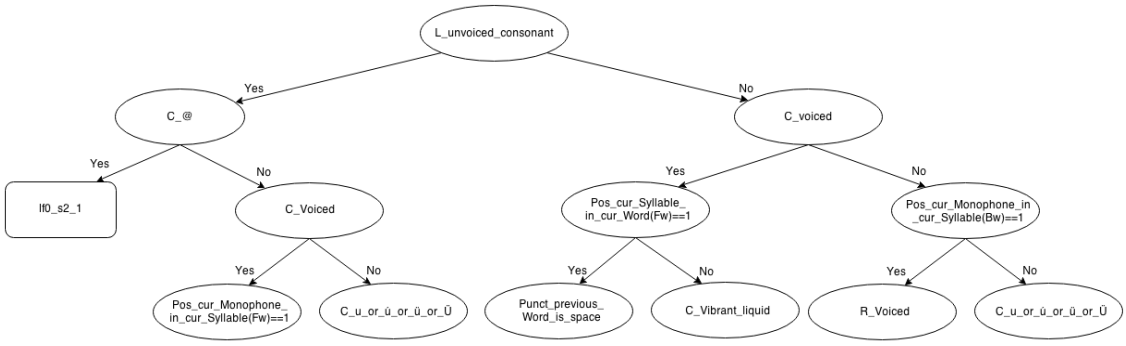



*Figura 3.1 - Exemplo de uma árvore de decisão*

Na figura acima está exemplificada uma árvore de decisão. Esta árvore foi obtida num treino de dados do HTS para o primeiro estado do HMM do tom. Os círculos representam nós da árvore cujo nome é o nome da questão presente no ficheiro de questões dado ao HTS. O quadrado à esquerda representa uma folha da árvore que diz que estado é o indicado para o caso do fone central ser um "@" (o C representa fone *central*) e o fone à esquerda ser uma consoante não vozeada (o L representa *left*), ou seja, um *cluster* apropriado para este fone com este contexto. A árvore acima

não está completa, pois uma árvore completa é enorme e neste caso tinha 640 nós possíveis. Correndo a árvore decisão é possível encontrar um *cluster* adequado para o fone a ser analisado.

A partir do ficheiro de questões o sistema cria um outro ficheiro que especifica as perguntas que melhor dividem os modelos por conjuntos e indica os nós apropriados conforme a resposta às perguntas. Para isso é usado um critério de MDL (*minimum discription length*) [13], ou seja, para cada modelo é usado o menor número de perguntas que seja possível para encontrar um *cluster* que melhor o define.

Os ficheiros de questões servem, então, como indicação de que perguntas o sistema de treino do HTS pode fazer uso para divisão dos modelos HMM por grupos cujos parâmetros dos modelos constituintes sejam aproximadamente iguais.

```
QS  "L_Anterior_and_high_vowel"       {*^i-*,*^í-*,*^ï-*,*^Ï-*}
QS  "L_Anterior_and_middle_vowel"     {*^e-*,*^ê-*,*^E-*,*^É-*,*^ë-*,*^Ë-*}
QS  "L_Anterior_and_closed_vowel"     {*^e-*,*^ê-*,*^i-*,*^í-*,*^ë-*,*^Ë-*,*^ï-*,*^Ï-*}
QS  "L_Anterior_and_open_vowel"       {*^a-*,*^á-*,*^E-*,*^É-*}
QS  "L_Anterior_and_reduced_vowel"    {*^a-*,*^á-*,*^i-*,*^í-*}
QS  "L_Anterior_and_oral_vowel"       {*^a-*,*^á-*,*^E-*,*^É-*,*^e-*,*^ê-*,*^i-*,*^í-*}
QS  "L_Anterior_and_nasal_vowel"      {*^ë-*,*^Ë-*,*^ï-*,*^Ï-*}
QS  "L_Central_and_open_vowel"        {*^ª-*,*^ã-*,*^Ã-*}
QS  "L_Central_and_nasal_vowel"       {*^ª-*,*^ã-*,*^Ã-*,*^@-*}
QS  "L_Posterior_and_high_vowel"      {*^ª-*,*^â-*,*^ú-*,*^ü-*,*^Ü-*}
QS  "L_Posterior_and_middle_vowel"    {*^O-*,*^Ó-*,*^o-*,*^ô-*,*^õ-*,*^Õ-*}
QS  "L_Posterior_and_closed_vowel"    {*^o-*,*^ô-*,*^ª-*,*^â-*,*^ú-*,*^õ-*,*^Õ-*,*^ü-*,*^Ü-*}
QS  "L_Posterior_and_reduced_vowel"   {*^ª-*,*^â-*,*^ú-*}
```

*Figura 3.2 - Exemplo de um ficheiro de questões para as árvores de decisão*

Na figura acima podemos ver na segunda coluna o nome da pergunta, ou seja, qual o parâmetro do contexto do fone está a ser analisado, e na terceira coluna o conjunto de respostas que se enquadram nesse contexto. Na primeira coluna apenas estão os caracteres "QS" que na leitura do ficheiro indicam a presença de uma pergunta nessa linha.

### 3.3 - Locuções

As locuções que são usadas no sistema de treino para criar a "voz" são uma parte essencial do processo.

O áudio capturado deve ter boa qualidade e o locutor deve ter uma boa dicção e controlo da voz para evitar entoar em demasia durante o processo de gravação por forma a ter locuções da mesma palavra ou fone iguais ou quanto muito semelhantes. Se o objetivo é extrair médias e variâncias de segmentos de fala, uma variação de tom da locução do mesmo fone em sítios diferentes gera uma discrepância e pode levar a erros no treino. Este fator generalizado à maior parte dos fonemas pode ser crítico na obtenção dos parâmetros dos HMM.

Para uma boa base de treino é importante que sejam gravadas várias frases que sejam ricas foneticamente, para se ter a maior combinação de fonemas possível, e que se possua um total de tempo de gravação aceitável.

De acordo com estudos feitos previamente, uma base de dados aceitável de voz para estes sistemas tem de ter, pelo menos, uma hora de locuções. Numa tese feita anteriormente foi estudada a classificação de frases ricas em termos fonéticos que consistem num conjunto de frases que garantem o número de combinações de trifones (conjunto de três fones consecutivos) o maior possível e que tornam o processo de treino mais abrangente e fiável. Baseado nisso têm-se 540 frases para o processo de treino, as quais geram um pouco mais de uma hora de locuções que podem ser usadas para o treino, já após a edição e processamento destas.

Temos, então, três conjuntos de locuções: duas masculinas e uma feminina. Com base nos critérios apresentados anteriormente pode-se concluir que o conjunto de locuções masculino (**CB**) é o que tem melhor qualidade, tendo as 540 frases gravadas e apresentando locuções mais constantes e sem grande diferença de entoação. O outro conjunto de locuções masculino (**RM**) não tem as 540 frases completas, estando apenas disponíveis 300 dessas frases. À semelhança do conjunto de locuções feminino (**DG**), este último conjunto de locuções masculino apresenta variações de tom ao longo das frases.

Na fase do alinhamento das locuções foram ainda corrigidos alguns erros nas frases e adaptaram-se algumas frases para estarem de acordo com as locuções. Estes erros tinham a ver, essencialmente, com corte nas locuções ou troca de palavras em relação às locuções.

### 3.4 - Criação dos ficheiros de etiquetas de monofones

Para preparar os ficheiros de treino foi necessário alinhar os ficheiros de áudio das locuções com a transcrição fonética das frases gravadas num processo denominado de alinhamento de monofones ou de etiquetagem dos fonemas.

Começou-se por transcrever as palavras das frases gravadas a partir de dicionários existentes para o efeito. Para os casos pontuais em que o dicionário não continha a transcrição adicionou-se uma entrada neste por forma a completar o processo. De notar que o dicionário de transcrições era muito completo, mas as frases podem conter, ainda assim, palavras não existentes ou o plural de palavras existentes no dicionário, como foi o caso da maior parte destas ocorrências.

Tendo completado o processo de transcrição das palavras segue-se o alinhamento das transcrições com o áudio. Para isso utilizaram-se ferramentas do HTK: *HDMan*, *HHEd* e *HVite*.

Neste processo a função da ferramenta *HDMan* é de criar uma lista de trifones, ou seja, conjuntos de possíveis sequências de 3 fones a partir do dicionário criado da lista de palavras das frases das locuções.

De seguida, é usada a ferramenta *HHEd* que, partindo da lista de trifones e de um conjunto de modelos previamente gerados para o reconhecimento de fala no projeto "TecnoVoz" [17], gera um ficheiro com os modelos HMM dos trifones das frases treinadas.

Tendo os modelos dos trifones feitos usa-se a ferramenta *HVite* para fazer o alinhamento propriamente dito usando o algoritmo de Viterbi e guarda o resultado do alinhamento num ficheiro MLF (*master label file*) que contém o alinhamento de todas as frases gravadas com o áudio, ou seja, contém o tempo inicial e final de cada fone de acordo com os modelos de trifones criados pela ferramenta *HHEd*.

Após executado este processo dividiu-se o MLF em ficheiros de etiquetas de monofones individuais para cada frase e usando o software *Transcriber* [18] analisou-se o alinhamento para despistar erros das locuções como diferenças entre o discurso e a frase escrita ou erros na gravação das locuções como um corte da locução a meio da frase. Além disso, confirmou-se o tempo dos fones em cada locução e fizeram-se algumas correções mínimas nos tempos de alguns fones.

Na dissertação anterior deste tema foi usada a coarticulação entre palavras para o alinhamento fonético. Na presente dissertação experimentou-se deixar de lado esse aspeto e verificou-se que o alinhamento ficou igual ou ligeiramente melhor devido também à melhoria das ferramentas, mas principalmente devido à grande melhoria dos modelos de trifones.

De seguida apresenta-se o resultado do alinhamento de algumas frases em relação à dissertação anterior da mesma área [2] que usou a coarticulação entre palavras para este passo.

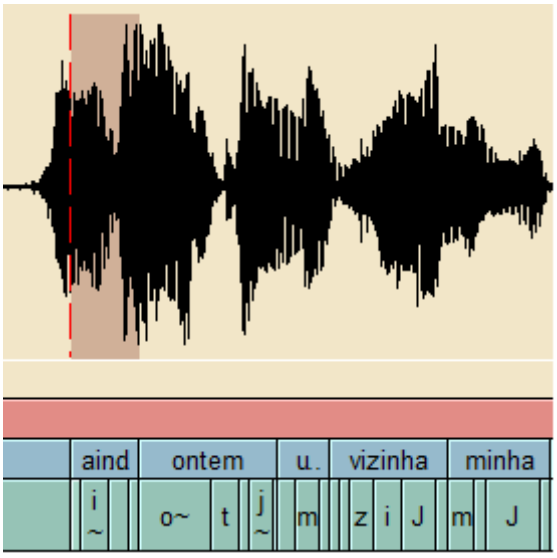



*Figura 3.3 – Alinhamento anterior da frase 1*


*Figura 3.4 - Alinhamento atual da frase 1*

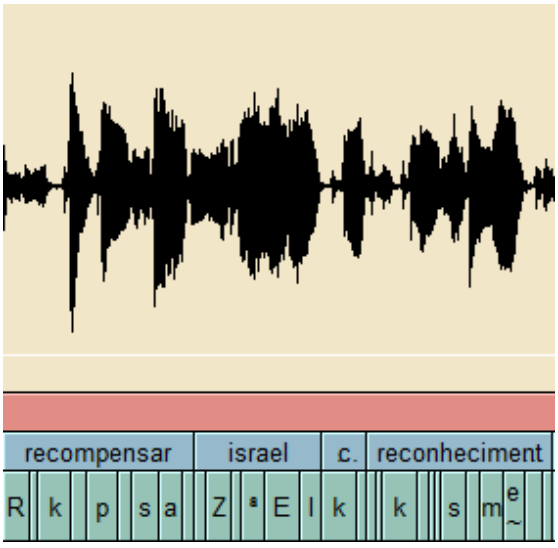



*Figura 3.5 - Alinhamento anterior da frase 70*

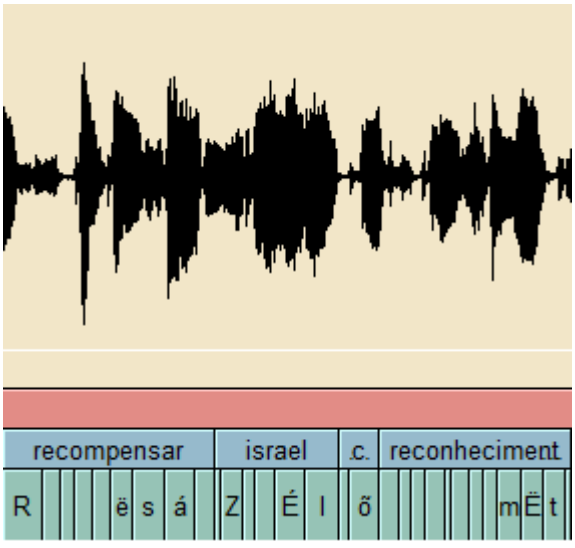



*Figura 3.6 - Alinhamento atual da frase 70*

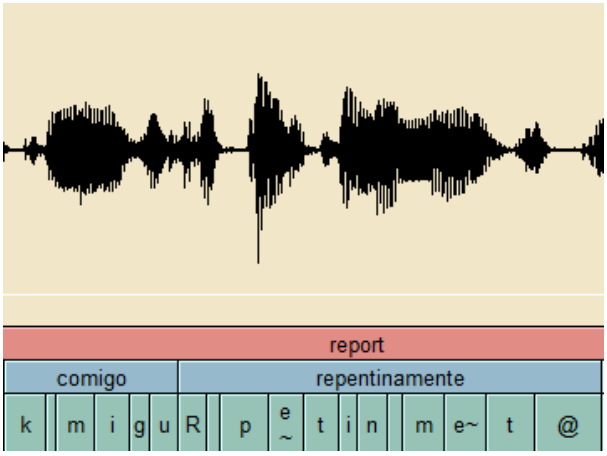



*Figura 3.7 - Alinhamento anterior da frase 489*

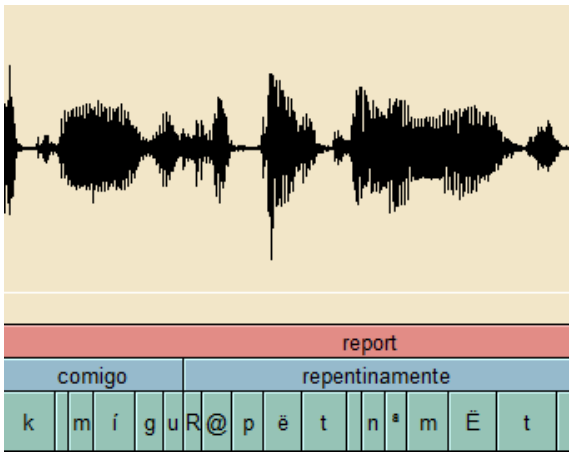



*Figura 3.8 - Alinhamento atual da frase 489*

Como se pode ver pelos três pares de figuras acima, o alinhamento atual é muito semelhante ao alinhamento efetuado na dissertação anterior a esta [2], vendo uma melhoria na frase 1 em que a palavra "ainda" que é a primeira palavra da frase não está exatamente bem alinhada com o áudio. Tendo em conta que o alinhamento anterior foi feito usando coarticulação entre palavras e no atual foram apenas usados os modelos mais recentes pode-se ver uma evolução nas ferramentas usadas.

### 3.5 - Primeira etapa do treino: extração dos parâmetros espectrais e de tom

O primeiro passo do treino é o cálculo dos coeficientes cepstrais da locução. Estes parâmetros resultam da análise do espectro de potência do som. Neste sistema é usada a representação do espectro de potência do som na escala de mel que é uma escala não linear que aproxima a representação das frequências à capacidade auditiva do ser humano tornando-se mais apropriada para o objetivo do sistema em questão.

Para este processo são utilizadas ferramentas da biblioteca SPTK. Na dissertação anterior foi usada uma frequência de amostragem de 16kHz para extração dos parâmetros a partir do áudio, mas neste momento o próprio HTS já permite o uso de frequências de amostragem de 48kHz que, apesar de tornar o processo de treino mais demorado (passa de 7-8 horas para 12-13 horas num computador normal) aumenta a qualidade do áudio produzido. De acordo com este fator são analisadas tramas de 25ms (1200 amostras), a cada uma é aplicada uma janela de *Hamming* e calculada a DFT de 2048 pontos, para posterior cálculo dos coeficientes MGC de ordem 35, ou seja, são calculados 35 coeficientes. Posteriormente são calculadas as derivadas de cada coeficiente, os delta e delta-delta.

De seguida faz-se a extração do tom do áudio. Este parâmetro é muito importante na síntese de fala natural e inteligível, pois vai melhorar a prosódia. Uma extração de tom sem sucesso torna a fala muito pouco natural e pouco fluida. O sistema HTS implementa, por defeito, o algoritmo de extração do tom da biblioteca *Snack*. Foi testado também o algoritmo usado na ferramenta *Praat* que tem mostrado bons resultados na execução deste passo [16]. O tom refere-se à determinação dos excertos vozeados e não vozeados do áudio gravado e à determinação da frequência fundamental do som (F0).

Após a determinação do tom são calculados os parâmetros delta e delta-delta de cada valor usando as fórmulas (2.3) e (2.4), respetivamente.

Por forma a dividir os parâmetros MGC e do tom o HTS usa o conceito de *streams*. Isto significa que para cada locução é gerado um *stream*, ou um bloco contíguo, que contém os valores dos MGC e os delta e delta-delta e outro que contém os valores do tom e dos seus delta-delta.

Por fim é gerado um ficheiro que resulta da concatenação dos *streams* dos valores de MGC e do tom.

Após o cálculo destes parâmetros é que se inicia o treino, propriamente dito.

O treino pode-se dividir em duas etapas: estimação de parâmetros dos HMM a partir dos monofones e posterior estimação dos HMM finais a partir dos pentafones.

### 3.6 - Segunda etapa do treino: treino a partir dos Monofones

O número de combinações de fones numa língua é tremendo e, por isso, é normal que nas frases de treino não estejam todas essas combinações. Por esta razão é que se treinam os modelos a partir dos monofones para se obter a primeira aproximação dos modelos e depois treinam-se os modelos de fones com contexto a partir destes para a geração de modelos mais precisos.

É necessário começar pela análise dos fones sem contexto (monofones) para se obter os modelos HMM iniciais.

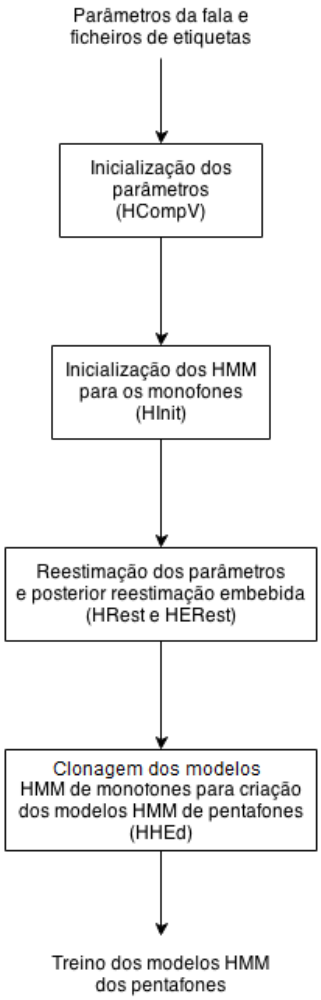



*Figura 3.9 - Diagrama da fase de treino de monofones do HTS. Editado de [13]*

A figura anterior descreve a sequência do script de treino durante a fase de treino dos monofones.

Primeiramente, definem-se protótipos de HMM que servem de base para a criação dos HMM a serem gerados durante o treino.

Depois calcula-se a variância mínima dos parâmetros do áudio a partir dos protótipos usando a ferramenta do HTS *HCompV*. Trata-se de um *flat start*, ou seja, inicia os estados de HMM todos iguais e define modelos iniciais de HMM baseados nestes protótipos.

De seguida, cria HMM para os todos os monofones e inicializa os parâmetros de cada um dos HMM correspondente a cada fonema usando o *HInit*.

O passo a seguir é executar a ferramenta *HRest* que faz a reestimação dos parâmetros dos HMM dos monofones usando o algoritmo de otimização de Baum-Welch, ou seja, faz uma reestimação independente do contexto do fonema. O algoritmo de Baum-Welch tem como objetivo calcular máxima verosimilhança, através do método da máxima expectativa (*expectation-*

*maximization*), da sequência de estados da cadeia de Markov dada uma observação da cadeia de Markov e um conjunto de parâmetros de um HMM. Após este passo é feita ainda a reestimação embebida dos parâmetros que melhora os resultados do passo anterior, usando a ferramenta *HERest*.

### 3.7 - Terceira etapa do treino: treino a partir dos Pentafones

Tendo os modelos dos monofones é altura de estimar os modelos HMM dos fones com a adição de contexto.

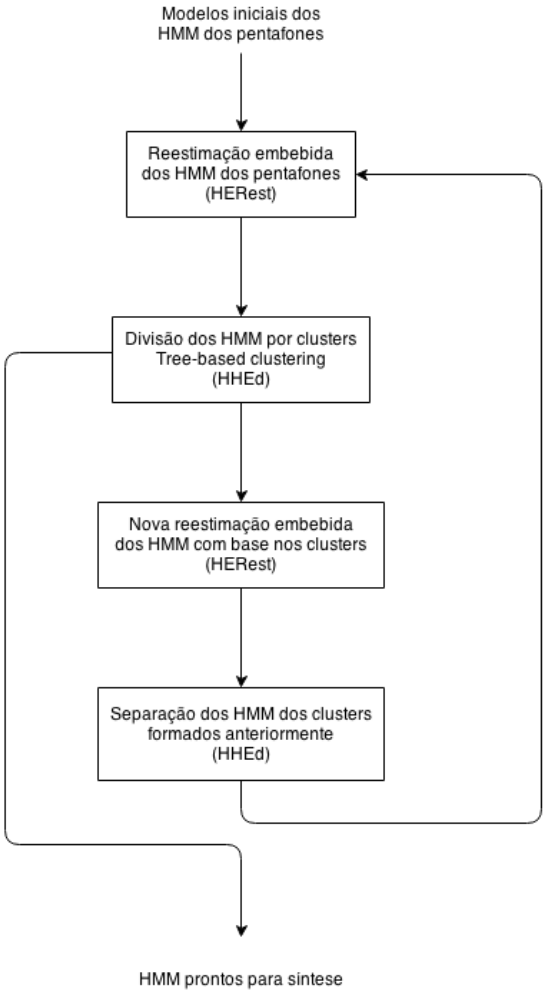



*Figura 3.10 - Diagrama da fase de treino de pentafones do HTS. Editado de [13]*

A figura acima descreve a sequência do script de treino dos pentafones.

Usando a informação dos monofones, faz-se a clonagem dos HMM dos monofones para os modelos iniciais dos pentafones.

Terminada a inicialização dos pentafones, o sistema de treino faz nova estimação dos parâmetros dos HMM para os pentafones usando a ferramenta *HERest*.

Nesta altura usa-se a ferramenta *HHEd* para construir a árvore de decisão para agrupar os modelos em *clusters* (*Tree-based clustering*), ou seja, agrupar os modelos em grupos de contexto semelhantes.

Construída a árvore de decisão, os modelos são agrupados de acordo com o seu contexto. O sistema cria *clusters* (ou grupos) em que junta fones que se assemelhem e cuja diferença de parâmetros entre eles seja muito reduzida. É usado um processo de *tree-based clustering*, em que a partir de perguntas presentes no ficheiro de questões se dividem os modelos até se reduzirem a *clusters* cujo conteúdo não se pode dividir mais. Para isso são usadas árvores binárias em que cada ramo é o resultado da resposta afirmativa ou negativa à pergunta que mais modelos consegue dividir. Tendo finalizada esta divisão de modelos fazem-se cinco iterações de reestimação de parâmetros de HMM de cada *cluster*.

Para melhorar o processo de treino e divisão de modelos por grupos, é feita a separação dos modelos e os parâmetros são reestimados (*HERest*) individualmente usando as definições de HMM de cada *cluster*. Após isto, é criada nova árvore de decisão e os modelos são distribuídos, novamente, por novos *clusters*.

Finalmente, fazem-se cinco iterações do processo de reestimação dos parâmetros de HMM destes novos grupos de modelos.

Neste passo é ainda estimado o modelo de duração para cada pentafone. Como a duração de um estado HMM nunca é fixa, pois a sequência de estados nunca é conhecida, o HTS implementa um novo HMM para a representação da duração com 3 estados em que apenas o intermédio é emissor e contém uma PDF Gaussiana tridimensional calculada pela observação dos histogramas de duração de todos os estados usando o algoritmo de Viterbi.

```
{*}[2].stream[2,3,4]
{
   0 L_Unvoiced_consonant                          -1        -8
  -1 C_Voiced                                      -2        -3
  -2 Pos_cur_Monophone_in_cur_Syllable(Bw)==1      -13        -4
  -3 Pos_cur_Syllable_in_cur_Word(Fw)==1            -5        -6
  -4 R_Voiced                                      -14       -12
  -5 C_Vibrant_liquid                              -10       -42
  -6 Punct_previous_Word_is_space                  -19        -7
  -7 C_i                                           -16       -24
  -8 C_@                                           -9          "lf0_s2_1"
  -9 C_Voiced                                      -82       -15
 -10 C_@                                           -11       -86
 -11 Pos_cur_Syllable_in_cur_Phrase(Bw)<=2         -18       -31
 -12 C_Unvoiced_fricative                          -29       -37
 -13 C_u_or_ú_or_ü_or_Ü                            -26       -43
 -14 C_u_or_ú_or_ü_or_Ü                            -22       -78
 -15 Pos_cur_Monophone_in_cur_Syllable(Fw)==1      -17       -34
```

*Figura 3.11 - Ficheiro de definição da árvore de decisão da figura 3.1*

Na figura anterior verifica-se que para o primeiro modelo do estado 2 da cadeia de Markov a primeira pergunta é se o fonema à esquerda do atual (L) é uma consoante vozeada ou não. Em caso negativo segue-se na árvore para o nó 1 que pergunta se o fonema central (C) é vozeado ou não e em caso afirmativo segue para o nó 8 em que se pergunta se o fone central é o "@". Para cada caso a árvore é corrida até chegar a uma folha que indique o modelo apropriado.

Para isso é usado o ficheiro de questões editado de acordo com os parâmetros de contexto definidos no decorrer desta dissertação (Anexo A) e o algoritmo calcula quais as perguntas que separam o maior número de modelos e a partir desse critério constrói a árvore de decisão.

### 3.8 - Etapa final do treino: cálculo da variância global (GV)

De seguida é calculada a variância global dos parâmetros (*global variance*). O uso da variância global dos parâmetros serve para aumentar a qualidade da voz e a expressividade do áudio gerado ao aumentar a variância do espectro e da geração do tom. A utilização dos coeficientes dinâmicos suaviza as transições entre modelos, como por exemplo as transições entre as formantes constituintes do tom que são por definição valores de pico em transições abruptas da representação do tom em frequência. Isto pode levar a suavização em demasia e a fala produzida pode perder a sua naturalidade, a sua semelhança à voz humana. O uso da variância global resume-se à execução de um algoritmo que maximiza a função objetivo

$$\mathcal{F}_{GV}(\boldsymbol{c}) = \boldsymbol{w} \log P(\boldsymbol{Wc} \mid \boldsymbol{q}, \lambda) + \log \mathrm{P}(v(\boldsymbol{c})|\lambda_V) \tag{3.1}$$

onde **c** representa vetor de coeficientes para os quais se calcula a variância global, sejam de tom, espectrais ou de duração, λ representa o estado do modelo da cadeia de Markov calculados anteriormente, **q** é a sequência de estados determinada pela distribuição de duração de cada estado,

**W** é a matriz dos pesos dos coeficientes vizinhos para o cálculo dos coeficientes dinâmicos (delta e delta-delta) dos coeficientes centrais, **w** é o peso da probabilidade para a saída de cada estado, v(**c**) é a variância global do vetor de coeficientes e $\lambda_V$ são os parâmetros da distribuição da variância global [13].

Podemos dividir a equação em duas partes: a probabilidade logarítmica de um estado HMM e a probabilidade da variância global de um vetor de parâmetros sabendo as propriedades da distribuição da variância global. A segunda parte serve como penalização do uso dos coeficientes dinâmicos e impede que as transições sejam suavizadas em demasia, cumprindo o objetivo do algoritmo [13].

### 3.9 - Resultado do treino

Tendo executado estes passos, acaba-se o processo de treino da voz e teremos como resultado uma "voz" sob a forma estatística usada na síntese.

Esta voz contém os HMM para os fonemas analisados nas locuções. Ao executar a síntese de um texto arbitrário o sistema vai procurar nos *clusters* criados no treino por um modelo que seja aproximado ao modelo pretendido de um fone no contexto em que se insere. Por isso, além dos modelos são criadas, também, as árvores de decisão que guiam o sistema de síntese na escolha do modelo certo para um dado fone num determinado contexto. Temos, tal como no caso dos modelos, três árvores de decisão: para o tom, para a duração e para os coeficientes espectrais.

É necessário, ainda, a criação de ficheiros que definam os pesos dos coeficientes de variação dos MGC, do tom e dos coeficientes do filtro de síntese, ou seja, os coeficientes dinâmicos. Os coeficientes do filtro têm sempre peso unitário, tal como os coeficientes dos MGC e do tom. No entanto, os coeficientes de variação dos MGC e do tom não têm peso unitário: os coeficientes de variação de primeira ordem ($\Delta$ ou "delta") têm peso de 0.5 para o coeficiente anterior e -0.5 para o seguinte do central e unitário para o central, e os de segunda ordem ($\Delta^2$ ou "delta-delta") têm peso unitário para os coeficientes vizinhos do central e peso -2 para o coeficiente central.

Estes coeficientes dinâmicos são usados para melhorar a qualidade da fala sintetizada através da amenização das transições entre modelos, impedindo transições bruscas de tom que tornem a fala pouco natural.

Finalmente, são gerados modelos e árvores para a variância global (*global variance*) do tom, do espectro e da duração.

Com estes parâmetros definidos constitui-se, então, uma “voz” que vai ser usada para a síntese de fala.

# Capítulo 4 - Refinamentos do sistema

## 4.1 - Alterações nas definições do contexto dos fones (pentafones)

Para o processo de síntese é necessário um ficheiro que descreva a fala a ser produzida em termos fonéticos para a criação das árvores de decisão.

O sistema de HTS usa ficheiros de etiquetagem dos pentafones (*fullcontext*), ou seja, um ficheiro que descreve um conjunto de cinco fones consecutivos segundo vários parâmetros. Estes ficheiros contêm o contexto de cada fone, ou seja, informam o sistema dos dois fonemas anteriores e seguintes ao fone a ser processado atualmente.

Além deste aspeto principal, o sistema de HTS dá-nos alguma liberdade de escolher os parâmetros que queremos usar para descrever o contexto do fone em termos de sílaba, palavra, frase e texto no global, bem como introduzir novos critérios para o uso nas árvores de decisão e poder investigar novas formas de melhorar a prosódia da fala produzida.

Perante este facto foram feitas alterações às definições dos ficheiros de etiqueta na tentativa de melhorar a prosódia da fala sintetizada. As definições dos parâmetros têm por base as definições do português brasileiro com algumas alterações implementadas no decorrer desta dissertação.

As principais alterações prenderam-se com o registo da pontuação na frase a ser sintetizada. Estas alterações e os seus efeitos serão discutidos na secção seguinte com mais detalhe.

O tipo de palavra associado ao fonema que o sintetizador analisa indica outro tipo de contexto. Os pronomes, determinantes e artigos podem indicar uma variação de coarticulação, pois interligam outros tipos de palavras na frase. Por isso este tipo de palavras foi classificado como *LINK* (*LINKER*) e o resto das palavras como conteúdo geral (*CONT* ou *CONTENT*). Como interessa saber se a palavra anterior e seguinte da palavra a ser processada são *LINK* regista-se, na criação das etiquetas de pentafones, se a palavra anterior e seguinte são deste tipo ou de conteúdo geral.

Nos parâmetros temos três níveis de definição: ao nível dos fones, da sílaba, da palavra, da frase e do texto completo.

Ao nível dos fones (secção **M** do ficheiro de definição das etiquetas de pentafones) temos os fones adjacentes ao central (dois para a direita e dois para a esquerda), além deste, e a posição do fonema dentro da sílaba (posição ascendente e descendente). Estes parâmetros são a fonte primária de informação sobre o contexto do fone a ser analisado.

Ao nível da sílaba (secção **S**) existem três indicadores de tónica que informam se a sílaba atual, anterior e seguinte são silabas tónicas, o número de fonemas nessas sílabas, as posições das sílabas nas palavras e nas frases (ascendente e descendente), o número de sílabas tónicas antes e depois da sílaba atual, o número de sílabas desde a sílaba tónica anterior e seguinte até à sílaba anterior e a vogal (fonema) da sílaba atual. Estes parâmetros atribuem um contexto a nível de sílaba que permitem prever o aparecimento de uma sílaba tónica e a consequente acentuação dessa sílaba ou o declínio da acentuação da frase proveniente da sílaba anterior ser tónica, por exemplo.

Ao nível da palavra (secção **W**) define-se o tipo de palavra (POS) da palavra atual, anterior e seguinte, o número de sílabas nestas três palavras, a posição da palavra atual dentro da frase (ascendente e descendente), o número de palavras com conteúdo importante (*LINK*) antes e depois da palavra atual, o número de palavras entre a palavra atual e a palavra anterior ou seguinte classificada como *LINK* e a pontuação das palavra atual e adjacentes. Estes parâmetros são importantes pois referem-se à transição entre palavras e aproximam-se da definição da entoação consoante o fim de frase. Como já foi explicado, a indicação da pontuação dá informação sobre possíveis pausas entre palavras ou diferença de entoação entre palavras.

Ao nível da frase (secção **P**) registam-se o número sílabas e palavras da frase atual, anterior e seguinte, a posição da frase no texto completo (ascendente e descendente) e a pontuação no fim da frase. De acordo com as locuções treinadas, não foi usado o contexto entre frases, pois as frases treinadas encontram-se separadas e cada locução apenas tem uma frase, não havendo informação de áudio para incluir este parâmetro de posição da frase no texto completo. Os parâmetros ao nível da frase devem dar informação sobre a longevidade do áudio e da entoação consoante a posição na frase do fonema a ser analisado.

Ao nível do texto (secção **U**) temos a informação do número sílabas do texto, do número de palavras e do número de frases. Mais uma vez, é de notar que o número de frases não foi usado devido ao contexto das locuções gravadas. Estes parâmetros acrescentam alguma informação aos parâmetros da frase quando usados em locuções individuais com várias frases.

A tabela que estabelece os parâmetros usados na construção destes ficheiros de etiquetas de pentafones encontra-se no **Anexo A**.

O algoritmo de criação de pentafones foi integralmente implementado em C++ no âmbito deste trabalho.

### 4.2 - Criação da árvore de decisão de prosódia

Como já foi dito, as árvores de decisão vão ser construídas com base nas possíveis perguntas que podem dividir os modelos de HMM por grupos cujos parâmetros dos modelos constituintes sejam aproximadamente iguais.

O sistema do HTS dá alguma liberdade na escolha das perguntas e, por isso, podemos adaptar o sistema à nossa definição do contexto de um fone (pentafones). Por forma a melhorar a prosódia gerada na síntese foram feitas algumas alterações para as árvores de decisão usadas nos HMM do tom levarem em conta a pontuação da frase.

É sabido que uma vírgula introduz uma ligeira pausa entre palavras e que a seguir aos dois pontos deve começar uma enumeração e a variação da prosódia tem como causa estes fatores, por exemplo. É claro, também, que a pontuação de fim de frase é importante para uma boa qualidade prosódica: a entoação no fim de uma frase interrogativa é diferente da entoação de uma frase declarativa e exclamativa. Estas variações de entoação definem a prosódia de uma frase.

A prosódia [15] é o estudo da variação do ritmo, entoação e ênfase de uma sílaba ou de um fone. Esta área da fonética traduz os fatores acústicos do discurso que a transcrição ortográfica não consegue definir. A melhor definição da prosódia na síntese de fala melhora a qualidade do sistema.

Foi definido, por isso, o registo da pontuação de final de frase, substituindo a *flag* de interrogação, a pontuação anterior e posterior à palavra em análise, pois a presença de, por exemplo, parêntesis antes da palavra indica uma ligeira pausa no discurso, à semelhança de uma vírgula, e uma vírgula após a palavra indica, também, uma pausa na fala. Além destes sinais de pontuação, todos os outros foram considerados e cada um pode trazer novo contexto ao fonema em análise e uma variação na prosódia.

Decidiu-se registar a pontuação após a palavra anterior à atual, que não é, necessariamente, a mesma pontuação que está antes da palavra atual, e a pontuação após as duas palavras seguintes à palavra atual em análise por forma a prever o fim de frase e a normal variação do tom característica do fim de frase. No caso de uma frase declarativa o tom normalmente desce em contraste com as frases exclamativas e interrogativas que vêm o seu tom subir.

Foram definidas perguntas (**Anexo B**) que permitem ao sistema melhorar a qualidade da prosódia ao identificar nos modelos HMM a pontuação das palavras.

# Capítulo 5 - O sistema de síntese de fala

Foram criadas três versões da aplicação desenvolvida: duas para execução em consola e uma com ambiente gráfico para demonstração das funcionalidades da aplicação. A diferença entre as duas versões de consola é que uma é destinada à integração numa página web e o modo de retorno dos dados será diferente. Ambas têm a mesma forma de serem executadas e têm versões compatíveis com ambiente Linux e Windows. A aplicação gráfica é que tem uma interface mais *user-friendly* e que se apresenta da seguinte forma:


*Figura 5.1 - Janela inicial da aplicação*

Na janela inicial o utilizador apenas tem uma breve mensagem de apresentação e a opção e avançar para o passo seguinte para síntese da fala.

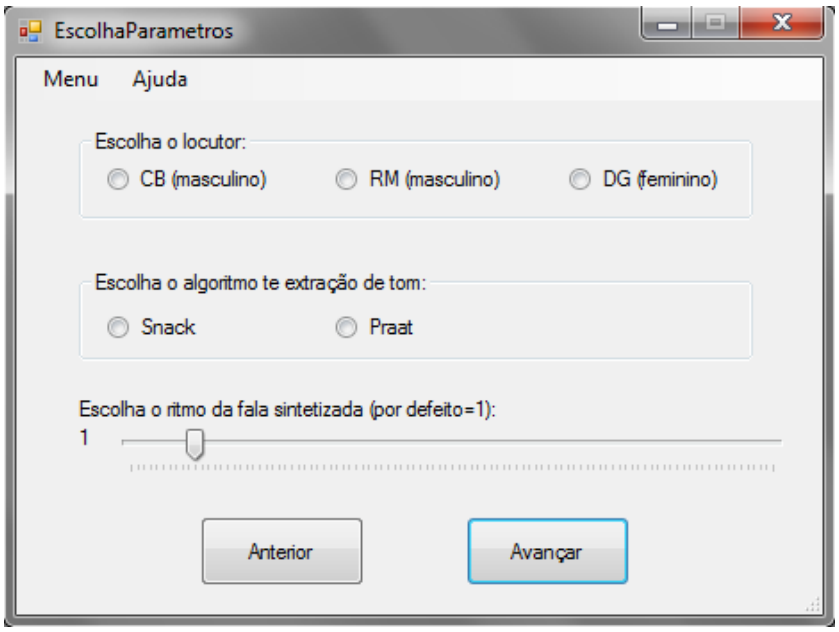



*Figura 5.2 - Janela para escolha dos parâmetros de áudio*

Nesta janela o utilizador pode escolher um dos três locutores presentes nas bases de dados de treino (CB, RM e DG) e o algoritmo de extração de tom usado no treino (*Snack* ou *Praat*).

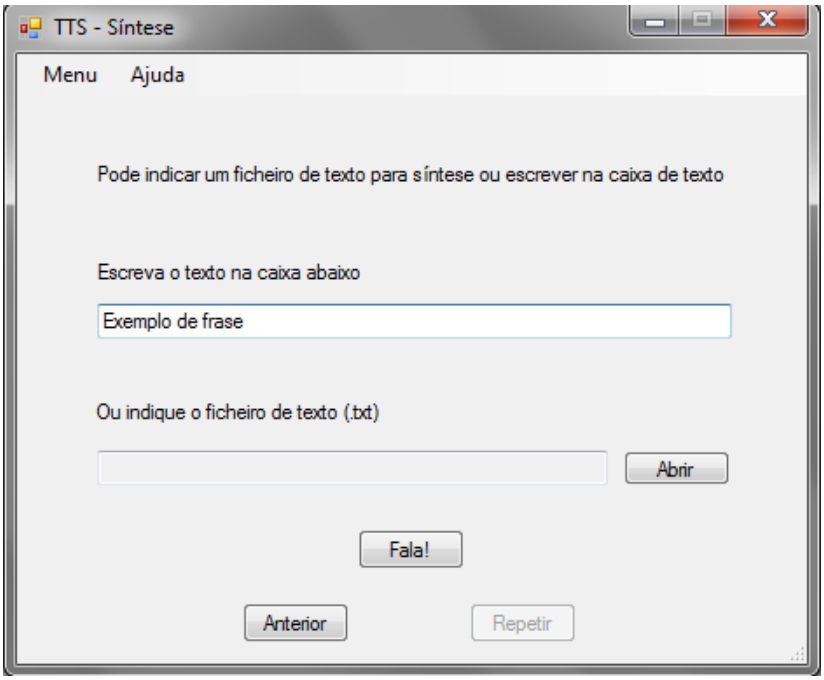



*Figura 5.3 - Janela final da síntese de fala*

Na janela final da síntese o utilizador deve indicar um texto para síntese na primeira caixa de texto ou escolher um ficheiro de texto (.txt) presente no computador como texto de entrada do sintetizador.

De seguida deve prosseguir carregando no botão “Fala!” e após isto é reproduzido o produto final, ou seja, a fala sintetizada. É ainda possível repetir a reprodução do último resultado da síntese, sendo que esta opção só aparece após a primeira síntese de fala, obviamente.

Esta aplicação contém ainda uma ferramenta para criação dos ficheiros de etiquetas de pentafones para uso no treino da base de dados. Para o seu uso remete-se o utilizador para a opção “Criar Pentafones” no menu “Ferramentas” na janela principal.

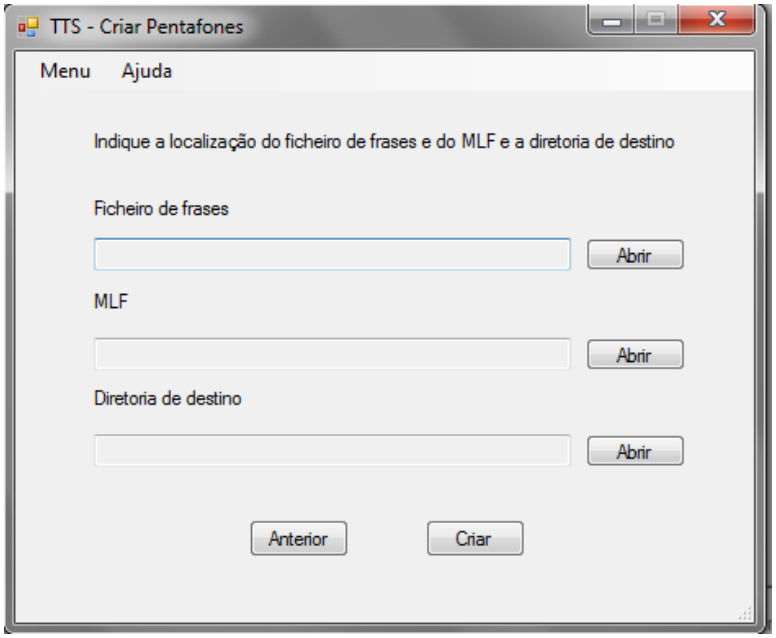



*Figura 5.4 - Janela de criação de ficheiros de etiqueta de pentafones*

Nesta janela o utilizador seleciona um ficheiro de texto (.txt) onde estejam as frases que vão ser utilizadas no processo de treino da base de dados do sistema, o ficheiro MLF dos monofones e uma diretoria para a criação dos ficheiros de etiquetas de pentafones individuais de cada frase.

A aplicação está presente também no website do Instituto de Telecomunicações como o sintetizador na página do projeto “Grafone” [11].


*Figura 5.5 - Website com o sintetizador embutido*

Para o website decidiu-se apenas disponibilizar como base para síntese a base de dados treinada do locutor CB por apresentar os melhores resultados e numa página web deste género ser usual simplificar a interação com o utilizador e sintetizar diretamente.

Ao implementar a aplicação no website foi preciso adaptar alguns métodos desta para funcionar em pleno com o *javascript* que comanda os processos internos da página.

O processo de síntese feito pelo website consiste em invocar a aplicação criada e enviar como parâmetro a frase introduzida pelo utilizador. Após a síntese a aplicação apenas envia a informação de áudio para o website que reproduz a fala sintetizada com recurso à funcionalidade de reprodução de áudio do HTML5.

# Capítulo 6 - Análise de Resultados

Para a avaliação dos resultados deve-se comparar o produto final, ou seja, a fala sintetizada, com as frases originais por forma a fazer uma comparação subjetiva dos resultados.

Segue-se uma análise dos efeitos das mudanças efetuadas nesta dissertação na definição do contexto de um fone.

Para verificar a influência dos parâmetros introduzidos no âmbito desta dissertação, cingimo-nos apenas ao locutor CB que apresenta melhores resultados a nível de síntese no global. Para as análises seguintes são usadas frases que não estão no conjunto de frases de treino.

## 6.1 - Part-of-Speech (POS)

Para o teste da importância da inclusão do *POS* pode-se usar qualquer frase, pois este parâmetro está presente em qualquer texto. Escolheu-se a frase de teste "*Um dia estava sol e no outro estava nublado."*

Após análise da síntese da frase usando uma base de dados de treino que levou em conta todos os parâmetros e outra à qual foram retiradas todas as perguntas para a construção da árvore de decisão relativas ao *POS* verificou-se que este parâmetro não teve grande influência no resultado pelo que as frases sintetizadas nos dois casos são iguais.

## 6.2 - Pontuação

Para o teste da influência da pontuação como parâmetro dos pentafones é necessário usar uma frase que contenha pontuações em grande número. Por isso escolheu-se a frase de teste "*um, dois, três, quatro. É uma frase de teste. Teste da pontuação. Fim*".

Seguindo a mesma lógica anterior usou-se uma base de dados de treino normal, com todos os parâmetros, e outra sem perguntas para construção da árvore de decisão relativas à pontuação.

Nesta situação notou-se uma diferença significativa nas pausas e articulação entre as palavras divididas por pontuação. O ponto final inclui pausas mais percetíveis que diferenciam os dois resultados.

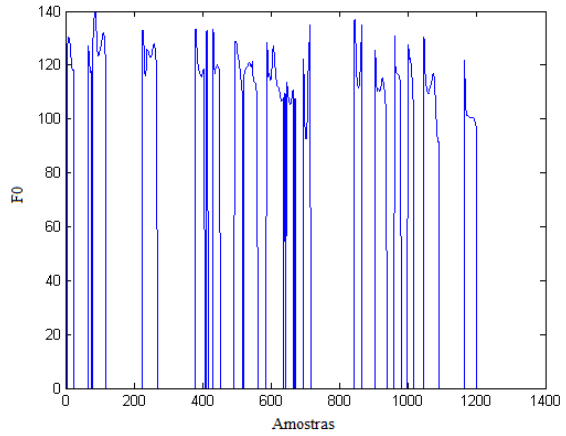



*Figura 6.1 - Tom extraído da frase sintetizada com indicação da pontuação*

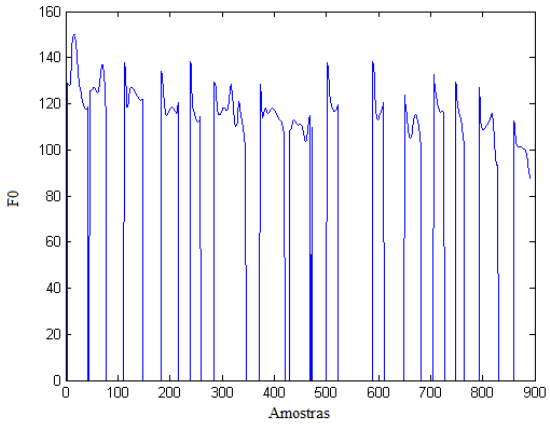



*Figura 6.2 - Tom extraído da frase sintetizada sem indicação da pontuação*

Pelas imagens pode-se verificar que a indicação da pontuação introduz as pausas que as vírgulas e os pontos finais causam e que são a razão de serem usados na escrita. Isto mostra que este parâmetro tem influência na definição da prosódia.

## 6.3 - Tónica

Tal como no caso do *POS*, pode-se usar qualquer frase de teste, pois a tónica está presente em qualquer texto. Escolheu-se a frase de teste "*Sabendo o que sei agora, tinha seguido o mesmo caminho.*"

Este parâmetro é o mais fundamental dos que são testados pois a comparação da síntese da frase com uma base de dados de treino que inclua na árvore de decisão perguntas sobre a tónica e outra que não tenha leva à conclusão da importância deste parâmetro. Não havendo informação de tónica nas sílabas, algumas sílabas acentuadas não são associadas ao modelo mais correto para a sua representação estatística e notou-se na palavra "agora" que o "o" não aparece acentuado ouvindo-se "agára" e a palavra perde o seu sentido. Isto pode acontecer pois não tendo informação de tonicidade o contexto do "o" em posição tónica fica igual ao contexto do "a" em posição tónica.

# Capítulo 7 - Conclusão e melhorias

O objetivo desta dissertação era implementar um sistema de síntese de voz para uso num website e esse objetivo foi cumprido. A aplicação garante uma interpretação satisfatória do texto introduzido em língua portuguesa e produz fala com qualidade e inteligível.

As principais melhorias centram-se no aumento da base de dados de locuções, na melhoria dos algoritmos de extração de tom na fase de treino e da diminuição do tempo do treino da voz.

Se o aumento da base de dados de locuções e a melhoria dos algoritmos de extração de tom focam-se na qualidade do produto final do sistema, ou seja, a fala sintetizada, a diminuição do tempo de treino de voz é uma melhoria que serve mais os responsáveis pelo sistema de síntese pois nesta situação o treino de áudio a 48kHz demora muito tempo.

Pode-se ainda incluir mais ferramentas adicionais na aplicação que façam parte do processo de treino e etiquetagem dos fonemas. A inclusão de uma ferramenta de alinhamento do áudio e do texto é um exemplo disso. A inclusão do próprio sistema de treino do HTS pode ser um objetivo de trabalho futuro.

A fala sintetizada apresenta boa qualidade usando como base o locutor CB, ainda que não tenham sido feitos testes auditivos para se ter uma análise objetiva do sistema. Neste caso a fala produzida é sempre inteligível e como a extração do tom é feita de forma mais eficaz pela consistência do tom do locutor ao longo da locução.

Concluindo, o trabalho produziu resultados interessantes, o objetivo principal foi conseguido e ainda foram criadas variantes da aplicação e uma ferramenta crucial na preparação do treino da base de dados do sistema de síntese. O sistema criado está preparado para uma grande variedade de combinações de texto de forma muito satisfatória.

## Bibliografia


[1] P. Taylor, “Text-to-Speech Synthesis”, Cambridge, Cambridge University Press, 2009

[2] João Carlos Cunha Gomes, "Treino de Modelos para um Sistema de Síntese de Fala em Português", Faculdade de Ciências e Tecnologia - Universidade de Coimbra, Coimbra, Dissertação de Mestrado, 2011.

[3] Nagoya Institute of Technology. (2014, Março) HTS [Online] http://hts.sp.nitech.ac.jp/

[4] Cambridge University Engineering Department. (2014, Março) Hidden Markov Toolkit (HTK) [Online] http://htk.eng.cam.ac.uk/

[5] Nagoya Institute of Technology. (2014, Março) Speech Signal Processing Toolkit (SPTK) http://sp-tk.sourceforge.net/

[6] (2014, Março) SoX - Sound eXchange. [Online] http://sox.sourceforge.net/

[7] (2014, Março) Praat. [Online] http://www.fon.hum.uva.nl/praat/

[8] Nagoya Institute of Technology. (2011, Janeiro) hts_engine API [Online]. http://hts-engine.sourceforge.net/

[9] (2014, Julho) SAMPA [Online]. http://pt.wikipedia.org/wiki/SAMPA

[10] (2014, Julho) IPA [Online]. http://en.wikipedia.org/wiki/International_Phonetic_Alphabet

[11] (2014, Julho) Grafone [Online]. http://www.co.it.pt/~labfala/g2p/

[12] HTS Working Group, HTS Slides, 2011, http://hts.sp.nitech.ac.jp/

[13] J. Yamagishi, “An Introduction to HMM-Based Speech Synthesis”, 2006.

[14] K. Tokuda, T. Mausko, N. Miyazaki, T. Kobayashi, “Multi-space probability distribution HMM”, IEICE Trans. Inf. & Syst., vol.E85-D, no.3, pp.455-464, 2002

[15] (2014, Março) Prosódia [Online]. http://pt.wikipedia.org/wiki/Pros%C3%B3dia

[16] Luís Miguel Belo da Silva, "Algoritmos de Determinação de Tom da Fala," Faculdade de Ciências e Tecnologia - Universidade de Coimbra, Coimbra, Dissertação de Mestrado 2010.

[17] Alexandre Maciel, Arlindo Veiga, Carla Lopes Cláudio Neves Fernando Perdigão José David Lopes, Luís de Sá, “A ROBUST SPEECH COMMAND RECOGNIZER FOR EMBEDDED APPLICATIONS”, SIGMAP 2008, 2008.

[18] (2014, Março) Transcriber http://trans.sourceforge.net/en/presentation.php

[19] (2014, Março) Hidden Markov Models [Online].
http://en.wikipedia.org/wiki/Hidden_Markov_model

[20] (2014, Março) Algoritmo de Viterbi [Online].
http://en.wikipedia.org/wiki/Viterbi_algorithm

# Anexo A

## An example of contextdependent label format for HMMbased speech synthesis in European Portuguese

Tiago Ferreira
June 16, 2014

$m_1$^$m_2$-$m_3$+$m_4$=$m_5$ /M2:$m_6$_$m_7$
/S1:$s_1$_@$s_2$$s_3$_@$s_4$+$s_5$_@$s_6$ /S2:$s_7$_$s_8$ /S3:$s_9$_$s_{10}$ /S4:$s_{11}$_$s_{12}$ /S5:$s_{13}$_$s_{14}$ /S6:$s_{15}$
/W1:$w_1$_#$w_2$-$w_3$_#$w_4$+$w_5$_#$w_6$ /W2:$w_7$_$w_8$ /W3:$w_9$_$w_{10}$ /W4:$w_{11}$_$w_{12}$ /W5:$w_{13}$
/W6:$w_{14}$_$w_{15}$ /W7:$w_{16}$_$w_{17}$
/P1:$p_1$_!$p_2$$p_3$_!$p_4$+$p_5$_!$p_6$ /P2:$p_7$_$p_8$ /P3:$p_9$
/U:$u_1$_$$u_2$_&$u_3$

| | |
|---|---|
| $m_1$ | the phoneme identity before the previous phoneme |
| $m_2$ | the previous phoneme identity |
| $m_3$ | the current phoneme identity |
| $m_4$ | the next phoneme identity |
| $m_5$ | the phoneme after the next phoneme identity |
| $m_6$ | position of the current phoneme identity in the current syllable (forward) |
| $m_7$ | position of the current phoneme identity in the current syllable (backward) |
| $s_1$ | whether the previous syllable stressed or not (0: not stressed, 1: stressed) |
| $s_2$ | the number of phonemes in the previous syllable |
| $s_3$ | whether the current syllable stressed or not (0: not stressed, 1: stressed) |
| $s_4$ | the number of phonemes in the current syllable |
| $s_5$ | whether the next syllable stressed or not (0: not stressed, 1: stressed) |
| $s_6$ | the number of phonemes in the next syllable |
| $s_7$ | position of the current syllable in the current word (forward) |
| $s_8$ | position of the current syllable in the current word (backward) |
| $s_9$ | position of the current syllable in the current phrase (forward) |
| $s_{10}$ | position of the current syllable in the current phrase (backward) |
| $s_{11}$ | the number of stressed syllables before the current syllable in the current phrase |
| $s_{12}$ | the number of stressed syllables after the current syllable in the current phrase |
| $s_{13}$ | the number of syllables, counting from the previous stressed syllable to the current syllable in this utterance |
| $s_{14}$ | the number of syllables, counting from the current syllable to the next stressed syllable in this utterance |
| $s_{15}$ | name of the vowel of the current syllable |
| $w_1$ | part-of-speech classification of the previous word |
| $w_2$ | the number of syllables in the previous word |
| $w_3$ | part-of-speech of classification of the current word |
| $w_4$ | the number of syllables in the current word |
| $w_5$ | part-of-speech classification of the next word |
| $w_6$ | the number of syllables in the next word |
| $w_7$ | position of the current word in the current phrase (forward) |

| | |
|---|---|
| $w_8$ | position of the current word in the current phrase (backward) |
| $w_9$ | the number of content words before the current word in the current phrase |
| $w_{10}$ | the number of content words after the current word in the current phrase |
| $w_{11}$ | the number of words counting from the previous content word to the current word in this utterance |
| $w_{12}$ | the number of words counting from the current word to the next content word in this utterance |
| $w_{13}$ | punctuation after the previous word |
| $w_{14}$ | punctuation before the current word |
| $w_{15}$ | punctuation after the current word |
| $w_{16}$ | punctuation after the next word |
| $w_{17}$ | punctuation after the second next word |
| $p_1$ | the number of syllables in the previous phrase |
| $p_2$ | the number of words in the previous phrase |
| $p_3$ | the number of syllables in the current phrase |
| $p_4$ | the number of words in the current phrase |
| $p_5$ | the number of syllables in the next phrase |
| $p_6$ | the number of words in the next phrase |
| $p_7$ | position of the current phrase in this utterance (forward) |
| $p_8$ | position of the current phrase in this utterance (backward) |
| $p_9$ | ending punctuation of the current phrase |
| $u_1$ | the number of syllables in this utterance |
| $u_2$ | the number of words in this utterance |
| $u_3$ | the number of phrases in this utterance |

# Anexo B

## Secção do ficheiro de perguntas relativa à pontuação do texto:

QS "Punct_previous_Word_is_space"{*/W5:space/W6:*}

QS "Punct_previous_Word_is_comma"{*/W5:comma/W6:*}

QS "Punct_previous_Word_is_semicolon"{*/W5:semicolon/W6:*}

QS "Punct_previous_Word_is_colon"{*/W5:colon/W6:*}

QS "Punct_previous_Word_is_quotation_mark"{*/W5:quote/W6:*}

QS "Punct_previous_Word_is_bracket"{*/W5:bracket/W6:*}

QS "Punct_before_curr_Word_is_space"{*/W6:space_*}

QS "Punct_before_curr_Word_is_comma"{*/W6:comma_*}

QS "Punct_before_curr_Word_is_semicolon"{*/W6:semicolon_*}

QS "Punct_before_curr_Word_is_colon"{*/W6:colon_*}

QS "Punct_before_curr_Word_is_quotation_mark"{*/W6:quote_*}

QS "Punct_before_curr_Word_is_bracket"{*/W6:bracket_*}

QS "Punct_after_curr_Word_is_ending_point"{*_point/W7:*}

QS "Punct_after_curr_Word_is_space"{*_space/W7:*}

QS "Punct_after_curr_Word_is_comma"{*_comma/W7:*}

QS "Punct_after_curr_Word_is_semicolon"{*_semicolon/W7:*}

QS "Punct_after_curr_Word_is_colon"{*_colon/W7:*}

QS "Punct_after_curr_Word_is_exclamation_mark"{*_exclamation/W7:*}

QS "Punct_after_curr_Word_is_question_mark"{*_question/W7:*}

QS "Punct_after_curr_Word_is_quotation_mark"{*_quote/W7:*}

QS "Punct_after_curr_Word_is_bracket"{*_bracket/W7:*}

QS "Punct_next_Word_is_ending_point"{*/W7:point_*}

QS "Punct_next_Word_is_space"{*/W7:space_*}

QS "Punct_next_Word_is_comma"{*/W7:comma_*}

QS "Punct_next_Word_is_semicolon"{*/W7:semicolon_*}

QS "Punct_next_Word_is_colon"{*/W7:colon_*}

QS "Punct_next_Word_is_exclamation_mark"{*/W7:exclamation_*}

QS "Punct_next_Word_is_question_mark"{*/W7:question_*}

QS "Punct_next_Word_is_quotation_mark"{*/W7:quote_*}

QS "Punct_next_Word_is_bracket"{*/W7:bracket_*}

QS "Punct_next_next_Word_is_ending_point"{*_point/P1:*}

QS "Punct_next_next_Word_is_space"{*_space/P1:*}

QS "Punct_next_next_Word_is_comma"{*_comma/P1:*}

QS "Punct_next_next_Word_is_semicolon"{*_semicolon/P1:*}

QS "Punct_next_next_Word_is_colon"{*_colon/P1:*}

QS "Punct_next_next_Word_is_exclamation_mark"{*_exclamation/P1:*}

QS "Punct_next_next_Word_is_question_mark"{*_question/P1:*}

QS "Punct_next_next_Word_is_quotation_mark"{*_quote/P1:*}

QS "Punct_next_next_Word_is_bracket"{*_bracket/P1:*}

QS "No_punctuation_previous_word" {*/W5:x*}

QS "No_punctuation_next_word"  {*/W7:x*}

QS "No_punctuation_next_next_word" {*_x/P1:*}