

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO
DEPARTAMENTO DE ENGENHARIA ELECTROTÉCNICA E DE COMPUTADORES

ETIQUETAGEM AUTOMÁTICA DE SINAIS DE FALA
SEGMENTAÇÃO E CLASSIFICAÇÃO FONÉTICA

LUÍS FILIPE MARTINS PINTO COELHO

Licenciado em Engenharia Electrotécnica e de Computadores
pela Faculdade de Engenharia da Universidade do Porto

Dissertação submetida para satisfação parcial dos requisitos do grau de
Mestre em Engenharia Electrotécnica e de Computadores

Dissertação realizada sob a supervisão de
Professor Doutor Diamantino Rui da Silva Freitas,
do Departamento de Engenharia Electrotécnica e de Computadores
da Faculdade de Engenharia da Universidade do Porto

PORTO, FEVEREIRO DE 2005

ETIQUETAGEM AUTOMÁTICA DE SINAIS DE FALA SEGMENTAÇÃO E CLASSIFICAÇÃO FONÉTICA

Dissertação submetida para satisfação parcial dos requisitos do grau de
Mestre em Engenharia Electrotécnica e de Computadores

Para a Daniela

Resumo

Nesta dissertação estuda-se o problema da etiquetagem (anotação) fonética aplicada às bases de dados de fala. A existência de corpora anotados de elevada qualidade é um requisito absolutamente essencial para o desenvolvimento e construção de sistemas de síntese ou reconhecimento de fala. Porém a tarefa de anotação manual de uma base de dados é fastidiosa e muito demorada sendo portanto de grande interesse possuir um sistema que execute esta operação automaticamente. Deste modo é possível criar rapidamente anotações para novas bases de dados de fala que possibilitam a transformação e adaptação de sistemas que utilizam esta informação.

É assim objectivo deste trabalho desenvolver um sistema totalmente automático de anotação de corpora de fala. A avaliação do sistema será feita por comparação com anotações manuais do mesmo conteúdo da base de fala gravada.

Inicialmente investigaram-se outros trabalhos na área e compararam-se técnicas e métodos e respectiva evolução com a intenção de reunir as melhores práticas que conduziram a um desempenho óptimo. Realizou-se também um estudo e uma actualização para a construção do estado da arte.

Uma vez que os corpora de fala ocupam um papel central neste trabalho efectuou-se também uma análise pormenorizada da sua concepção. Como é recolhido o sinal de voz, quais as características que os textos devem possuir, a anotação e respectivas métricas de comparação, foram problemas que se analisaram a fundo. Fez-se ainda uma revisão dos corpora FEUP/IPB e TIMIT utilizados para o desenvolvimento e teste do sistema desenvolvido.

Seguidamente e para fundamentar todos os passos dados na construção da ferramenta de anotação estudaram-se pormenores da análise de sinais de fala e técnicas de etiquetagem. Para a análise de sinais abordam-se várias características que podem ser extraídas e que permitem evidenciar diversos comportamentos e apresentam-se ainda modelos de representação de sinais no tempo. Aqui investigam-se essencialmente as técnicas de predição linear (domínio AR, MA e ARMA) e a representação alternativa por frequências de linhas espectrais (LSF).

Ainda como base para o trabalho apresentam-se as principais técnicas de segmentação utilizadas: *Dynamic Time Warping* (DTW) e *Hidden Markov Models* (HMM).

Depois de fundamentada a teoria iniciou-se a descrição de duas abordagens distintas que foram experimentadas. A primeira técnica é baseada apenas na análise directa do sinal em que, a partir de comparações sucessivas, se efectua a segmentação da fala. São utilizadas redes neurais artificiais (ANN) que intervêm posteriormente no processo para classificar cada um dos segmentos identificados. Na segunda técnica, mais comum e utilizada em outros trabalhos, construiu-se um sistema inteiramente baseado em HMMs. Em ambas as situações apresentam-se os resultados obtidos e sua análise crítica.

A dissertação é terminada com a apresentação das principais conclusões e com a referência às possibilidades de evolução e melhoria que se colocam.

Abstract

In this dissertation the problem of automatic annotation at the phonetic level applied to speech databases is studied. The existence of high quality annotated speech corpora is an absolute and essential requirement for the development and construction of speech synthesis and recognition systems. However, the task of manually annotating a database is tedious and time consuming thus being of great interest the availability of a system that can automatically execute this operation. In this manner it is possible to quickly create new corpora annotations that enable the adaptation of systems that use this kind of information.

The objective of this work is to develop a totally automatic system for annotating speech corpora. The performance evaluation will be made by comparison with other manually generated annotations of the same audio data.

Initially other works in the area were researched and studied. Techniques and methods and their relative evolutions were compared with the purpose of finding the best practices that could enable the achievement of optimal results. The most recent papers and scientific work were analysed and a clear picture of the state of the art in this area was created.

Speech corpora production, due to the main role in this dissertation, has also been highly considered. How the speech signal is acquired, what are the key elements in the definition of the texts, the annotation and comparison measures were all deeply studied. In addition, the main characteristics of the FEUP/IPB and TIMIT speech databases, used in the development and benchmarking of the system, were also presented.

To establish the basis of all the subsequent steps in the construction of the annotation tool, relevant signal processing and annotation techniques were studied. For signal processing a special focus has been made in speech features and what kind of information can be provided by them. Linear prediction models (AR, MA and ARMA) and the alternative line spectral frequencies (LSF) representation were also analysed.

Still on the theoretical concepts the most widely used techniques for this task are presented: *Dynamic Time Warping* (DTW) and *Hidden Markov Models* (HMM).

After explaining the whole theory and introducing some preliminary experimental work the description of two distinct ways that were followed is initiated. A first technique uses direct analysis of the speech signal and generates the time divisions by successive comparisons in time. Artificial neural networks (ANN) are used for classifying the segments. The second technique, a more common one, follows an HMM based annotation approach. In both cases practical results and their related analysis were presented.

The dissertation ends with the presentation of the main conclusions and some comments to the evolution and improvement possibilities that emerged during the work.

Resumée

Dans cette dissertation on étudie le problème de l'annotation phonétique appliquée aux bases de données de parole. L'existence de corpora de parole annotés de grande qualité est une condition absolument essentielle pour le développement et construction de systèmes de synthèse ou de reconnaissance de la voix. Cependant, la tâche d'annotation manuelle d'une base de données est fastidieuse et très lente, et c'est pourquoi qu'il est très nécessaire d'avoir un système qui fasse cette opération automatiquement. Ainsi c'est possible de créer, d'une façon rapide, des annotations pour nouvelles bases de données de parole qui permettent la transformation et adaptation de systèmes qui utilisent cette information.

L'objectif de ce travail est alors de développer un système d'annotation de corpora de parole tout à fait automatique. L'évaluation du système sera faite par comparaison avec des annotations manuelles du même matériel d'audio.

Initialement, on a recherché d'autres travaux concernant le même sujet, et on a comparé des techniques et des méthodes et sa respective évolution afin de réunir les meilleures pratiques qui conduisent à un dégagement optimale. On a fait aussi une étude de l'actualité pour la construction de l'état de l'art.

Une fois que les corpora de parole jouent un rôle central dans ce travail, on a fait aussi une analyse détaillée de sa conception. L'enregistrement du signal de parole, les propriétés que les textes doivent avoir, l'annotation et respectives métriques de comparaison ont été quelques problèmes qui ont été analysés en détail. On a fait une révision des corpora FEUP/IPB et TIMIT utilisés pour le développement et pour le test du système qu'on présente.

Ensuite, pour appuyer tous les procédés qui seront faits dans la construction de l'outil d'annotation, on a étudié des détails de l'analyse des signaux de parole et techniques d'annotation. Pour l'analyser des signaux on fait référence à plusieurs caractéristiques qui peuvent être obtenues et qui permettent de souligner divers comportements. On présente aussi des modèles de représentation des signaux dans le temps, où on a fait la recherche des techniques de prédiction linéaire (AR, MA et ARMA) et sa représentation alternative par fréquences de lignes spectrales (LSF).

Comme base pour ce travail on a présenté les principales techniques de segmentation utilisées: *Dynamic Time Warping* (DTW) et *Hidden Markov Models* (HMM).

Après avoir fondamenté la théorie on a fait la description des deux approches distinctes qui ont été expérimentées. La première technique est appuyée seulement dans l'analyse directe du signal dans lequel, après des comparaisons successives, on a effectué la segmentation de la parole. Les réseaux neuronaux artificiels (ANN) interviennent après dans le procédé pour classer chacun des segments identifiés. Avec la deuxième technique, plus fréquente et utilisée dans pluspart des autres travaux, on a construit un système complètement basé en HMMs. Dans les deux situations, on présente des résultats et son analyse critique.

La dissertation termine avec la présentation des principales conclusions et avec la référence aux possibilités d'évolution et d'amélioration pertinentes.

Agradecimentos

Agradeço aos meus pais, à minha irmã e aos meus avós por todo o incentivo e pelo bom ambiente que sempre me proporcionaram.

À minha esposa por todo o apoio e companhia em mais uma viagem académica.

Ao professor Diamantino pelos seus comentários sempre enriquecedores e pela paciência que teve comigo.

Aos colegas do Laboratório de Sinais e Sistemas com quem aprendi muito e espero continuar a aprender. Enalteço o grande companheirismo e o espírito aberto de partilha de conhecimento que permite que a experiência de cada um exista como um todo. Abraço especial para o António Moura, Filipe Moreira, Fernando Lopes e Maria João.

Aos colegas Kamil Ekstein e Tomas Pavelka, da República Checa, que contribuíram com boas sugestões e mostraram-se sempre disponíveis para ajudar.

E finalmente deixo um agradecimento aos meus colegas Marco Alves, Nuno Lago, Osvaldo Azevedo e Pedro Oliveira pela grande amizade e pelos momentos que desde há muito temos partilhado.

Conteúdo

1	Introdução	1
1.1	Objectivos	1
1.2	Motivações	1
1.3	Estudo das Línguas	2
1.4	Fonética e Fonologia	4
1.5	Articulação	5
1.5.1	Anatomia e Fisiologia do Sistema de Produção de Fala . .	6
1.5.2	Sons Consonânticos e sua Classificação	8
1.5.3	Sons Vocálicos e sua Classificação	9
2	Corpora de Fala	13
2.1	Recolha dos Dados	13
2.2	Riqueza Fonética e Acústica	14
2.3	Anotação	15
2.4	Aplicações Informáticas para Anotação	17
2.4.1	Speech Filling System	17
2.4.2	Praat	18
2.4.3	WaveSurfer	18
2.4.4	Transcriber	19
2.5	Corpus FEUP/IPB	19
2.5.1	Texto	20
2.5.2	Segmentação e Classificação	20
2.5.3	Estatísticas Fonéticas	21
2.6	Corpus TIMIT	23
2.6.1	Texto	24
2.6.2	Segmentação e Classificação	24
2.7	Sistemas de Segmentação Automática	25
2.7.1	Indicadores de Desempenho	25
2.7.2	Anotação Manual	29
2.7.3	Utilização de DTW	30
2.7.4	Anotação Automática com HMM	31
2.7.5	Outros Sistemas de Anotação	33
2.7.6	Comparações	35
2.7.7	Fenómenos Linguísticos	35

3	Análise de Sinais de Fala	39
3.1	Conceitos Básicos	39
3.2	Amplitude Média Deslizante	42
3.3	Energia Média Deslizante	42
3.4	Passagens por Zero e Classificação de Segmentos	43
3.5	Modelos de Predição Linear	45
3.5.1	Modelo Auto-Regressivo	46
3.5.2	Modelo de Média Movente	51
3.5.3	Modelo Auto-Regressivo e de Média Movente	54
3.5.4	Considerações	58
3.6	Line Spectral Frequencies	59
3.6.1	Cálculo	59
3.6.2	Desempenho das LSF	62
3.7	MFCC	63
3.8	Distâncias	65
3.8.1	Distância de Minkowski	65
3.8.2	Distância Euclideana	66
3.8.3	Distância de Mahalanobis	66
3.8.4	Distância de Itakura-Saito	66
3.8.5	Distância de Kullback-Leibler	67
3.8.6	Alinhamento com DTW	67
4	Modelo LSF/ANN	71
4.1	Abordagem 1: Alinhamento Fonético	71
4.2	Redes Neurais	72
4.2.1	Neurónio Básico	73
4.2.2	Ligações	73
4.2.3	Treino	74
4.3	Abordagem 2: Segmentação e Classificação	77
5	Anotação com HMMs	81
5.1	HMM - Hidden Markov Models	81
5.1.1	Modelos de Markov	83
5.1.2	Modelos Escondidos de Markov	85
5.1.3	Problema da Avaliação (<i>Forward Algorithm</i>)	87
5.1.4	Problema da Decodificação (Algoritmo de Viterbi)	88
5.1.5	Problema da Aprendizagem (Algoritmo de Baum-Welch)	90
5.1.6	Limitações dos HMMs	91
5.2	Hidden Markov Model Toolkit	92
5.2.1	Motivações	92
5.2.2	Funcionamento	93
5.2.3	Ferramentas de Preparação	94
5.2.4	Ferramentas de Treino	95
5.2.5	Ferramentas de Segmentação e Anotação	96
5.2.6	Ferramentas de Análise	97
5.3	Nova Ferramenta Denominada HTKFace	98
5.4	Corpora Utilizado	103
5.4.1	Corpus FEUP-IPB	103
5.4.2	Corpus TIMIT	104
5.5	Procedimento e Resultados	104

5.5.1	Avaliação	104
5.5.2	Preparação	105
5.5.3	Treino	113
5.5.4	Segmentação e Classificação	116
5.6	Resultados	119
5.6.1	Ensaio gerais	120
5.6.2	Refinamento	121
6	Conclusões e Trabalho Futuro	125
6.1	Trabalho Futuro	128
A	Divisão e Multiplicação de Polinómios por um Binómio	129
B	Determinação dos Parâmetros AR	131
B.1	Método da Autocorrelação	133
B.2	Método da Covariância	134
C	Tabelas IPA e SAMPA para o Português Europeu	135
D	Características da Base de Dados FEUP/IPB	137
E	Símbolos Utilizados	141
F	Acrónimos e Abreviaturas	145

Lista de Figuras

1.1	Evolução do reconhecimento da fala	2
1.2	Aparelho fonador: desenho científico	6
1.3	Aparelho fonador: esquema	7
1.4	Impulsos glotais artificiais	7
1.5	Triângulo acústico das vogais tónicas para o Português Europeu	11
2.1	Ocorrências de vogais na base de dados FEUP-IPB	23
2.2	Situações de erro na comparação de duas sequências de anotação	27
2.3	Resultados de tarefas de alinhamento manual	30
2.4	Resultados de tarefas de alinhamento baseadas em HMM	33
2.5	Resultados globais para as várias técnicas de alinhamento	35
3.1	Sequência típica para a aquisição e análise de um sinal	39
3.2	Resposta em frequência de um filtro FIR de primeira ordem para vários valores de α	40
3.3	Amplitude média deslizando	42
3.4	Taxa de passagens por zero	44
3.5	Domínio de decisão para a caracterização de sinais de voz tendo por base a energia média deslizando e a taxa de passagens por zero	44
3.6	Modelo genérico para a produção de sinais de fala	46
3.7	Sinal de voz foneticamente anotado correspondente à palavra "pataca" e quadrado do resíduo de predição respectivo para $p = 12$	48
3.8	Exemplo de variação do erro médio de predição em função do número de coeficientes AR	48
3.9	Variação do número de coeficientes AR (domínio do tempo)	49
3.10	Variação do número de coeficientes AR (domínio das frequências)	50
3.11	Variação do sinal de excitação dos modelos AR de 12 coeficientes para a palavra "pataca"	51
3.12	Estrutura <i>lattice</i> para um sistema AR(p) a transformar	52
3.13	Comportamento do modelo MA com excitação por impulsos	53
3.14	Comportamento do modelo MA com excitação por ruído	54
3.15	Diagrama para determinação dos parâmetros AR num modelo ARMA	55
3.16	Modelo do término do trato vocal com a cavidade oral fechada	57
3.17	Erro médio de predição em função dos valores de p e q	58
3.18	Resultado da degradação dos coeficientes AR	59
3.19	Localização dos zeros dos polinómios $P(z)$, $Q(z)$ e $A(z)$ no plano de Argand	62

3.20	Representação da palavra "pataca" por coeficientes LSF	63
3.21	Filtros para o cálculo de MFCCs	65
3.22	Coeficientes de correlação para a comparação de uma métrica de distância espectral e o resultado de um ouvinte humano	67
3.23	Alinhamento utilizando a técnica de <i>Dynamic Time Warping</i>	68
3.24	Procedimento genérico para anotação com DTW.	69
4.1	Procedimento para a segmentação baseada na acústica do sinal	72
4.2	Modelo computacional de um neurónio	73
4.3	Topologias mais comuns para redes neuronais.	74
4.4	Procedimento alternativo para a segmentação baseada na acústica do sinal e posterior classificação	77
4.5	Abordagens para a marcação de fronteiras.	79
5.1	Processamento computacional de mensagens de fala	82
5.2	Modelo de uma cadeia de Markov com 3 estados	84
5.3	Funcionamento de um HMM	85
5.4	Funcionamento do algoritmo de Viterbi aplicado ao reconheci- mento de fala com um modelo de Markov	90
5.5	Arquitectura das ferramentas HTK	94
5.6	Utilização da ferramenta HInit com a aplicação HTKFace	99
5.7	Definições principais para os HMMs com a aplicação HTKFace	101
5.8	Definição das características a utilizar com a aplicação HTKFace	101
5.9	Comparação de dois ficheiros com etiquetas acompanhados do sinal de voz correspondente na aplicação HTKFace	102
5.10	Topologias HMM.	110
5.11	Procedimento para treino dos HMMs	113
5.12	Taxa de acerto (TA) na tarefa de segmentação ao longo de várias iterações	115
5.13	Exemplo de um modelo de linguagem	117
5.14	Situações de erro encontradas com frequência nas anotações ge- radas automaticamente	122
5.15	Procedimentos para a eliminação de fones	123
5.16	Procedimentos para a inserção de fones	123
A.1	Esquema exemplificativo para o divisão de polinómios	130

Lista de Tabelas

1.1	Matriz fonológica para o português europeu	5
1.2	Classificação das consoantes quanto ao ponto de articulação . . .	9
1.3	Características acústicas das consoantes de acordo com o modo e o ponto de articulação.	9
1.4	Classificação das vogais orais	10
1.5	Classificação das vogais nasais	10
2.1	Símbolos utilizados para etiquetar a base de dados FEUP/IPB .	21
2.2	Ocorrências de fonemas para a base de dados FEUP/IPB	22
2.3	Durações médias e respectivo desvio padrão para os fonemas da base de dados FEUP/IPB. Valores em milisegundos.	23
2.4	Material disponível no corpus TIMIT	24
2.5	Exemplos de deslizos dialectais para o Português Europeu carac- terísticos da região do Porto	36
2.6	Exemplos de reduções na articulação.	36
3.1	Variação do número de coeficientes de um modelo ARMA com excitação por impulsos	57
3.2	Variação do número de coeficientes de um modelo ARMA com excitação por ruído	57
5.1	Formatos de armazenamento de etiquetas	100
5.2	Ficheiros utilizados da base de dados FEUP/IPB	103
5.3	Resultados da classificação (PC%/EX%) e segmentação (TA%) variando a dimensão da janela e o passo de avanço.	107
5.4	Resultados da classificação e segmentação variando o número de coeficientes mel-cepstrais.	108
5.5	Resultados da classificação e segmentação variando o vector de características.	108
5.6	Resultados da classificação e segmentação variando a topologia dos modelos.	110
5.7	Resultados da classificação e segmentação variando o número de estados dos modelos.	111
5.8	Resultados da classificação e segmentação variando o número de misturas gaussianas por estado.	111
5.9	Resultados da classificação e segmentação variando o léxico. . . .	113
5.10	Resultados da classificação e segmentação considerando ou não o contexto de cada fone.	116

5.11	Resultados da classificação e segmentação utilizando ou não uma gramática.	118
5.12	Resultados da classificação e segmentação para os corpora utilizados.	120
5.13	Resultados da classificação e segmentação otimizados para a base de dados FEUP/IPB.	121
C.1	Símbolos IPA e SAMPA para o português europeu	136
D.1	Ocorrências de fonemas por cada track da base de dados FEUP/IPB I	138
D.2	Ocorrências de fonemas por cada track da base de dados FEUP/IPB II	139
D.3	Ocorrências de difones na base de dados FEUP/IPB	140
E.1	Conjunto base de HMMs para a base de dados FEUP/IPB . . .	142
E.2	Conjunto alargado de HMMs para a base de dados FEUP/IPB .	143
E.3	Conjunto de HMMs para a base de dados TIMIT	144

Capítulo 1

Introdução

Inicialmente é feita uma pequena introdução ao estudo das línguas onde são apresentados vários conceitos utilizados pela linguística e que são importantes para o estudo do problema da identificação de fronteiras¹.

Seguidamente é também apresentada uma breve resenha histórica sobre a evolução dos sistemas de segmentação automáticos terminando-se com a descrição do estado da arte actual.

1.1 Objectivos

O objectivo desta dissertação é estudar e construir um sistema de anotação automático para sinais de fala e, em particular, para bases de dados de fala. De um modo mais simples, pretende-se desenvolver um sistema que, a partir de áudio de fala, gere um ou vários ficheiros que contenham uma lista de símbolos representativos de uma determinada ocorrência acústica. No trabalho a desenvolver cada símbolo corresponderá essencialmente a um fonema e trabalhar-se-á com todo o conjunto dos fonemas existentes para uma língua. Cada unidade ou símbolo identificado deverá estar acompanhado de informação temporal relativa ao seu instante de início e de fim.

1.2 Motivações

Para o desenvolvimento de sistemas de síntese ou reconhecimento ou simplesmente para a investigação na área da voz e fala é indispensável a existência de uma base de dados que sirva de suporte aos estudos a realizar. Normalmente, esta base de dados é constituída por um conjunto de dados áudio acompanhados por um ou vários níveis de etiquetagem² que podem passar pela simples transcrição fonética até à identificação de marcas e contornos prosódicos, ou mesmo anotações pragmáticas e semânticas.

¹Conceito que se tornará claro mais tarde mas que pretende designar valores temporais que limitam um determinado segmento de áudio de fala.

²A tarefa de etiquetagem consiste, de um modo genérico para este contexto, na segmentação temporal de uma informação de áudio e posterior classificação. As unidades a isolar e identificar estão de acordo com o nível de etiquetagem desejado.

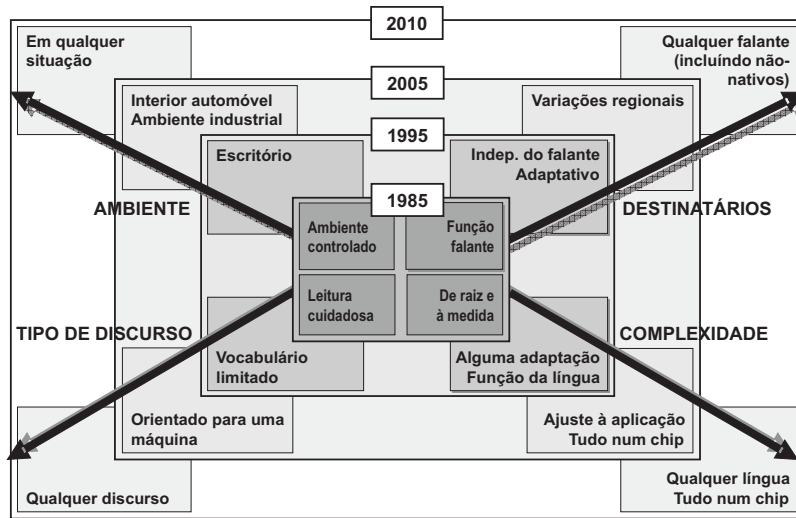


Figura 1.1: Evolução do reconhecimento da fala

A tarefa de etiquetagem quando efectuada à mão é de execução bastante demorada e aborrecida. Ao nível fonético, mais interessante para a maioria das aplicações, é possível demorar, dependendo da dificuldade do trecho, vários dias para etiquetar apenas 1 minuto de fala. Além disto esta tarefa por constituir a base de desenvolvimento dos sistemas de voz deve ser executada com o máximo de rigor. O anotador deve ser alguém com formação específica e de preferência com experiência na área. Deste modo para além do longo tempo necessário existem ainda custos elevados com pessoal especializado.

A importância da existência de um sistema que efectue a tarefa de anotação automaticamente fica assim comprovada e com estes motivos decidiu-se levar a cabo o desenvolvimento deste trabalho.

1.3 Estudo das Línguas

O Homem, como animal social, desenvolveu desde muito cedo métodos de comunicação que permitiam a interacção com outros membros do grupo. Inicialmente, aproveitando a capacidade auditiva aliada à utilização do aparelho fonador³, ambos pouco desenvolvidos, a comunicação⁴ seria essencialmente realizada por meio de grunhidos ou outros sons básicos mas sempre com o objectivo de transmissão de uma mensagem⁵. Na sua essência mais básica estas mensagens seriam alarmes em relação a situações de perigo, o que aumentava a segurança global do grupo. De um modo mais evoluído, a utilização de códigos mais sofisticados permitia por exemplo coordenar vários elementos numa caçada. Uma boa

³Conjunto dos órgãos do corpo humano que permitem a produção de sons, nomeadamente sons de fala [23].

⁴Processo de transmissão e recepção de ideias, informações ou mensagens [58].

⁵Conceito da teoria da comunicação que designa uma sequência portadora de informação enviada por um emissor para um receptor que, em geral, assenta num código compartilhado por ambos [32].

utilização da comunicação contribuía nesses tempos para uma maior longevidade daqueles que a dominavam. No entanto, a comunicação oral pressupõe um emissor e, pelas suas características intrínsecas, conduz a mensagens efémeras, localizadas no tempo e não repetíveis. O registo da informação assume uma importância vital e a necessidade de dar à mensagem recebida um carácter mais persistente e definitivo é fulcral. Criam-se então técnicas de registo em pedra, em madeira, em folhas, e em outros materiais e com elas códigos de símbolos mais ou menos elaborados. Surgem desenhos representativos de cenas do quotidiano, registos de informações importantes ou alertas para um ou outro perigo. Seguem-se códigos pictográficos/ideográficos que, obedecendo à mesma filosofia de representação de uma imagem/ideia, possuem um conjunto de símbolos fixo e limitado com o qual a mensagem é construída. Destes sistemas têm-se como exemplos bem divulgados os hieróglifos egípcios, mais antigos, e os caracteres japoneses, utilizados na actualidade e com cerca de 20.000 símbolos.

Da evolução destes sistemas surgem códigos simplificados em que cada símbolo representa apenas uma sílaba, silabários⁶, ou apenas um som, alfabeto⁷. Nestes existe uma correspondência directa entre símbolos e sons, sendo um alfabeto, à luz dos objectivos que levaram à sua criação, uma representação fonética dos sons de uma linguagem⁸. No vale do Nilo, no Egipto, datadas de 1900 a.C., existem inscrições referentes a transacções comerciais que estão representadas utilizando o que se considera o primeiro alfabeto. O seu desenvolvimento é atribuído aos semitas, conhecidos pela sua organização e pelos grandes comerciantes. Este primeiro alfabeto era constituído por 22 símbolos distintos correspondentes apenas às consoantes. As vogais, apenas três, eram consideradas como implícitas. Em 1000 a.C. os Gregos adoptaram a variante Fenícia do alfabeto Semita e este foi-se disseminado em todo o Mediterrâneo, aceite como código em todas as actividades mercantis. Deste, surge o alfabeto Romano que é espalhado pela força do império e que ainda hoje persiste.

Para além das evoluções gráficas de representação da linguagem surgem também evoluções no significado original dos símbolos. Devido a fenómenos de simplificação da linguagem por economia linguística (assimilações, sonorizações, elisões, etc.) a oralidade vai-se distanciando gradualmente da grafia. Como extensões ao alfabeto básico surgem acentos e outras marcas diacríticas que permitem representar variantes de alguns sons. Do distanciamento dos dois códigos surgem dificuldades. Actualmente são necessário muitos anos de prática para conseguir falar correctamente uma língua pois a um mesmo carácter, de acordo com o contexto, podem corresponder vários sons. Por exemplo, no português europeu, o carácter "a" surge duas vezes na palavra "pata" e possui correspondências sonoras distintas.

As discrepâncias referidas são mais acentuadas nos alfabetos mais antigos. O alfabeto coreano, com apenas alguns séculos de existência, é por esta razão um dos que mais se aproxima da oralidade. Numa outra perspectiva existem os silabários japoneses, *hiragana* e *katakana*, que possuem também uma grande

⁶Conjunto dos sinais componentes de uma escrita silábica ou por sílabas.

⁷Da união de alfa e beta, as duas primeiras letras do alfabeto Grego. Diz respeito a um conjunto de símbolos escritos que representam um ou vários sons e que são combinados para formar as múltiplas palavras de uma linguagem [58].

⁸Referente, em sentido estrito, à linguagem verbal utilizada por dois falantes que comunicam entre si uma mensagem e que estão inseridos num contexto específico. De um modo mais lato designa todos os sistemas de sinais, signos ou símbolos escritos utilizados segundo uma convenção no seio de uma comunidade [62].

proximidade entre os dois campos.

Para o estudo científico das linguagens em geral, os alfabetos tradicionais tornaram-se limitados e criaram-se, com a filosofia inicial, novos alfabetos em que a cada som corresponde bi-univocamente um símbolo: alfabetos fonéticos. Actualmente e com maior expressão existem o IPA⁹ (*International Phonetic Alphabet*), utilizado essencialmente por profissionais da Linguística, e o SAMPA¹⁰ [93], orientado para o processamento computacional das línguas [69]. O primeiro, criado em finais do séc. XIX e já revisto por algumas vezes, é considerado como referência internacional. Utiliza os símbolos do alfabeto latino e alguns símbolos extra específicos para a representação de alguns sons. O segundo, mais recente e adaptado para a representação em computadores, baseia-se em símbolos universais existentes nos teclados com que procura a mesma flexibilidade e abrangência do primeiro. No apêndice C apresentam-se em pormenor, para o Português Europeu, os alfabetos referidos e dão-se exemplos de ocorrências em palavras.

Importa agora introduzir três conceitos fundamentais que serão intensivamente utilizados ao longo deste documento: fonema, fone e alofone. Um fonema é uma entidade abstracta que representa um som mínimo com significado dentro de uma língua. É com estes sons mínimos que, por associação, se constróem palavras. A origem da palavra vem do grego *phónema* que significa som da voz [58]. Os fonemas representam-se tradicionalmente entre duas barras oblíquas. Um fone é uma realização prática de um fonema. Representam-se tradicionalmente dentro de parêntesis rectos bem como as transcrições fonéticas das palavras. Para clarificar, considere-se por exemplo a palavra <vaga>, esta pronuncia-se em Lisboa como "vaga" (em Sampa [bag6]) e no Porto como "baga" (em Sampa [bag6]) sem qualquer perda de significado desde que esteja dentro de um contexto adequado. Ao fonema /v/, primeiro da palavra, correspondem portanto os fones [v] e [b]. Por último surge o conceito de alofone que é referente às várias realizações possíveis de um fonema devido a interacções de contexto com fones vizinhos ou variações livres.

O número de sons diferentes é variável consoante a língua e, dentro desta, consoante as variações dialectais. No caso do português europeu distinguem-se habitualmente 37 fonemas, sendo 14 vogais e 23 consoantes, podendo estes ser classificados segundo vários aspectos [23].

1.4 Fonética e Fonologia

A linguística é a área científica que se dedica ao estudo da linguagem e das línguas naturais. Possui vários ramos de onde se destacam para os nossos objectivos a Fonética e a Fonologia. A Fonética [26] concentra-se no estudo dos sons da fala humana e do modo como são produzidos e percebidos pelo locutores/auditores envolvidos. Subdivide-se em três grandes ramos: fonética articulatória, estudo da produção dos sons pelo aparelho fonador; fonética acústica: estudo dos fenómenos físicos dos sons de fala; fonética perceptiva, estudo do relacionamento dos órgãos da audição com as zonas e mecanismos cerebrais de percepção [23]. A Fonologia [56] procura caracterizar o sistema sonoro de uma língua, as regras subjacentes à combinação desses sons e o modo como

⁹<http://www.arts.gla.ac.uk/IPA/index.html>

¹⁰<http://www.phon.ucl.ac.uk/home/sampa/portug.htm>

	i	e	ɛ	a	ɔ	o	u	j	w	l	λ	r	R	m	n	ɲ	p	b	t	d	k	g	f	v	s	z	ʃ	ʒ
Soante	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-
Silábico	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Consonântico	-	-	-	-	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Alto	+	-	-	-	-	+	+	+	-	+	+	-	+	-	-	+	-	-	-	-	+	+	-	-	-	-	+	+
Baixo	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Recuado	-	-	+	+	+	+	-	+	-	-	-	+	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-
Arredondado	-	-	-	-	+	+	+	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Anterior	-	-	-	-	-	-	-	-	-	+	-	+	-	+	+	+	+	+	+	+	-	-	+	+	+	+	-	-
Coronal	-	-	-	-	-	-	-	-	-	+	-	+	-	+	-	+	-	-	+	+	-	-	-	-	-	-	+	+
Contínuo	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	-	-	-	-	-	-	-	-	+	+	+	+	+
Sonoro	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	-	+	-	+	-	+	-	+	-	+	-
Lateral	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Nasal	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	+	-	-	-	-	-	-	-	-	-	-	-	-

Tabela 1.1: Matriz fonológica para o português europeu

esse sons exprimem distinções de significado [62]. Trubetzkoy [85] é o principal responsável por esta distinção:

"Daremos à ciência dos sons da fala o nome de fonética e à ciência dos sons da língua o nome de fonologia."

A fonologia efectua uma classificação dos sons baseada no conceito de traço distintivo que é a unidade mínima de significado dentro de um fonema que é responsável por oposições fonológicas distintivas. Surgiram duas propostas, uma primeira, em 1952, de base acústica [41] e outra em 1968 de base articulatória [19]. Esta última foi mais bem aceite e foi adaptada ao Português por Mateus [57]. Na tabela 1.1 apresenta-se a matriz fonológica e os respectivos traços distintivos.

Este tipo de classificação é utilizado por alguns autores para realizar tarefas de reconhecimento de fala e segmentação [43].

1.5 Articulação

A articulação diz respeito à forma como o aparelho fonador se comporta para a produção de cada um dos sons e está inserida no campo de estudo da fonética articulatória. A classificação da articulação dos fonemas pode ser feita em relação ao modo de articulação, que especifica a natureza da obstrução que ocorre em determinado local do tracto vocal, e em relação ao ponto de articulação, que especifica o local onde ocorre a obstrução. Considera-se ainda importante a caracterização do vozeamento¹¹ ou não vozeamento.

Analise-se assim os constituintes do aparelho fonador e como se desenrola todo o processo que permite a produção dos sons de fala e contribui para isso.

¹¹Constitui a primeira distinção fonológica ao longo do aparelho fonador. É produzida pela vibração das cordas vocais resultante da pressão sub-glótica que elas exercem à passagem do ar egressivo pela laringe. Ou seja, se as cordas vocais não vibrarem está-se perante um som surdo ou não-vozeado se as cordas vocais vibrarem está-se perante um som sonoro ou vozeado.

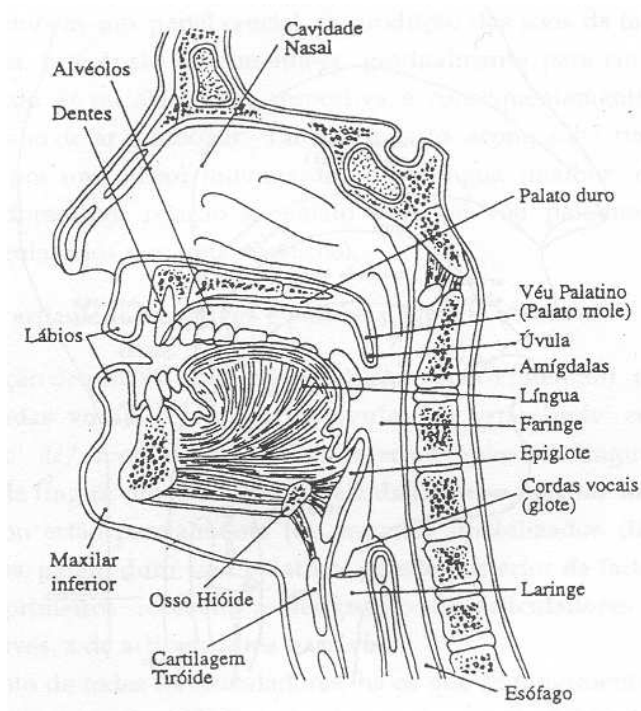


Figura 1.2: Aparelho fonador: desenho científico

1.5.1 Anatomia e Fisiologia do Sistema de Produção de Fala

Para a completa compreensão dos assuntos relacionados com o processamento da fala importa, para além do conhecimento e domínio das múltiplas técnicas e ferramentas, perceber a natureza da produção dos próprios sons. O aparelho fonador, possuindo apenas um órgão específico, laringe, partilha vários com os sistemas respiratório e digestivo. Pode ser decomposto em três secções principais: pulmões, fontes de energia do processo como fornecedores de um fluxo de ar contínuo, laringe e cordas vocais, que criam uma vibração e cavidades supralaríngeas, responsáveis pela articulação e ressonância [100]. Nas figuras 1.2 e 1.3 surgem duas representações do tracto vocal, uma primeira mais realista onde se podem ver claramente os pormenores fisiológicos e a segunda, esquemática, onde ressaltam os elementos que têm sido referidos [6]

A respiração, processo do aparelho respiratório, compreende para este estudo dois momentos essenciais controlados pela contracção e distensão do diafragma. Na inspiração, o ar, por um mecanismo de diferença de pressões, atravessa um percurso que se inicia na boca ou nariz e termina nos pulmões. Na expiração, fase em são produzidos os sons da maioria das línguas, o fluxo de ar efectua o trajecto inverso até ao exterior. Depois de sair dos pulmões, alvéolos e brônquios, passa pela traqueia e seguidamente pela laringe onde se situam as cordas vocais. Estas são constituídas por vários ligamentos e músculos e ao seu espaçamento chama-se glote. As cordas vocais podem estar completamente unidas fechando a glote e contrariando a circulação do ar, podem estar abertas, com vários graus de

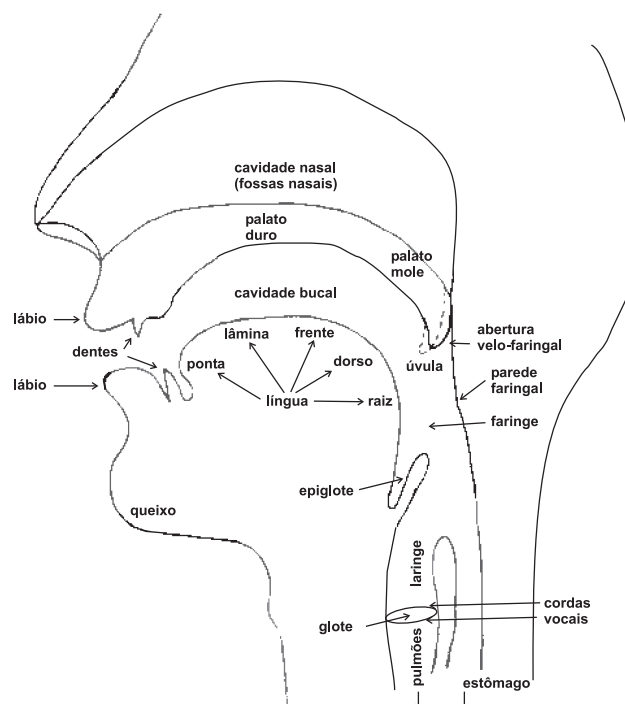
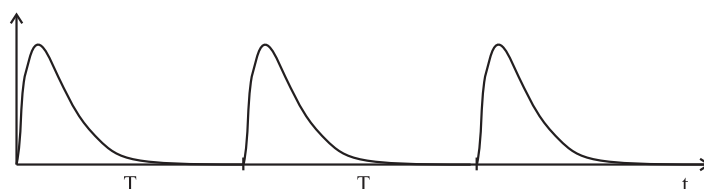


Figura 1.3: Aparelho fonador: esquema

abertura, permitindo que o ar circule e podem efectuar movimentos repetidos de abertura e fecho que produzirão um caudal de ar pulsado. Estes impulsos de ar, designados por impulsos glotais, serão depois modulados pelas cavidades superiores, movimentos da língua e lábios até chegarem ao exterior com a forma de sons vozeados ou sonoros. Quando não existe vibração das cordas vocais os sons são considerados não vozeados ou surdos.

Os impulsos glotais possuem uma forma bastante característica que se apresenta ilustrada na figura 1.4, sendo também assinalado o período fundamental T . Em sistemas artificiais utiliza-se com frequência um modelo que se designa por impulso glotal sintético, dado pela expressão 1.1, e que foi trabalhado por Rowden [73]. O impulso sintético é utilizado em modelos artificiais para a geração de sinais de voz que simulam o comportamento do aparelho fonador humano.

Figura 1.4: Impulsos glotais artificiais ($a=0.9$)

$$G(z) = \frac{-a.e.ln(a).z^{-1}}{(1 - a.z^{-1})^2} \quad (1.1)$$

Assim a cada período da figura 1.4 corresponde um ciclo de abertura e fecho das cordas vocais. A duração de cada ciclo denomina-se período fundamental e o seu inverso será a frequência fundamental.

Após atravessar as cordas vocais o ar prossegue o trajecto através da faringe que se localiza imediatamente antes da cavidade oral e se pode considerar decomposto nas zonas laríngea, oral e nasal. A divisão entre as duas últimas é realizada pelo véu palatino. Este é um músculo ligado às paredes laterais posteriores da cavidade oral e ao palato duro. Possui uma extremidade solta designada por úvula. O véu palatino, pelas suas características musculares, possui a capacidade de se movimentar para cima ou para baixo orientando o fluxo de ar apenas para a cavidade oral ou em simultâneo para as cavidades oral e nasal respectivamente.

Aqui faz-se outra importante divisão na classificação dos sons distinguindo os sons nasalados, quando existe passagem de ar pela cavidade nasal, e sons não-nasalados quando a passagem do ar se efectua exclusivamente pela cavidade oral.

Depois de definido o trajecto inicial do ar durante o processo de fonação, distinguindo os sons em vozeados e não vozeados e em nasais e orais importa clarificar qual o papel das cavidades supraglotais nesse processo e como se forma cada um dos sons da língua.

Designam-se por supraglotais as cavidades orais e nasais e a laringe. A modulação final dos sons, no momento da saída do ar para o exterior, é essencialmente efectuada na cavidade oral devido às várias zonas móveis que possui. É também a partir destas zonas que se efectua uma classificação em relação aos pontos que influenciam e contribuem para a articulação.

Devido às suas várias diferenças realizam-se classificações diferentes para os sons consonânticos e para os sons vocálicos. Nos primeiros, analisados no próximo ponto, o ar encontra obstáculos importantes até à sua saída para o exterior enquanto que nos segundos existe uma quase livre circulação do ar.

1.5.2 Sons Consonânticos e sua Classificação

Em relação ao ponto de articulação, cuja noção foi atrás introduzida, e apenas para os sons consonânticos, distinguem-se, para a língua portuguesa, essencialmente sete locais a que correspondem as classificações bilabial, labiodental, dental, alveolar, pré-palatal, palatal e velar. Na tabela 1.2 apresentam-se com clareza as categorias, seguidas dos órgãos articuladores envolvidos e alguns exemplos.

Em relação ao modo de articulação a classificação é efectuada entre oclusivas, fricativas e líquidas que se dividem em laterais e vibrantes.

As oclusivas possuem a particularidade de serem articuladas em duas fases existindo inicialmente a oclusão, onde o ar é completamente impedido de circular para o exterior com um consequente aumento de pressão e finalmente a explosão ou plosão, em que a pressão acumulada é rapidamente libertada e o som é produzido. O ponto de oclusão é variável.

Nas fricativas existe um fluxo de ar contínuo no tracto vocal que dependendo da obstrução e da velocidade do ar irá provocar diferentes sons.

Classe	Articuladores	Exemplo
Bilabial	Ambos os lábios	b, p, m
Labiodental	Lábio inferior com os dentes incisivos	v, f
Dental	Ponta da língua com os dentes incisivos	d, t, z
Alveolar	Ponta da língua com os incisivos superiores	l, n, r
Pré-Palatal	Lâmina da língua com o pré-palato	z, x
Palatal	Lâmina da língua com o palato	lh, nh
Velar	Zona posterior da língua com o véu palatino	g, k, R

Tabela 1.2: Classificação das consoantes quanto ao ponto de articulação

Clas.	Articul.	Características Acústicas
Ponto	Bilabial	F_2 e F_3 comparativamente baixas
	Dental	F_2 perto de 1600-1800Hz; F_3 perto de 2000-3050Hz
	Alveolar	F_2 perto de 1700-1800Hz
	Velar	F_2 alto; origem comum das transições de F_2 e F_3
Modo	Líquida	Descida de F_3 e F_4
	Oclusiva	Pico saliente no início da estrutura de formantes
	Fricativa	Padrão aleatório dependente do ponto de articulação
	Nasal	Estrutura de formantes semelhante à das vogais com formantes perto de 250,2500,3250Hz
	Laterais	Estrutura de formantes semelhante à das vogais com formantes perto de 250,1200,2400Hz

Tabela 1.3: Características acústicas das consoantes de acordo com o modo e o ponto de articulação.

As líquidas são articuladas com uma obstrução completa ou quase completa da cavidade oral seguida de um escoamento livre do ar. Quando o ar se escoar através dos dois lados da língua temos as laterais. Quando surge uma vibração dos articuladores temos as vibrantes.

No apêndice C apresenta-se uma tabela geral para a classificação das consoantes e outras tabelas complementares.

Em alguns trabalhos procuram criar-se relações acústico-articulatórias que poderão fornecer pistas importantes para as tarefas de processamento da fala que envolvam a extracção de informação partindo do sinal. Num dos trabalhos mais recentes David Chappell [18] propõe as relações que se apresentam na tabela 1.3.

1.5.3 Sons Vocálicos e sua Classificação

Como já foi referido, no caso das vogais existe uma livre circulação do ar para o exterior existindo sempre a intervenção das cordas vocais. Deste modo todas as vogais serão necessariamente vozeadas e não é possível definir um ponto específico de articulação.

A classificação possui assim categorias distintas das consoantes e é realizada segundo os critérios de lugar ou zona de articulação, abertura e labialização. O primeiro diz respeito à posição da língua, o segundo ao grau de abertura da

Abertura	Zona de Articulação			Língua
	Anteriores	Médias	Posteriores	
Abertas		[a] - pato		Baixa
Semi-abertas	[ɛ] - terra		[ɔ] - gola	Baixa
Semi-Fechadas	[e] - Pedro	[α] - manual	[o] - poço	Média
Fechadas	[i] - livro	[ə] - secar	[u] - pular	Alta

Tabela 1.4: Classificação das vogais orais

Abertura	Zona de Articulação			Língua
	Anteriores	Médias	Posteriores	
Semi-Fechadas	[ɛ̃] - dente	[ã] - manto	[õ] - fonte	Média
Fechadas	[ĩ] - pinto		[ũ] - mundo	Alta

Tabela 1.5: Classificação das vogais nasais

cavidade oral e o último refere-se à intervenção dos lábios. Nos quadros 1.4 e 1.5 apresentam-se as vogais da língua portuguesa e a respectiva classificação, repartidas em função da posição da úvula, ou seja, em função da distinção entre oral e nasal.

Para o reconhecimento de voz, a identificação de vogais é uma das tarefas mais simples. Está já estudado e provado por vários autores [25, 66] que é possível caracterizar completamente uma vogal apenas através da análise das duas primeiras formantes¹² F_1 e F_2 . Estas são fornecidas respectivamente pelas ressonâncias da cavidade oral e da faringe e surgem depois de F_0 , frequência fundamental, definida pela vibração das cordas vocais. Em 1988, Delgado Martins [25] propõe para o Português o triângulo acústico que se apresenta na figura 1.5. A configuração das escalas de frequência é um pouco invulgar mas resulta da intenção de facilitar a visualização da posição da língua nos vários pontos quando o triângulo se sobrepõe à imagem 1.3 do trato vocal.

Resta ainda referir as glides ou semi-vogais, fonemas /j/ e /w/ não incluídos nas tabelas, que são sons vocálicos e próximos das vogais fechadas. Estes sons surgem normalmente acoplados a outras vogais na formação de ditongos.

Na articulação dos ditongos os órgãos articuladores movem-se suavemente do primeiro som para o som final sendo por vezes difícil identificar a fronteira entre os dois. Por este motivo os ditongos podem ser consideradas unidades independentes e autónomas.

¹²O termo formante é um adjetivo atribuído às frequências que, somadas, constituem um determinado som. Deste modo, cada som é caracterizado por um conjunto de formantes. As formantes principais, geralmente designadas por F_1 e F_2 , surgem em frequências mais favorecidas pelas cavidades supra-glóticas, as frequências de ressonância do trato vocal.

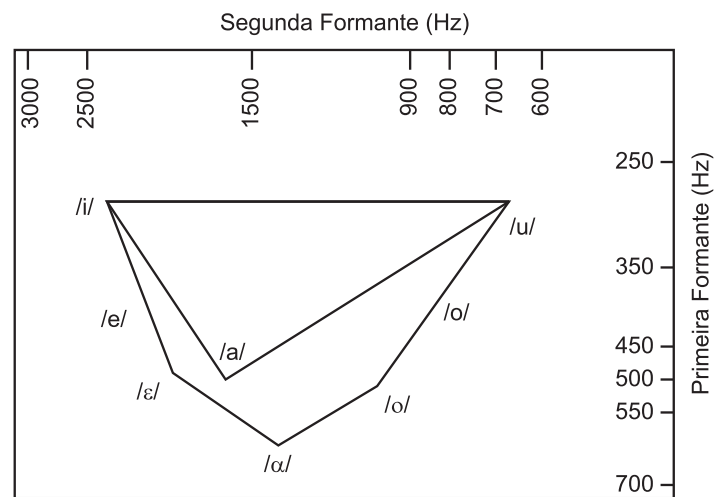


Figura 1.5: Triângulo acústico das vogais tónicas para o Português Europeu

Capítulo 2

Corpora de Fala

Todos os sistemas de síntese e reconhecimento de voz utilizam, no mínimo durante o seu desenvolvimento, base de dados com múltiplas informações que permitem treinar os modelos em que se baseiam. Assim, para além da qualidade intrínseca dos modelos, as informações que lhes irão dar uma expressão prática possuem um papel preponderante no desempenho e na qualidade dos resultados finais. Quais as informações que devem constar de uma base de dados para uma aplicação de processamento de voz? O que define uma boa base de dados? As respostas não são fáceis pois muitas vezes dependem fortemente da aplicação a que se destinam. Existem porém pontos comuns que ajudam a delimitar a extensão das variáveis envolvidas. Nas próximas linhas tentar-se-á abordar os vários pontos que devem ser considerados no desenvolvimento de uma base de dados de fala e que conduzirão a uma maior ou menor adequação e qualidade para uma determinada aplicação.

2.1 Recolha dos Dados

Antes de qualquer outra explicação é importante começar pela criação da própria base de dados que muitas vezes vê alguns aspectos serem negligenciados. A voz, para que possa ser processada por um computador, tem de ser submetida a um processo cuidadoso de recolha para que o seu registo se aproxime o mais possível da realidade.

Logo à partida e como condição base para a gravação dos dados deverá ser escolhida uma sala com boas características de insonorização em relação ao ruído exterior e com um nível reduzido de reverberação interior. Os equipamentos em funcionamento devem reduzir-se aos absolutamente necessários à tarefa, para a eliminação de perturbações de ventoinhas e outros ruídos.

O sinal de voz, variações de pressão do acústica, é captado por um transdutor que produz um sinal eléctrico correspondente. Aqui deve ser considerado um bom microfone, com largura de banda reduzida e de preferência com mecanismos activos de redução de perturbações ambientais. Seguidamente o sinal é conduzido para uma carta de conversão analógico-digital com os devidos cuidados de redução de perturbações electromagnéticas. Na carta, o sinal analógico é convertido num sinal digital, processo de amostragem, obedecendo a uma frequência fixa e posteriormente quantizado num determinado número

de níveis. As duas últimas variáveis são escolhidas em função da aplicação e a sua incorrecta escolha pode ditar a inviabilização da informação recolhida.

A informação digital é registada num suporte físico e armazenada. O suporte poderá ser um CD, DVD, uma cassete DAT, etc. Existem diferenças entre eles mas o CD, por comodidade, é frequentemente o escolhido. O armazenamento pode ser feito sem compressão de dados, com compressão sem perda de informação ou com compressão e perdendo alguma informação considerada negligenciável. Depois de realizada alguma organização que facilite a manipulação, a recolha dos dados está concluída.

2.2 Riqueza Fonética e Acústica

As bases de dados de fones possuem, para além da qualidade do áudio resultante do cuidado colocado na gravação, outras características que permitem compará-las e afirmar se esta ou aquela se adequa mais ou menos a determinada tarefa. Obviamente o ideal seria que todas as características que aqui se vão descrever pudessem ser encontradas numa única base de dados tornando-a universal. No entanto, na fala, as situações a considerar são de tal modo vastas que é praticamente impossível reunir numa única colecção todas as possibilidades desejáveis.

A riqueza fonética de uma base de dados diz respeito ao número de diferentes fones nela existentes. A maioria dos registos existentes são sempre restringidos a apenas uma língua ou a uma variante da língua limitando a generalização dos sistemas desenvolvidos e a comparação de resultados inter-línguas. Deste modo, a situação ideal seria que todos os fones previstos na representação fonética IPA, que se considera universal, pudessem ser encontrados num determinado corpus cobrindo várias línguas e dentro destas todas as variações possíveis. Também importante para a riqueza fonética é o esforço colocado na selecção dos textos que deverão conduzir a fala foneticamente diversificada.

Do ponto de vista acústico existem também várias características a considerar. Tanto para o treino de modelos para o reconhecimento como para a síntese de fala, a existência de múltiplas ocorrências de fones com durações diferentes é importante. Na síntese concatenativa, por exemplo, utilizam-se técnicas para ajustar a dimensão temporal de cada fone ao desejado para a palavra a sintetizar. O resultado final consegue ser bom mas seria indubitavelmente melhor se existisse uma unidade com a dimensão adequada que não sofresse qualquer tratamento.

Também importante é a multiplicidade de falantes. Cada pessoa possui características físicas particulares concernentes ao tracto vocal que influenciam o timbre da sua voz. A dicção, referente à forma de articulação, poderá também conduzir a sons melhor ou pior definidos. A compreensão humana é afectada por estes factores e do mesmo modo é a das máquinas. Assim, quando maior número de falantes puderem ser incluídos mais possibilidades de utilização de fones surgirão. Esta característica é crucial para o desenvolvimento de sistemas de reconhecimento de fala independentes do falante.

Devido à natureza contínua da articulação e dos sinais de voz a vizinhança de um fone, à esquerda e à direita, altera as suas características junto às fronteiras. A existência do maior número possível de vizinhanças distintas de um certo fone aumenta também a riqueza da base de dados. Aqui os ditongos merecem uma nota especial pois devido à grande dificuldade que existe em determinar com

exactidão o ponto de corte das duas vogais muito autores defendem o seu registo como unidade independente.

2.3 Anotação

Um outro factor que também influencia a qualidade de uma base de dados fonética é a qualidade da anotação e o número de níveis ou características a que diz respeito.

A tarefa de anotação é constituída por duas outras distintas: uma, designada por segmentação, que consiste na identificação das fronteiras que delimitam cada fonema e outra, a classificação fonética, que cria correspondências entre o sinal de voz e um conjunto finito de símbolos constituintes de um inventário fonético. Estas tarefas podem ser realizadas de forma independente, o que acontece em muitos sistemas automáticos, ou de uma única vez, como sucede com frequência quando se realiza de forma manual. Para efectuar a anotação é necessário definir um conjunto de símbolos finito representativos dos fonemas que poderão surgir. Normalmente utilizam-se para o efeito o alfabeto SAMPA ou, quando este é limitativo, variações que incluem outros símbolos que se adequem à tarefa. Para além dos símbolos fonéticos definem-se também outros para assinalar a ocorrência de fenómenos com interesse (silaba tónica, oclusão, silêncio, respiração, etc.).

A anotação manual é feita geralmente a partir de uma análise conjunta da imagem da forma de onda do sinal nos tempos e do espectrograma com a audição do som correspondente. As capacidades de percepção visual e acústica dos anotadores são sempre diferentes e daí resultam opiniões diferentes. A tarefa de anotação é também bastante variada e apresenta muitas vezes situações ambíguas que obrigam a tomar decisões nem sempre consensuais. Com o objectivo de minimizar as discrepâncias que inevitavelmente surgiriam do trabalho de vários anotadores desenvolveram-se vários documentos que procuram criar regras e definir estratégias para a anotação. Um destes documentos mais amplamente utilizado é o *CSLU Labeling Guide* produzido no *Oregon Graduate Institute* [45].

O nível de anotação mais baixo localiza-se geralmente ao nível do fone[55]. Aqui, o número de fronteiras a assinalar é bastante grande e nem sempre fácil de definir. Esta tarefa apresenta os melhores resultados quando é realizada manualmente por profissionais treinados e com experiência. Um destes profissionais conseguirá demorar em média pouco menos de 1 minuto para identificar as fronteiras e classificar um determinado fone. Considerando que uma base de dados possui várias horas de som facilmente se percebe que o tempo necessário para efectuar a etiquetagem fonética completa será bastante longo. A título de curiosidade pode-se referir que os laboratórios ATR no Japão utilizam para o treino dos seus sistemas de reconhecimento de voz uma base de dados de 750 horas totalmente etiquetada à mão. No entanto, este processo só é possível quando estão disponíveis grandes recursos financeiros e humanos, sendo noutros casos inviável. Por este motivo é frequente utilizar-se um sistema de segmentação automática de boa qualidade e posteriormente proceder a uma revisão das anotações para corrigir eventuais erros. Este processo reduz consideravelmente o tempo necessário para efectuar a anotação e conduz igualmente a bons resultados.

Podem também surgir níveis de anotação onde se incluem difones e trifones que são unidades com relevância para várias tarefas. No entanto, a anotação destas unidades em particular é normalmente dispensada optando-se por procurar conjunto de fones ao nível do fone.

Os níveis superiores são geralmente os da sílaba, palavra e da frase. Aqui assinalam-se os inícios e fins e no caso das frase outras marcas que possam caracterizar o tipo de frase. No caso das palavras podem também utilizar-se anotações de classificação morfológica. Actualmente começam também a surgir anotações semânticas e prosódicas que permitem analisar outras situações.

Uma classificação mais detalhada e que complementa a apresentada é descrita por Martins et al. [55] que se baseia nas recomendações EAGLES [31]. Resumidamente sugerem-se os seguintes níveis:

1. **Guião de Gravação.** Conjunto dos textos que serviram de guião para a gravação do áudio.
2. **Transcrição Ortográfica.** Representação ortográfica do que foi na realidade pronunciado pelos locutores.
3. **Morfológico, Sintático, Semântico e Pragmático.** Relativo aos níveis linguísticos acima do fonema.
4. **Citação Fonémica.** Corresponde a sequências fonéticas de cada uma das palavras pronunciadas e consideradas isoladamente. Utilização as regras de conversão grafema-fone adequadas à língua.
5. **Transcrição Fonética Larga.** Baseado no nível anterior onde, a partir de regras fonológicas se consideram fenómenos de inserção, eliminação ou substituição de fones. Os símbolos utilizados possuem um significado idêntico ao dos fonemas. Alguns autores defendem que este tipo de transcrição só pode ser correctamente realizado após a audição do sinal de voz.
6. **Transcrição Fonética Estreita.** Surge da transcrição do nível anterior onde, após a inspecção e análise do sinal, se acrescentam outros símbolos, não necessariamente fonéticos, para a representação de fenómenos como alofones, vozeamento, nasalização, etc.
7. **Transcrição Acústico-Fonética.** Refinamento do nível anterior considerando outros fenómenos acústicos como a divisão de uma oclusiva nos seus momentos de oclusão e explosão.
8. **Físico.** Utilizado para a separação de várias fontes de informação simultâneas (ex: palatografo) quando existentes.
9. **Prosódico.** Para a marcação de eventos prosódicos. Não existem ainda recomendações rígidas para o tipo de símbolos ou notação a utilizar sendo a mais frequente baseada no sistema ToBI (*Tone and Break Indices*) [75].

Para além dos níveis de anotação referidos, que são os mais frequentes e os que permitem uma utilidade mínima da base de dados, é ainda possível, paralelamente, efectuar outras anotações com qualquer informação que seja relevante. Aqui incluem-se, por exemplo, informações de frequência, marcações de *pitch*, mudanças de falante, etc.

2.4 Aplicações Informáticas para Anotação

As dificuldades inerentes à anotação de um *corpus* são bem conhecidas e sentidas por todos os que têm de executar esta tarefa em larga ou pequena escala. As informações analisadas para realizar esta tarefa manualmente, em geral sinal de fala e espectrograma correspondente, são calculadas e apresentadas por um computador. Faz então sentido que existam aplicações de *software* que para além de apresentarem as informações referidas permitam também a manipulação e registo das etiquetas subjacentes à anotação.

Nesta secção serão abordadas as ferramentas mais populares para a execução desta tarefa e analisar-se-ão algumas funcionalidades extra que facilitam a realização de outras operações associadas. Os programas apresentados oferecem possibilidades que lhes são comuns e possuem outras que são exclusivas. Por esta razão e para tirar todo o partido das ferramentas é frequentemente necessário utilizar vários dos programas em conjunto ou alternadamente. Para as operações mais comuns a escolha dependerá da familiaridade ou gosto pela *interface* de cada aplicação.

2.4.1 Speech Filling System

Desenvolvido por Mark Huckvale do *University College of London*, o *Speech Filling System* (SFS)¹ é talvez a aplicação mais popular para a análise de sinais de fala. Possui uma *interface* bem organizada e simples, disponibiliza um conjunto alargado de funções e oferece boas possibilidades de integração com outras aplicações e sistemas.

As principais características apresentam-se:

- **Plataformas:** Windows, Linux, MS-Dos
- **Análise de fala:** espectrograma, *pitch*, formantes, processamento de laringográficos, predição linear.
- **Síntese de fala:** por formantes, por filtros, a partir do *pitch*, formantes e intensidade. Síntese articulatória.
- **Anotação:** etiquetagem de intervalos e marcas temporais em vários níveis, utilização de alfabetos fonéticos, anotação automática, comparação de alinhamentos.
- **Manipulação de áudio:** Aquisição e manipulação directa de áudio, alteração dos contornos de *pitch* e duração, filtragem, *resampling*.
- **Compatibilidade:** suporte para vários formatos de áudio (WAV, AU, AIFF, ILS, HTK, etc), suporte para vários formatos de anotação (exportação de imagens em formato EPS e GIF, compatibilidade com HTK).
- **Programação:** *scripts* SML (*Speech Measurement Language*), SPC (*Speech Pascal*), C-SPAN e K-SPAN.

Pelas facilidades de integração oferecidas este foi, dentro das funcionalidades apresentadas, o *software* mais utilizado ao longo dos trabalhos desta dissertação.

¹Disponível em <ftp://ftp.phon.ucl.ac.uk/pub/sfs>

2.4.2 Praat

Apresentado em 1995 por Paul Boersma e David Weenink, do Instituto de Ciências Fonéticas da Universidade de Amesterdão, na Holanda, o Praat² é, em paralelo com o SFS, extensivamente utilizado para a análise de sinais de fala. A aplicação bem como todo o código são distribuídos gratuitamente sob uma licença GNU.

As principais funcionalidades passam-se a descrever:

- **Análise de fala:** espectrograma, *pitch*, formantes, intensidade, *jitter*, *shimmer*, cocleograma e padrões de excitação.
- **Síntese de fala:** a partir do *pitch*, formantes e intensidade. Síntese articulatória.
- **Ensaio de audição:** testes de identificação e distinção.
- **Anotação:** etiquetagem de intervalos e marcas temporais em vários níveis, utilização de alfabetos fonéticos, ficheiros de áudio até 2Gb.
- **Manipulação de áudio:** Alteração dos contornos de *pitch* e duração, filtragem
- **Algoritmo de aprendizagem:** Redes neuronais e teorias de optimização.
- **Estatísticas:** análise PCA (Principal Component Analysis), de discriminantes e multidimensional.
- **Compatibilidade:** suporte para a maioria dos formatos áudio, exportação de imagens em EPS.
- **Programação:** utilização de scripts.

2.4.3 WaveSurfer

O WaveSurfer³ é uma aplicação menos ambiciosa do que as duas apresentadas antes e poderá servir de complemento para realizar algumas operações. Devido ao facto de ser mais recente, os autores Kare Sjolander e Jonas Beskow do *Royal Institute of Technology* em Estocolmo, apresentam frequentemente novas actualizações. O código fonte é distribuído gratuitamente sob uma licença BSD.

As funções mais importantes são:

- **Plataformas:** Linux, Windows 95/98/NT/2K/XP, Macintosh, Sun Solaris, HP-UX, FreeBSD e SGI IRIX;
- **Análise de fala:** espectrograma, *pitch* e formantes;
- **Anotação:** suporte para os formatos (HTK e MLF, TIMIT, ESPS/Waves+ e Phondat);
- **Manipulação de áudio:** Aquisição e manipulação directa de áudio; , alteração dos contornos de *pitch* e duração, filtragem, *resampling*.

²Praat homepage: <http://www.fon.hum.uva.nl/praat/>

³WaveSurfer homepage: <http://www.speech.kth.se/wavesurfer/index.html>

- **Compatibilidade:** suporte para vários formatos de áudio sem limite de duração (WAV, AU, AIFF, MP3, CSL, SD, Ogg/Vorbis e NIST/Sphere).
- **Programação:** *scripts* e funcionamento modular.

Esta aplicação apesar de não ser ainda comparável ao SFS e ao Praat possui já um interessante conjunto de funcionalidades ao nível da manipulação do áudio e respectivas etiquetas. Na presente dissertação foi utilizado para a partição de ficheiros de áudio longos em outros mais pequenos.

2.4.4 Transcriber

Uma alternativa menos específica, orientada apenas para a anotação de áudio em geral, é o *software* Transcriber⁴ cuja primeira versão remonta a 1998 [4]. Este utiliza em várias situações um conjunto de bibliotecas de funções de distribuição gratuita e, à semelhança do que acontece com outras aplicações, tem também disponível todo o seu código fonte sem qualquer custo sob uma licença GNU.

As principais características são:

- **Anotação:** Vários níveis de segmentação e possibilidade organização hierárquica, manipulação de fronteiras. Gestão de uma lista de falantes e tópicos de anotação. Possibilidade de alteração do tipo de letra e cores para cada nível de anotação.
- **Manipulação de áudio:** suporte para áudio com várias horas de duração, zoom horizontal e vertical, visualização do mesmo sinal em vários pontos diferentes.
- **Compatibilidade:** suporte para a maioria dos formatos áudio, suporte para anotações em formato texto e XML

O Transcriber não está orientado para a análise do sinal mas sim para a sua anotação propriamente dita. Assim, e pelas características apresentadas, esta aplicação será mais adequada para tarefas de anotação manual de grande complexidade mas que não exijam o cálculo de características particulares do sinal.

2.5 Corpus FEUP/IPB

A gravação do corpus FEUP/IPB, desenvolvido pelo Laboratório de Sinais e Sistemas da Universidade do Porto, surge de uma colaboração entre a Faculdade de Engenharia da Universidade do Porto e a Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Bragança. A principal motivação para a sua criação foi a inexistência de uma base de dados de fala de elevada qualidade que pudesse ser utilizada para o desenvolvimento de sistemas de síntese de fala em português europeu.

Toda a gravação foi efectuada num estúdio da RDP (Rádio Difusão Portuguesa) com um tratamento acústico de elevada qualidade. A voz utilizada é de um locutor profissional, adulto, do sexo masculino e com cerca de 30 anos de

⁴Transcriber homepage: <http://www ldc.upenn.edu/mirror/Transcriber/>

idade. As sessões de gravação foram preparadas com grande cuidado. Foram feitos vários ensaios onde se simularam sessões de gravação. Os textos foram sempre lidos previamente pelo locutor e foram dadas indicações sobre a intenção de determinada leitura. O áudio foi gravado utilizando equipamento de estúdio de utilização comum e após algumas operações de edição (corte e união) foram produzidos cerca de 100 minutos de material áudio organizados em múltiplos ficheiros com duração entre 2 e 3 minutos cada. Com estes dados foi produzido um CD em formato CDA e um conjunto de ficheiros em formato *wav* com uma taxa de amostragem de 44,1 KHz, 16 bits por cada amostra, mono.

2.5.1 Texto

O corpus de texto é constituído por 9 excertos de diferentes artigos publicados no jornal português diário de maior tiragem em Novembro de 1999, 2 textos adicionais de um outro jornal, uma entrevista publicada no jornal português semanal de maior tiragem no mesmo mês, 2 conjuntos de frases do tipo interrogativo especialmente preparadas para o efeito, com e sem pronomes (quem, qual, quantos, como, onde, etc.), e um conjunto de logátomos⁵, também especialmente preparados, com todos os difones do Português e vários trifones em contexto congruente. Alguns dos textos, devido à sua extensão estão divididos em várias partes.

O conjunto de logátomos consiste em vogais orais, nasais e ditongos, lidos continuamente numa alternância entre sons vocálicos ou entre sons vocálicos e sons consonânticos. Este conjunto está dividido em três ficheiros. O objectivo da gravação dos logátomos prende-se com a criação de alguns difones de rara ocorrência no léxico ou mesmo nele inexistentes, para efeitos de sínteses de fala, pronunciados da forma mais monótona possível.

2.5.2 Segmentação e Classificação

Cada um dos ficheiros de som foi cuidadosamente examinado e segmentado utilizando a aplicação SFS da UCL [39]. A aplicação Cool Edit foi também utilizada para qualquer eventual manipulação do áudio necessária e a aplicação PRAAT [8] para a visualização de algumas características do sinal. A segmentação e classificação dos ficheiros relativos a textos foi efectuada ao nível do fonema, palavra e frase, tendo sido registadas num *logbook* todas as situações que ao longo do processo se consideraram especiais por alguma razão. Os símbolos utilizados para a classificação de cada segmento são baseados no alfabeto SAMPA e apresentam-se na tabela 2.1. São na totalidade 42 símbolos ao nível do fone, 38 representam fonemas da língua portuguesa, 1 símbolo único para a representação da zona de oclusão para todas as oclusivas e três símbolos adicionais para o silêncio (X), a aspiração (XX) e a marcação de sílaba tónica ("). Este último símbolo é colocado imediatamente antes do primeiro símbolo de fonema da sílaba em causa. As informações de etiquetagem relativas a cada ficheiro de som e a cada nível foram guardadas em ficheiros tipo texto em que cada linha possui o tempo de início e o símbolo respectivo.

No nível de etiquetagem de palavras utilizaram-se apenas dois símbolos para início e para fim. Sempre que uma palavra começa imediatamente a seguir a

⁵Unidade fonética concebida para fins científicos ou experimentais sendo constituída por uma sequência consoante-vogal, vogal-consoante ou consoante-vogal-consoante.

Símbolo	Significado
Ao nível do fonema	
p, t, k, b, d, g	Explosões das oclusivas em SAMPA
!	Oclusão comum a todas as oclusivas
f, v, s, z, S, Z	Fricativas em SAMPA
m, n, J	Nasais em SAMPA
L, l, R, r	Líquidas em SAMPA
l*	l no final das sílabas (l velar)
i, e, E, a, 6, O, o, u, @	Vogais em SAMPA
i~, e~, 6~, o~, u~, w~, j~	Vogais nasais em SAMPA
w, j	Semi-vogais em SAMPA
X	Silêncio
XX	Aspiração
”	Início de sílaba tónica
Ao nível da palavra	
i	Início da palavra
f	Fim da palavra
Ao nível da frase	
i	Início da frase
.	Fim da frase
, ! () - ; : ... ”	Todas as marcas de pontuação do texto

Tabela 2.1: Símbolos utilizados para etiquetar a base de dados FEUP/IPB

outra sem qualquer interrupção acústica utiliza-se apenas o símbolo de marcação de início. O mesmo princípio foi utilizado para o nível de frase.

Ao nível da frase utilizaram-se dois símbolos para a marcação de início e fim e ainda os símbolos tradicionais para as marcas de pontuação do texto. A marcação a este nível é útil para o estudo da prosódia e tem sido utilizado em vários trabalhos para este efeito por Braga et al. [11, 9, 12] e Teixeira [82].

O trabalho de etiquetagem para cerca de metade do tempo total foi efectuado manualmente por foneticistas profissionais e estimou-se um tempo médio para a tarefa de 1 dia para cada minuto de fala. O restante material foi pré-etiquetado automaticamente utilizando um sistema desenvolvido na universidade de Gent, por Vorstermans [90] e posteriormente foi revisto manualmente obedecendo às regras que foram utilizadas na primeira metade.

2.5.3 Estatísticas Fonéticas

Para avaliar a riqueza da base de dados foram realizadas diversas estatísticas que se centraram essencialmente nos níveis fonéticos e de palavra. Entre outras ferramentas utilizou-se o Matlab, uma aplicação dedicada de divisão silábica para o português europeu desenvolvida por Gouveia et al. [35] e outra, do mesmo autor, para a conversão grafema-fonema. Outro trabalho para a tarefa de conversão grafema-fonema para o Português pode ser encontrado em [17].

Foi analisada a totalidade dos ficheiros com uma duração total de 1 hora e 10 minutos e cerca de 45000 fonemas. Foi encontrada uma taxa média de fala de 12,3 fonemas/s.

#	Símbolo	Ocorrências	#	Símbolo	Ocorrências
1	p	1327	20	b	447
2	t	2371	21	d	2484
3	k	1833	22	g	499
4	f	465	23	v	457
5	s	1721	24	z	483
6	S	2099	25	Z	469
7	m	1478	26	n	1146
8	J	125	27	l	1029
9	l*	308	28	L	121
10	r	2901	29	R	307
11	i	2811	30	e	1928
12	E	578	31	a	1943
13	6	4245	32	O	276
14	o	1446	33	u	3027
15	@	1428	34	i~	250
16	e~	639	35	6~	1173
17	o~	519	36	u~	249
18	w~	320	37	j~	316
19	w	688	38	j	850
Total: 44756 fonemas					

Tabela 2.2: Ocorrências de fonemas para a base de dados FEUP/IPB

Para cada fonema foram consideradas duas variáveis estatísticas: o número de ocorrências e a duração média. O número de ocorrências é um indicador absoluto e dá-nos uma ideia da riqueza fonética da base de dados ou de cada ficheiro independentemente. Uma vez que os textos foram recolhidos de artigos de jornais é também legítimo considerar, com alguma margem de erro, que o número relativo de ocorrências de cada fonema na base de dados será idêntico ao da língua Portuguesa [24].

Estas estatísticas, relativas ao número de ocorrências, foram comparadas com outras efectuadas a partir de textos disponíveis na base de dados on-line Linguatca [48]. Os resultados foram idênticos.

Muitas vezes associadas às estatísticas sobre a riqueza fonética de uma base de dados surge o conceito de foneticamente balanceado ou foneticamente equilibrado. Em relação a este conceito não existe grande clareza no seu significado e surgem mesmo desacordos entre autores. O equilíbrio fonético é considerado por alguns como a existência de um número idêntico de ocorrências para cada fonema, outros, definem-no como um número de ocorrências idêntico ao existente na língua. A primeira situação é a mais desejável do ponto de vista da criação de modelos para um sistema de reconhecimento, no entanto, é também a mais artificial e surgirá de uma colecção propositada de alguns fonemas. A segunda situação é mais natural mas poderá levar a que surja uma quantidade de dados bastante reduzida e que limitará o rigor de determinados modelos fonéticos.

Apenas a título de curiosidade e uma vez que a sua importância na compreensão da língua é grande apresenta-se na figura 2.1 um gráfico com os pesos de cada vogal no seu conjunto.

Para as durações dos fonemas efectuou-se também um estudo semelhante cujos resultados se apresentam na tabela 2.3 onde surgem os fonemas, suas

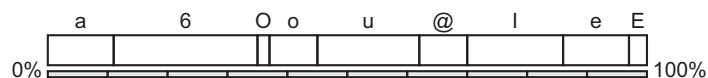


Figura 2.1: Ocorrências de vogais na base de dados FEUP-IPB

#	Símbolo	Duração	DP	#	Símbolo	Duração	DP
1	p	86,7	0,0172	20	b	44,1	0,0060
2	t	51,5	0,0126	21	d	43,7	0,0051
3	k	79,5	0,1982	22	g	45,7	0,0080
4	f	86,7	0,0319	23	v	77,1	0,1884
5	s	77,5	0,1282	24	z	68,5	0,0263
6	S	62,4	0,0079	25	Z	74,7	0,0218
7	m	65,5	0,0374	26	n	86,2	0,0239
8	J	67,1	0,0141	27	l	67,5	0,0112
9	l*	89,6	0,0184	28	L	88,3	0,0250
10	r	61,4	0,0064	29	R	72,8	0,0304
11	i	66,1	0,0102	30	e	51,5	0,0269
12	E	70,4	0,0268	31	a	50,6	0,0328
13	6	55,3	0,0174	32	O	52,6	0,0188
14	o	52,3	0,0086	33	u	60,3	0,0264
15	@	52,5	0,0098	34	i~	71,5	0,0322
16	e~	59,9	0,0161	35	6~	61,7	0,0279
17	o~	40,0	0,0905	36	u~	62,0	0,0270
18	w~	81,4	0,0090	37	j~	78,4	0,0100
19	w	61,3	0,0171	38	j	71,5	0,0139

Tabela 2.3: Durações médias e respectivo desvio padrão para os fonemas da base de dados FEUP/IPB. Valores em milissegundos.

durações médias e desvio padrão. Neste caso os valores apresentados não podem ser generalizados pois a duração é bastante dependente do falante e é afectada por muitas variáveis não facilmente controláveis (por exemplo nervosismo, sonolência ou mesmo outras características que sejam intrínsecas do falante). A duração é ainda uma das características da prosódia e pode também ser alterada pelo tipo de discurso.

Num trabalho de Teixeira et al.[83] podem ser encontradas outras informações com estatísticas sobre a ocorrência de alguns fenómenos fonéticos e ainda num trabalho de Braga et al.[10] surge um estudo sobre o comportamento de alguns fones do Português.

2.6 Corpus TIMIT

A base de dados TIMIT com fala resultante da leitura de textos foi criada com o objectivo de estudar características acústico-fonéticas da língua Inglesa, variante Americana e para o desenvolvimento e avaliação de sistemas automáticos de reconhecimento de voz. A base de dados resulta de um projecto conjunto que envolveu várias instituições e foi patrocinado pelo respectivo governo através da *Defense Advanced Research Projects Agency - Information Science and Techno-*

Tipo	#Frases	#Oradores	Total	#Frases/Orador
Dialectal (SA)	2	630	1260	2
Compacta (SX)	450	7	3150	5
Diversa (SI)	1890	1	1890	3
Total	2342		6300	10

Tabela 2.4: Material disponível no corpus TIMIT

logy Office (DARPA-ISTO). O audio foi gravado pela TI (*Texas Instruments*) e a anotação totalmente manual foi realizada pelo MIT (*Massachusetts Institute of Technology*). Toda a informação foi tratada, verificada e preparada para distribuição em CD-ROM pelo *National Institute of Standards and Technology* (NIST).

2.6.1 Texto

As principais organizações envolvidas e que definiram os textos onde se baseia o corpus são o *Massachusetts Institute of Technology*, o *Stanford Research Institute* (SRI) e a *Texas Instruments*.

O texto do corpus é constituído por duas frases dialectais escolhidas pelo SRI, 450 frases com um conjunto de fonemas reduzido definidas pelo MIT e 1890 frases com elevada diversidade fonética desenvolvidas pela TI. As frases dialectais (assinaladas com a referência SA) têm por objectivo expôr as variações no dialecto de cada orador e foram lidas por todos os oradores. As frases com conjunto de fonemas reduzido (assinaladas com a referência SX) foram escolhidas de forma a cobrir a maior diversidade possível de pares de fones e vizinhanças possíveis. Foi também colocado especial cuidado na criação de ocorrências de contextos fonéticos pouco frequentes ou com um interesse particular. As frases foneticamente diversas (com a referência SI) foram retiradas de fontes já existentes como o *Brown Corpus* (Kuchera e Francis, 1967) e *Playwrights Dialogs* (Hultzen et al., 1964). O critério de selecção foi a da maximização da variedade de contextos alofónicos. O conjunto do material recolhido no corpus TIMIT está sumariado na tabela 2.4.

Assim, o corpus TIMIT é constituído por 63000 frases, 10 frases pronunciadas por 630 oradores diferentes. Os oradores são originários das 8 principais regiões dialectais dos Estados Unidos da América e a sua atribuição a cada região dialectal foi realizada tendo em conta o local de morada durante a infância.

Foram ainda criadas sub-divisões do material com determinados objectivos. Existe assim um conjunto de teste e um conjunto de treino etiquetado ao nível do fone, palavra e frase que é constituído por 4620 frases correspondentes a 462 falantes.

2.6.2 Segmentação e Classificação

As leituras TIMIT foram etiquetadas manualmente utilizando, como habitualmente, um conjunto de símbolos particular. Foram definidos símbolos para 61 fones diferentes e outros adicionais para a marcação de eventos acústicos particulares [97]. Existem autores que defendem uma redução para 48 fones

ou mesmo para apenas 39 fones baseando-se na proximidade aos fones existentes habitualmente nas línguas ocidentais e tendo por objectivo a redução das possibilidades para os decodificadores fonéticos [71]. Na maioria dos estudos publicados baseados nesta base de dados adopta-se o conjunto de 48 fones e está provado que a utilização do conjunto mais alargado não é necessariamente vantajoso [46].

A etiquetagem contemplou para todas as oclusivas, sonoras e surdas, de modo independente, o momento da oclusão e da explosão. Têm-se assim em complemento das etiquetas *p*, *t*, *k*, *b*, *d* e *g* os símbolos das oclusões respectivas *pcl*, *tcl*, *kcl*, *bcl*, *dcl* e *gcl* onde a sigla *cl* vem do inglês *closure*.

Existem ainda outros símbolos de importância menor para o Português pois não fazem parte do léxico fonético da língua na variante europeia. Nestes incluem-se o *h* com vozeamento (por exemplo em *inherited*), as consoantes africadas *jh* e *ch* (que existem na variante do Português do Brasil), para as quais foi seguida uma metodologia idêntica à utilizada para as oclusivas, as *flaps dx* (*muddy*) e *nx* (*dinner*) que surgem nas consoantes duplicadas e outras que não se irão referir exaustivamente.

Foram ainda utilizados os símbolos *h#* para os silêncios iniciais e finais, *pau* para as pausas e *epi* para os silêncios epentéticos⁶.

2.7 Sistemas de Segmentação Automática

A necessidade de possuir um corpora de fala foneticamente anotado é fundamental para qualquer investigação ou desenvolvimento de sistemas baseados em voz. Sendo esta tarefa, quando realizada manualmente, altamente fastidiosa e demorada, apesar de ser realizada com relativa facilidade por anotadores treinados, rapidamente surge o desejo de desenvolver um sistema que automaticamente execute o trabalho ou pelo menos ofereça algum auxílio na sua execução.

2.7.1 Indicadores de Desempenho

Num sistema do tipo em estudo nesta dissertação, como já se referiu anteriormente, existem duas tarefas distintas a efectuar: uma, designada por segmentação, que consiste na identificação das fronteiras que delimitam cada fonema e outra, a classificação fonética, que cria correspondências entre o sinal de voz e um conjunto finito de símbolos constituintes de um léxico ou inventário fonético.

Importa agora definir como avaliar o desempenho na realização de cada uma das tarefas e ao mesmo tempo criar indicadores que forneçam resultados quantitativos que permitam apreciar a qualidade dos resultados, que facilitem a comparação entre sistemas e obviem discrepâncias.

A avaliação isolada de cada tarefa é sempre feita por confrontação com os resultados de uma anotação manual, efectuada por profissionais, que se considera de grande rigor. Daqui, e com os indicadores utilizados, podem efectuar-se comparações com outras situações ou com outros sistemas.

Na tarefa de segmentação, partindo do princípio que a sequência fonética já se encontra ajustada, ou seja, a sequência em análise é igual à referência, o

⁶Relativo à epêntese que é um fenómeno fonético que consiste no acrescento de fonema ou sílaba no meio de palavra. [62].

problema resume-se a um alinhamento das fronteiras limítrofes. Por esta razão a tarefa é também designada por alinhamento fonético. Para obter indicadores quantitativos basta definir-se um intervalo de erro que ditará a conformidade ou não de duas fronteiras e efectuar as respectivas contagens. A avaliação de ocorrência de uma determinada situação é realizada por comparação com fronteiras definidas manualmente, dentro de um intervalo temporal. O desvio admitido para ser considerado o acerto para a fronteira no intervalo deve ser pré-especificado e constante ao longo de toda a avaliação. Os erros encontrados na segmentação possuem quase sempre um erro de classificação associado e reciprocamente pois uma análise recai sobre valores temporais consecutivos (fronteiras ou limites) ao passo que a outra recai sobre os segmentos enquadrados pelos valores anteriores.

O valor de $\pm 20\text{ms}$ para o intervalo de concordância entre fronteiras é aceite como norma na maioria das situações sendo também frequente encontrar-se um limite de $\pm 10\text{ms}$ (serão estes os intervalos a utilizar ao longo deste trabalho). A dimensão destes intervalos de conformidade pode ser qualquer outra mas alguns autores [21, 86], depois de realizarem comparações de segmentações realizadas por múltiplos anotadores, verificaram que são frequentes erros até $\pm 10\text{ms}$ e que com um intervalo de 20ms se reúnem a maioria dos erros.

Os indicadores mais básicos são obtidos por uma relação entre o número de fronteiras conformes (quando uma fronteira se encontra dentro do intervalo temporal em redor da fronteira de referência) e o número total de fronteiras. Deste modo, tem um indicador que se designa por taxa de acerto ou taxa de concordância e que se calcula pela expressão 2.1 onde surgem a totalidade de fronteiras (N) e o número de fronteiras correctamente identificadas (A). Os nomes a atribuir a cada métrica não encontram ainda consenso entre autores apesar de em geral terem a mesma representatividade. Tentar-se-á utilizar uma nomenclatura em harmonia com os termos mais frequentes nos trabalhos científicos da área.

$$\text{Acerto}_{\text{front}}(\%) = \frac{A}{N} \times 100\% \quad (2.1)$$

No entanto, o indicador apresentado não considera situações de eliminação e inserção de fronteiras que surgem frequentemente e não ficam assim claramente identificadas. Por esta razão utilizam-se numa análise de maior complexidade algoritmos de programação dinâmica que ao comparar duas sequências consideram a possibilidade de ocorrência das situações referidas.

As situações a considerar para a avaliação são, com excepção da última, semelhantes para fronteiras e fones e elencam-se rapidamente:

- Concordância. Quando uma fronteira/fone identificados forem assinalados no local correcto
- Eliminação. Quando uma fronteira/fone existente não for assinalada (D)
- Inserção. Quando forem assinaladas fronteiras/fones não existentes (I)
- Substituição. Quando for assinalado um fone existente no local errado (S)

A Substituição não faz sentido quando se faz referência a uma fronteira pois esta é indiferenciada. Os três últimos itens correspondem a situações de erro

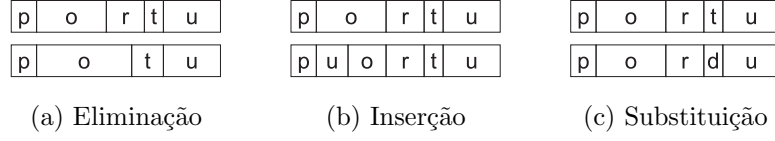


Figura 2.2: Situações de erro na comparação de duas sequências de anotação (em cima a sequência original e em baixo a hipotética sequência gerada automaticamente).

que devem ser contabilizadas na taxa de erro global. Na figura 2.2 ilustram-se com clareza os casos apresentados.

Tendo em conta estas considerações tem-se um novo indicador para a percentagem de fronteiras correctas dado pela expressão 2.2 na qual surgem como variáveis a totalidade de fronteiras na base de referência (N), o número de erros de eliminação (D) e o número de erros de inserção (I).

$$\text{Exactidão}_{front}(\%) = \frac{N - D - I}{N} \times 100\% \quad (2.2)$$

Do cálculo deste último indicador podem resultar valores negativos quando o número de erros ocorridos for superior ao número total de ocorrências da referência.

Para além destes indicadores surgem outros que podem fornecer outro tipo de resultados. Acrescente-se por exemplo a taxa de erro dada pela expressão 2.3 e que considera apenas os erros cometidos pelo sistema.

$$\text{Erro}(\%) = \frac{D + I}{N} \times 100\% \quad (2.3)$$

Este último indicador possui uma variação inversa à dos anteriores o que se pode revelar interessante para algumas aplicações.

Para a tarefa de classificação fonética, analogamente ao que se passa com a segmentação, é aceite como indicador básico a percentagem de fonemas correctamente identificados face à totalidade dos fonemas a identificar. A expressão de cálculo será idêntica e designa-se também por taxa de acerto.

Analogamente ao que se passa na segmentação, também aqui é possível utilizar indicadores mais fiáveis onde se consideram erros de inserção e eliminação e se ainda os de substituição.

Considerando então todas as situações que podem surgir, incluindo agora as substituições, pode definir-se um outro indicador, dado pela expressão 2.4, para a percentagem de fonemas correctos. Neste, as variáveis são semelhantes às utilizadas para o indicador homónimo da segmentação considerando ainda uma variável para a contabilização dos erros de substituição (S).

$$\text{Exactidão}_{fones}(\%) = \frac{N - D - S - I}{N} \times 100\% \quad (2.4)$$

Também frequente, mas que não considera os erros de inserção é a "Percentagem Correcta" dada por:

$$\text{Percentagem Correcta}_{fones}(\%) = \frac{N - D - S}{N} \times 100\% \quad (2.5)$$

É também possível definir outros indicadores, como a Taxa de Erro, mas considerando sempre os erros de substituição que são frequentes na classificação fonética automática. Quando se realiza apenas o alinhamento fonética estes erros não fazem sentido.

A análise das situações de erro referidas, tanto para a segmentação como para a classificação, é feita com recurso a algoritmos dinâmicos relativamente simples. Uma possibilidade para a contabilização das situações assinaladas [38] apresenta-se nos passos seguintes para a classificação por considerar mais possibilidades.

1. **Inicialização.** Considerem-se duas sequências fonéticas compostas por n e m fones respectivamente. Considerem-se também as matrizes $R[i, j]$ e $B[i, j]$ com $0 \leq i \leq n$ e $0 \leq j \leq m$.

$$\begin{cases} R[0, 0] = 0 \\ B[0, 0] = 0 \end{cases} \quad (2.6)$$

2. **Iteração.** Serão percorridas ambas as sequências do início ao fim e utiliza-se $R[i, j]$ para armazenar o custo acumulado. A contabilização das várias situações é feita com os ciclos:

Com $i=1, \dots, n$ executar { Com $j=1, \dots, m$ executar {

$$R[i, j] = \min \begin{bmatrix} R[i-1, j] + 1 & \text{(eliminação)} \\ R[i-1, j-1] & \text{(concordância)} \\ R[i-1, j-1] + 1 & \text{(substituição)} \\ R[i, j-1] + 1 & \text{(inserção)} \end{bmatrix} \quad (2.7)$$

$$B[i, j] = \begin{cases} 1 & \text{se eliminação} \\ 2 & \text{se inserção} \\ 3 & \text{se concordância} \\ 4 & \text{se substituição} \end{cases} \quad (2.8)$$

} }

3. **Retro-Propagação e Conclusão.** Em $R[n, m]$ tem-se o alinhamento ótimo.

$$\text{Taxa de erro} = 100\% \times \frac{R(n, m)}{n} \quad (2.9)$$

$$\text{Caminho ótimo} = (s_1, s_2, \dots, 0) \quad (2.10)$$

$$s_1 = B[n, m] \quad (2.11)$$

$$s_t = \begin{bmatrix} B[i-1, j] & \text{se } s_{t-1} = 1 \\ B[i, j-1] & \text{se } s_{t-1} = 2 \\ B[i-1, j-1] & \text{se } s_{t-1} = 3 \vee s_{t-1} = 3 \end{bmatrix} \quad (2.12)$$

Um aspecto de elevada importância para a presente dissertação é que o algoritmo descrito poderá também ser aplicado, com ligeiras alterações, às fronteiras da segmentação.

2.7.2 Anotação Manual

Os primeiros ensaios realizados com experiências de segmentação e classificação colocam à prova várias sequências produzidas por diferentes anotadores sobre um mesmo corpus. Nesta situação uma das anotações é tido como correcta, sendo utilizada como referência, e as restantes são avaliadas em relação a esta. Esta forma de avaliação é aceite e utilizada pela maioria dos autores.

Em 1984, Leung e Zue [47] compara o alinhamento manual efectuado por dois anotadores sobre 5 frases em Inglês Americano retiradas da "Harvard List of Phonetically Balanced Sentences". Comunicou uma média de 30s para o ajuste das fronteiras de cada fonema, cerca de 80% de concordância para um intervalo de 10ms, 87% para um intervalo de 15ms e 93% para um intervalo de 20ms.

Em 1991, Cosi et al. [21] avalia o alinhamento manual para 10 frases de fala contínua em Italiano efectuado por três anotadores diferentes. Foi encontrado um desvio médio nas fronteiras de 6ms, cerca de 55% de concordância para o intervalo de ± 5 ms e 94% para o intervalo de 20ms.

Cole et Al. [20], em 1994, efectua uma comparação baseada num corpora multilingue (OGI Multi Language Corpus [60]) constituído por frases de conversas telefónicas em Inglês Americano, Alemão, Mandarim e Espanhol anotadas por falantes nativos e falantes não-nativos. Para o Inglês Americano, anotadas por dois nativos, anuncia 79% de concordância para um intervalo de 11ms. A mesma comparação para o Alemão, com condições semelhantes, produziu os mesmos valores sendo no entanto 4% superior para o Mandarim. Na comparação entre falantes nativos e não-nativos anuncia 81% de concordância para o mesmo intervalo.

Cosi [21], Torkkola [84] e Van Erp [86] pronunciam-se sobre a elevada dificuldade em conseguir resultados óptimos de concordância absoluta entre alinhamentos realizados por diferentes alinhadores humanos. Isto devido à grande variabilidade das capacidades de percepção visual e acústica de cada indivíduo e também devido à inexistência frequente de regras de segmentação claras e bem definidas que constituam uma estratégia a utilizar para cada situação.

Tendo em conta estas afirmações, Wesenick e Kipp [94] comparam, em 1996, o alinhamento efectuado por três anotadores sobre frases em língua Alemã. Os anotadores eram estudantes universitários de fonética e receberam formação específica para a realização da tarefa existindo um conjunto de regras comuns a utilizar. Encontram uma concordância de 73% para um intervalo de 5ms, 87% para um intervalo de 10ms e 96% para um intervalo de 20ms. Dos trabalhos analisados surge aqui a melhor taxa de concordância.

Em 1997, Ljolje et. al [50] comparam dois anotadores que trabalharam 100 frases em Italiano. Comunicam para o intervalo de 10ms, concordância em 80% dos casos e para o intervalo de 20ms, concordância em 93% dos casos.

Da breve análise dos trabalhos de alguns autores aqui realizada verifica-se a existência de uma grande coerência na tarefa de alinhamento manual com uma dependência reduzida da língua, dos anotadores e do ano. Na figura 2.3 apresentam-se claramente os resultados obtidos por cada um dos autores. Os

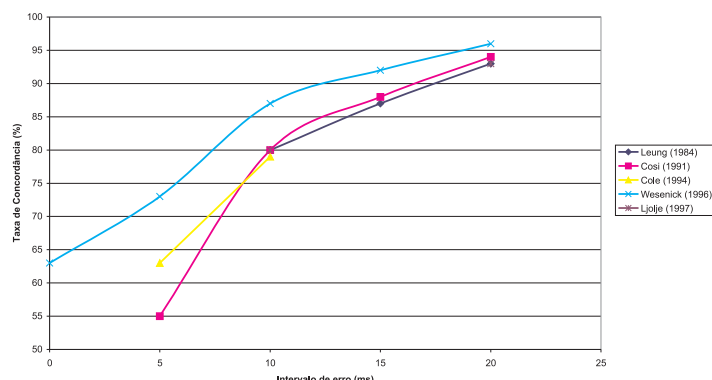


Figura 2.3: Resultados de tarefas de alinhamento manual

valores apresentados para as discrepância entre diferentes anotadores poderão ser utilizados como tecto para a tarefa de anotação automática. Para além disto prova-se [21, 86] que diferentes anotadores humanos discordam em mais de 20ms em 10% das situações.

Infelizmente nenhum dos trabalhos apresentados se centra na base de dados TIMIT que serve com frequência para efectuar comparações.

Em 2004, Ulrike [36] analisa várias anotações manuais e conclui que a qualidade da anotação está correlacionada com a complexidade do conjunto de símbolos escolhidos para a tarefa.

2.7.3 Utilização de DTW

Uma das primeiras técnicas utilizadas para a tarefa de alinhamento é a que utiliza DTW (*Dynamic Time Warping*). Baseada num algoritmo de programação dinâmica (como o é também o algoritmo de Viterbi e o algoritmo Forward-Backward), procura alinhar duas sequências minimizando as diferenças entre características do sinal. As distâncias entre características são medidas utilizando uma qualquer métrica que se convencie. A utilização desta técnica não obriga à existência de um corpus de treino.

Em 1981, Wagner [91] apresenta um dos primeiros sistemas baseados nesta técnica. Em 1987, Svendsen e Soong [80] utilizam DTW para alinhamento fonético e anunciam 32% de concordância para um intervalo de 15ms e 72% para um intervalo de 30ms. Falavigna, em 1990, com objectivos semelhantes, comunica uma concordância em 61% dos casos avaliados para um intervalo de 20ms.

Mais recentemente, em 1998, Malfrere et al. [54] utilizam técnicas DTW refinadas para efectuar o alinhamento de um sinal de voz proveniente de um sistema TTS (*Text-to-Speech*) e comparam o seu desempenho com um sistema baseado em modelos escondidos de Markov. Anunciam 82,1% de concordância para um intervalo de 20ms para o sistema DTW o que torna este resultados nos melhores encontrados para esta tecnologia. Para o sistema baseado em modelos de Markov obtêm 84,0% no mesmo intervalo.

O desempenho inferior da técnica DTW, já bastante madura e explorada, que está presente nos resultados de Malfrere et al. é confirmado pela abandono quase completo desta técnica para efectuar o alinhamento fonético. Os avanços na tecnologia trouxeram máquinas que permitiram utilizar algoritmos de base probabilística que se revelaram superiores.

2.7.4 Anotação Automática com HMM

Incontornáveis na área do processamento de voz, os modelos escondidos de Markov (*HMM - Hidden Markov Models*) permitem entre outras aplicações o desenvolvimento de sistemas de alinhamento fonético através de um processo conhecido por alinhamento forçado. Os HMM fazem uma descrição da evolução temporal de um processo que no caso da voz é constituído por sequências de características⁷ do sinal em estudo. Os algoritmos utilizados são globalmente mais exigentes do ponto de vista computacional que os utilizados no DTW. No caso de uma tarefa de reconhecimento de voz⁸ genérica todos os HMMs existentes, cada um correspondente a um tipo de sequência acústica, são testados com o objectivo de procurar o que melhor se ajusta a um determinado sinal. No alinhamento forçado a sequência de fonemas é já conhecida e portanto também o é o HMM correspondente, restando apenas efectuar o ajuste temporal. Esta tarefa é efectuada normalmente pelo algoritmo de Viterbi que para além de fornecer as fronteiras de cada fonema permite também ter uma ideia da qualidade do ajuste sob a forma de uma probabilidade. A pesquisa Viterbi na sequência de características é baseada em critérios de máxima verosimilhança e em probabilidades de transição.

Em 1990, Svendsen e Kvale [79] apresentam um sistema de alinhamento automático parcialmente baseado em HMMs e efectua testes no corpus EUROM0 constituído por um pequeno número de frases em várias línguas europeias. Para o Inglês-Britânico é anunciada uma taxa de concordância de 82.3% para um intervalo de 20ms. O sistema utiliza modelos mono-fonema com três estados e uma mistura gaussiana.

Em 1991, Ljolje e Riley [51] desenvolvem um sistema que utiliza para a modelização dos fonemas três HMMs distintos que são escolhidos em função da disponibilidade de dados para o seu treino. Todos os modelos possuem três estados. Se existirem dados suficientes utiliza-se um modelo de trifone completo que considera o contexto à esquerda e à direita, se não existirem dados suficientes constrói-se um "quasi trifone" em que se vão criando independência sucessivas do contexto, se ainda assim os dados não forem suficientes utiliza-se um modelo completamente independente do contexto. O sistema foi testado na base de dados TIMIT e para um intervalo 15ms conseguiu-se uma taxa de concordância de 80%.

Brugnara et al. [13, 14], em 1993, comunicam os resultados da ferramenta de alinhamento desenvolvida. Na base de dados TIMIT obtêm 75,3% para um intervalo de 10ms, 84,4% para um intervalo de 15ms e 88,9% para um intervalo

⁷Esta palavra surge da tradução do inglês *features* e que designa vectores que contêm um conjunto de valores que descrevem de algum modo uma parte de um sinal de voz. Frequentemente utilizam-se como características MFCCs, LPCs ou outras. Este tema será abordado mais à frente neste texto.

⁸Tarefa que consiste na identificação de uma sequência de símbolos (fonemas, difones, palavras, etc.) dentro de um léxico finito a partir de um sinal de fala.

de 20ms. A utilização de características de variação espectral permitiu reduzir o erro em 2% em relação a um primeiro ensaio com coeficientes cepstrais.

Em 1995, Rapp [70], utiliza o HTK Toolkit da Entropics [96] para o alinhamento fonético num corpus em língua alemã. Anuncia um concordância de 84,4% para um intervalo de 20ms.

Em 1996, Kipp et al. [44] surgem com um sistema para transcrição automática ao nível da palavra. Avaliado com base no corpus em língua alemã PHONDAT-II o sistema apresentou uma taxa de concordância de 84% para um intervalo de 20ms.

Ljolje et al. [50] apresentam, em 1997, numa abordagem diferente do problema, uma taxa de concordância de 78.1% para um intervalo de 20ms. Neste caso utilizam um sistema baseado em HMMs com três estados mono-fonema e utilizam um corpus em língua Italiana.

O conhecido sistema *The Aligner* é apresentado em 1997 por Wightman e Talkin [95]. Este sistema de alinhamento é construído utilizando as ferramentas disponibilizadas pelo HTK Toolkit [96]. Foram utilizados modelos mono-fonema com cinco misturas e para o treino foi feita uma distinção entre as oclusões dos fonemas surdos e sonoros. No corpus TIMIT o sistema apresentou uma taxa de concordância de 80% para intervalos de 20ms.

Em 1998, Malfreire et al. [54] comparam um sistema de alinhamento baseado em HMMs com um sistema de alinhamento baseado em DTW (*Dynamic Time Warping*). Para o sistema baseado em HMMs foram utilizados modelos para os fonemas com 16 misturas e o sistema foi desenvolvido treinado e avaliado sobre corpus em língua francesa. Anunciam uma taxa de concordância de 84% para um intervalo de 20ms.

Em Portugal, Amaral et al. [2] apresenta em 1999, um alinhador baseado em HMMs que utiliza as ferramentas do HTK Toolkit [96]. Foi utilizado o corpus multilingue EUROM1 que compila frases em Português Europeu, Espanhol e Grego. Para a tarefa de alinhamento anuncia uma taxa de concordância de 78,9% para um intervalo de 10ms. Estes resultados são ligeiramente melhorados por Carvalho [16]. Este último autor utiliza uma técnica baseada em HMMs para efectuar o alinhamento fonético e aborda vários problemas relacionados com discrepâncias entre transcrições largas e estreitas que dificultam a melhoria dos resultados.

Em 2000, Hosom [37] desenvolve um sistema de alinhamento baseado em HMMs e redes neuronais (*ANN - Artificial Neural Networks*) que utiliza um conjunto alargado de informação acústica e fonética. O sistema é extensivamente testado em 14 corpora diferentes em várias línguas. Para o base de dados TIMIT, termo de comparação frequente, anuncia 92,6% de concordância para um intervalo de 20ms. Este é o autor que comunica os melhores resultados para a tarefa de alinhamento fonético.

Mais recentemente, em 2002, Sethy e Narayanan [74], propõem um sistema de refinamento para um seu sistema de alinhamento por HMMs dependentes do contexto. O refinamento conduz a um ganho relativo de 10 a 12% para a concordância avaliada para intervalos de 20ms. Demuynck [28] compara ainda alternativas para o algoritmo Viterbi e propõe alterações que levam a um ganho relativo de 7% em relação a trabalhos seus anteriores.

Os resultados apresentados podem ser mais facilmente comparados na figura 2.4. Os valores anunciados por Hosom destacam-se como os melhores para a tarefa do alinhamento e o seu sistema é testado num grande variedade de corpus.

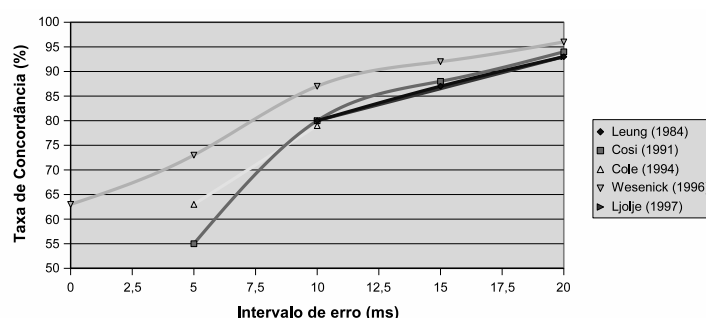


Figura 2.4: Resultados de tarefas de alinhamento baseadas em HMM

Com excepção deste último autor a maioria dos sistemas propostos é testado apenas num corpus específico e na língua dos autores o que introduz algum grau de erro na comparação dos resultados.

2.7.5 Outros Sistemas de Anotação

As tecnologias apresentadas, DTW e HMM, não são obviamente as únicas existentes e muito menos as únicas possíveis para realizar a tarefa de alinhamento. Existem diversas outras técnicas que com maior ou menor sucesso conseguem cumprir este objectivo. Apresentam-se aqui algumas referências a trabalhos que utilizaram outras abordagens para o problema e que mereceram a atenção da comunidade científica.

Em 1984, Leung e Zue [47] anunciam um processo de alinhamento fonético em três fases. Inicialmente o sinal é classificado por uma árvore de decisão em seis grandes categorias fonéticas. Depois, através de um algoritmo de programação dinâmica são seleccionadas sub-unidades da primeira segmentação que se pretende já corresponderem a fonemas. Por fim, utilizam-se regras heurísticas para refinar o processo. Foi comunicada uma taxa de concordância de 75% para intervalos de 10ms e 90% para um intervalo de 20ms. Estes resultados são excelentes na época em que foram publicados.

Vorstermans et al. [90], em 1996, utilizam uma técnica de alinhamento baseada em segmentos/redes neuronais com uma base estatística. Fazem uma extensa fundamentação matemática do problema da anotação e desenvolvem mecanismos probabilísticos formais para a linguagem e para características acústicas do sinal. Utilizam para o treino do sistema um corpus em língua flamenga e com pequenas adaptações efectuam ensaios na base de dados TIMIT, no corpus EUROM0 e em outros corpus em várias línguas. Para a primeira anuncia uma concordância em 81% dos casos para um intervalo de 20ms.

Em 1997, Amaral et al. [3], apresentam um sistema de anotação para a base de dados TELEFALA constituído por locuções dos dígitos de 0 a 9 em língua Portuguesa variante Europeia. Testam para a segmentação um algoritmo baseado em MLR (*Maximum Likelihood Ratio*) e outro baseado em KLT (*Karhunen-Loeve Transform*). Do sinal extraem características baseadas na percepção através de um conjunto de filtros alinhados na escala de Bark. Utilizando este método de segmentação não é necessário efectuar qualquer treino prévio do sistema sendo possível adaptá-lo a outras línguas. A classificação é

realizada por um processo baseado em lógica difusa. Anunciam uma taxa de concordância de 88.7% para um intervalo de 20ms.

Em 1998, Gholampour e Nayebi [33] apresentam um sistema baseado em dendogramas e em comparações de máxima verosimilhança. Anunciam para a base de dados TIMIT uma taxa de concordância de 77,4% para um intervalo de 20ms.

No mesmo ano, Keller [43] apresenta um sistema de segmentação com redes neurais e utiliza uma abordagem original baseada apenas na extração dos traços fonológicos de cada fonema existente na língua.

Em 1999, Van Santen e Sproat [87] utilizam para a tarefa de alinhamento a energia do sinal e características do domínio espectral em diferentes bandas de frequência. O sistema foi testado num corpus de apenas um falante e possui uma taxa de concordância de 90% para um intervalo de 20ms. Quando avaliado sobre os dados de treino comunica-se uma taxa de 95% para um intervalo de 6ms.

Num trabalho recente, Wang et al. [92] utilizam uma técnica para efectuar a segmentação de um sinal de fala e classificar cada segmento identificado como pertencente a uma das classes vogal (V), consoante (C) ou pausa (P). A identificação das pausas utiliza uma técnica adaptativa que se ajusta a diferentes níveis de ruído de fundo e a identificação das fronteiras fonéticas utiliza uma suavização dos parâmetros. Pela sua simplicidade e interesse apresenta-se com brevidade o algoritmo desenvolvido:

1. Sinal de fala segmentado em janelas de 20ms sem sobreposição e são extraídas a taxa de passagens por zero, energia e informação de *pitch*.
2. As curvas de energia e *pitch* são suavizadas.
3. Calcula-se a média (MedE) e o desvio padrão (DPE) da energia para obter uma estimativa do nível de energia do ruído de fundo (NRF). Este será dado por $NRF = MedE - 0.75DPE$. Estima-se também um nível de *threshold* para a taxa de passagens por zero (ThresTPZ) que será dada por $ThresTPZ = MedTPZ + 0.5DPTPZ$.
4. Cada janela é classificada grosseiramente como vogal, consante ou pausa de acordo com as regras seguintes:
 - Se $TPZ > ThresTPZ$ então sinal da janela é consoante;
 - Se $Energia < NRF$ então sinal da janela é pausa;
 - Se não se verificar nenhuma das situações anteriores então o sinal da janela é vogal.
5. Actualização do nível de ruído (NRF) como a enrgia média ponderada das janelas em cada fronteira de vogal e os segmentos de pausa
6. Reclassificar os segmentos utilizando o procedimento do passo 4 com o novo nível de ruído. As pausas são concatenadas através da remoção de consoantes soltas de curta duração. As vogais serão partidas quando no vale de energia correspondente se a sua duração for elevada.

Com este algoritmo os autores afirmam conseguir sobre o corpus *HUB4 English Broadcast News* uma taxa de 82.3% na segmentação de frases. Na classificação obtem-se um acerto de 86.7% não sendo porém fornecida informação

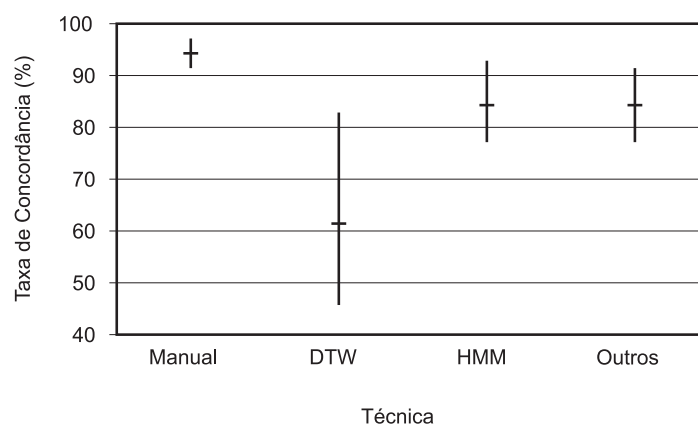


Figura 2.5: Resultados máximo, médio e mínimo para cada um das técnicas de alinhamento para um intervalo de concordância de ± 20 ms

sobre a qualidade da segmentação fonética. Uma abordagem semelhante será utilizada mais à frente num dos ensaios realizados.

2.7.6 Comparações

Os resultados apresentados pelos vários autores e para as diferentes técnicas estão resumidos no gráfico da figura 2.5 e podem ser facilmente comparados.

Os valores observados para a segmentação manual são os mais regulares e simultaneamente os melhores na globalidade. Isto mostra que existe uma elevada coerência entre os vários anotadores e vem legitimar a utilização das fronteiras fonéticas colocadas manualmente como referência para a avaliação do desempenho dos sistemas automáticos.

De entre as várias técnicas automáticas os HMMs são os que melhores resultados apresentam e serão por isso utilizados neste trabalho. A reduzida diferença entre os valores máximo e mínimo é também indicadora da adequação desta tecnologia à tarefa. No caso do DTW está bem patente a variabilidade dos resultados demonstrando que existem muitas outras variáveis a considerar para além do algoritmo de alinhamento propriamente dito.

O breve resumo aqui apresentado apenas versa alguns dos autores mais conhecidos e os trabalhos que possuem maior representatividade no panorama científico nacional e internacional não incluindo exaustivamente tudo o que na área foi feito. Os vários resultados apresentados permitem dar um termo de comparação ao trabalho que neste texto se irá descrever e ao mesmo tempo apontam objectivos de desenvolvimento que de outro modo correriam o risco de ser menos ousados.

2.7.7 Fenómenos Linguísticos

Quando se realiza a conversão grafema-fonema de uma dada palavra utilizam-se regras baseadas na pronúncia padrão da língua. No entanto existem alterações devidas a variedades dialectais ou mesmo outros fenómenos relacionados com a

Exemplo	Norma	Alteração
doutores	/o/	/ow/ ditongação
hoje	/o/	/oj/, /je/ ditongação
ele	/e/	antes de consoante palatal
regressou	/R/	/r/ vibrante alveolar múltipla
embora	/e~/	/6~j/

Tabela 2.5: Exemplos de deslizes dialectais para o Português Europeu característicos da região do Porto

Reduções	[@]	[u]
No início	<explorado> [<i>Splu</i> "radu]	n/a
Ao meio	<decisão> [<i>dsi</i> "z6 ~ w]	<português> [<i>prt</i> "geS]
No fim	<deve> ["dEv]	<Porto> ["port]

Tabela 2.6: Exemplos de reduções na articulação.

oralidade que fazem com que a o equivalente fonético difira do que se considera norma.

Os fenómenos mais comuns para a zona do Porto apresentam-se segundo Teixeira [83]:

- **Deslizes dialectais.** Causados por hábitos de articulação relaxada que e são assinaláveis na maioria dos falantes nativos da língua podendo surgir mesmo em locutores profissionais. Alguns exemplos na tabela 2.5.
- **Supressões ou Reduções.** Alterações devidas ao contexto que provocam a não articulação de fonemas. Por exemplo as vogais [@] e [u] são frequentemente omitidas em várias posições no contexto da palavra excepto quando se encontram em posição tónica. Alguns exemplos na tabela 2.6.
- **Transformação da qualidade das vogais.** Surge quando duas vogais de qualidades (ou timbres) diferentes se associam. Duas situações podem resultar: na primeira as duas vogais fundem-se e sofrem uma mudança vocálica (ex.: <fica admirado> [fikadmiradu]; <contra o> [ko~trO]), na segunda a vogal fechada [@] ou [i] reduz-se e transforma-se numa semivogal resultado daí um ditongo (ex.: <se aprende> [sj6pre~d]; <na idade> [n6jdad]). Estes fenómenos são conhecidos há muito pelos poetas que deles se aproveitam para o ajuste de métricas e ritmos nas suas criações.
- **Adições.** Nos pares de consoantes oclusiva-líquida, oclusiva-lateral (pl, tl, kl, bl, dl, gl) e oclusiva-vibrante (pr, tr, kr, br, dr, gr), surgem pequenos sons vocálicos designados por "schwas" (ex.: <branco> [b@ra~ku]).
- **Mudanças fonéticas.** Os fenómenos de co-articulação e os mecanismos de compensação devido à natureza contínua de funcionamento do tracto

vocal muitas vezes provocam erros na articulação mas que ainda assim não destrõem a inteligibilidade da mensagem. Aqui existem um grande número de fenómenos. Um exemplo habitual é a transmissão das características de vozeamento de uma vogal à consoante não vozeada procedente (ex.: <ao contrário> [awgo ~ "trariu]; <quarenta> [kware ~ d6]). Nesta situação surge uma sonorização.

Na anotação de um ficheiro de voz deve existir um especial cuidado na identificação dos fenómenos apresentados. Esta tarefa é bastante difícil e para que não se cometam erros, num corpus que se quer que esteja de acordo com o padrão da língua, o melhor será escolher um locutor (ou vários) que possuam a melhor articulação possível.

A anotação incorrecta poderá levar a um treino erróneo de modelos de fala que deterioram o desempenho de sistemas de síntese e de reconhecimento.

Capítulo 3

Análise de Sinais de Fala

Antes de qualquer consideração sobre o trabalho de segmentação e alinhamento desenvolvido ao longo desta dissertação convém apresentar sucintamente as ferramentas e técnicas utilizadas para o seu desenvolvimento.

Ao longo deste capítulo serão revistos inicialmente os conceitos básicos onde se fundamenta a análise dos sinais de fala e seguidamente apresentar-se-ão características dos sinais e respectivas técnicas que conduzem à respectiva obtenção. Cada uma destas características será enquadrada no trabalho desenvolvido na dissertação.

3.1 Conceitos Básicos

Para se proceder à análise de um sinal é necessário possuir previamente o objecto da análise. Num sistema real o sinal de fala é adquirido através de um microfone e convertido através de um processo de amostragem e quantificação numa sequência numérica que será a sua representação digital. A amostragem é feita com uma determinada frequência que normalmente varia entre os 8 KHz e os 44 KHz (qualidade de CD) e a quantificação utiliza 8 ou 16 bits. No desenvolvimento de sistemas em laboratório esta informação está em geral disponível em bases de dados. Na figura 3.1 apresenta-se a sequência com os passos típicos envolvidos na aquisição e análise de um sinal sendo o primeiro passo o que se acabou de descrever.

Seguidamente são realizadas eventuais filtrações quando o sinal estiver contaminado com algum tipo de ruído e efectua-se a pré-ênfase. Esta consiste em passar o sinal por um filtro de primeira ordem, passa-alto, nivelando a representação espectral do sinal [63]. A função de transferência utilizada é a de um

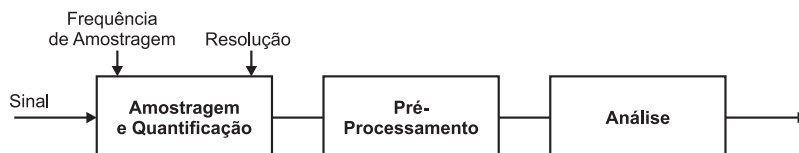


Figura 3.1: Sequência típica para a aquisição e análise de um sinal

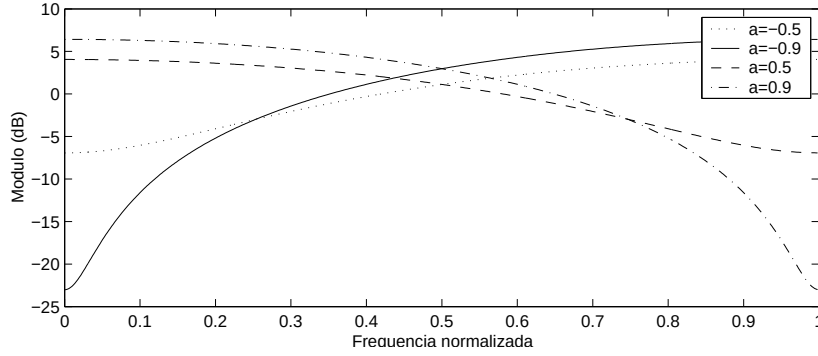


Figura 3.2: Resposta em frequência de um filtro FIR de primeira ordem para vários valores de α

filtro FIR de primeira-ordem:

$$y[n] = x[n] + \alpha x[n - 1] \quad (3.1)$$

com transformada Z:

$$H(z) = 1 + \alpha z^{-1} \quad (3.2)$$

Este filtro é não uniforme para qualquer valor de α com exceção de 0. A amplitude e fase da sua resposta em frequência são dadas por:

$$\begin{aligned} |H(e^{j\omega})| &= |1 + \alpha(\cos \omega - j \sin \omega)|^2 \\ &= (1 + \alpha \cos \omega)^2 + (\alpha \sin \omega)^2 \\ &= 1 + \alpha^2 + 2\alpha \cos \omega \end{aligned} \quad (3.3)$$

ou, em (dB),

$$10 \log |H(e^{j\omega})|^2 = 10 \log [(1 + \alpha)^2 + 2\alpha \cos \omega] \quad (3.4)$$

Para a fase tem-se:

$$\theta(e^{j\omega}) = -\arctan \left(\frac{\alpha \sin \omega}{1 + \alpha \cos \omega} \right) \quad (3.5)$$

A resposta em frequência para vários valores de α apresenta-se na figura 3.2. No caso do filtro de pré-ênfase escolhe-se $-0.9 < \alpha < -0.99$ o que levará a um filtro que irá colocar alguma *ênfase* nas frequências mais altas.

Os sinais de voz são altamente não lineares mas possuem uma variação bastante suave no tempo. Por esta razão, após a filtragem procede-se à partição das sequências numéricas representativas do sinal em blocos idênticos e com uma determinada duração. Esta operação é designada por janelamento e possibilita, para cada bloco, a criação de condições de quasi-estacionariedade que permitem desenvolver padrões e estabelecer modelos descritivos. A duração de cada bloco é normalmente múltipla da frequência fundamental não ultrapassando um factor de 4.

O tipo de janelamento que se sugeriu passa apenas pela simples truncatura da sequência numérica anulando todos os valores abaixo e acima dos limites estabelecidos. Esta janela, designada rectangular, conduz a elevadas fugas espectrais que deterioram a análise. É comum por esta razão utilizarem-se funções de janelamento alternativas que reduzam este efeito.

As funções de janelamento mais comuns apresentam-se abaixo para, salvo excepção, os intervalos $0 \leq n \leq M$ sendo 0 para os restantes valores:

- Rectangular

$$w[n] = 1 \quad (3.6)$$

- Hanning

$$w[n] = 0.5 - 0.5 \cos\left(\frac{2\pi n}{M}\right) \quad (3.7)$$

- Hamming

$$w[n] = 0.54 - 0.46 \cos\left(\frac{2\pi n}{M}\right) \quad (3.8)$$

- Bartlett ou Triangular

$$w[n] = \begin{cases} \frac{2n}{M}, & 0 \leq n \leq M/2 \\ \frac{2-2n}{M}, & M/2 \leq n \leq M \\ 0, & \text{outros valores} \end{cases} \quad (3.9)$$

Devido à atenuação nos extremos da janela causado pelo uso de funções de janelamento diferentes da rectangular é habitual que o passo de avanço na análise seja inferior à duração de uma janela produzindo uma sobreposição entre janelas consecutivas com a finalidade de utilizar convenientemente todas as regiões do sinal. É comum utilizar uma janela de Hamming com uma dimensão correspondente a dois períodos fundamentais e um passo de avanço (ou simplesmente avanço) de meia janela.

Terminadas as operações de pré-processamento, já com o sinal dividido em blocos e devidamente janelado, é possível começar a extrair informação do mesmo. A análise de um qualquer sinal pode ser realizada no domínio do tempo, directamente sobre a forma de representação tradicional, ou no domínio das frequências, depois de uma transformação frequencial. Menos populares mas com a mesma legitimidade de ferramentas de análise existem também a análise por ondulas (*wavelets*) [65] e a distribuição de Wigner [1], de aplicação crescente [5], sendo ambas realizadas num domínio simultâneo de tempo e frequência.

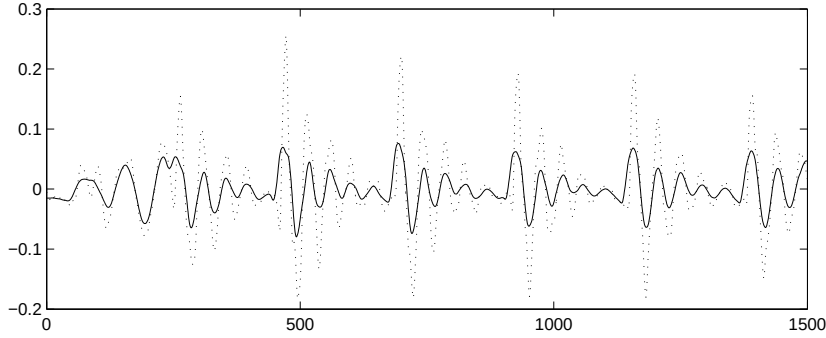


Figura 3.3: Amplitude média deslizante calculada utilizando uma janela com dimensão de 30 amostras.

3.2 Amplitude Média Deslizante

A amplitude média deslizante é calculada efectuando, bloco a bloco, a média das amplitudes de cada bloco, do início ao fim do sinal.

Para um único bloco a expressão é dada por:

$$A_{med}[n] = \frac{1}{M} \sum_{i=n-M+1}^n |x[i]| \quad (3.10)$$

Na totalidade do sinal e a partir da convolução com uma função de janelamento w :

$$A_{med}[n] = \frac{1}{M} \sum_{i=-\infty}^{\infty} |x[m]| \cdot w[n-i] \quad (3.11)$$

A amplitude média deslizante cujo único parâmetro de controlo é a dimensão M da janela tem um efeito de suavização do sinal atenuando as variações bruscas. Quando o comprimento da janela é cerca de um período fundamental os valores resultantes do cálculo da amplitude média deslizante terão correspondências com o próprio período fundamental. Quando o comprimento tem uma dimensão superior conseguem-se pistas importantes para a identificação de zonas de vozeamento e não vozeamento[81].

Na figura 3.3 mostra-se um sinal de fala, a pontilhado, e a respectiva amplitude média, a traço cheio, ficando claro o efeito suavizador desta operação. Os valores máximo e mínimo resultantes tendem a aproximar-se do valor médio do sinal em análise à medida que a dimensão da janela vai abrangendo um maior número de amostras.

3.3 Energia Média Deslizante

A energia de um sinal é obtida a partir do cálculo da variância da sequência numérica representativa desse mesmo sinal, ou seja, é dada pelo quadrado da diferença entre os valores da sequência numérica e a sua média. Para os sinais de fala que possuem geralmente valor médio nulo a energia será dada pela média

dos quadrados da cada valor. Considerando o sinal e efectuando os cálculos para uma janela w_n de dimensão M tem-se:

$$E(n) = \frac{1}{M} \sum_{i=-\infty}^{\infty} [x[n] \cdot w[n-i]]^2 \quad (3.12)$$

A energia média deslizando é utilizada como indicador sobre as zonas com vozeamento, não vozeamento e silêncio. Quando comparada com a amplitude média deslizando em relação a estes aspectos, a energia média deslizando fornece melhores resultados devido à maior selectividade de amplitudes da função quadrática utilizada. O cálculo num sistema computacional é também efectuado sobre um conjunto de janelas sucessivas e com alguma sobreposição.

3.4 Passagens por Zero e Classificação de Segmentos

As passagens por zero ou a taxa de passagens por zero são, como o nome deixa pressupor, indicadores da frequência com que a amplitude do sinal atravessa o eixo dos tempos e muda de sinal. As passagens por zero podem ser dadas pela expressão:

$$Z[n] = \frac{1}{2M} \sum_{i=-\infty}^{\infty} |\text{sign}(x[i]) - \text{sign}(x[i-1])| \cdot w[n-i] \quad (3.13)$$

na qual,

$$\text{sign}(a) = \begin{cases} 1, & a \geq 0 \\ 0, & a < 0 \end{cases} \quad (3.14)$$

Para a detecção de fala, Rowden [73] recomenda a utilização do cálculo da energia média deslizando acompanhado do cálculo das passagens por zero. Estes dois indicadores permitem, de modo simples e com razoável fiabilidade, distinguir zonas com vozeamento de zonas sem vozeamento. Na figura 3.4 apresenta-se um exemplo onde se representam as passagens por zero sobre o espectrograma do sinal de fala correspondente à palavra "pataca". Os pontos relativos às passagens por zero, um pouco menos perceptíveis, estão assinaladas na zona mais inferior da figura e são mais evidentes para as vogais.

Segundo Teixeira [81] os melhores resultados obtêm-se aplicando a taxa de passagens por zero à derivada discreta do sinal que se calcula pela expressão 3.15 e considera um conjunto de amostras igualmente espaçadas (como em princípio acontece):

$$d[n] = x[n+1] - x[n] \quad (3.15)$$

Ainda segundo o trabalho de Teixeira, onde se podem encontrar diversos ensaios com estes dois indicadores, os sinais vozeados são caracterizados por uma energia elevada e uma taxa de passagens por zero reduzida ao passo que nos sons não vozeados a energia é elevada e uma taxa de passagens por zero é também elevada. Na figura 3.5 apresenta-se um domínio de decisão resultante dos estudos do autor referido em que as escalas se baseiam nos valores de energia e taxa de passagens por zero. O seu cálculo obedece a um algoritmo não apresentado.

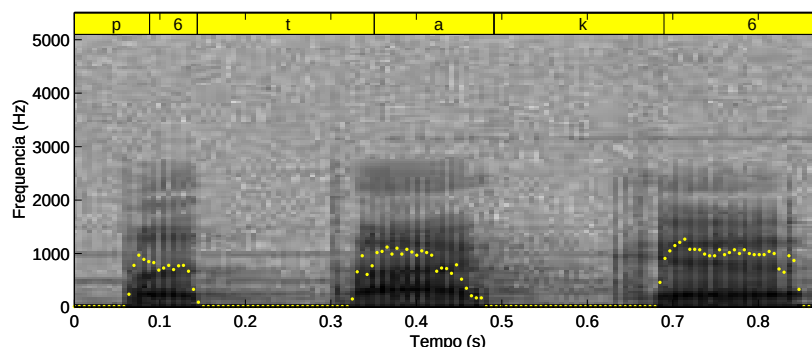


Figura 3.4: Taxa de passagens por zero representada sobre o espectrograma do sinal correspondente à palavra "pataca".

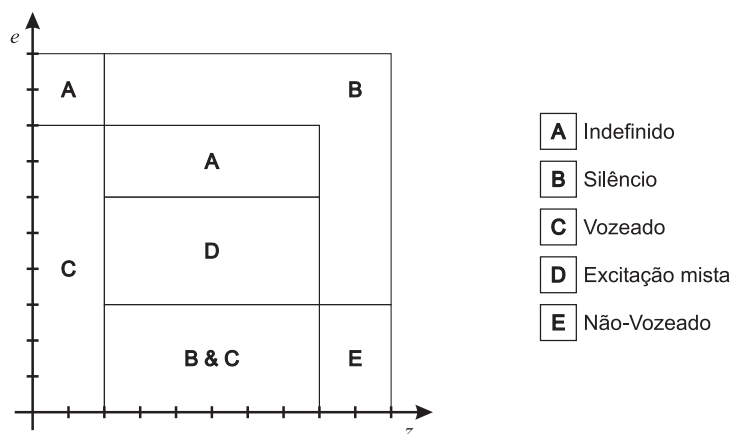


Figura 3.5: Domínio de decisão para a caracterização de sinais de voz tendo por base a energia média deslizante e a taxa de passagens por zero

3.5 Modelos de Predição Linear

A técnica da predição linear tem um conjunto de aplicações extremamente amplo sendo utilizado em Engenharia, Economia, Biologia, etc.. No caso do processamento de voz esta técnica é particularmente interessante pois permite modelizar de forma bastante satisfatória o comportamento do trato vocal e do sinal de voz em geral. Para além disto o esforço computacional para obter um modelo de boa qualidade é relativamente reduzido e facilmente implementável em computadores ou outros sistemas de computação digital. Por estas razões, a predição linear serve de base a muitas outras técnicas, mais recentes e elaboradas, para a análise de sinais de fala.

Inicialmente referido por Makhoul [53] e mais divulgado por Rabiner [68], o princípio básico da codificação por predição linear (*Linear Predictive Coding* ou LPC na literatura internacional na área do processamento de sinal) considera que uma qualquer amostra de um sinal pode ser determinada a partir de uma combinação linear de p amostras anteriores e q amostras de um sinal de excitação externo ao sistema.

Este princípio pode ser representado matematicamente por uma equação às diferenças:

$$s(n) = \sum_{k=1}^p a_k s(n-k) + G_0 \cdot \sum_{l=0}^q b_l u(n-l) \quad (3.16)$$

Na expressão 3.16, a_k e b_l são constantes e G_0 é um factor de ganho. Estes são os parâmetros a definir para a determinação do modelo de predição linear. O sinal é representado por $s(n)$ e a excitação do modelo por $u(n)$, sendo n um determinado instante de tempo discreto.

No domínio Z tem-se:

$$S(z) = \sum_{k=1}^p a_k S(z) z^{-k} + G_0 \cdot \sum_{l=0}^q b_l U(z) z^{-l} \quad (3.17)$$

Ou, numa representação mais tradicional para a análise de sistemas:

$$H(z) = \frac{S(z)}{U(z)} = G \frac{1 + \sum_{l=1}^q b_l z^{-l}}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (3.18)$$

com

$$G = G_0 \cdot b_0$$

Os coeficientes com índice zero, a_0 e b_0 , apresentam-se geralmente com valor unitário por uma questão de normalização.

Destas expressões podem distinguir-se três casos distintos. Um primeiro, constituído exclusivamente por pólos, conhecido na literatura internacional por *all-pole model* ou *auto-regressive (AR) model*, em que os coeficientes $b_l = 0$ com $l \neq 0$. Um outro, constituído apenas por zeros, *moving-average (MA) model*, em que os coeficientes $a_k = 0$ ($k \neq 0$), e por último, correspondente às expressões gerais, constituído por pólos e zeros, o modelo *auto-regressive moving average (ARMA)*.

Na figura 3.6 apresenta-se um modelo genérico para a produção de sinais de voz baseado nas técnicas de predição linear.

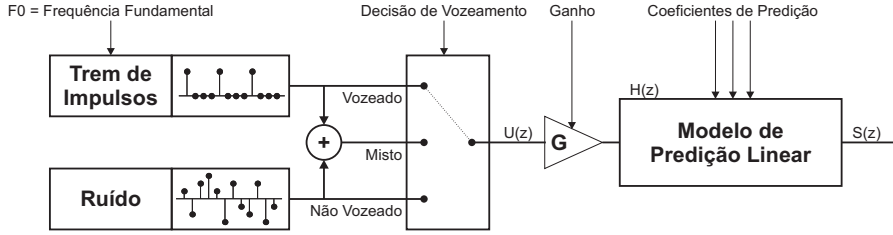


Figura 3.6: Modelo genérico para a produção de sinais de fala

As particularidades de cada um dos modelos referidos serão alvo de análise detalhada.

3.5.1 Modelo Auto-Regressivo

Considere-se então a equação 3.16 aplicada ao modelo auto-regressivo.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)} \quad (3.19)$$

Neste caso o modelo fica inteiramente determinado pelos coeficientes a_k e o ganho G . O numerador $A(z)$ é designado como filtro de predição inverso.

Os métodos que se utilizam tradicionalmente para a obtenção dos parâmetros dos modelos são os da auto-correlação e da covariância sendo ambos descritos em mais pormenor no anexo B.

Pelo método da autocorrelação efectua-se uma dedução onde, em determinado ponto efectuando uma substituição utilizando a função de auto-correlação que dá nome ao método, se obtém um conjunto de equações que se apresenta abaixo na forma matricial:

$$\begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \vdots & \vdots & & \vdots \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix} * \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \quad (3.20)$$

A matriz principal é uma matriz de auto-correlações. A resolução do sistema levará à obtenção dos coeficientes do modelo.

O erro de predição vem também como:

$$E_n = R_n(0) - \sum_{k=1}^p \alpha_k R_n(k) \quad (3.21)$$

O método da covariância baseia-se em princípios semelhante mas leva a uma outra matriz, desta vez de covariâncias:

$$\begin{bmatrix} C(1,1) & C(1,2) & \dots & C(1,p) \\ C(2,1) & C(2,2) & \dots & C(2,p) \\ \vdots & \vdots & & \vdots \\ C(p,1) & C(p,2) & \dots & C(p,p) \end{bmatrix} * \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} C(1,0) \\ C(2,0) \\ \vdots \\ C(p,0) \end{bmatrix} \quad (3.22)$$

Este método, devido a características particulares da matriz principal, utiliza algoritmos diferentes para a resolução do sistema. Os resultados obtidos podem diferir ligeiramente dos conseguidos pelo método da auto-correlação.

Outros Métodos

Os métodos existentes para o cálculo dos coeficientes LPC são vários e já se apresentaram os da autocorrelação e da covariância que constituem a abordagem clássica e mais popular. Existem ainda outros que procuram obter o resultado por outras vias ou de um modo mais eficiente. Relata-se então a existência do método *lattice* [68], o método da filtragem inversa, método da estimação espectral, método da máxima verosimilhança e o método do produto interno que não serão abordados neste documento.

Ganho do Modelo AR

Resta ainda calcular o ganho para determinar completamente o modelo de predição que tem vindo a ser desenvolvido. Este cálculo baseia-se quase sempre na igualdade de energias do sinal original e do sinal produzido pelo modelo de predição. Assim, utilizando esta ideia e partindo das equações anteriores deduz-se facilmente:

$$G^2 = R_n(0) - \sum_{k=1}^p \alpha_k R_n(k) = E_n \quad (3.23)$$

Desempenho do Modelo AR

Depois de apresentados os detalhes matemáticos para o cálculo dos parâmetros dos modelo de predição baseados em pólos apresentam-se alguns resultados demonstrativos do seu desempenho.

Uma das características que é alvo de particularização no modelo LPC é o facto de o sinal de voz a estimar ser vozeado ou não vozeado. Na figura 3.7 apresenta-se um sinal de voz e o resíduo resultante de um modelo de predição baseado em 12 pólos analisado com janelas de *Hamming* com duração de 20ms e um *overlap* de meia-janela.

O resíduo de predição corresponde à função de excitação ideal, ou seja, utilizando o resíduo como sinal de entrada, o sinal produzido pelo modelo será exactamente igual ao sinal original. Esta observação fica bastante clara comparando as equações B.3 e B.5.

As zonas vozeadas, correspondentes aos fones [a] e [6], estão bem evidenciadas no sinal residual, com maior energia, e os impulsos correspondentes aos pulsos glotais, marcados pelos picos mais elevados, bem salientados. Aqui fica também patente a necessidade de, sempre que possível, utilizar sinais de excitação distintos de acordo com a existência ou não de vozeamento.

Analise-se agora o efeito da variação do número de coeficientes de predição. Na figura 3.8 apresenta-se o erro médio de predição para um modelo de um sinal vozeado, desenvolvido para uma vogal [a], excitado por impulsos.

O decréscimo mais acentuado que se pode observar na figura termina com a utilização de 13 coeficientes sendo as reduções no erro, depois de um ligeiro

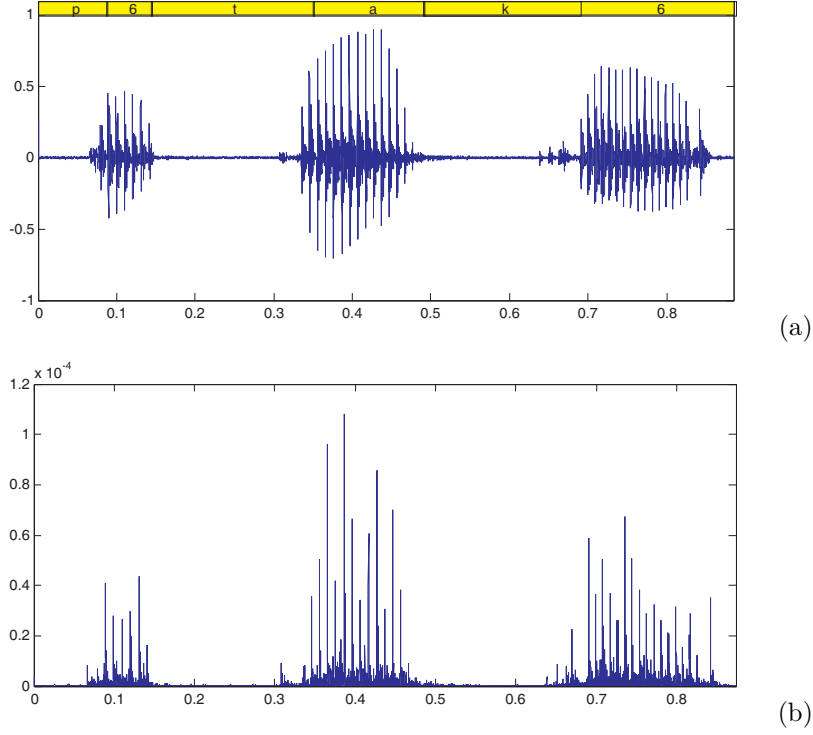


Figura 3.7: Sinal de voz foneticamente anotado correspondente à palavra "pataca" e quadrado do resíduo de predição respectivo para $p = 12$. Tempo em segundos.

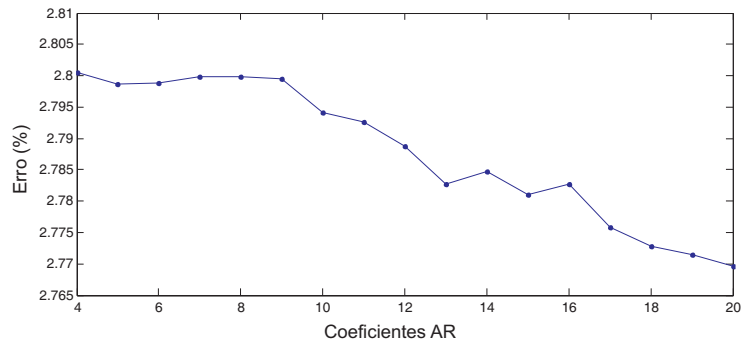


Figura 3.8: Exemplo de variação do erro médio de predição em função do número de coeficientes AR

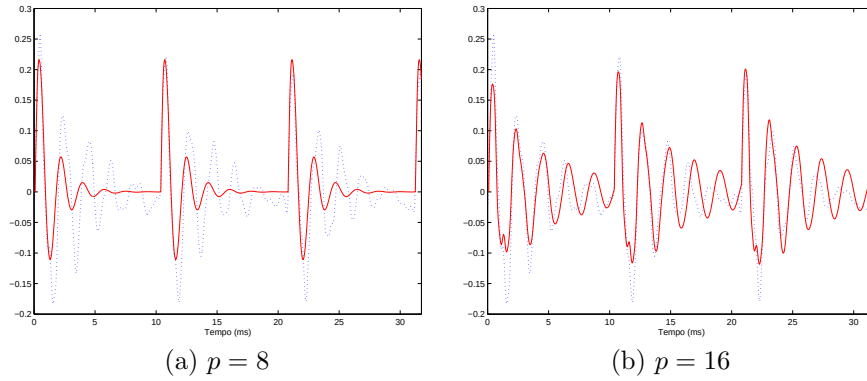


Figura 3.9: Variação do número de coeficientes AR (domínio do tempo)

aumento, mais ligeiras. Este resultado vem de encontro aos estudos de Rabiner [68] que afirma que um modelo AR baseado em 12 pólos possui a melhor relação desempenho/esforço computacional e oferece um bom comportamento na maioria das situações.

Na curva da figura 3.8 surgem por vezes algumas evoluções positivas que contrariam a tendência descendente global. Isto deve-se ao facto de a adição de um pólo extra nem sempre significar um modelo com melhor desempenho. Por outro lado, a adição de um par de pólos, permite obter quase sempre melhores resultados.

Portanto, a tendência descendente da curva deve ser generalizada com cuidado pois diz apenas respeito a uma ocorrência do fone utilizado e que está relacionada com a capacidade de ajuste do modelo AR. Porém, é esperado que nas condições apresentadas, qualquer sinal vozeado com boa qualidade, devido ao seu comportamento mais cíclico, seja eficazmente descrito por um modelo AR.

A variação do número de coeficientes AR traduz-se numa variação do grau do denominador na equação do filtro de predição. O grau deste polinómio dá uma indicação do número de oscilações que o seu gráfico poderá ter. Para clarificar o relacionamento dos conceitos matemáticos com uma aplicação real apresentam-se na figura 3.9 as diferenças que surgem, no domínio do tempo, na predição de um sinal utilizando um modelo de 8 e 16 coeficientes. Escolheram-se aproximadamente três ciclos do primeiro fone [6] da palavra "pataca" já utilizada noutros exemplos. A traço cheio surge o sinal produzido pelo modelo e a tracejado o sinal original.

Como seria esperado a saída do modelo aproxima-se do sinal original quando o número de coeficientes de estimação aumenta. Um polinómio de grau superior possui uma maior capacidade descritiva sendo deste modo o modelo respectivo mais exacto.

A partir da comparação das figuras é possível observar que as maiores oscilações são correctamente modelizadas mesmo utilizando apenas 8 coeficientes e que a utilização de 16 coeficientes apenas contribui para melhorar o ajuste em zonas com oscilações de menor amplitude. Fica assim evidente que o aumento do grau do modelo, reflectindo-se apenas em zonas do sinal com menor energia,

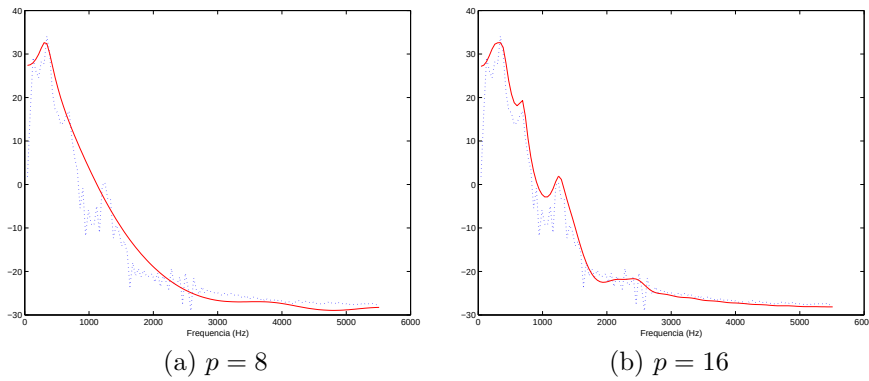


Figura 3.10: Variação do número de coeficientes AR (domínio das frequências)

terá, por consequência, uma menor importância.

No domínio das frequências, rico em outras informações, pode ser feita uma análise semelhante. Na figura 3.10 tem-se a cheio o periodograma do sinal de saída do modelo e a tracejado o periodograma correspondente ao sinal original.

Os picos do periodograma a cheio correspondem às várias frequências formantes ou de ressonância que caracterizam este sinal de fala. Neste caso, o modelo mais simples possui claramente um periodograma mais suave perdendo-se alguma informação frequencial. Numa utilização real, esta suavização frequencial é notória mas geralmente e para sinais de fala não resulta numa diminuição grave da inteligibilidade da mensagem. Importante ainda acrescentar que muitas vezes, por este processo, se utiliza a técnica da análise LPC para a obtenção de uma imagem do envelope espectral de um sinal.

Em relação ao tipo de excitação, e recordando o modelo da figura 3.6, deve-se utilizar um trem de impulsos síncrono com a frequência fundamental para os sinais vozeados e ruído branco para os sinais não-vozeados. Em casos particulares, nomeadamente nos casos das consoantes sonoras, pode ser feita a combinação de ambos obtendo-se um sinal de excitação misto.

Simplificando o modelo e ignorando as recomendações relativas ao sinal de entrada apresentam-se, na figura 3.11, dois sinais produzidos por um modelo AR de 12 coeficientes para a palavra "pataca" com uma excitação constituída apenas por um trem de impulsos ou apenas por ruído.

Na figura 3.11, para o modelo excitado por impulsos, estes ficam claramente visíveis ao passo que no resultado baseado numa excitação por ruído o sinal apresenta-se mais compacto.

Do ponto de vista da voz tem-se, para uma excitação completamente baseada em ruído, uma voz sussurrante. Por outro lado, utilizando apenas um trem de impulsos, obtém-se uma voz um pouco mais limpa, mas bastante metálica. Neste último caso deve ser ressaltada a importância do sincronismo dos impulsos com o trem de marcas relativas à frequência fundamental de forma a simular os pulsos produzidos pela glote.

A ordem do modelo a utilizar em cada situação é ajustada pela frequência de amostragem do sinal e é independente do método de cálculo escolhido. Geralmente os sinais de fala, devido às características do tracto vocal, possuem 2

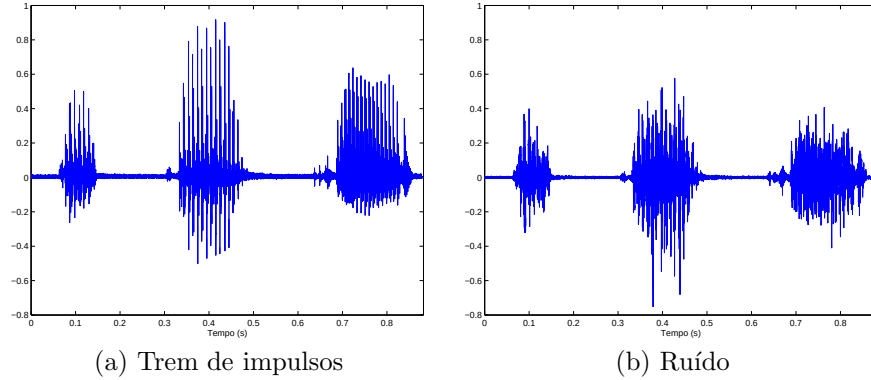


Figura 3.11: Variação do sinal de excitação dos modelos AR de 12 coeficientes para a palavra "pataca"

pólos por Khz, ou seja, 1 pólo complexo conjugado por Khz. Tanto a influência do sinal de excitação e como a da radiação podem ser modelizadas por 3 a 4 pólos. Assim, para um valor habitual de 22Khz para a frequência de amostragem, seriam aconselháveis para a construção de um modelo de desempenho satisfatório, 22/2+4 pólos.

Até agora, analisou-se uma técnica que permite desenvolver um modelo bastante fiel do comportamento do tracto vocal e que se baseia apenas em pólos. Porém, alguns fones, nomeadamente os nasalados e alguns não vozeados, possuem características espectrais que se associam mais facilmente a zeros e não são tão bem modelizados pela técnica apresentada.

3.5.2 Modelo de Média Movente

Outra abordagem menos frequente nos sistemas de processamento de fala é a que se baseia num sistema baseado apenas em zeros. Neste caso, recordando a equação 3.16, os coeficientes a_k têm valor nulo e o processo de média móvel ou *moving average (MA)* é descrito pela equação:

$$s(n) = G_0 \cdot \sum_{l=0}^q b_l u(n-l) \quad (3.24)$$

ou, no domínio Z ,

$$S(z) = G_0 \cdot \sum_{l=0}^q b_l U(z) z^{-l} \quad (3.25)$$

Neste caso não são consideradas amostras anteriores do sinal de saída. Toda a predição é baseada em amostras do sinal de excitação que deste modo terá um papel preponderante.

Mais uma vez, e à semelhança dos modelos AR, pretende-se calcular os coeficientes b_l que minimizem o erro de predição. A saída do modelo será $\hat{s}(n)$:

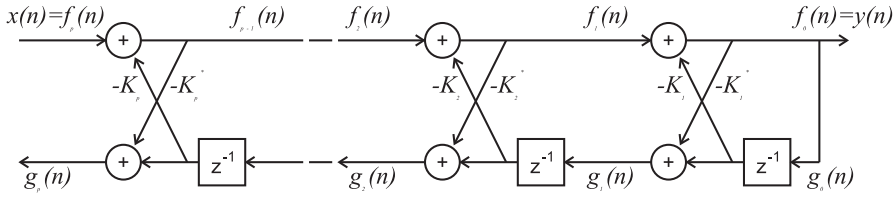


Figura 3.12: Estrutura *lattice* para um sistema AR(p) a transformar

$$\tilde{s}(n) = G_0 \cdot \sum_{l=0}^q \beta_l u(n-l) \quad (3.26)$$

O erro de predição $e(n)$ respectivo é a diferença entre ambos os sinais.

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{l=0}^q \beta_l u(n-l) \quad (3.27)$$

Optando pelo erro médio quadrático temos um problema idêntico ao dos modelos AR e que pode ser resolvido por técnicas semelhantes.

Considere-se o sistema baseado apenas em pólos com função de transferência:

$$H(z) = \frac{1}{1 + \sum_{k=1}^p a(k)z^{-k}} \quad (3.28)$$

Esta equação é idêntica à de um filtro IIR a que corresponde a equação às diferenças:

$$y(n) = - \sum_{k=1}^p a(k)y(n-k) + x(n) \quad (3.29)$$

Trocando $x(n)$ com $y(n)$, ou seja, invertendo os papéis da entrada e da saída, tem-se a equação às diferenças:

$$x(n) = - \sum_{k=1}^p a(k)x(n-k) + y(n) \quad (3.30)$$

ou, equivalentemente,

$$y(n) = x(n) + \sum_{k=1}^p a(k)x(n-k) \quad (3.31)$$

Observando a equação 3.31 surgem rapidamente as semelhanças com um sistema FIR com função $A(z)$. Então um filtro IIR baseado em pólos pode ser convertido num sistema baseado em zeros através da troca dos papéis da entrada com a saída. Ficam assim mais evidentes as semelhanças entre os processos AR e MA.

A título de exemplo, na figura 3.12 apresenta-se um modelo de um filtro AR versão *lattice* em que a entrada é $x(n) = f_p(n)$ e a saída é $y(n) = f_0(n)$. Então, para obter a versão MA basta efectuar a transformação:

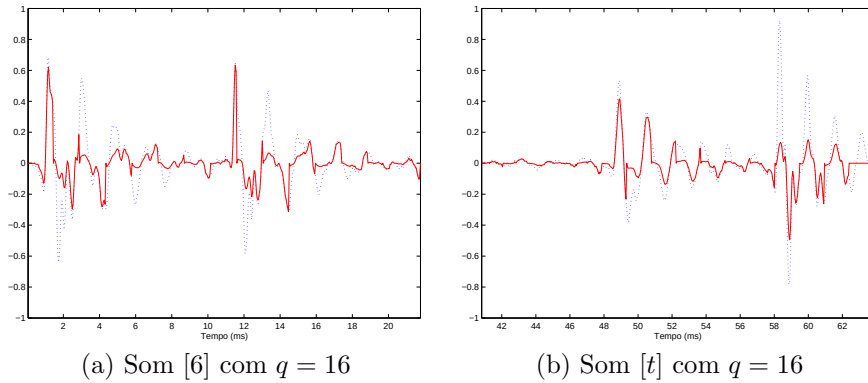


Figura 3.13: Comportamento do modelo MA com excitação por impulsos

$$\begin{aligned} x(n) &= f_0(n) \\ y(n) &= f_p(n) \end{aligned} \quad (3.32)$$

Os coeficientes de reflexão K_i são idênticos em ambas as situações.

Importa ainda afirmar que a predição baseada em zeros obriga ao conhecimento do sinal de excitação o que quase sempre é impossível. No entanto, devido às características intrínsecas do modelo, é fundamental a utilização do sinal adequado no momento adequado.

Desempenho do Modelo MA

Na figura 3.13 apresentam-se parte dos comportamentos de dois modelos MA desenvolvidos para o fone [6] e para o fone [t], baseados em 16 coeficientes MA, com excitação por impulsos. Para o fone [6] mostram-se aproximadamente 2 ciclos fundamentais, para o fone [t] mostra-se a zona transitória da oclusão para a explosão. O sinal original surge a pontilhado enquanto a traço cheio se tem o sinal produzido pelo modelo.

No resultado obtido para o fone [6] é nítido o inferior desempenho quando comparado com o dos modelos AR. O comportamento aproximadamente cíclico de [6], que possui alguma previsibilidade, não é aproveitado pelo modelo, e mesmo as energias de ambos os sinais apresentam grandes discrepâncias.

Por outro lado, para o fone não-vozeado [t], observa-se uma maior similaridade com o sinal original. As sequências com súbitas inversões de sentido, ciclos pouco evidentes e múltiplas frequências envolvidas são modelizadas com boa qualidade e as energias de ambos os sinais aproximam-se. Neste tipo de ocorrências, com sinais de comportamento mais aleatório e menos periódico, os modelos baseados em zeros proporcionam melhor desempenho comparativamente com os modelos baseados em pólos.

Optando por uma excitação do modelo com ruído os resultados não são tão animadores. Na figura 3.14 apresentam-se novamente parte dos comportamentos de dois modelos MA desenvolvidos para os fone [6] e [t], também baseados em 16 coeficientes MA, mas com excitação por ruído. As restantes condições são idênticas às da figura 3.13.

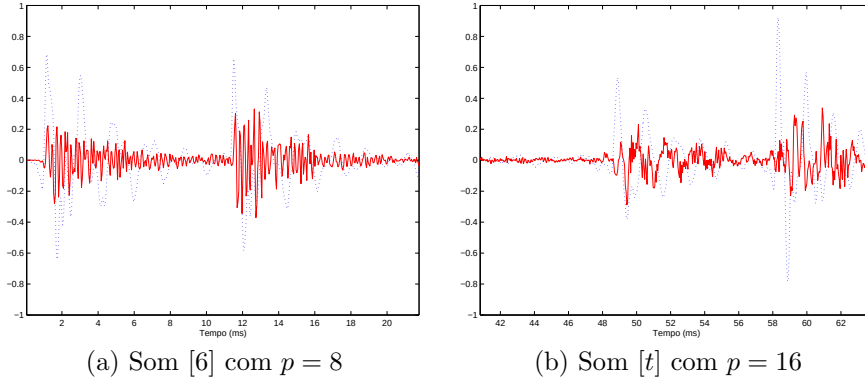


Figura 3.14: Comportamento do modelo MA com excitação por ruído

Mais um vez, o comportamento do modelo no caso não-vozeado é bastante melhor, com uma boa similaridade entre sinal original e o sintético. Para o fone [6], cujo ensaio se apresenta como curiosidade e que nas condições em que foi efectuado desrespeita os pressupostos da predição linear, tem-se uma grande quantidade de ruído extra que reduz a qualidade e clareza da onda. Do ponto de vista auditivo este efeito não passa despercebido e diminui a inteligibilidade.

3.5.3 Modelo Auto-Regressivo e de Média Movente

O último caso a analisar no estudo dos modelos de predição linear utiliza simultaneamente pólos e zeros para a modelização do tracto vocal. Assim, e recordando a equação geral 3.16, tem-se para os modelos ARMA (*Auto-Regressive Moving Average*), $p > 0$ e $q > 0$.

$$s(n) = \sum_{k=1}^p a_k s(n-k) + G_0 \cdot \sum_{l=0}^q b_l u(n-l)$$

No domínio Z e incluindo o factor de ganho G_0 nos parâmetros b_l tem-se:

$$H(z) = \frac{S(z)}{U(z)} = \frac{\sum_{l=0}^q b_l z^{-l}}{1 - \sum_{k=1}^p a_k z^{-k}} \quad (3.33)$$

O modelo fica totalmente determinado pelos parâmetros a_k e b_l que são no total $p+q+1$. No cálculo destes coeficientes, a utilização do método dos mínimos quadrados, resulta num sistemas de equações não lineares de difícil resolução. Os métodos existentes para a resolução do problema, levam a soluções sub-óptimas, mas que ainda assim resultam em boas aproximações.

Método de Prony

O algoritmo de *Prony* [61], que se passa a descrever, é uma opção popular que produz resultados satisfatórios quando o sinal não possui perturbações de ruído acentuadas.

Para a determinação dos parâmetros do modelo, este método, pressupõe uma excitação $u(n) = \delta(n)$. Assim, a resposta do sistema será:

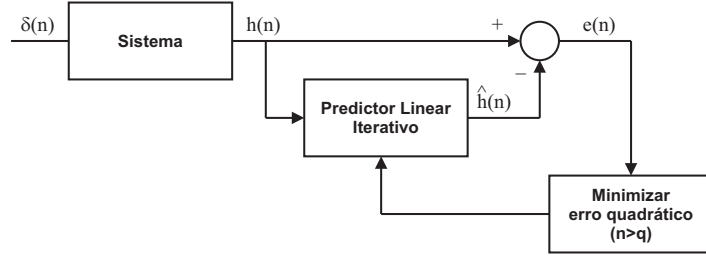


Figura 3.15: Diagrama para determinação dos parâmetros AR num modelo ARMA

$$s(n) = h(n) = \sum_{k=1}^p a_k h(n-k) + \sum_{l=0}^q b_l \delta(n-l), \quad n \geq 0 \quad (3.34)$$

Uma vez que a função $\delta(n-k)$ é sempre nula excepto em $n=k$, pode-se escrever a equação 3.34 num outro formato.

$$h(n) = \sum_{k=1}^p a_k h(n-k) + b_n, \quad 0 \leq n \leq q \quad (3.35)$$

$$h(n) = \sum_{k=1}^p a_k h(n-k), \quad n > q \quad (3.36)$$

A partir da equação 3.36 podem ser calculados os parâmetros a_k que minimizam o erro $e(n)$. Este, nas mesmas condições é dado por:

$$e(n) = h(n) - \tilde{h}(n) = h(n) - \sum_{k=1}^p a_k h(n-k), \quad (3.37)$$

Aplicando o tradicional erro quadrático, para um conjunto de amostras $N \gg p$, tem-se a expressão:

$$E = \sum_{n=q+1}^N e^2(n) = \sum_{n=q+1}^N \left[h(n) - \sum_{k=1}^p a_k h(n-k) \right]^2 \quad (3.38)$$

Na figura 3.15 apresenta-se um esquema representativo do processo até aqui desenvolvido.

Aplicando agora um raciocínio análogo ao exposto para os processos AR, calculam-se as derivadas parciais da equação 3.38, igualam-se a zero os resultados e, partindo das semelhanças com a função de correlação, tem-se:

$$\sum_{k=1}^p a_k \Phi(i, k) = \Phi(i, 0), \quad i = 1, 2, \dots, p \quad (3.39)$$

O cálculo dos coeficientes a_k torna-se agora imediato. Substituindo estes valores em 3.35 chega-se rapidamente a um valor para b_n .

$$b_n = h(n) - \sum_{k=1}^p a_k h(n-k), \quad 0 \leq n \leq q \quad (3.40)$$

Método de Steiglitz-McBride

A alternativa ao método exposto proposta por Steiglitz e McBride [77] leva geralmente a melhores resultados tendo porém uma maior tendência para criar um modelo instável quando a sua ordem é elevada. Este é um método iterativo rápido que calcula simultaneamente os coeficientes do numerador e denominador da função de transferência do modelo através da minimização do erro entre a sua saída e a saída dada [52]. A convergência do método não é garantida mas surge geralmente ao fim de poucas iterações.

O filtro IIR ou sistema a modelizar baseia-se na expressão geral dos modelos ARMA com p pólos e q zeros que, sendo fiel à apresentação do autor mas adaptando à notação que tem sido utilizada, é dada pela expressão:

$$H(z) = \frac{b_0 + b_1 z^{-1} + \dots + b_q z^{-q}}{a_0 + a_1 z^{-1} + \dots + a_p z^{-p}} = \frac{B(z)}{A(z)}$$

O cálculo dos parâmetros do modelo a_i e b_j é feito a partir da sua resposta ao impulso $h(z)$ tentando minimizar o erro quadrático com o sinal $x(z)$.

$$\min_{a,b} \sum_{i=0}^{\infty} |x(i) - h(i)|^2 \quad (3.41)$$

Inicialmente é gerada uma primeira aproximação para os coeficientes a_i que pode ser dada por qualquer um dos métodos estudados para os modelos AR ou pelo método de Prony considerando nula a ordem do numerador. Tendo os primeiros coeficientes é feita uma filtragem de h e x utilizando $1/a(z)$ que resulta num sistema de equações lineares que permite calcular os coeficientes b_i . O processo é repetido quantas iterações forem desejadas ou até se atingir o limiar de erro exigido. A convergência do algoritmo não é garantida e quando é conseguida nem sempre o modelo é estável. Apesar destes problemas conhecidos o algoritmo de Steiglitz-McBride conduz geralmente a modelos estáveis e bastante fiéis sendo por isso uma boa opção.

Para além dos algoritmos apresentados existem ainda a aproximação de Padé, o método de Shank e outros. Quase todos são baseados em técnicas recursivas e computacionalmente bastante exigentes. Um estudo aprofundado de alguns destes algoritmos encontra-se nas referências [15], [61] e [64].

Desempenho do Modelo ARMA

Apresentado o modelo e as algumas técnicas de cálculo dos parâmetros respectivos, analise-se agora a capacidade descritiva proporcionada. O modelo AR, que pode ser deduzido a partir do modelo de tubos sem perdas já apresentado, pressupõe no seu desenvolvimento a inexistência da cavidade nasal. No entanto, na produção de sons nasais, surgem fenómenos de reflexão que levam ao aparecimento de anti-ressonâncias correspondentes a zeros. Esta situação ocorre quando, para determinadas frequências, a onda reflectida na oclusão oral cancela a onda produzida na faringe.

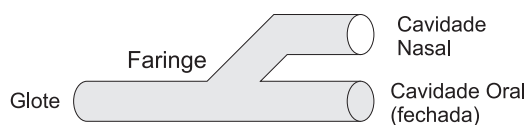


Figura 3.16: Modelo do término do trato vocal com a cavidade oral fechada

MA	AR	Erro Médio	Erro Máximo
16	4	0.53%	3.07%
4	16	1.51%	12.13%
16	16	0.19%	18.60%

Tabela 3.1: Variação do número de coeficientes de um modelo ARMA com excitação por impulsos

Nas tabelas 3.1 apresentam-se, para várias realizações do fone $[m]$, o valor do erro médio e erro máximo, em percentagem.

O erro é claramente inferior quando se opta por um modelo com maior número de zeros e menor número de pólos.

Para além dos sons nasais, que teoricamente são os mais adaptados e compatíveis, os modelos ARMA oferecem boas possibilidades em outras situações mais genéricas. Na figura 3.17 apresenta-se o erro médio de predição para vários modelos de um sinal vozeado $[a]$ e um sinal não vozeado $[s]$. O eixo horizontal apresenta uma escala crescente para o número de coeficientes AR. Cada curva representa a evolução do erro médio com o número de coeficientes AR para um número de coeficientes MA constante.

No caso do sinal vozeado, o aumento do número de coeficientes AR traduz-se numa clara melhoria dos resultados, bem patente nas curvas descendentes. O aumento dos coeficientes MA oferece benefícios modestos na qualidade do modelo. Por outro lado, para o sinal não vozeado, observa-se o comportamento inverso. Ou seja, a melhoria introduzida pelo aumento do número de pólos é quase imperceptível nos resultados finais, porém, à medida que o número de zeros é incrementado tem-se uma redução do erro acentuada e consistente, indicadora da compatibilidade do modelo.

Os resultados apresentados estão de acordo com o esperado depois do estudo independente realizado para os modelos AR e para os modelos MA.

MA	AR	Erro Médio	Erro Máximo
16	4	2.16%	31.14%
4	16	3.47%	54.51%
16	16	2.41%	18.60%

Tabela 3.2: Variação do número de coeficientes de um modelo ARMA com excitação por ruído

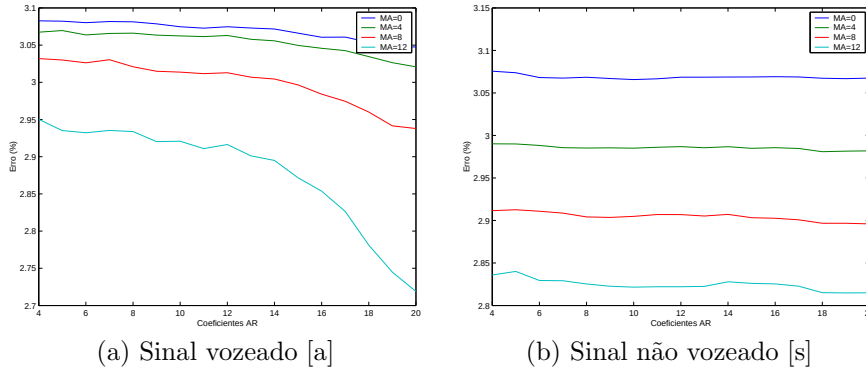


Figura 3.17: Erro médio de predição em função dos valores de p e q

3.5.4 Considerações

Foram assim apresentadas as várias possibilidades oferecidas pela teoria da predição linear e desenvolveu-se pormenorizadamente cada um dos modelos. Em cada caso foram analisados os efeitos da variação dos parâmetros envolvidos sempre acompanhados de representações gráficas dos resultados.

A predição linear é extensivamente utilizada pela sua qualidade global e eficiência. Para além das possibilidades referidas esta técnica permite também desenvolvimentos para o cálculo de formantes, estimação do espectro, entre outros.

Assim, e em tom de resumo, os modelos AR possuem uma boa capacidade descritiva para os sinais vozeados enquanto os modelos MA são mais adaptáveis a sinais não-vozeados. Quando existir capacidade computacional disponível pode-se optar por um modelo ARMA que combine as potencialidades de ambos. O número de coeficientes a escolher dará uma maior ou menor capacidade de descrição ao modelo, sendo função da frequência de amostragem como já se referiu. É frequente utilizarem-se entre 12 e 16 pólos e 1 a 2 zeros. Estas opções devem ser ponderadas caso a caso tendo em conta a relação capacidade descritiva/esforço computacional.

Normalmente a análise LPC é realizada após a pré-ênfase do sinal, que visa a elevação das amplitudes das frequências mais altas na banda da voz, e cada bloco de amostras é janelado com uma função de atenuação nas extremidades, muitas vezes uma janela de Hamming, que se impõe para não degradar os resultados dado que reduz as fugas espectrais (*spectral leakage*).

Como inconveniente apenas é apontada a grande sensibilidade dos coeficientes que, ao serem alvo de alguma perturbação, podem destruir por completo a validade do modelo calculado. Os pólos e zeros são uma representação numericamente menos sensível, excepto quando os pólos se aproximam do círculo unitário.

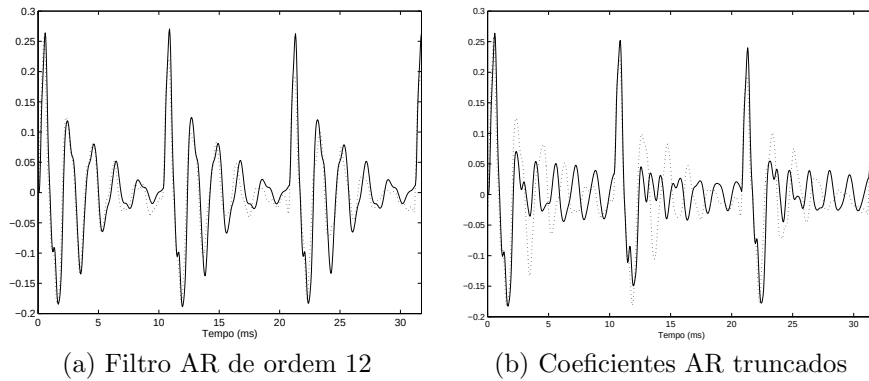


Figura 3.18: Resultado da degradação dos coeficientes AR

3.6 Line Spectral Frequencies

A técnica LPC, de acordo com as explicações já apresentadas, foi inicialmente aplicada na transmissão de sinais de voz em linhas digitais de baixo débito. O sinal é janelado, extraem-se os parâmetros do modelo LPC, enviam-se pela linha e procede-se no receptor à operação inversa, recuperando o sinal inicial com perdas no geral não significativas. No entanto, a transmissão directa dos parâmetros do modelo LPC, valores reais de elevada precisão, apresenta alguns problemas devido à necessidade de quantização e/ou ao ruído na comunicação.

Na figura 3.18 mostram-se as representações resultantes da modelização de um sinal por um filtro AR de 12 coeficientes. O sinal original a ponteadado, a cheio à esquerda o sinal produzido pelo modelo de ordem 12 e a cheio à direita o sinal produzido pelo mesmo modelo mas com os coeficientes de predição truncados na 7ª casa decimal. Este último resultado é bastante distante do original e a degradação de qualidade substancial. No sinal utilizado qualquer truncatura numa casa decimal inferior, 6ª por exemplo, conduz mesmo a um filtro instável.

Pelas razões apresentadas, não compatíveis com os níveis de qualidade desejados, desenvolveram-se representações alternativas como os coeficientes de reflexão, as razões logarítmicas de área (*log area ratios*) e os coeficientes LSF (*Line Spectral Frequencies*) ou LSP (*Line Spectral Pairs*). Inicialmente introduzidos por estudos de Itakura em 1975 [40] e popularizados no âmbito da compressão de sinal posteriormente por Soong e Juang [76] que apresentaram um trabalho de referência, os coeficientes LSF tornaram-se bastante populares e pelas características que se irão apresentar possuem um interesse relevante para este estudo.

3.6.1 Cálculo

De acordo com a definição de Itakura, os coeficientes LSF correspondem aos ângulos das raízes complexas de dois polinómios $P(z)$ e $Q(z)$ derivados de $A(z)$, filtro de predição inverso.

A expressão do filtro de predição inverso de ordem p , já apresentado durante a dedução do modelo AR, é:

$$A(z) = 1 - \sum_{k=1}^p a_k z^{-k} \quad (3.42)$$

Os coeficientes a_k caracterizam o modelo de predição linear.

Os polinômios $P(z)$ e $Q(z)$, de ordem $p + 1$, correspondentes, respectivamente, aos filtros simétrico e antisimétrico, podem ser obtidos por adição e subtração entre $A(z)$ e o seu conjugado (série invertida no tempo) [42].

$$P(z) = A(z) + z^{-(p+1)} A(z^{-1}) \quad (3.43)$$

$$Q(z) = A(z) - z^{-(p+1)} A(z^{-1}) \quad (3.44)$$

As raízes dos polinômios permitem determinar os coeficientes LSF. Estes polinômios são também representativos de uma estrutura recursiva cruzada de um filtro IIR de predição com grau $p + 1$. Ao filtro de ordem p é acrescentado em cascata um novo nível com coeficientes de reflexao $+1$ e -1 para se obter o comportamento dos polinômios $P(z)$ e $Q(z)$. Destes polinômios, o primeiro corresponde ao comportamento do trato vocal com a fonte glotal completamente fechada, ou seja, com coeficiente de reflexão $k_{p+1} = 1$, enquanto ao segundo, corresponde uma representação do tracto vocal com a fonte glotal completamente fechada, com coeficiente de reflexão $k_{p+1} = -1$.

A soma de $P(z)$ e $Q(z)$ multiplicada por 0.5 permite a recuperação do polinômio inicial.

$$A(z) = \frac{P(z) + Q(z)}{2} \quad (3.45)$$

Para explicitar algumas das características destes polinômios procedeu-se, para um caso genérico, à sua expansão:

$$\begin{aligned} P(z) = & 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_p z^{-p} - \dots \\ & - a_1 z^{-p} - a_2 z^{1-p} - \dots - a_p z^{-1} + z^{-(p+1)} \end{aligned} \quad (3.46)$$

$$\begin{aligned} P(z) = & 1 - (a_1 + a_p) z^{-1} - (a_2 + a_{p-1}) z^{-2} - \dots \\ & - (a_1 + a_p) z^{-p} + z^{-(p+1)} \end{aligned} \quad (3.47)$$

e

$$\begin{aligned} Q(z) = & 1 - a_1 z^{-1} - a_2 z^{-2} - \dots - a_p z^{-p} + \dots \\ & + a_1 z^{-p} + a_2 z^{1-p} + \dots + a_p z^{-1} - z^{-(p+1)} \end{aligned} \quad (3.48)$$

$$\begin{aligned} Q(z) = & 1 - (a_1 - a_p) z^{-1} - (a_2 - a_{p-1}) z^{-2} - \dots \\ & - (a_1 - a_p) z^{-p} - z^{-(p+1)} \end{aligned} \quad (3.49)$$

Desta formulação observa-se que o coeficiente do termo da maior ordem é idêntico ao de menor ordem, acontecendo, de acordo com a mesma lógica, o mesmo para os restantes. Desta forma, numa perspectiva mais prática e de implementação, apenas será necessário o cálculo de metade dos coeficientes.

Soong e Juang [76] mostraram que se $A(z)$ é um filtro estável então as raízes de $P(z)$ e $Q(z)$ estão no círculo unitário e são distintas. Foi também demonstrado que -1 e $+1$ são raízes respectivas.

Efectuando a divisão dos polinómios pelas respectivas raízes é possível simplificar a análise mas leva a obrigatoriedade de particularizar para os casos de p par e p ímpar o que é facilmente observável nas equações 3.47 e 3.49.

Tem-se assim, para p par:

$$\tilde{P} = \frac{P(z)}{1 - z^{-1}} \quad (3.50)$$

$$\tilde{Q} = \frac{Q(z)}{1 - z^{-1}} \quad (3.51)$$

E, para p ímpar:

$$\tilde{P} = P(z) \quad (3.52)$$

$$\tilde{Q} = \frac{Q(z)}{1 - z^{-1}} \quad (3.53)$$

Sejam R_P e R_Q o número de pares de zeros conjugados. Então, para clarificar a contribuição de zeros de cada polinómio tem-se, para p par:

$$\begin{cases} R_P = p/2 \\ R_Q = p/2 \end{cases} \quad (3.54)$$

E, para p ímpar:

$$\begin{cases} R_P = \frac{p+1}{2} \\ R_Q = \frac{p-1}{2} \end{cases} \quad (3.55)$$

Para encontrar as raízes desejadas existem múltiplos métodos. Soong e Juang procuram mudanças de sinal numa grelha de intervalos bem definidos depois de uma transformada de cosenos. Outros autores propõem a manipulação das equações para:

$$\begin{cases} P(z) = A(z) \left[1 + z^{-(p+1)} \frac{A(z^{-1})}{A(z)} \right] \\ Q(z) = A(z) \left[1 - z^{-(p+1)} \frac{A(z^{-1})}{A(z)} \right] \end{cases} \quad (3.56)$$

e definindo o filtro passa-tudo $H(z)$ com ganho unitário como:

$$H(z) = z^{-(p+1)} \frac{A(z^{-1})}{A(z)} \quad (3.57)$$

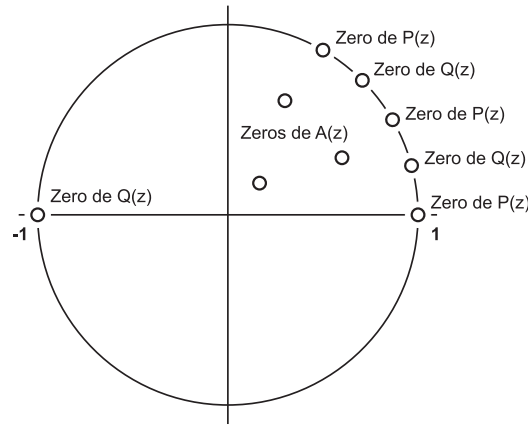


Figura 3.19: Localização dos zeros dos polinômios $P(z)$, $Q(z)$ e $A(z)$ no plano de Argand

É possível descobrir as raízes fazendo $H(z) = 1$ e $H(z) = -1$. Neste caso é analisada a resposta em frequência e os coeficientes LSF correspondem aos valores múltiplos de π .

As soluções apresentadas, matematicamente correctas, não possuem contudo as características desejáveis para uma implementação em máquinas digitais. Kabal e Ramachandran [42] e mais recentemente Rothweiler [72] apresentam formulações alternativas que sem perderem rigor permitem o cálculo dos coeficientes LSF de um modo mais expedito. A abordagem de Kabal utiliza polinômios de Chebyshev e a de Rothweiler, um aperfeiçoamento da primeira, com polinômios de Chebyshev de cálculo recursivo. Esta última é menos exigente computacionalmente e segundo o autor permite obter resultados mais precisos.

3.6.2 Desempenho das LSF

Os coeficientes LSF possuem algumas particularidades matemáticas interessantes. Uma delas é a tendência para se localizarem perto dos coeficientes LPC quando representados num plano de Argand.

Os ângulos de cada coeficiente são responsáveis pelos formantes no espectro de predição linear. A largura de banda de cada formante é definida pela proximidade dos LSFs, correspondendo a uma maior proximidade um pico mais aguçado para esse formante na representação espectral. Assim sendo a deterioração de um coeficiente LSF possui um impacto espectral localizado e limitado (esta característica não existe nos coeficientes LPC).

Assim sendo, para efeitos de representação de sinais de fala os coeficientes LSF possuem em relação aos LPC uma maior imunidade a pequenas variações. Ou seja, se devido a arredondamentos resultantes de um qualquer processamento intermédio, os coeficientes LSF apresentarem algum desvio em relação aos seus valores iniciais, as consequências nefastas para a reconstrução do sinal de fala serão inferiores às que resultariam de um processamento semelhante com coeficientes LPC [34].

Por estas razões esta representação alternativa tornou-se popular na codi-

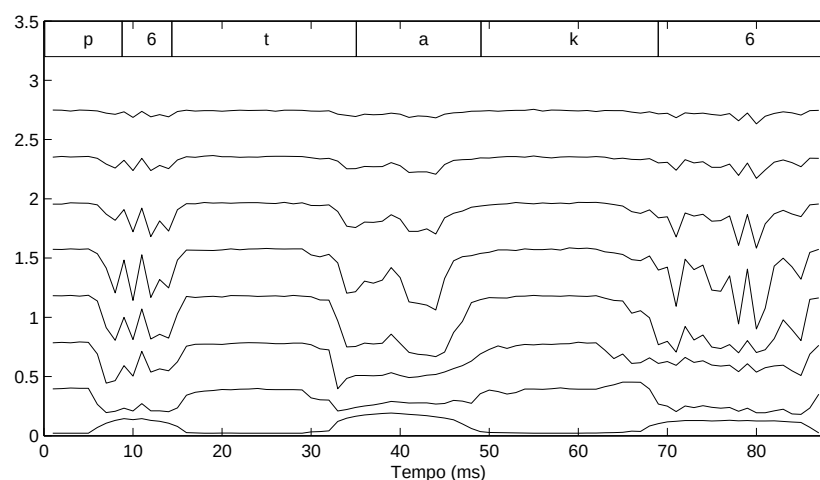


Figura 3.20: Representação da palavra "pataca" por coeficientes LSF

ficação de sinal para comunicações. Por razões semelhantes mas numa outra perspectiva será utilizada neste documento.

De acordo com o apresentado as frequência dadas pelos coeficientes LSF estão distribuídas no plano de Argand de um modo relativamente espaçado e dentro de uma gama de ângulos limitada. O sinal de voz, quando analisado por janelamentos sucessivos, possui no domínio das frequências variações bastante suaves. Pela relação íntima das LSF com as principais frequências formantes do espectro conclui-se que os coeficientes LSFs evoluirão também de modo suave (embora com perturbações ocasionais).

Como exemplo de uma codificação LSF apresenta-se na figura 3.20 uma representação gráfica da evolução temporal dos ângulos dados por 8 coeficientes para uma sequência de janelas de 20ms com sobreposição de meia janela. A palavra utilizada não é rica do ponto de vista fonético mas é a adequada para explicitar as características referidas.

Pela análise da figura vê-se claramente que existem diferenças nos ângulos bem marcadas entre os vários fonemas. Vê-se também que os primeiros coeficientes, ou seja, os que possuem ângulos mais reduzidos possuem maior sensibilidade enquanto que os mais afastados do ângulo nulo, na figura mais acima, mantêm-se praticamente inalterados. Isto é igualmente visível quando surgem as passagens das consoantes para as vogais onde se nota que nas linhas mais inferiores existe uma quase previsão da transição entre fonemas. Mostra-se ainda que a evolução dos ângulos ao longo de toda a palavra se faz de modo suave (não seria possível efectuar uma representação semelhante com os coeficientes LPC).

3.7 MFCC

A representação por coeficientes Mel cepstrais é uma representação baseada no comportamento do ouvido humano e é uma das mais utilizadas no domínio do reconhecimento da fala pelo seu desempenho mais favorável [22].

Os MFCC (*Mel Frequency Cepstral Coefficients*) são representativos do espectro real da FFT de um sinal analisado numa janela de curta duração. No entanto, é utilizada uma escala de frequência não linear que tenta reproduzir a forma como o ouvido humano percebe os sons em bandas de frequência diferentes. Assim, considerando a DFT do sinal de entrada:

$$X_a[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\pi nk/N}, \quad 0 \leq k < N \quad (3.58)$$

O cálculo dos coeficientes Mel é efectuado a partir de um conjunto de M filtros triangulares que permitem obter o espectro médio sobre uma frequência central. O banco de filtros podem ser constituído por M filtros cuja expressão é dada em 3.59 onde $m = 1, 2, \dots, M$:

$$H_m[k] = \begin{cases} 0, & k < f[m-1] \\ \frac{2(k-f[m-1])}{(f[m+1]-f[m-1])(f[m]-f[m-1])}, & f[m-1] \leq k \leq f[m] \\ \frac{2(f[m+1]-k)}{(f[m+1]-f[m-1])(f[m]-f[m-1])}, & f[m] \leq k \leq f[m+1] \\ 0, & k > f[m+1] \end{cases} \quad (3.59)$$

Em alternativa, existe uma representação normalizada para o banco de filtros na qual $\sum_{m=1}^M H'_m[k] = 1$, dada pelas expressões:

$$H'_m[k] = \begin{cases} 0, & k < f[m-1] \\ \frac{k-f[m-1]}{f[m]-f[m-1]}, & f[m-1] \leq k \leq f[m] \\ \frac{f[m+1]-k}{f[m+1]-f[m-1]}, & f[m] \leq k \leq f[m+1] \\ 0, & k > f[m+1] \end{cases} \quad (3.60)$$

Os valores das frequências na escala de Mel podem ser calculados a partir da escala linear assim como a operação inversa, pelas expressões:

$$B(f) = 1125 \cdot \ln \left(1 + \frac{f}{700} \right) \quad (3.61)$$

$$B^{-1} = 700 \cdot \left(\exp\left(\frac{b}{1125}\right) - 1 \right) \quad (3.62)$$

As frequências centrais de cada filtro são dadas, na escala de Mel, por:

$$f[m] = \left(\frac{N}{F_s} \right) b^{-1} \left(b(f_l) + m \frac{B(f_h) - B(f_l)}{M+1} \right) \quad (3.63)$$

Nesta expressão, f_l e f_h são as frequências (em Hz) mais baixa e mais elevada do banco de filtros, F_s é a frequência de amostragem do sinal (em Hz), M é o

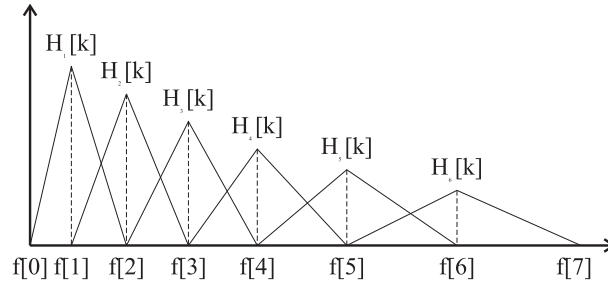


Figura 3.21: Banco de filtros baseado na percepção utilizado para o cálculo dos MFCCs

número de filtros e N é a dimensão da FFT. Na figura 3.21 apresentam-se um banco de 6 filtros de Mel onde é visível a redução da contribuição das frequências mais elevadas que é característica do ouvido humano.

Os coeficientes Mel cepstrais por serem calculados no domínio das frequências fornecem uma descrição do sinal superior ao coeficientes baseados no tempo. No entanto, a evolução temporal do espectro dá um contributo não negligenciável para a tarefa de reconhecimento para além de ser uma característica de funcionamento da audição humana. Por estes motivos é frequente utilizarem-se, para além dos próprios coeficientes Mel-Cepstrais, os coeficientes delta, representativos da derivada discreta, e os coeficientes de aceleração ou delta-delta, correspondentes à derivada discreta da derivada anterior.

3.8 Distâncias

Em muitas situações no processamento de fala é necessário comparar dois sinais e saber em que medida estes são ou não semelhantes. As métricas devem possuir as seguintes propriedades:

- $d(S_a, S_b) \geq 0$;
- $d(S_a, S_b) = 0$, *s.s.s.* $S_a = S_b$;
- $d(S_a, S_b) \leq d(S_a, S_c) + d(S_c, S_b)$

Estes propriedades vão de encontro ao tradicional conceito de métrica e existem muitas possibilidades que cumprem os requisitos apresentados. A maioria das métricas que se utilizam em processamento de fala são casos particulares da distância de Minkowski ou estão relacionadas [27].

3.8.1 Distância de Minkowski

A distância de Minkowski de ordem a entre os vectores S_i e S_c é dada pela expressão abaixo:

$$d_{\text{Minkowski}}(S_i, S_c) = \sqrt[a]{\sum_{k=1}^N |S_{ik} - S_{ck}|^a} \quad (3.64)$$

As variáveis S_{ik} e S_{ck} representam as componentes de ordem k dos vectores em análise S_i e S_c . Sendo l_x a norma do vector S como:

$$|S|_a = \sqrt[a]{\sum_{k=1}^N |S|^a} \quad (3.65)$$

Então a distância de Minkowski é igual à norma do vector distância entre os vectores em análise. A distância Euclideana será o caso particular mais imediato desta distância para $a = 2$.

3.8.2 Distância Euclideana

A distância Euclideana é já sobejamente conhecida e é dada pela expressão:

$$d_{\text{Euclideana}}(S_i, S_c) = \sqrt{\sum_{k=1}^n (S_{ik} - S_{ck})^2} = \sqrt{(S_i - S_c)^T (S_i - S_c)} \quad (3.66)$$

Segundo um trabalho de Donovan [29], esta distância não é a que mais se compatibiliza com as diferenciações perceptuais humanas. Uma alternativa será utilizar uma matriz w que permita atribuir diferentes ponderações para cada uma das características. A nova expressão será dada por:

$$d_E(S_i, S_c) = \sqrt{(S_i - S_c)^T w^{-1} (S_i - S_c)} \quad (3.67)$$

Esta métrica aproxima-se bastante da métrica de Mahalanobis.

3.8.3 Distância de Mahalanobis

Se não se considerar a operação de raiz quadrada utilizada na distância Euclideana modificada, o que não trará grandes alterações aos resultados do ponto de vista da diferenciação, tem-se a distância de Mahalanobis:

$$d_M(S_i, S_c) = (S_i - S_c)^T w^{-1} (S_i - S_c) \quad (3.68)$$

Se w^{-1} for a matriz identidade então esta distância será igual à distância Euclideana quadrática.

3.8.4 Distância de Itakura-Saito

A distância de Itakura-Saito é baseada nos polinómios de predição AR. De acordo com [67] a medida é dada pela expressão:

$$d_{IS}(A, B) = \int_{-\pi}^{\pi} \left[e^{V(w)} - V(w) - 1 \right] \frac{dw}{2\pi} \quad (3.69)$$

onde $V(w) = \log A(w) - \log B(w)$ e $A(w)$ e $B(w)$ são espectros potência. Esta métrica, mais pesada computacionalmente, tem como vantagem a sua sensibilidade às variações espectrais.

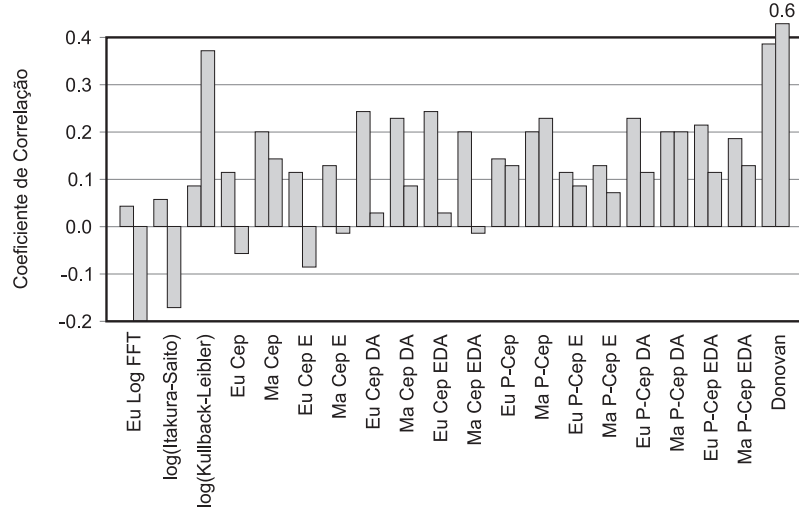


Figura 3.22: Coeficientes de correlação para a comparação de uma métrica de distância espectral e o resultado de um ouvinte humano

3.8.5 Distância de Kullback-Leibler

Uma outra medida bastante popular é a de Kullback-Leibler que, utilizando a mesma notação, é dada pela expressão [27, 29]:

$$d_{KL}(A, B) = \int_{-\pi}^{\pi} A(w) \log \left(\frac{A(w)}{B(w)} \right) \frac{dw}{2\pi} \quad (3.70)$$

Num trabalho de Donovan compararam-se várias distâncias 3.22 com a percepção humana onde se incluem as que aqui foram apresentadas. O mesmo autor propõe uma nova métrica que segundo os seus resultados é a que mais se aproxima das distinções do ouvido humano.

3.8.6 Alinhamento com DTW

Apesar de não ser utilizado nesta dissertação, o algoritmo DTW é incontornável na segmentação e reconhecimento de fala sendo uma das mais antigas e importantes técnicas em utilização neste campo [88, 40].

O funcionamento é simples e baseia-se na comparação directa do sinal ou na comparação de outras características representativas com outros sinais de referência e daqui obter a maior semelhança possível. Existem porém algumas dificuldades a ultrapassar. Em primeiro, palavras diferentes terão durações diferentes apesar das semelhanças acústicas. Este problema poderia ser contornado pela normalização dos segmentos em análise no caso de não se verificar um outro problema. A velocidade com que um sinal é produzido, que em fala se designa tipicamente por taxa de discurso, pode não ser constante. Ou seja, o alinhamento óptimo entre dois segmentos de sinal pode não ser linear. Para resolver estas situações pode-se utilizar eficazmente o algoritmo DTW.

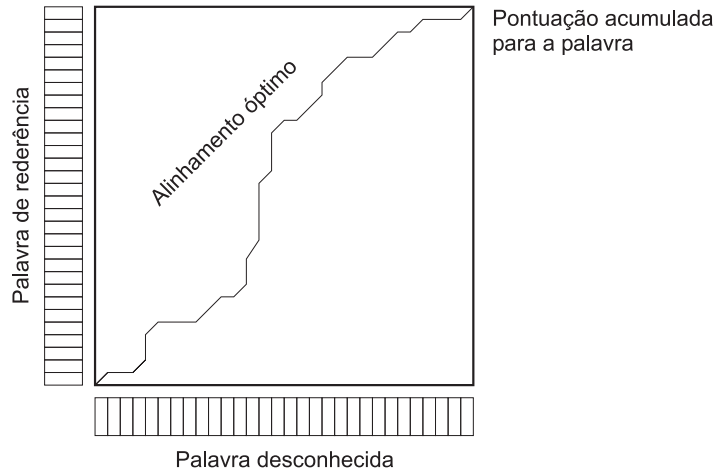


Figura 3.23: Alinhamento utilizando a técnica de *Dynamic Time Warping*

O DTW pertence a uma classe de algoritmos conhecidos por programação dinâmica. O algoritmo efectua uma única passagem ao longo de uma matriz de vectores de características com pontuações enquanto calcula segmentos optimizados localmente que formam o caminho de alinhamento total. Se a estimação da similaridade entre dois vectores pode-se usar uma qualquer das métricas referidas. Se $D(x, y)$ representar a distância Euclideana entre o segmento x da sinal em análise e o segmento y do sinal de referência, e se $C(x, y)$ for a pontuação acumulada ao longo um caminho óptimo de alinhamento que leva a (x, y) então:

$$C(x, y) = \min [C(x - 1, y), C(x - 1, y - 1), C(x, y - 1)] + D(x, y) \quad (3.71)$$

O cálculo é realizado desde o ponto $(0,0)$ até ao ponto limite (X,Y) do extremo oposto. No final possui-se um valor acumulado ao longo do caminho e o percurso seguido pode ser analisado fazendo o inverso. Quando se efectua esta análise para vários sinais, o que mais se aproximará será o que, no final, possuir um menor valor acumulado.

O algoritmo foi apresentado na sua forma mais simples existindo muitas variações possíveis. É por exemplo frequente utilizarem-se formas de cálculo da pontuação acumulada diferentes onde se consideram outros pontos vizinhos dentro da matriz. Este algoritmo é muito semelhante ao algoritmo de Viterbi que se analisará mais à frente.

De notar que a exposição feita se baseia no pressuposto da existência de dois sinais que podem ser comparados. No caso da segmentação existe inicialmente apenas um sinal, o que se pretende segmentar, sendo então necessário possuir um outro que permita a realização das comparações necessárias ao algoritmo DTW. Um procedimento genérico para a anotação baseada em DTW apresenta-se na figura 3.24.

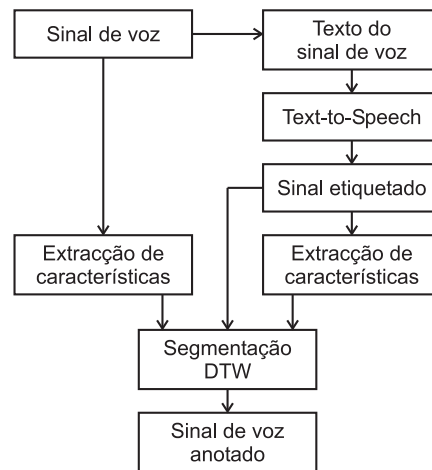


Figura 3.24: Procedimento genérico para anotação com DTW.

Capítulo 4

Modelo LSF/ANN

Para a etiquetagem de sinais de fala optou-se por utilizar numa primeira abordagem de menor complexidade uma técnica baseada na inspecção directa do sinal acústico. As tarefas de segmentação e classificação serão executadas separadamente o procedimento seguido será descrito. Uma vez que na classificação são utilizadas redes neuronais apresentar-se-á também uma breve descrição da teoria envolvida.

4.1 Abordagem 1: Alinhamento Fonético

A primeira tentativa de anotação de um sinal de fala é baseada num trabalho apresentado por Amaral et al. [3] onde se propõe um sistema para a segmentação e etiquetagem de uma base de dados de dígitos de fala telefónica. O autor realiza uma segmentação baseada na acústica do sinal e numa medida de verosimilhança. A classificação é depois realizada com um sistema baseado em lógica difusa.

A abordagem aqui proposta, por se destinar a uma utilização mais abrangente, foi ligeiramente modificada. Utilizou-se a base de dados FEUP/IPB e considerou-se que o sinal de fala vinha acompanhado de informação sobre a sequência fonética restando assim efectuar apenas a segmentação. Os passos da sequência de processamento constituída são descritos em pormenor abaixo e representam-se também num esquema apresentado na figura 4.1.

1. Inicialmente, após a normalização do sinal e pré-ênfase, fez-se a partição do sinal de fala em blocos de 128 amostras e com um passo de avanço de 40 amostras, pouco menos de um terço da duração da janela.
2. Extraí-se para cada bloco a energia e 16 coeficientes LSF e calculam-se as correspondentes características dinâmicas, delta e aceleração constituindo assim vectores de características de dimensão 51.
3. Calcula-se a distância euclidiana entre vectores de características sucessivos.
4. Analisam-se os valores das distâncias e sempre que estas são superiores a um determinado valor pré-especificado (nível de *threshold*) assume-se

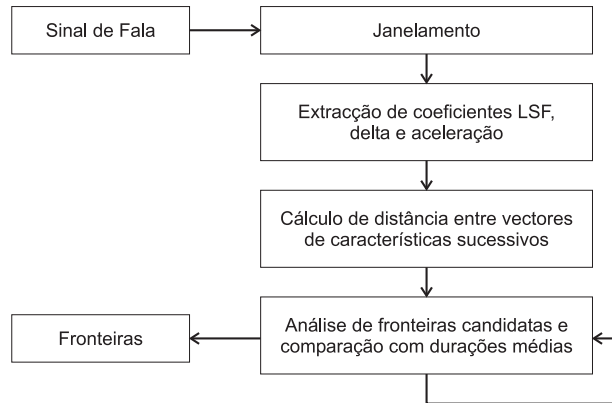


Figura 4.1: Procedimento para a segmentação baseada na acústica do sinal

a existência de uma candidata a fronteira. As durações dos segmentos encontrados são depois comparadas com outras da base de dados sendo o nível de *threshold* ajustado dentro de intervalos temporais que se vão ajustando. Os valores limite que indicam o aumento ou diminuição do nível de *threshold* são dados respectivamente por metade da duração média do fone em causa e pelo dobro da duração do fone em causa.

Para a avaliação deste sistema utilizou-se a métrica que atrás se designou de Exactidão que considera no seu cálculo situações de inserção e eliminação. Obteve-se um valor de 62.14% para um intervalo de 10ms e 69.33% para um intervalo de 20ms. Foram ainda ensaiadas outras possibilidades para a dimensão da janela de análise mantendo-se sempre a relação de cerca de um terço para o passo de avanço. Com 64 amostras os resultados pioraram consideravelmente surgindo vários demasiados erros de inserção que, na tentativa de ser corrigidos pelo ajuste do nível *threshold*, levavam ao aparecimento de erros de eliminação. Para 256 amostras os resultados foram pouco diferentes sendo mesmo em alguns casos ligeiramente superiores.

Todo o procedimento, implementado utilizando o *software* Matlab, baseia-se em múltiplos pressupostos que o afastam um pouco de situações reais e limitam a sua utilização. O ajuste do nível de *threshold* em função de durações médias pré-conhecidas obriga à posse de informações que podem não estar disponíveis e serem de difícil obtenção. Decidiu-se tentar uma alternativa.

4.2 Redes Neurais

Sendo as redes neuronais utilizadas na abordagem alternativa apresenta-se aqui uma pequena descrição do assunto.

As redes neuronais artificiais são estruturas computacionais constituídas por múltiplas unidades de processamento simples, designadas por neurónios ou nodos, que são interligadas de um modo inspirado nos neurónios do cérebro.

Numa rede neuronal, um grupo de neurónios de entrada recebe um dado conjunto de informações. Estes neurónios estão ligados a outros, escondidos, que

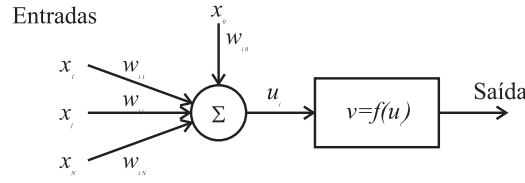


Figura 4.2: Modelo computacional de um neurónio

por sua vez estão ligados a outros. Existirá finalmente ligados aos restantes um conjunto de neurónios de saída. Cada ligação, também designada por sinapse, tem um peso associado. Dependendo do valor destes pesos, a saída será mais ou menos sensível a certos estímulos. O valor dos referidos pesos é ajustado num processo de treino obrigatório antes da rede ser utilizada.

4.2.1 Neurónio Básico

O neurónio é o elemento processador da rede neuronal (figura 4.2). Cada neurónio gera uma saída a partir da combinação linear dos sinais de entrada recebidos de outros neurónios aos quais está ligado ou a partir de sinais externos. O sinal de saída é o resultado da aplicação da função de transferência ou função de activação à combinação dos sinais de entrada. O estado do neurónio é representado pelo seu sinal de saída.

O sinal de entrada total do neurónio é obtido pela combinação linear dos sinais recebidos:

$$u_i = \sum_j w_{ij} v_j + \lambda_i \quad (4.1)$$

Na expressão u_i representa a entrada total do neurónio i , v_j representa a saída do neurónio j , w_{ij} representa o peso da ligação entre os neurónios i e j e, por último, λ é um factor de polarização.

O peso atribuído a cada ligação representativa das sinapses é uma forma de ponderação do sinal recebido. Em alguns modelos o factor de polarização é considerado como mais um sinal recebido podendo representar uma forma de limiar de estímulo. A função de activação é responsável por produzir a saída final do neurónio. Podem ser utilizadas diversas funções sendo as mais comuns a função sigmóide e a função degrau.

4.2.2 Ligações

As unidades de processamento dentro de uma rede neuronal estão organizadas de acordo com uma determinada topologia e através de uma série de ligações pesadas. Cada peso possui um valor real, tipicamente entre $-\infty$ e $+\infty$, sendo em alguns casos limitado a uma gama menor. O peso de cada ligação descreve a influência que uma determinada unidade terá na unidade vizinha; considera-se [49] que um peso positivo terá uma função de excitação da unidade vizinha ao passo que um peso negativo terá uma função inibidora da unidade vizinha. As ligações e respectivos pesos são geralmente unidireccionadas mas quando,

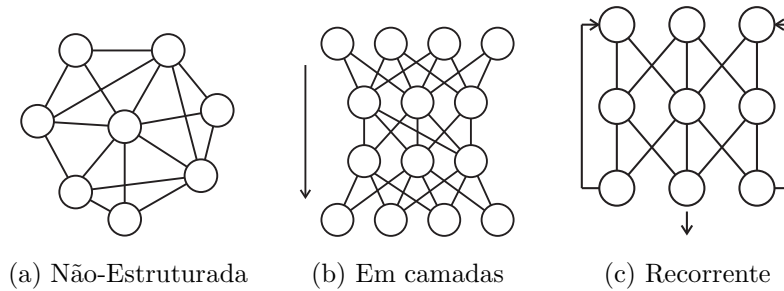


Figura 4.3: Topologias mais comuns para redes neuronais.

dependendo do tipo de rede, não existe uma distinção clara entre entrada e saída, podem surgir ligações bidireccionais.

Assim, o conjunto de todos os pesos determinará a reacção computacional da rede a um qualquer estímulo externo. Os pesos representam a memória ou o conhecimento da rede. O seu valor é ajustado como já se referiu por um processo inicial de treino.

As topologias de rede podem ser quaisquer existindo porém alguns formatos típicos que se apresentam representados na figura 4.3. Cada uma das topologias é mais adequada a um determinado tipo de tarefas. Por exemplo:

- **Não-Estruturada.** Adequada para tarefas onde seja necessário completar um padrão dada uma parte desse padrão.
- **Em Camadas.** Adequada para associações entre conjuntos de valores de entrada e conjuntos de valores de saída.
- **Recorrente.** Adequada para o sequenciamento de padrões.

As redes não-estruturadas podem ter ciclos internos que as tornem recorrentes; As redes em camada por ou não ser recorrentes.

4.2.3 Treino

Apresentados os elementos e princípios básicos das redes neuronais torna-se agora importante analisar as formas como uma rede neuronal se adapta a um determinado problema através do processo de treino.

De acordo com os dados disponíveis e com a topologia da rede podem utilizar-se vários métodos de treino:

- **Aprendizagem supervisionada.** Quando é possível fornecer à rede um conjunto de entradas/saídas conhecidos e a partir do cálculo do erro à saída se efectuem ajustes nos pesos das ligações inter-neuronais.
- **Aprendizagem semi-supervisionada.** Situação semelhante à anterior diferindo no facto de o erro não ser calculado explicitamente. É apenas indicado se a o resultado está ou não de acordo com as expectativas.
- **Aprendizagem não-supervisionada.** Quando a rede tem de encontrar por si própria padrões ou regularidades nos dados fornecidos.

Nas aplicações de processamento de voz é em geral possível calcular claramente o erro cometido pela rede sendo portanto aplicável a aprendizagem supervisionada. O treino nesta situação faz-se tradicionalmente pelo algoritmo de retro-propagação.

O algoritmo de treino da rede por *back-propagation* ou retropropagação é uma generalização de um outro algoritmo designado por algoritmo de erro quadrático médio ou regra delta. O seu funcionamento é bastante simples baseado-se numa pesquisa no sentido e direcção do gradiente descendente para uma função erro igual à diferença quadrática média entre as saídas esperadas e as geradas pela rede, para todos os padrões¹ do conjunto de treino.

O erro para cada padrão pode ser calculado pela expressão abaixo:

$$E_p = \frac{1}{2} \sum_{k=1}^{NS} (r_{pk} - v_{pk})^2 \quad (4.2)$$

Aqui tem-se E_p como o erro quadrático para o padrão p , r_{pk} como a saída esperada do neurónio k da camada de saída para o padrão p , v_{pk} como a saída calculada para o neurónio k da camada de saída para o padrão p , u_{pk} como a entrada total do neurónio k da camada de saída para o padrão p , w_l como o peso da ligação l da rede com $l=0,1,2 \dots$ número total de ligações e NS como o número total de neurónios da camada de saída.

Os pesos das ligações constituem as variáveis a ajustar. A partir do cálculo do gradiente de E_p , resultam derivadas parciais em relação a cada w_l e a partir destas são calculados os incrementos para os pesos. O algoritmo de treino por retro-propagação [27, 38] simplificado, para apenas duas camadas de neurónios, passa-se a descrever:

1. Inicialização dos pesos das ligações e respectivos factores de polarização com valores aleatórios (geralmente entre -0.5 e 0.5); definir velocidade de aprendizagem η ; inicializar erro total $E = 0$;
2. Utilização de um novo vector de dados para a entrada $x_p = (x_{p1}, x_{p2}, \dots, x_{pN})$ e, caso não existam mais na sequência de treino, saltar para o passo 12
3. Cálculo do somatório de entrada dos neurónios da camada oculta considerando NI o número de neurónios.

$$u_{pj}^e = \sum_{i=1}^{NI} w_{ji}^e x_{pi} + \lambda_j^e \quad (4.3)$$

4. Cálculo da saída da camada oculta

$$v_{pj}^e = f(u_{pj}^e) \quad (4.4)$$

5. Cálculo do somatório de entrada para cada neurónio k da camada de saída considerando NE o número de neurónios da camada oculta.

$$u_{pk}^s = \sum_{j=1}^{NE} w_{kj}^s v_{pj}^e + \lambda_k^e \quad (4.5)$$

¹Considera-se aqui como padrão um determinado conjunto de valores, a fornecer na entrada da rede, acompanhados de um outro conjunto com os valores a obter na saída da rede. Ou seja, no processo de treino são fornecidos à rede padrões, conjunto de entradas e respectivas saídas, com o objectivo de ajustar os parâmetros internos.

6. Cálculo da saída de cada neurónio da camada de saída

$$v_{pk}^s = f(u_{pk}^s) \quad (4.6)$$

7. Cálculo de E_p e componentes de erro associadas a cada neurónio k da camada de saída

$$\delta_{pk}^s = (r_{pk} - v_{pk}^s) f_{pk}(v_{pk}^s) \quad (4.7)$$

8. Cálculo das componentes de erro para cada neurónio j da camada oculta

$$\delta_{pj}^e = f_{pj}(v_{pj}^e) \sum_k \delta_{pk}^s w_{kj}^s \quad (4.8)$$

9. Actualização dos pesos das ligações entre a camada de saída e a oculta

$$w_{kj}^s(t+1) = w_{kj}^s(t) + \eta \delta_{pk}^s u_{pj}^e \quad (4.9)$$

10. Actualizar os pesos das ligações entre a camada oculta e a entrada da rede

$$w_{ji}^e(t+1) = w_{ji}^e(t) + \eta \delta_{pj}^e x_{pi} \quad (4.10)$$

11. Actualização do valor do erro total fazendo $E = E + E_p$ e regresso ao passo 2.

12. Se o valor de E for menor do que erro mínimo estipulado então o processo pode ser terminado, caso contrário, reinicializar lista de vectores de padrões e regressar ao passo 2.

Esta iteração deve ser feita ate que o erro E esteja dentro de um limite considerado aceitável. O valor da taxa de aprendizagem η afecta o desempenho na fase de treino. Em geral, η deve ser um valor positivo pequeno (0.05 a 0.25 são valores frequentes), de modo a assegurar a convergência. Quanto menor o valor de η maior será o número de iterações. Em algumas situações o valor de η pode ser alterando durante o processo de treino com o intuito de acelerar o processo de convergência e evitar oscilações.

Um outro modo de acelerar a convergência pode ser utilizado se se alterar ligeiramente o algoritmo apresentado. Substitui-se as equações de ajuste dos pesos das ligações entre a camada de saída e a oculta pela expressão abaixo:

$$w_{kj}^s(t+1) = w_{kj}^s(t) + \eta \delta_{pk}^s X_j + \alpha \Delta w_{kj}^s(t-1) \quad (4.11)$$

Considerando $\Delta w_{kj}^s(t-1)$ a variação do peso na iteração $t-1$ e α um coeficiente de momento. Deste modo, ao actualizar os pesos, uma fracção do ajuste anterior será incluída. Este acréscimo tenderá a manter constante o sentido de ajuste que foi utilizado na iteração anterior. Em geral, tem-se $\alpha < 1$. Uma equação semelhante deverá ser aplicada para a camada oculta.

O processo de treino pode resultar numa convergência para um mínimo local e não num global. Isto dependerá do valor inicial dos pesos, do número de neurónios utilizados e dos parâmetros de aprendizagem. Se a rede convergir para um mínimo local inaceitável em termos do erro, o processo de aprendizagem deve ser repetido com a alteração de alguns dos parâmetros citados. Num caso prático, apesar desta situação ser remota, chegar voltar a inicializar a rede com novos valores aleatórios. Se o processo convergir para um mínimo aceitável do ponto de vista do erro obtido não fará sentido a distinção entre o mínimo local ou global.

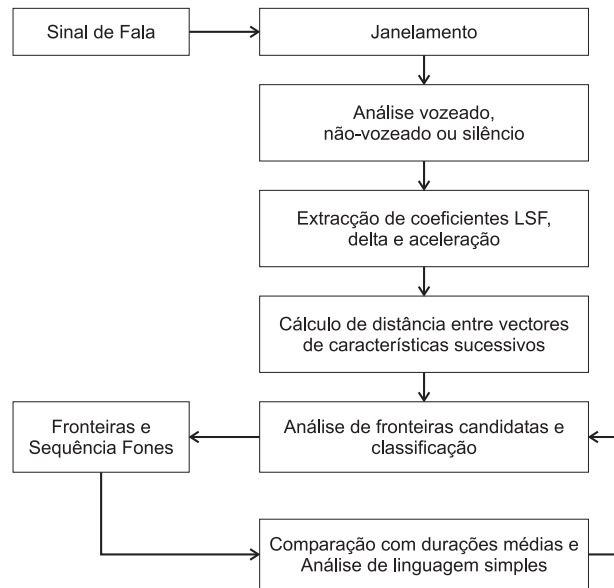


Figura 4.4: Procedimento alternativo para a segmentação baseada na acústica do sinal e posterior classificação

4.3 Abordagem 2: Segmentação e Classificação

Partiu-se então para um outro procedimento desta vez utilizando apenas como material de base o sinal de fala. Descreve-se então o novo procedimento e apresenta-se na figura 4.4 o esquema correspondente.

1. Inicialmente, após a normalização e pré-ênfase, fez-se a partição do sinal de fala em blocos de 128 amostras e com um passo de avanço de 40 amostras, pouco menos de um terço da duração da janela.
2. Extraí-se para cada bloco a energia e a taxa de passagens por zero e com estas informações efectua-se a classificação em vozeado, não-vozeado e silêncio. Aqui foram utilizadas várias ideias de Teixeira [81].
3. Extraí-se para cada bloco 16 coeficientes LSF e calculam-se as correspondentes características dinâmicas, delta e aceleração constituindo assim vectores de características de dimensão 51.
4. Calcula-se a distância euclidiana entre vectores de características sucessivos.
5. Analisam-se os valores das distâncias e sempre que estas são superiores a um determinado valor pré-especificado (nível de *threshold*) assume-se a existência de uma candidata a fronteira. O nível de *threshold* é sucessivamente aumentado até que não existam segmentos de duração superior a um dado valor. Ou seja, pretende-se reduzir ao mínimo a possibilidade de ocorrência de erros por eliminação. Os segmentos encontrados são depois

classificados por uma rede neuronal. Uma vez que se potenciou o aparecimento de erros por inserção, um pequeno módulo final junta ocorrências consecutivas de fones iguais num único fone e elimina fones de acordo com algumas regras gramaticais simples.

A rede neuronal construída era constituída por 4 camadas sendo uma de entrada, com 153 neurónios correspondentes a três vectores de características, duas camadas escondidas (também designadas por ocultas ou intermédias), com 200 e 100 neurónios e uma camada de saída, com 40 neurónios correspondentes ao símbolos principais utilizados na base de dados FEUP/IPB. As funções de activação escolhidas foram sigmóidais. Com a rede em funcionamento, após o treino, os resultados obtidos à saída eram comparados com a classificação de vozeado, não-vozeado ou silêncio, efectuada no segundo passo da sequência, com o objectivo de reduzir os erros cometidos.

Os resultados obtidos com este procedimento foram um pouco mais animadores mas ainda assim um pouco abaixo do esperado. Para a tarefa de segmentação obteve-se um valor de 62.81% para um intervalo de 10ms e 70.32% para um intervalo de 20ms. A tarefa de classificação, introduzida na sequência, foi também avaliada utilizando a métrica de Exactidão correspondente. Esta forneceu um valor global de 83.49%. Os sons vozeados e em particular as vogais, analisando os resultados à saída da rede neuronal, foram mais facilmente aprendidos pela rede. Consegui-se cerca de 86% de resultados correctos em relação ao número total de ocorrências e 97% para as vogais. Para os sons não-vozeados conseguiu-se o valor de 78% mais uma vez considerando os resultados correctos sobre a totalidade de ocorrências submetidas à análise da rede.

Uma alternativa aos processos apresentados é utilizada por Gholampour [33] que sugere que se utilizem três vectores para a comparação, um fixado temporalmente e outros dois, sucessivos, que se vão deslocando no tempo. Os vectores, A, B e C, são inicialmente adjacentes, sendo C o mais adiantado no tempo. É calculada a distância entre A e B e entre B e C. Os vectores B e C vão avançando no eixo dos tempos e em cada passo são calculadas as novas distâncias de B a A e a C. Quando a diferença/distância entre B e A é maior que a diferença entre B e C marca-se uma candidata a fronteira e o vector A, que estava fixo, é deslocado para a posição de B. O processo é repetido ao longo de todo o sinal de fala sendo os resultados processados por dois outros módulos que utilizam dendogramas (não estudados nesta dissertação). É utilizada uma métrica de comparação baseada na distância de Kullback-Leibler.

O procedimento inicial seguido por Gholampour foi também experimentado e resultou num maior número de erros de inserção existindo porém a vantagem de não ser necessário pré-especificar um nível de *threshold*.

A dificuldade de especificação de um nível de *threshold* adequado e o excesso ou défice de fronteiras que daqui podem resultar é consequência da análise de algum modo rígida que tem sido efectuada e que não se compatibiliza totalmente com os sinais de fala. Considere-se a evolução do sinal acústico ou de características representativas ao longo de uma transição de um fone para um outro adjacente. Do ponto de vista do primeiro fone, a distância que o separa do segundo tenderá a diminuir com a evolução do tempo, no entanto, devido às características intrínsecas da fala, esta aproximação poderá não ser monótona. Ao se considerar uma linha de fronteira rígida entre os dois supostos fones surgirão inevitavelmente situações não desejadas resultantes de uma certa *aproximação*

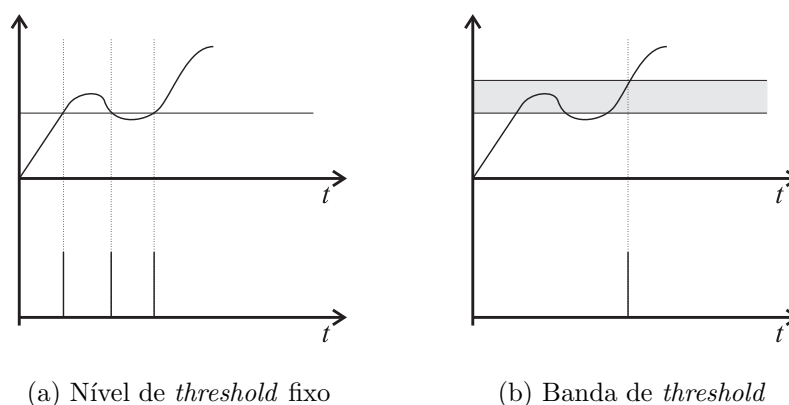


Figura 4.5: Abordagens para a marcação de fronteiras.

hesitante entre os dois.

Para a resolução destes problemas sugere-se um novo processamento um pouco mais complexo. A análise das distâncias, no lugar de ser efectuada directamente sobre as características extraídas, passará a ser feita sobre um conjunto de características cujas transições/trajectórias foram suavizadas. Procura-se com isto criar trajectórias de evolução do sinal mais monótonas e sustentadas. Em relação à fixação do nível de *threshold* sugere-se uma abordagem dinâmica com um comportamento que possua alguma histerese. Ou seja, passar-se-á a possuir uma banda de *threshold* com um funcionamento semelhante a um sistema de controlo *on/off*. Na figura 4.5 esquematiza-se este funcionamento. A linha curva nos gráficos superiores pretende modelizar uma trajectória entre dois sons. Na zona inferior marcam-se as fronteiras resultantes da intersecção da curva com os limites de *threshold*. No exemplo apresentado utiliza-se um espaço bidimensional mas que na realidade será um espaço multidimensional ajustado aos vectores de características. Então, na figura 4.5(a) mostra-se o funcionamento que tem sido implementado e, na figura 4.5(b), mostra-se o funcionamento alternativo proposto. Assim, executando a operação sugerida, será supostamente reduzido o número de fronteiras hipotéticas gerado pelo algoritmo.

O algoritmo alternativo aqui apresentado foi imaginado para dar resposta aos problemas que se encontraram não tendo, no entanto, sido implementado. Apesar do interesse da abordagem considerou-se que esta constituía um percurso menos promissor e adequado para os objectivos a atingir. Deste modo devido às restrições temporais subjacentes a este trabalho decidiu-se abandonar o aprofundamento desta técnica e procurar-se-á retoma-lo numa outra ocasião. As análises efectuadas e os algoritmos implementados com o nível de desenvolvimento conseguido poderão ser ainda úteis como complemento para outras técnicas.

Capítulo 5

Anotação com HMMs

A utilização de modelos escondidos de Markov é uma técnica largamente utilizada e que permite efectuar em simultâneo as tarefas de segmentação e anotação. Comparando diversas técnicas na sua forma mais pura, ou seja, sem que se utilizem sistemas de refinamento dos resultados, os modelos de Markov apresentam o melhor desempenho. Em relação ao esforço computacional, que não é um ponto demasiado importante pois a anotação não tem de ser em tempo-real, também não existem grandes desvantagens sendo as várias técnicas muito idênticas para resultados semelhantes.

Neste capítulo analisar-se-á em pormenor a construção de um sistema de etiquetagem automático baseado em modelos escondidos de Markov. Em cada fase do processo serão apresentadas as variáveis que influenciam o desempenho e serão feitos diversos ensaios com que se procura identificar a configuração mais favorável.

5.1 HMM - Hidden Markov Models

Os modelos escondidos de Markov são um técnica de base estocástica que permite descrever no tempo a evolução de uma sequência de valores discretos. Esta técnica para além de possibilitar o desenvolvimento de modelos paramétricos fiáveis possui também flexibilidade para que se utilizem técnicas de programação dinâmica para a segmentação e classificação de padrões resultantes das variações temporais dos dados. O essencial desta teoria deve-se a Baum et al. [7] e remonta a 1967. Com a evolução da tecnologia que permitiu explorar ao máximo todas as potencialidades e resolver eficientemente alguns dos problemas intrínsecos, os modelos de Markov tornaram-se uma das principais técnicas para a modelização de sinais de fala (reconhecimento de voz, síntese, segmentação, modelos de linguagem, etc.) e encontram aplicações numa vasta gama de áreas.

Os sistemas de processamento de sinais resultantes da aquisição de um sinal de fala assumem que este corresponde a uma sequência de símbolos bem definidos e com fronteiras delimitadas no tempo. Na figura 5.1 ilustra-se de modo esquemático este funcionamento. Para identificar tanto a sequência como os símbolos e suas fronteiras, o sinal é segmentado, com janelas de dimensão fixa e a intervalos constantes e seguidamente calculam-se, através de uma parametrização, vectores de características correspondentes a cada janela. Dentro de

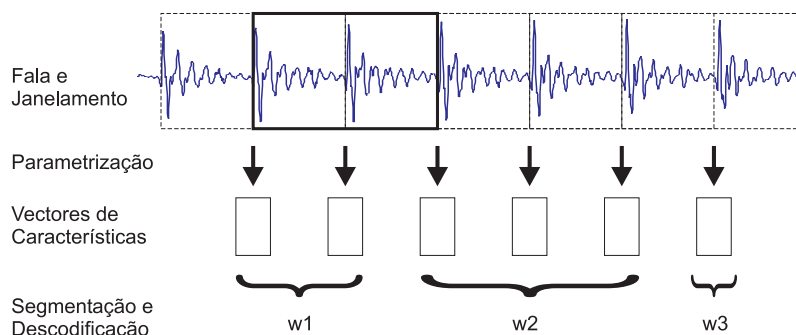


Figura 5.1: Processamento computacional de mensagens de fala

cada janela pode-se, com um erro em geral pequeno, partir do princípio que o sinal de fala é estacionário e que o vector que lhe corresponde descreve com rigor o seu comportamento.

Após a obtenção da sequência de vectores descritivos do sinal o principal objectivo é efectuar uma correspondência entre esta e uma sequência de símbolos final. Daqui surgem dois problemas: a classificação dos vectores de características e a segmentação por blocos que façam sentido do ponto de vista da saída do sistema.

O primeiro problema prende-se com a criação de uma correspondência entre uma determinada ocorrência acústica e um símbolo que o identifica. A dificuldade é grande pois o sinal acústico é extremamente variado e facilmente influenciado pelo ruído ambiente, orador, estado de espírito, etc. Para além disto podem existir sinais muito semelhantes com sons/símbolos correspondentes diferentes.

No caso de um sistema de reconhecimento automático de fala pode-se considerar que a principal questão a resolver é a da classificação, terminando as suas funções quando é encontrada uma sequência de símbolos representativos da mensagem acústica. Ainda que existam erros de classificação resultantes de um reconhecimento menos correcto, estes podem ser posteriormente corrigidos por sistemas baseados em regras (ex.: gramática) ou bases de dados (ex.: conjunto de nomes de cidades) e que assim melhoram o desempenho global do reconhecimento. No caso de um sistema de anotação automática em que apenas é fornecido o sinal de voz (ou os vectores de características correspondentes) surge ainda um segundo problema a resolver. É necessário efectuar a segmentação, ou seja, definir com precisão as fronteiras que limitam cada uma das palavras do léxico de trabalho. Por palavra entende-se, neste caso, um sinal que possui um símbolo associado que o representa podendo ser uma palavra propriamente dita, um fone ou qualquer outra unidade que se convencie. Nos sistemas de anotação automáticos, que constituem o tema a abordar aqui, as palavras correspondem a fones e estes possuem um símbolo fonético que os representa.

Muitas vezes a tarefa de anotação é simplificada pelo fornecimento da sequência de símbolos fonéticos a que corresponde o sinal de fala em conjunto com o próprio sinal. Deste modo elimina-se o primeiro problema e é apenas necessário efectuar a segmentação que neste caso se designa por alinhamento fonético.

5.1.1 Modelos de Markov

Para iniciar a descrição dos processos de Markov partir-se-á do princípio que se está a realizar o reconhecimento de palavras isoladas não se colocando o problema de segmentação ou de identificação de limites temporais. Considere-se então uma sequência de vectores de características a que se chamará temporariamente observações de forma a compatibilizar a explicação com a base probabilística da teoria. A cada observação O , constituída por uma sequência de vectores de características o_t surgidos nos instante t , corresponde uma palavra.

$$O = o_1, o_2, \dots, o_T \quad (5.1)$$

A solução do problema do reconhecimento passa então por encontrar a palavra w_i do vocabulário que mais se aproxima da ocorrência em análise, ou seja:

$$\arg \max_i \{P(w_i|O)\} \quad (5.2)$$

Esta probabilidade não pode ser resolvida directamente mas fica simplificada pela aplicação da regra de Bayes:

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)} \quad (5.3)$$

Assim, conhecendo à partida um conjunto de probabilidades $P(w_i)$ para cada palavra, seria possível unicamente a partir de $P(O|w_i)$ efectuar o cálculo completo. No entanto O é constituído por vários vectores de características e portanto multi-dimensional passando o cálculo a efectuar a ser o de uma probabilidade conjunta $P(o_1, o_2, \dots | w_i)$ que não é facilmente resolúvel nas condições apresentadas.

É aqui que os processos de Markov proporcionam uma ajuda inegável substituindo o cálculo da probabilidade pelo ajuste simples dos parâmetros de um modelo. Considere-se então um sistema que num instante de tempo discreto t está num estado S_i pertencente a um conjunto limitado de estados $1 \leq i \leq N$. O sistema pode mudar de um estado i para um outro estado j , que poderá ser o próprio estado, obedecendo a um conjunto de probabilidades a_{ij} . Os estados S_i possuem também uma probabilidade inicial π , ou seja, uma probabilidade de serem os primeiros estados sem predecessores de uma sequência. Estes pressupostos são os necessários para definir uma cadeia Markoviana de primeira ordem. Na figura 5.2 representa-se um modelo de Markov exemplificativo com apenas 3 estados.

Assim, com base num modelo λ qual será então a probabilidade de uma observação O ? Assumindo que o estado actual é representado por o_i tem-se:

$$P(O|\lambda) = P(o_1) \cdot \prod_{i=2}^T P(o_i|o_{i-1}) \quad (5.4)$$

Ou seja, a probabilidade de uma determinada sequência é dada pela multiplicação das probabilidades de transição entre estados. O primeiro estado não tendo nenhum precedente surge sozinho.

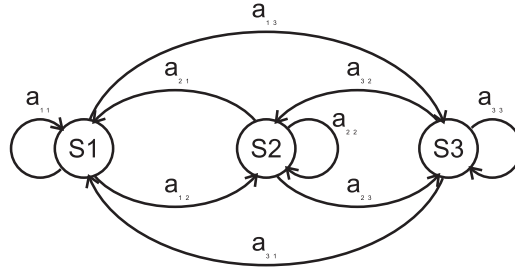


Figura 5.2: Modelo de uma cadeia de Markov com 3 estados

Para clarificar apresenta-se o exemplo do artigo clássico de Rabiner [66] que utiliza os estados $S_1 = \text{chuva}$, $S_2 = \text{nuvens}$ e $S_3 = \text{sol}$. A matriz de probabilidades de transição é dada abaixo (note-se que a soma das probabilidades de cada linha é 1):

$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.4 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix} \quad (5.5)$$

Sabendo que está Sol no primeiro dia qual será a probabilidade do tempo nos próximos 7 dias ser "sol, sol, chuva, chuva, sol, nuvens, sol"? De um modo mais formal, qual a probabilidade da observação corresponder à sequência $O = \{S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3\}$?

$$\begin{aligned}
 P(O|Modelo) &= P(S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3|Modelo) \\
 &= P(S_3) \cdot P(S_3|S_3) \cdot P(S_3|S_3) \cdot P(S_1|S_3) \cdot \dots \\
 &\quad \dots \cdot P(S_3|S_1) \cdot P(S_2|S_3) \cdot P(S_3|S_2) \\
 &= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23} \\
 &= 1 \cdot 0.8 \cdot 0.8 \cdot 0.1 \cdot 0.4 \cdot 0.3 \cdot 0.1 \cdot 0.2 \\
 &= 1.536 \times 10^{-4}
 \end{aligned}$$

Assim, e em tom de resumo, um modelo de Markov é um autómato de estados finito com transições de base estocástica em que a sequência de estados é uma cadeia de Markov. Um modelo de Markov fica totalmente definido a partir de:

- $A = \{a_{ij}\}$ - Matriz quadrada com as probabilidades de transição do estado i para o estado j , ou seja, $a_{ij} = P(S_t = j | S_{t-1} = i)$, $1 \leq i, j \leq N$
- $\Pi = \{\pi_i\}$ - Matriz coluna com as probabilidades iniciais do modelo começar no estado i , ou seja, $\pi_i = P(S_1 = i)$, $1 \leq i \leq N$

O número de estados fica implicitamente definido pela dimensão das matrizes referidas. Os modelos tem por objectivo a identificação de um padrão no tempo e baseia-se nos pressupostos de que o tempo é discreto bem como os estados e as probabilidades de transição de estado são constantes durante a vida do modelo.

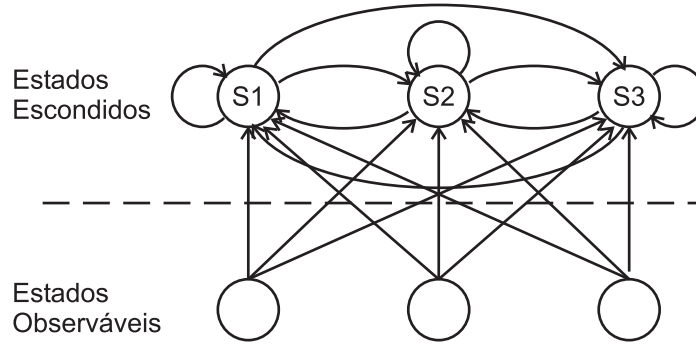


Figura 5.3: Funcionamento de um HMM

5.1.2 Modelos Escondidos de Markov

Utilizando o exemplo do tempo dado antes, imagine-se a situação de um alguém que está fechado num quarto sem acesso ao mundo exterior e apenas possui uma televisão onde pode ver a meteorologia. Neste caso, para quem está dentro do quarto, não existe a evidência de que o estado do tempo é o afirmado pela meteorologia, apenas existe uma possibilidade de o ser mas poderá ser qualquer outro estado. Se a esta possibilidade se puder associar uma base estatística é então possível definir um modelo escondido de Markov. Na figura 5.3 ilustra-se a situação referida.

Assim, nos modelos de Markov cada saída do modelo corresponde a um evento determinístico ao passo que nos modelos escondidos de Markov cada saída corresponde a uma função de densidade de probabilidade.

Um modelo escondido de Markov é então definido por:

- $A = \{a_{ij}\}$ - Matriz quadrada com as probabilidades de transição do estado i para o estado j , ou seja, $a_{ij} = P(S_t = j | S_{t-1} = i)$, $1 \leq i, j \leq N$
- $B = \{b_i(k)\}$ - Matriz confusão com a probabilidade de ocorrência de uma saída o_k quando se está no estado i , ou seja, $b_i(k) = P(X_t = o_k | S_t = i)$ em que $X = X_1, X_2, \dots, X_T$ é a sequência observada
- $\Pi = \{\pi_i\}$ - Matriz coluna com as probabilidades iniciais do modelo começar no estado i , ou seja, $\pi_i = P(S_1 = i)$, $1 \leq i \leq N$

É necessário ainda que se verifiquem as condições:

$$a_{ij} \geq 0, \quad b_i(k) \geq 0, \quad \pi_i \geq 0, \quad \forall i, j, k$$

$$\sum_{j=1}^N a_{ij} = 1, \quad \sum_{k=1}^M b_i(k) = 1, \quad \sum_{i=1}^N \pi_i = 1$$

A notação que se utiliza tradicionalmente para a representação de um modelo escondido de Markov Φ é:

$$\Phi = (A, B, \Pi) \quad (5.6)$$

Em resumo, um HMM é idêntico a um modelo de Markov tradicional cujos estados deixaram de ser observáveis mas que em contrapartida recebeu um conjunto de estados associados e observáveis que possuem uma relação probabilística com os que foram escondidos.

Os HMMS possuem para cada estado uma ou mais funções de densidade de probabilidade e podem segundo esta perspectiva classificar-se como discretos, contínuos e semi-contínuos.

Os discretos possuem na relação dos estados observáveis com os escondidos uma probabilidade discreta. Este tipo tem sido utilizado até aqui na apresentação dos HMMS pois permite mais facilmente criar analogias com situações da realidade. Para estes, e considerando já a possibilidade de processar informação de várias fontes (processamento *multistream*), a probabilidade $b_j(o_t)$ do estado j produzir a observação o_t é dada por:

$$b_j(o_t) = \prod_{s=1}^S \{P_{js}[v_s(o_{st})]\}^{\gamma_s} \quad (5.7)$$

Na expressão $v_s(o_{st})$ é a saída do quantizador vectorial para o *stream* s dado um vector de entrada o_{st} e $P_{js}[v]$ é a probabilidade do estado j gerar o símbolo v para o *stream* s .

No caso contínuo a função densidade de probabilidade é contínua optando-se quase sempre por uma distribuição gaussiana. A opção por este tipo de distribuição prende-se com o facto de a partir desta ser possível aproximar qualquer outra função densidade de probabilidade contínua. Quando se utilizam várias funções deste tipo é frequente utilizar-se a expressão mistura gaussianas (*gaussian mixtures*). As distribuições contínuas podem ser vistas como uma generalização das distribuições discretas. A mesma probabilidade calculada acima é neste caso, considerando M misturas, dada por:

$$b_j(o_t) = \prod_{s=1}^S \left[\sum_{m=1}^M c_{jms} N(o_{st}; \mu_{jms}, \Sigma_{jms}) \right]^{\gamma_s} \quad (5.8)$$

Aqui, as M misturas são relativas ao estado j para o *stream* s , c_{jms} é o peso da mistura m e $N(\cdot; \mu_{jms}, \Sigma_{jms})$ é uma distribuição Gaussiana multivariável com um vector média μ e uma matriz covariância Σ . Esta distribuição é:

$$N(o; \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} e^{-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)} \quad (5.9)$$

em que n é a dimensão de o .

Estes são potencialmente os mais precisos [59] pois não existe o erro associado à quantização do espaço de características de entrada. Uma outra vantagem é o facto de qualquer distribuição poder ser suavemente ajustada pelo aumento ou diminuição do número de componentes utilizadas nas misturas permitindo assim a utilização de distribuições gaussianas simples para modelos não ligados e com uma quantidade de dados de treino reduzida [59, 99].

Os semi-contínuos possuem distribuições que são comuns a vários estados [59] e segundo Strom [78] permitem que com um baixo número de características,

aproximadamente ortogonais se faça uma boa representação das propriedades estatísticas do modelos e dos métodos de treino. Assim e comparando com os HMMs contínuos, os semicontínuos permitem libertar parâmetros e reduzir a complexidade computacional devido à partilha de informação entre as funções densidade de probabilidade. Este tipo de HMMs possui a melhor relação de compromisso entre modelização e treinabilidade.

Da definição dos HMMs surgem três problemas implícitos [66, 38]:

- Problema da Avaliação, em que dado um modelo Φ e uma sequência de observações $X = (X_1, X_2, \dots, X_T)$ se procura calcular a probabilidade $P(X|\Phi)$, ou seja, a probabilidade de o modelo gerar as observações (idêntico ao reconhecimento de voz).
- Problema de Decodificação, em que dado um modelo Φ e uma sequência de observações $X = (X_1, X_2, \dots, X_T)$ se procura identificar qual a sequência de estados $S = (S_1, S_2, \dots, S_T)$ mais provável no modelo que produziu as observações.
- Problema da Aprendizagem, em que dado um modelo Φ e uma sequência de observações $X = (X_1, X_2, \dots, X_T)$ se procura ajustar os parâmetros do modelo (estimação de modelos a partir de dados de treino).

Todos estes problemas têm de ser resolvidos em qualquer sistema que utilize a teoria dos modelos escondidos de Markov. Cada um dos problemas possui múltiplas formas de abordagem e resolução, aqui abordar-se-ão os algoritmos mais populares e que encontram implementação em vários pacotes de software para processamento de fala.

5.1.3 Problema da Avaliação (*Forward Algorithm*)

Para resolver o primeiro problema, ou seja, para efectuar o cálculo de $P(X|\Phi)$ dada uma sequência de observações $X = (X_1, X_2, \dots, X_T)$, o método mais imediato é enunciar todas as sequências S de comprimento T e somar todas as suas probabilidades. A operação resume-se então à soma das probabilidades de todas as sequência de estado possíveis:

$$P(X|\Phi) = \sum_S P(S|\Phi)P(X|S, \Phi), \quad \forall S \quad (5.10)$$

A probabilidade da sequência S é o produto da probabilidade da sequência de estados pela probabilidade conjunta da saída ao longo da sequência.

O princípio de Markov (*Markov assumption*) para os modelos de primeira ordem dado pela equação 5.11 indica que a probabilidade de um estado apenas depende do último estado.

$$P(s_t|s_1^{t-1}) = P(s_t|s_{t-1}), \quad s_1^{t-1} = s_1, s_2, \dots, s_{t-1} \quad (5.11)$$

Aplicando este princípio, o primeiro factor da equação 5.10 pode ser reescrito na forma:

$$\begin{aligned} P(S|\Phi) &= P(s_1|\Phi) \prod_{t=2}^T P(s_t|s_{t-1}, \Phi) \\ &= \pi_{s_1} a_{s_1 s_2} \dots a_{s_{T-1} s_T} \end{aligned} \quad (5.12)$$

Do mesmo modo, o segundo factor fica:

$$\begin{aligned} P(X|S, \Phi) &= P(X_1^T | s_1^T, \Phi) = \prod_{t=1}^T P(X_t | s_t, \Phi) \\ &= b_{s_1}(X_1) b_{s_2}(X_2) \dots b_{s_T}(X_T) \end{aligned} \quad (5.13)$$

A substituição na equação 5.10 dos desenvolvimentos apresentados leva à escrita de:

$$P(X|\Phi) = \sum_S \pi_{s_1} b_{s_1}(X_1) a_{s_1 s_2} b_{s_2}(X_2) \dots a_{s_{T-1} s_T} b_{s_T}(X_T), \quad \forall S \quad (5.14)$$

Aqui vê-se claramente que a operação a efectuar é bastante simples mas obriga a um esforço computacional exigente e de crescimento exponencial. O algoritmo *forward* efectua este cálculo de um modo mais eficiente através do armazenamento de valores intermédios utilizados num ciclo recursivo.

Considere-se a probabilidade *forward* como :

$$\alpha_i(t) = P(X_1^t, s_t = i | \Phi) \quad (5.15)$$

Esta representa a probabilidade do modelo Φ com N estados no estado i no instante t ter gerado a observação parcial X_1^t em que $t \leq T$. O algoritmo *forward* é então constituído por três passos, inicialização, indução e conclusão, que se apresentam pela ordem respectiva:

$$\begin{aligned} \alpha_i(1) &= \pi_i b_i(X_1), \quad 1 \leq i \leq N \\ \alpha_j(t) &= \left[\sum_{i=1}^N \alpha_i(t-1) a_{ij} \right], \quad 2 \leq t \leq T; 1 \leq j \leq N \\ P(X|\Phi) &= \sum_{i=1}^N \alpha_i(T) \end{aligned}$$

Este algoritmo possui complexidade $O(N^2T)$ o que consegue ser bastante inferior na maioria das situações à complexidade $O(N^T)$ da formulação inicial que não explora a recursividade permitida.

5.1.4 Problema da Descodificação (Algoritmo de Viterbi)

O segundo problema da teoria dos modelos de Markov procura encontrar a sequência S^* mais provável dada uma determinada observação X dado um modelo Φ . O algoritmo *forward* calcula a probabilidade de um modelo gerar uma determinada sequência de observação através do cálculo das probabilidades de todas as sequências de estado (ou caminhos) possíveis. No caso da descodificação pretende-se calcular o melhor caminho. Este é um problema típico do reconhecimento contínuo de fala onde se pretende, com base num conjunto de modelos, encontrar a melhor sequência fonética a partir da observação de características acústicas. A técnica tradicional para resolver eficientemente este problema recorre ao algoritmo de Viterbi [89] baseado em programação dinâmica.

Considere-se $V_i(t)$ como a probabilidade do melhor caminho ou sequência de estados no tempo t ter resultado na observação X_1^t e ter terminado no estado i :

$$V_i(t) = P(X_1^t, S_1^{t-1}, s_t = i | \Phi) \quad (5.16)$$

O algoritmo de Viterbi resume-se em quatro passos descritos pelas equações:

- Inicialização

$$\begin{aligned} V_i(1) &= \pi_i b_i(X_1), \quad 1 \leq i \leq N \\ B_i(1) &= 0 \end{aligned}$$

- Indução

$$\begin{aligned} V_j(t) &= \max_{1 \leq i \leq N} [V_i(t-1) a_{ij}] b_j(X_t), \quad 2 \leq t \leq T; 1 \leq j \leq N \\ B_j(t) &= \arg \max_{1 \leq i \leq N} [V_i(t-1) a_{ij}], \quad 2 \leq t \leq T; 1 \leq j \leq N \end{aligned}$$

- Conclusão

$$\begin{aligned} \text{melhor} &= \max_{1 \leq i \leq N} [V_i(t)] \\ s_T^* &= \arg \max_{1 \leq i \leq N} [B_i(T)] \end{aligned}$$

- Recuo

$$\begin{aligned} s_t^* &= B_{s_{t+1}^*}(t+1), \quad t = T-1, T-2, \dots, 1 \\ S^* &= (s_1^*, s_2^*, \dots, s_T^*) \end{aligned}$$

A melhor sequência ou o melhor caminho é dado por S^* . À semelhança do algoritmo *forward* a complexidade do algoritmo de Viterbi é $O(N^2T)$.

Na figura 5.4 ilustra-se o funcionamento do algoritmo apresentado para um modelo com 6 estados. O modelo está adaptado para uma utilização com o software HTK no qual os primeiro e último estados são designados por *non-emitting states*. O espaço parametrizado está organizado em duas dimensões em que no eixo dos xx, com significação temporal, surgem as janelas sucessivas de fala (ou os seus vectores de características representativos do sinal) e no eixo dos yy têm-se os vários estados do HMM. Cada um dos pontos da matriz representa a probabilidade de observação daquela janela naquele instante (dada por b_s) e cada arco representa a probabilidade de transição de um estado para outro (dada por $a_{s_1 s_2}$). Devido à topologia do modelo associada ao eixo vertical e à dimensão temporal do eixo horizontal apenas se permite que a pesquisa do melhor percurso se efectue da esquerda para a direita e de baixo para cima. Em casos práticos todas as probabilidades referidas surgem na forma logarítmica pois deste modo é possível, explorando as propriedades dos logaritmos, substituir multiplicações por adições o que obviamente simplifica e acelera o desempenho dos algoritmos de cálculo.

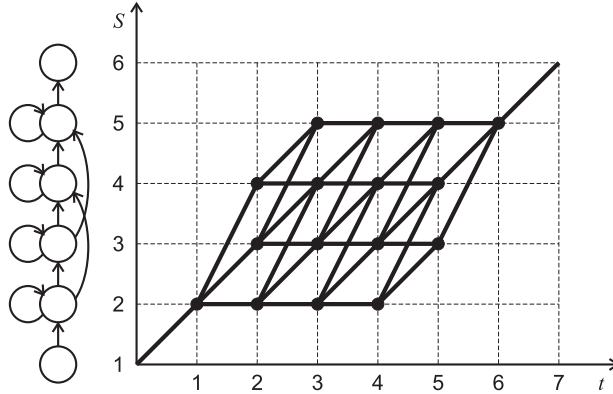


Figura 5.4: Funcionamento do algoritmo de Viterbi aplicado ao reconhecimento de fala com um modelo de Markov (adaptado de: HTK Book [98])

5.1.5 Problema da Aprendizagem (Algoritmo de Baum-Welch)

Resta agora dar resposta ao último e mais complexo problema intrínseco à teoria dos HMMs. Neste caso pretende-se estimar os parâmetros A , B e Π de um modelo Φ de forma a que este descreva o melhor possível as sequências de observações.

Para solucionar este problema não é conhecido nenhum método analítico que possa ser equacionado e que permita maximizar as probabilidades conjuntas resultantes da análise dos dados de treino. No entanto o cálculo pode ser realizado pelo algoritmo iterativo de Baum-Welch também conhecido como algoritmo *forward-backward*.

Defina-se a probabilidade *backward* analogamente ao que se fez para a probabilidade *forward*:

$$\beta_i(t) = P(X_{t+1}^T, s_t = i | \Phi) \quad (5.17)$$

Aqui $\beta_i(t)$ é a probabilidade do modelo Φ no estado i no instante t ter gerado a observação parcial X_{t+1}^T (de $t+1$ até ao final). Esta probabilidade pode ser calculada recursivamente em dois passos:

- Inicialização

$$\beta_i(T) = 1/N, \quad 1 \leq i \leq N \quad (5.18)$$

- Indução

$$\beta_i(t) = \left[\sum_{j=1}^N a_{ij} b_j(X_{t+1}) \beta_j(t+1) \right], \quad t = T-1, \dots, 1; 1 \leq i \leq N \quad (5.19)$$

Defina-se também $\gamma_t(i, j)$ (note-se que aqui, por uma questão de organização, se alterou ligeiramente a notação que tem vindo a ser utilizada passando o tempo

a ser representado em índice) que é a probabilidade de levar a transição do estado i para o estado j no instante t dado o modelo Φ e a sequência observada X .

$$\begin{aligned}\gamma_t(i, j) &= P(s_{t-1} = i, s_t = j | X_1^T, \Phi) \\ &= \frac{P(s_{t-1} = i, s_t = j, X_1^T | \Phi)}{P(X_1^T | \Phi)} \\ &= \frac{\alpha_{t-1}(i) a_{ij} b_j(X_t) \beta_t(j)}{\sum_{k=1}^N \alpha_T(k)}\end{aligned}$$

Uma vez que o objectivo é efectuar o refinamento dos parâmetros do modelo ao longo de várias iterações define-se ainda um novo modelo $\hat{\Phi} = (\hat{A}, \hat{B}, \hat{\Pi})$ que resulta do modelo de Φ .

Com isto é possível formular as equações que permitem criar uma estimativa para o modelo:

$$\hat{a}_{ij} = \frac{\frac{1}{P(X|\Phi)} \sum_{t=1}^T P(X | s_{t-1} = i, s_t = j | \Phi)}{\frac{1}{P(X|\Phi)} \sum_{t=1}^T P(X | s_{t-1} = i | \Phi)} = \frac{\sum_{t=1}^T \gamma_t(i, j)}{\sum_{t=1}^T \sum_{k=1}^N \gamma_t(i, k)} \quad (5.20)$$

$$\hat{b}_j(k) = \frac{\frac{1}{P(X|\Phi)} \sum_{t=1}^T P(X | s_t = j | \Phi) \delta(X_t, o_k)}{\frac{1}{P(X|\Phi)} \sum_{t=1}^T P(X | s_t = j | \Phi)} = \frac{\sum_{t \in X_t = o_k} \sum_i \gamma_t(i, j)}{\sum_{t=1}^T \sum_i \gamma_t(i, j)} \quad (5.21)$$

A equação 5.20 representa basicamente a relação entre o número esperado de transições do estado i para o estado j e o número esperado de transições a partir do estado i . A equação 5.21 é o quociente entre o número esperado de vezes que os dados observados originários do estado j são o símbolo o_k e o número esperado de vezes que os dados observados são originários do estado j .

Importante também é a definição da função auxiliar Q , utilizada por [38], e que é de um modo simples uma forma de avaliar a proximidade entre dois modelos.

É agora possível descrever o algoritmo de Baum-Welch que se resume em quatro passos:

1. Encontrar uma estimativa inicial para o modelo Φ
2. Calcular a função $Q(\Phi, \hat{\Phi})$
3. Calcular os parâmetros do novo modelo $\hat{\Phi}$ de acordo com as equações de estimação 5.20 e 5.21
4. Caso não se tenha atingido a convergência desejada, fazer $\Phi = \hat{\Phi}$ e regressar ao passo 2.

5.1.6 Limitações dos HMMs

Os modelos escondidos de Markov já demonstraram que são uma boa técnica para a modelização de sinais de fala pois permitem, com bom desempenho, descrever simultaneamente a variabilidade no tempo e no espaço de características.

O facto de se poderem treinar e ajustar de forma a refinar a qualidade do modelo e a sua base estocástica que possibilita a combinação de diversas fontes de informação são também factores importantes que contribuem para a grande implantação dos HMMS.

Porém, a teoria dos modelos de Markov possui duas características menos interessantes que acabam por limitar a sua utilização. A primeira limitação, concernente aos HMMS de primeira ordem, refere-se aos pressupostos de independência condicional que indicam que a passagem para um determinado estado apenas depende do estado anterior. Esta não é uma limitação dramática pois é possível definir HMMS de ordem superior que aumentam o grau de dependência em relação a estados anteriores. Este aumento de ordem tem o inconveniente de aumentar também as exigências computacionais de forma bastante superior. A segunda limitação é mais forte e prende-se com o facto de os HMMS serem apenas função de uma variável independente como o tempo ou um espaço unidimensional. A solução para este problema, mais difícil de resolver, passa muitas vezes pela utilização de uma transformada que permita combinar numa única dimensão informações de várias dimensões (Esta técnica é utilizada por exemplo para o reconhecimento óptico de caracteres).

5.2 Hidden Markov Model Toolkit

Inicialmente desenvolvido em 1989 pelo *Speech Vision and Robotics Group* da Universidade de Cambridge para utilização própria, o HTK, depressa se mostrou de interesse público e começou a ser distribuído sob licença às faculdades e instituições de ensino. Devido ao grande aumento do número de utilizadores foi criada a Entropics, empresa parceira da universidade, que passou a comercializar parte das ferramentas. Pelo interesse que o software mostrava ter, em 1999, a Microsoft adquiriu a Entropics mas o *core-business* da empresa foi-se alterando sucessivamente. Em Setembro de 2000, a Microsoft decide tornar novamente livre a utilização do software e o código fonte passou a estar disponível sítio web do Departamento de Engenharia da Universidade de Cambridge¹.

Actualmente, as referências ao pacote de *software* HTK (*Hidden Markov Model Toolkit*) são incontornáveis na área do processamento de voz e o seu mérito é amplamente reconhecido. Este conjunto de aplicações disponibiliza de forma modular ferramentas que permitem implementar rapidamente sistemas baseados em modelos escondidos de Markov. Existem módulos para a criação de vectores de características, treino dos modelos, pesquisa Viterbi e outros que cobrem a maioria das necessidades.

5.2.1 Motivações

A utilização do HTK não seria um opção obrigatória para o desenvolvimento do sistema que aqui se propõe. Existem outras ferramentas e pacotes de software, desenvolvidos por empresas ou laboratório académicos, que poderiam servir de base ao etiquetador automático. O software Matlab², já com vários algoritmos implementados, foi uma opção considerada. Os conjunto de rotinas disponíveis

¹CUED - <http://htk.eng.cam.ac.uk>

²Mathworks: <http://www.mathworks.com>

no Instituto de Tecnologia de Nagoya³ e no CSLU⁴ (*Center for Spoken Language Understanding*), algumas complementares ou variações do HTK, foram também analisadas. No entanto a utilização do HTK mostrou-se sempre mais adequada e favorável.

O HTK foi desenvolvido especificamente para o processamento de sinais de fala e possui a maioria das ferramentas necessárias para a implementação completa de um sistema deste tipo. O facto de ser baseado em módulos independentes que correm na linha de comando oferece uma grande flexibilidade e facilita a sua integração com outras aplicações de nível superior. Os algoritmos mais utilizados são implementados por programadores profissionais e recorrendo a diversas técnicas de optimização que permitem um desempenho muito superior. Para além disto, o código fonte completo está disponível na internet o que permite a criação de novas ferramentas sem ser necessário o desenvolvimento de raiz. Este desenvolvimento, completo, sem recorrer a ferramentas já existentes, seria demorado, consumidor de tempo e sem dúvida um desvio dos objectivos deste trabalho.

Ainda importante é a possibilidade de comparação de resultados com outros sistemas idênticos. Devido à grande popularidade e difusão do HTK uma boa parte dos resultados que surgem em comunicações científicas é baseada em sistemas que utilizam este pacote de ferramentas. Deste modo, existindo a garantia de que a base de trabalho é exactamente igual, é possível comparar com maior rigor os trabalhos que vão surgindo.

Assim e pelas razões apresentadas utilizaram-se para o desenvolvimento do etiquetador aqui proposto as ferramentas disponibilizadas pelo HTK versão 3.2.1.

5.2.2 Funcionamento

A maioria das ferramentas do HTK não possuem interface gráfica e a sua utilização é realizada a partir da linha de comandos o que permite o armazenamento de sequências de comandos e a criação de *scripts* (realizadas por exemplo em Perl) para a automatização de procedimentos. Cada ferramenta possui um conjunto de parâmetros obrigatórios e permite a alteração do seu modo de funcionamento a partir de alguns parâmetros opcionais.

Um comando típico poderia ser:

```
Hxyz -T 1 -f 12.3 -a -s ficheiro1 ficheiro2
```

A primeira palavra da linha será o nome da ferramenta a utilizar que começa sempre pelo carácter "H" (no exemplo dado o nome da ferramenta não tem significado) e vem seguida de um conjunto de parâmetros. Os parâmetros opcionais são identificados por um hífen antes de um carácter identificador e, quando necessário, são procedidos dos novos valores a atribuir às variáveis separados por um espaço. Existem parâmetros específicos que apenas dizem respeito a uma determinada ferramenta e outros que possuem funcionalidades comuns às várias ferramentas. Estes últimos são identificados sempre pelo mesmo carácter maiúsculo. Para reduzir a dimensão das linhas de comando e facilitar uma utilização mais intensiva é possível, através da opção -C, especificar um ficheiro

³SPTK (*Speech Signal Processing Toolkit*): <http://kt-lab.ics.nitech.ac.jp/tokuda/SPTK/>

⁴Oregon Graduate Institute-CSLU: <http://cslu.cse.ogi.edu>

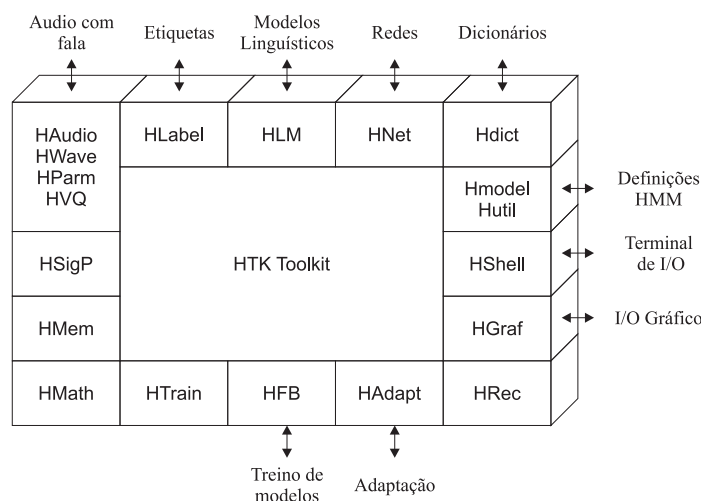


Figura 5.5: Arquitectura das ferramentas HTK (fonte: HTK Book [98])

externo onde estão armazenados conjuntos frequentes de parâmetros de configuração.

A arquitectura base por trás de cada ferramenta HTK surge esquematizada na figura 5.5. Todas as ferramentas partilham um conjunto de rotinas que possuem funcionalidades comuns e fornecem assim um comportamento global homogéneo. As interacções com o sistema operativo são geridas pelo HShell e as operações de memória recorrem ao HMem. As operações matemáticas estão a cargo do HMath e o processamento de sinal utilizam o HSigP. Existem ainda módulos para o processamento de cada tipo de ficheiro, HLabel para os ficheiros com anotações, HLM para os modelos de linguagem, HModel para os modelos de HMMs, etc. O comportamento de cada uma destas rotinas pode também ser alterado pela manipulação de parâmetros de configuração.

Será agora feita uma breve apresentação das ferramentas fundamentais do HTK guiada pelas orientações dos seus autores [98] e seguindo uma sequência constituída por quatro fases: preparação dos dados, treino dos modelos, segmentação e anotação e finalmente análise.

5.2.3 Ferramentas de Preparação

Antes de qualquer operação de processamento é necessário organizar os dados que permitirão treinar e testar o sistema. Para o treino, estes são constituídos geralmente por ficheiros de áudio contendo fala e em anexo ficheiros com as anotações respectivas. No caso de não estarem disponíveis estas informações é possível utilizar o HSLab que permite efectuar a gravação do áudio e a etiquetagem manual do mesmo. Neste trabalho utilizaram-se exclusivamente bases de dados já prontas não sendo por isso necessário efectuar esta tarefa.

Com os dados reunidos passa-se para a conversão destes em vectores de características para que possam ser utilizados pelas ferramentas seguintes. Para o efeito existe a ferramenta HCopy que para além das operações tradicionais de

janelamento, divisão em blocos e concatenação permite ainda o cálculo das *features* mais populares como os MFCCs, PLPs, LPCs e os respectivos parâmetros dinâmicos. Todos os pormenores podem ser ajustados pela configuração dos parâmetros apropriados.

A ferramenta HList permite visualizar o conteúdo dos ficheiros gerados pelo HCopy. Aqui é possível analisar os vectores de características, o seu conteúdo e a forma como foi efectuado o janelamento e o armazenamento dos valores calculados. Com esta análise breve é possível detectar eventuais erros antes de processar um grande volume de dados.

Para a manipulação dos ficheiros com anotações existe o HLed. Este é um poderoso editor de etiquetas baseado em *scripts* simples e que permite por exemplo a conversão entre conjuntos de *labels*. Esta operação é quase sempre necessária pois o HTK não permite grande flexibilidade nos nomes atribuídos às etiquetas e efectua a contagem do tempo em unidades de $100ns$, bastante invulgar se comparado com outros *softwares*. A ferramenta HLed é também importante na manipulação dos ficheiros de anotação e permite realizar importantes tarefas na conversão de um sistema de modo a considerar o contexto. Os ficheiros gerados podem ser independentes ou compilados num ficheiro global, com extensão MLF (*Master Label File*), que oferece algumas vantagens no processamento posterior.

Por último existe o HLStat que produz estatísticas a partir dos ficheiros com as etiquetas e o HQuant que permite a construção de *codebook* para a construção de HMMs discretos.

5.2.4 Ferramentas de Treino

Tendo toda a informação recolhida e pronta para ser utilizada é possível iniciar a preparação do treino dos modelos. É necessário definir a topologia de cada HMM o que pode ser feito num qualquer editor de texto. Nesta fase, criam-se apenas HMMs protótipo, cujos valores especificados são ignorados sendo apenas aproveitada a sua presença como forma de definição da arquitectura. A excepção é feita aos valores das transições de estado cujos valores serão analisados mas que podem, inicialmente, possuir valores idênticos (mas cuja soma seja 1).

Para o processo de treino podem ser seguidos dois caminhos distintos. Se existirem, para além do áudio, ficheiros etiquetados contendo as fronteiras temporais de cada palavra então pode utilizar um processo de *bootstrap*. Neste caso utilizam-se as ferramentas HInit e HRest. A primeira isola todas as ocorrências de dados correspondentes ao modelo a ser treinado e calcula iterativamente uma primeira aproximação aos parâmetros respectivos através de um procedimento *segmental k-means*⁵. Na primeira iteração é feita uma segmentação uniforme, cada estado do modelo é comparado com os segmentos de dados correspondentes e com esta informação calculam-se estimativas da média e variância. No caso de se utilizarem modelos com misturas gaussianas é utilizada uma técnica semelhante mas com algumas modificações. Nas iterações seguintes a segmentação é substituída por um alinhamento de Viterbi. Os parâmetros iniciais calculados pelo HInit são depois re-estimados pelo HRest. Aqui são utilizados novamente os dados de *bootstrap* mas o procedimento *segmental k-means* é substituído pela re-estimação Baum-Welch. Um outro caminho pode ser seguido quando não

⁵Consiste em classificar n ocorrências em q conjuntos minimizando a soma do quadrado das distâncias de cada ocorrência de cada conjunto ao centro do conjunto

existe qualquer etiquetagem dos ficheiros de áudio. Neste caso opta-se por um processo de *flatstart* no qual todos os modelos possuem parâmetros idênticos que resultam das médias e variâncias calculadas para o áudio global. Para esta situação pode-se recorrer à ferramenta HCompV que no trabalho presente não foi utilizada por não se verificarem as condições assinaladas.

Possuindo um conjunto inicial de valores para os modelos é possível recorrer ao HERest para realizar o treino embebido utilizando a totalidade dos dados de treino. Esta ferramenta efectua uma re-estimação Baum-Welch de todo o conjunto de modelos simultaneamente. Para cada ocorrência, os modelos correspondentes são concatenados e seguidamente é utilizado o algoritmo *forward-backward* para acumular valores estatísticos representativos para cada HMM na sequência. Após o processamento de todos os dados de treino as estatísticas acumuladas são utilizadas para re-estimar os parâmetros dos modelos. Esta é a ferramenta central do HTK estando projectada para processar grandes volumes de dados e para realizar processamento paralelo com várias máquinas em rede.

O HTK foi desenvolvido com o objectivo de permitir o refinamento gradual do conjunto de modelos melhorando assim sucessivamente o desempenho do sistema. Inicialmente começa-se com um conjunto de modelos sem misturas e independentes do contexto e, por iterações, atinge-se um sistema com modelos dependentes do contexto e com várias misturas gaussianas com um desempenho optimizado. A ferramenta HHed é um editor de parâmetros dos HMMS que permite efectuar a clonagem dos modelos isolados para modelos dependentes do contexto. É também possível criar diversos tipos de ligações entre modelos e incrementar o número de misturas. Isto é feito habitualmente pela repetição sucessiva de alterações com o HHed seguida de re-estimação com o HERest. A ferramenta HHed utiliza também *scripts* com comandos próprios para a execução das operações. Para melhorar o desempenho em aplicações orientadas para um determinado utilizador existem as ferramentas HEAdapt e HVite que permitem criar modelos de boa qualidade quando existem poucos dados de treino.

Para a construção de modelos dependentes do contexto é necessário possuir uma maior quantidade de dados de treino devido à maior complexidade do sistema e maior número de situações diferentes envolvidas. O aumento da complexidade de um modelo obriga a que exista uma maior quantidade de informação de treino para conseguir um conjunto robusto de parâmetros. Como os dados de treino são sempre limitados deve-se procurar conseguir obter um compromisso entre quantidade e qualidade que vá ao encontro dos resultados esperados. Para sistemas que utilizem distribuições contínuas é possível ligar alguns parâmetros de modo a partilhar valores que serão também estimados com qualidade. Para sistema com misturas ou com probabilidades discretas é também possível realizar a ligação e partilha de parâmetros. Neste caso, a insuficiência de dados é atenuada por uma suavização das distribuições realizada com o HSmooth.

5.2.5 Ferramentas de Segmentação e Anotação

Para estas tarefas existe apenas a ferramenta HVite que se baseia na técnica Viterbi para efectuar uma tarefa que genericamente será a de reconhecimento de fala. Para o efeito é carregada uma rede que descreve as sequências de palavras permitidas, um dicionário que contém a forma como as palavras são

pronunciadas (no caso das palavras a identificar serem fonemas a pronúncia será o próprio símbolo fonético) e um conjunto de HMMs. A partir daqui pode ser efectuado o reconhecimento a partir de ficheiros ou directamente a partir de áudio directo. Existem várias possibilidades de configuração do sistema podendo ser realizado por exemplo um alinhamento forçado.

As redes de palavras necessárias utilizadas pelo HVite são geralmente *loops* simples de palavras, em que a seguir a uma palavra pode ocorrer uma qualquer outra, ou podem ser baseadas em gramáticas definidas por grafos que limitam as possibilidades de escolha e às quais se associam probabilidades *bi-gram*. As redes ou gramáticas são escritas num ficheiro de texto e obedecem a uma especificação própria do HTK. Para auxiliar no desenvolvimento destas existe a ferramenta HBuild que possibilita a criação de sub-redes e a sua utilização dentro de redes de mais alto nível. Deste modo evitam-se duplicações pois utiliza-se o mesmo baixo nível de notação. Existe ainda a possibilidade de aproveitar os resultados gerados pelo HLStats, estatísticas de palavras que podem constituir um modelo linguístico designado por *backed-off bi-gram*, para modificar as transições dos *loops* de palavras.

Alternativamente está disponível a facilidade de interpretação de gramáticas especificadas utilizando uma notação BNF estendida (*Extended Backus Naur Form*) de mais alto nível e mais fácil de compreender.

Para analisar a rede criada existe a ferramenta HSGen que produz exemplos de linguagem a partir das ligações definidas na gramática. A ferramenta HSGen carrega a rede e percorre-a aleatoriamente, baseado-se nas ligações e probabilidades e produz cadeias de palavras. A perplexidade empírica da tarefa pode também ser avaliada deste modo.

Para dicionários de grandes dimensões existe ainda a ferramenta HDMan que facilita a gestão de várias fontes e permite compilações e cortes.

5.2.6 Ferramentas de Análise

Por último, depois de o sistema realizar todo o processamento que lhe compete, é necessário avaliar o seu desempenho. Para o efeito utiliza-se um conjunto de dados independente do utilizado para o treino mas igualmente etiquetado. A avaliação é realizada pela ferramenta HResult que efectua a comparação da etiquetagem já existente com a produzida pelo sistema em análise. Esta ferramenta utiliza um algoritmo de programação dinâmica que alinha ambos os conjuntos de etiquetas e efectua a contabilização dos casos correctos e das situações de erro (substituições, inserções e eliminações). Para além das métricas de desempenho características do HTK é ainda possível que os resultados sejam apresentados segundo as regras do *US National Institute of Standards and Technology (NIST)*, comuns na literatura científica. Existem ainda outras métricas parciais ou outras análises como o cálculo de figuras de mérito (*figures of merit - FOM*) ou *Receiver Operating Curves (ROC)*.

Esta tarefa é realizada apenas para as palavras pois o HTK destina-se essencialmente ao reconhecimento. Para a avaliação da qualidade das fronteiras geradas desenvolveu-se um outro *software (HTKFace)* que oferece, entre outras funcionalidades, a possibilidade de contabilizar o mesmo tipo de erros encontrados pelo HResults.

5.3 Nova Ferramenta Denominada HTKFace

A utilização frequente de ferramentas de software inteiramente baseadas na linha de comandos para a realização de vários ensaios com múltiplas configurações pode ser bastante demorada. Para além disto é necessário realizar algum tipo de memorização dos caracteres representativos dos parâmetros mais comuns para que se possam escrever os comandos com alguma eficiência. O HTK, como já se referiu, permite a utilização de ficheiros externos onde se podem armazenar os parâmetros de configuração utilizados. Ainda assim, quando se estão a testar várias configurações em várias ferramentas com parâmetros distintos é necessário consultar com frequência a ajuda. Os ecrans de ajuda das ferramentas do HTK possuem uma apresentação sempre semelhante e um exemplo, para o HVite, mostra-se abaixo:

```
D:\HTK>hinit
```

```
USAGE: HInit [options] hmmFile trainFiles...
```

Option		Default
-e f	Set convergence factor epsilon	1.0E-4
-i N	Set max iterations to N	20
-l s	Set segment label to s	none
-m N	Set min segments needed	3
-n	Update hmm (suppress uniform seg)	off
-o fn	Store new hmm def in fn (name only)	outDir/srcfn
-u mvwt	Update m)ean v)ar s)ig)hts t)rans	mvwt
-v f	Set minimum variance to f	1.0E-2
-w f	set mix wt/disc prob floor to f	0.0
-A	Print command line arguments	off
-B	Save HMMs/transforms as binary	off
-C cf	Set config file to cf	default
-D	Display configuration variables	off
-F fmt	Set source data format to fmt	as config
-G fmt	Set source label format to fmt	as config
-H mmf	Load HMM macro file mmf	
-I mlf	Load master label file mlf	
-L dir	Set input label (or net) dir	current
-M dir	Dir to write HMM macro files	current
-S f	Set script file to f	none
-T N	Set trace flags to N	0
-V	Print version information	off
-X ext	Set input label (or net) file ext	lab

Após a consulta frequente destas ajudas, tendo em conta a forma organizada de apresentação da ajuda e sabendo quais os parâmetros específicos das ferramentas e os parâmetros comuns (distinguidos por maiúsculas) decidiu-se desenvolver uma aplicação que facilite a utilização das ferramentas do HTK. Esta aplicação, baptizada de HTKFace, foi integralmente desenvolvida em ambiente Windows utilizando a linguagem Borland Delphi 5. A caixa de diálogo utilizada para todas as ferramentas é apenas uma, sendo as opções correspondentes lidas e colocadas na interface automaticamente a partir da análise da ajuda dessa ferramenta. Um exemplo, também para o HInit, mostra-se na figura 5.6. As

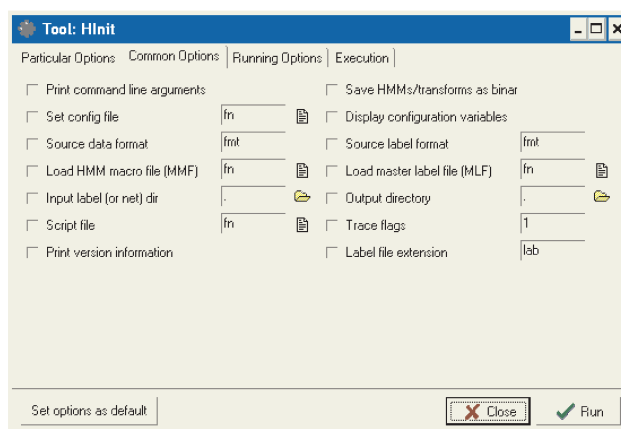


Figura 5.6: Utilização da ferramenta HInit com a aplicação HTKFace

opções são divididas em particulares e comuns, é analisada a existência ou não de parâmetros, que levará à criação ou não de uma caixa para introdução de dados, o conteúdo desta caixa será igual ao existente na coluna *default*. Sempre que surge a palavra *dir* é adicionado um botão que permite a selecção de uma pasta do sistema de arquivos, para as palavras *mmf*, *mlf* e *ext* é adicionado um botão que permite a selecção de um ficheiro e para as restantes palavras nenhum botão é acrescentado.

O conjunto de opções pode deste modo ser facilmente ajustado uma vez que todas as possibilidades se encontram visíveis acompanhadas das respectivas descrições. A existência de botões permite a referência rápida a ficheiro e directórios. Todas as opções que são configuradas podem também ser guardadas num ficheiro para serem posteriormente analisadas ou re-utilizadas. A activação de opções da interface ou o preenchimento de campos reflecte-se numa linha de texto onde vai sendo construído o comando que será executado pelo sistema operativo. No final, depois de parametrizado o comportamento desejado, possui-se um comando completo construído automaticamente que pode ser executado pela activação do botão *Run* no separador *Execution*. A execução da mesma ferramenta múltiplas vezes com a variação de apenas um parâmetro é também possível, no separador *Running Options*, pela introdução de uma lista de valores e pela selecção do parâmetro a variar.

A aplicação desenvolvida permite a utilização das ferramentas do HTK de modo bastante simples e facilitado e foi útil para a maioria das situações necessárias. Para além das funcionalidades apresentadas foram acrescentadas outras que se enquadram também no desenvolvimento do sistema de etiquetagem automática com HMMs e que surgiram de várias necessidades a que foi necessário dar resposta.

Um dos primeiros problemas a ser identificado foi a diferença entre a forma como as informações de etiquetagem estavam armazenadas nos corpora utilizados e o formato utilizado pelo HTK. Para o corpus FEUP/IPB, por exemplo, foram detectadas discrepâncias na unidade de representação dos tempos, na organização e quantidade de informação por linha de texto e no tipo de nomes de etiquetas admitidos. O mesmo aconteceu para a base de dados TIMIT que,

FEUP / IPB	TIMIT	HTK
0.00320 XX	512 10150 h#	32000 6344000 sp
0.63440 "s	10150 11245 s	6344000 7028000 s
0.70280 E	11245 13210 eh	7028000 8256500 em
0.82565 !	13210 14762 tcl	8256500 9226500 ocl
0.92265 t	14762 17323 t	9226500 10827000 t
1.08270 X	17323 21784 h#	10827000 13615000 sil
1.36150 XX	21784 26699 zz	13615000 16687000 asp
1.66870 u	26699 27766 u	16687000 17354000 u

Tabela 5.1: Formatos de armazenamento de etiquetas

apesar de possuir a mesma informação por linha que o HTK, utiliza números de amostras como medida de posicionamento no áudio o que implica também alguma conversão. Os nomes atribuídos aos fonemas também não são os mais adequados para utilizar com o HTK. Na tabela 5.1 apresentam-se três pequenos exemplos que espelham claramente as diferenças entre os formatos de registo das informações de etiquetagem (existe correspondência entre os valores para o posicionamento dentro do áudio mas o mesmo não se passa para os nomes das etiquetas que são apenas indicativos).

O criação de um formato de etiquetagem adequado ao HTK foi uma das primeiras questões que se colocou e que levou ao desenvolvimento de uma aplicação. Para este efeito poderia ter sido utilizada a ferramenta HLed mas ainda assim optou-se por desenvolver um executável específico. A linguagem de programação a utilizar foi ponderada. C++ e Java são linguagens bastante comuns e o código poderia ser re-utilizado noutras situações. Matlab, seria um pouco mais complicado mas facilitaria a integração com outras ferramentas disponíveis para cálculo estatístico e a realização de gráficos. Por último analisou-se o Delphi que acabou por ser a linguagem de eleição. Pela sua simplicidade na programação, pela rapidez com que se constroem interfaces gráficas de boa qualidade (que foram também aqui previstas) e por existirem já disponíveis rotinas específicas para o processamento de sinal esta foi a escolha final.

Em Delphi 5 desenvolveu-se um conjunto de rotinas que permitem a conversão entre vários formatos de etiquetas bastando para isso especificar o formato inicial, o formato final e uma tabela de equivalências dos símbolos fonéticos. Devido a problemas de limitação de memória causados pela grande dimensão dos ficheiros de áudio da base de dados FEUP/IPB foi também acrescentada a possibilidade de partir um ficheiro de etiquetas e o respectivo áudio em vários sub-ficheiros. Neste caso o processo é realizado sobre um único formato de dados e basta especificar os pontos de corte em segundos ou em amostras. O áudio é segmentado bem como as etiquetas respectivas. Nestas é retirado a cada valor de tempo o valor inicial de modo que a contagem temporal em cada sub-ficheiro comece sempre a partir do zero. O conversor de formatos e segmentador de etiquetas foram assim as primeiras ferramentas do que viria a ser o HTKFace.

Uma outra necessidade que surgiu foi a de criação de modelos para os HMMS. Estes são armazenados em ficheiros de texto e as regras de definição são bastante simples. A construção dos modelos é igualmente simples mas obriga a definir vários parâmetros que têm de ser frequentemente repetidos um número específico de vezes. A alteração manual destes parâmetros é bastante tediosa

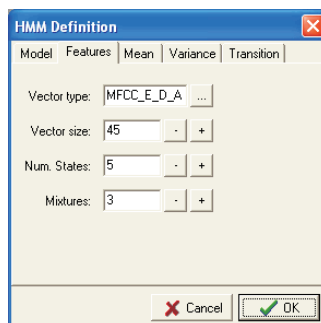


Figura 5.7: Definições principais para os HMMs com a aplicação HTKFace

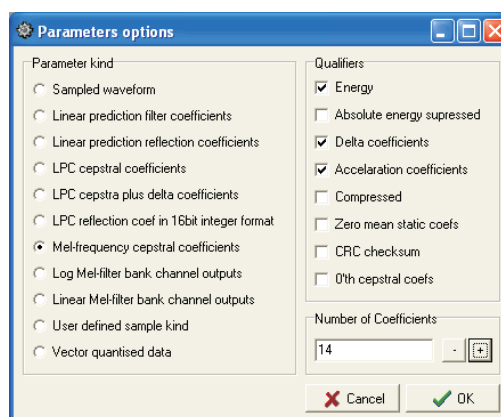


Figura 5.8: Definição das características a utilizar com a aplicação HTKFace

quando se realizam vários ensaios. Além disso, à medida que cresce a complexidade do modelo cresce também o número de parâmetros. Assim, e procurando facilitar também esta tarefa, desenvolveu-se uma pequena *interface* com muito pouco código extra que permite desenvolver rapidamente um modelo à medida das necessidades. Na figura 5.7 apresenta-se o aspecto da caixa de diálogo que permite aceder às funcionalidades referidas. É possível especificar completamente as configurações mais habituais manipulando facilmente o número de estados, número e tipo de misturas, alteração dos vectores de características e ajuste de médias, variâncias, matriz de transição, etc. Adicionalmente foi ainda acrescentada a possibilidade de criar de uma só vez vários modelos com a mesma configuração (situação bastante comum).

Por fim, e já numa fase mais avançada do desenvolvimento do sistema de etiquetagem, foi necessário criar novas ferramentas de análise de resultados que permitissem o cálculo de alguns indicadores que se consideraram importantes e que ao mesmo tempo permitissem visualizar mais informação em simultâneo. Nasceu então uma nova caixa de diálogo que permite carregar um sinal em formato WAV e n ficheiros de etiquetas que serão depois apresentados simultaneamente. A visualização possui os habituais comandos de deslocamento no tempo e *zoom*. Para a análise das tarefas de segmentação e anotação criaram-se vários procedimentos que calculam as métricas mais frequentes e que apresen-

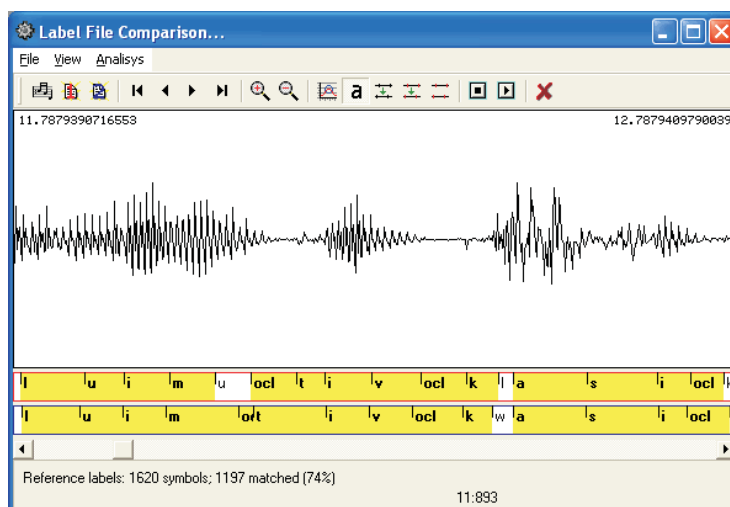


Figura 5.9: Comparação de dois ficheiros com etiquetas acompanhados do sinal de voz correspondente na aplicação HTKFace

tam graficamente as várias situações. A caixa de diálogo principal surge na figura 5.9.

Foram ainda adicionadas funcionalidades gerais que permitem a criação de *scripts*, a organização dos vários ficheiros em projectos e que se ajustam a uma estrutura de directórios própria, e outras, menores, que foram sendo desenvolvidas à medida das necessidades que surgiam. O produção de uma interface gráfica para cada funcionalidade foi sempre ponderada, pesando-se o benefício conseguido e o esforço e tempo a despendar. Em muitas situações e sempre que não foram encontradas mais valias significativas foi utilizado o HTK directamente na linha de comandos ou numa janela que permite esta facilidade criada no HTKFace.

Para além do referido existem ainda outras características importantes. Para não alargar a descrição do sistema resumem-se rapidamente as principais funções da aplicação desenvolvida HTKFace:

- **Plataformas:** Windows e facilmente portátil para Linux
- **Análise de fala:** espectrograma, energia e frequência fundamental.
- **Anotação:** compatível com vários formatos de etiquetas, etiquetagem de intervalos e marcas temporais em vários níveis, utilização de alfabetos fonéticos, comparação de alinhamentos utilizando várias métricas sob as perspectivas de segmentação e classificação e com apresentação de resultados numéricos e gráficos;
- **Manipulação de áudio:** Manipulação directa de áudio sem limite de duração, *zoom* e deslocamento temporal, selecção de blocos e audição;
- **Compatibilidade:** Suporte para os formatos de áudio WAV e TIMIT, compatibilidade com HTK.

Tarefa	Tracks	#Fonemas	Duração
Treino	2 a 8	12465	20m57s
Teste	1	2452	3m28s

Tabela 5.2: Ficheiros utilizados da base de dados FEUP/IPB

- **Interacção HTK:** utilização das ferramentas do HTK independentemente da versão, *scripts* compatíveis, visualização de resultados em linha de comando integrada, gravação de parâmetros ao longo de múltiplas utilizações, geração de HMMs independentes ou em grupo com manipulação dos seus parâmetros;
- Funções de editor de texto e gestor de ficheiros.

As funcionalidades principais do software HTKFace foram apresentadas e terminam assim as referências às ferramentas de trabalho utilizadas para o desenvolvimento do sistema de etiquetagem automático. Passar-se-á à descrição dos procedimentos utilizados para o tratamento dos corpora já abordados.

5.4 Corpora Utilizado

Todos os modelos baseados em dados estatísticos têm de ser sujeitos a um processo de treino antes da sua utilização. Posteriormente e antes do sistema final é necessário realizar alguns testes, preferencialmente com dados diferentes dos de treino, para confirmar a validade dos modelos encontrados. Assim, para o desenvolvimento de um sistema de etiquetagem importa, para além da boa qualidade dos dados propriamente ditos, definir que dados serão utilizados em cada situação. Para o objectivo a que se propõe este trabalho, esta separação foi feita de modo a que certas métricas estatísticas de cada sub-conjunto sejam idênticas às obtidas para a globalidade da base de dados.

5.4.1 Corpus FEUP-IPB

A base de dados FEUP/IPB é constituída por 13 *tracks* num total de cerca de 100 minutos de fala. Para a selecção das *tracks* efectuaram-se algumas análises estatísticas e consideraram-se essencialmente a frequência relativa de ocorrências e a duração média de cada fone. Com valores idênticos à totalidade dos dados, utilizou-se para o teste do sistema a *track* 1 e para treino *tracks* de 2 a 8. Na tabela 5.2

Todos os ficheiros de áudio utilizados foram re-amostrados sendo feita a conversão de 44.1KHz para 16.0KHz. As restantes características, resolução 16 bits e mono-canal, foram mantidas conforme o original. Estes valores foram considerados suficientes para os objectivos e são habituais quando comparados com outros sistemas idênticos.

No treino do sistema, durante a refinação dos parâmetros dos modelos, surgiram algumas dificuldades. A ferramenta HERest produzia erros frequentes devido a problemas de limitação de memória (o PC utilizado possui 512Mb de memória). Os ficheiros de áudio, que em média possuem cerca de 8Mb cada,

foram então segmentados em ficheiros mais pequenos. A unidade utilizada para a selecção dos pontos de corte foi a frase e tentou-se que os ficheiros finais tivessem-se um tamanho sempre inferior a 1Mb.

5.4.2 Corpus TIMIT

A base de dados TIMIT tem já vários anos e possui múltiplos falantes. Por outro lado, pela sua larga divulgação, é uma das referências mais habituais para a comparação do desempenho de sistemas de anotação e reconhecimento. Assim sendo decidiu-se também realizar alguns testes sobre esta base de dados.

As anotações disponíveis baseiam-se num conjunto de 61 etiquetas, que podem ser consultadas no anexo E, e que estão adaptadas aos sons da língua inglesa. De forma a aproximar este conjunto de etiquetas do considerado para a base de dados FEUP/IPB optou-se por efectuar uma redução do inventário fonético. Utilizando as sugestões dadas por Lee [46] chegou-se a um novo conjunto de 48 etiquetas.

Apesar da compatibilidade com as ferramentas HTK todas as informações disponíveis neste *corpus* foram convertidas para formatos mais comuns e semelhantes ao utilizados para a base de dados FEUP/IPB.

A duração do áudio contido em cada um dos ficheiros da base de dados é bastante pequena não sendo necessário qualquer tipo de transformação a este nível. As frases estão também bastante bem organizadas.

5.5 Procedimento e Resultados

Passar-se-á a descrever o procedimento seguido para o desenvolvimento do sistema de anotação automático. A sequência seguida compreende os quatro passos principais que já foram enumerados quando se referiram as ferramentas utilizadas. Nesta secção serão colocados em segundo plano os meios e será dada particular ênfase à exposição dos parâmetros implícitos num sistema de anotação automático baseado em HMMs e à forma como estes influenciam o seu desempenho. Os resultados da variação de um determinado parâmetro deveriam ser apresentados apenas no final de todo o procedimento concernente à avaliação do sistema. No entanto, por se considerar mais oportuno, os resultados serão apresentados imediatamente no ponto onde cada referência for efectuada. A ligeira quebra de sequência resultante da antecipação do final será com certeza minimizada pela eliminação da expectativa de resultados práticos. Deste modo será então descrita em avanço a metodologia seguida para a análise do desempenho.

5.5.1 Avaliação

A avaliação do desempenho de um sistema de anotação automático com as características aqui definidas é feita separadamente sobre as tarefa de segmentação e classificação. A referência a este assunto foi já realizada na secção 2.7.1.

Para a tarefa de classificação vão ser utilizadas as métricas "Percentagem Correcta" (PC) e "Exactidão" (EX). Para a tarefa de segmentação será considerada uma métrica que se designará por "Taxa de Acerto" (TA) dada pela expressão:

$$\text{Taxa de Acerto (\%)} = \frac{N - D - I}{N} \times 100\% \quad (5.22)$$

Para as variáveis tem-se N , como a totalidade de fronteiras na referência, D como número de fronteiras eliminadas e I como número de fronteiras inseridas (sempre em relação à referência).

A aplicação destas métricas de comparação recai sobre duas sequências fonéticas que serão, como habitualmente, a anotação manual e a anotação automática realizada pelo sistema. No sistema desenvolvido valida-se como correcta uma determinada ocorrência de fronteira quando esta se encontra a menos de 10ms da fronteira real (intervalo de erro de $\pm 10\text{ms}$). Num segundo caso, valida-se como correcta para uma distância menor que 20ms da fronteira real a fim de obter os valores das métricas para esse intervalo (intervalo de erro de $\pm 20\text{ms}$).

Com estas pequenas notas iniciar-se-á agora a descrição do processo de anotação automática desde o início.

5.5.2 Preparação

A preparação dos dados passa inicialmente pela organização do grande número de ficheiros envolvidos. Foram criadas pastas para armazenar ficheiros de áudio, ficheiros de etiquetas, HMMs protótipo e respectivas iterações e ficheiros com resultados e análise.

A primeira tarefa realizada foi a transformação dos ficheiros de etiquetas para um formato interpretável pelo HTK. Esta tarefa foi já descrita e passa apenas por uma conversão dos valores de posicionamento dentro do áudio e dos nomes das etiquetas. O corpus FEUP/IPB utiliza o tempo em segundos e o corpus TIMIT utiliza o número da amostra. Em ambos os casos foi feita a conversão para a unidade de 100ns usada pelo HTK utilizando o HTKFace.

A partir dos ficheiros de áudio são produzidos outros com os vectores de características correspondentes. Foram utilizados essencialmente MFCCs e energia e respectivas características dinâmicas, delta e delta-delta (aceleração).

O ficheiro de configuração que se mostra abaixo, gerado pelo HTKFace, possui os parâmetros principais que foram sendo ajustados.

```
SOURCEKIND = WAVEFORM
SOURCEFORMAT = WAV
SOURCERATE = 625
TARGETKIND = MFCC_E_D_A
TARGETRATE = 100000.0
SAVECOMPRESSED = F
SAVEWITHCRC = F
WINDOWSIZE = 250000.0
ZMEANSOURCE=T
USEHAMMING = T
PREEMCOEF = 0.97
NUMCHANS = 26
CEPLIFTER = 22
NUMCEPS = 12
ENORMALISE = F
```

Indicam-se o tipo de dados de entrada (WAV a 16KHz), o tipo de dados à saída (MFCC e número de filtros) e a forma como a análise é efectuada (dimensão das janelas, tipo de janela, etc.).

No HTK o cálculo de algumas características é realizado de modo um pouco diferente do tradicional. Para referência, a energia é calculada como o logaritmo da energia do sinal:

$$E = \log \sum_{n=1}^N s_n^2 \quad (5.23)$$

Os coeficientes Mel-Cepstrais são especificados num banco de filtros logarítmico de amplitudes m_j através de uma DCT:

$$c_i = \sqrt{\frac{2}{N}} \sum_{j=1}^N m_j \cos \left(\frac{\pi i}{N} (j - 0.5) \right) \quad (5.24)$$

Onde N representa o número de filtros que no ficheiro de configurações é indicado pelo parâmetro NUMCHANS. O parâmetro NUMCEPS especifica o número de coeficientes cepstrais que neste caso não tem de ser necessariamente o mesmo que o número de canais do banco de filtros.

Os coeficientes delta são calculados por:

$$d_t = \frac{\sum_{\theta=1}^{\Theta} \theta (c_{t+\theta} - c_{t-\theta})}{2 \sum_{\theta=1}^{\Theta} \theta^2} \quad (5.25)$$

Cada um dos parâmetros apresentado foi variado individualmente varrendo uma zona de interesse e o seu efeito nos resultados analisado. Avaliaram-se também separadamente as tarefas de classificação e segmentação pois os seus diferentes objectivos obrigam à construção de sistemas com parametrizações distintas.

Para se compreender a contribuição de cada parâmetro define-se como configuração base a apresentada anteriormente e varia-se cada um deles independentemente. Para além do especificado antes, utilizaram-se 41 fones da base de dados FEUP/IPB e uma gramática sem regras (onde a cada fone do léxico pode seguir-se qualquer outro). Realizaram-se para cada ensaio as re-estimações necessárias até que a redução do erro entre duas iterações sucessivas seja inferior a 1%. Com o sistema construído deste modo obteve-se, como valores de referência, para a tarefa de classificação, 68.60% na *Percentagem Correcta*, 52.75% na *Exactidão*, para a tarefa de segmentação, 61.14% na *Taxa de Acerto*.

Inicialmente, efectuou-se um estudo genérico onde se utilizaram modelos de topologia idêntica para todos os fonemas e, num fase de refinação, analisou-se cada uma das variáveis, de forma independentemente, sobre um conjunto de fonemas pertencentes a uma mesma categoria fonética.

Dimensão da Janela e do Passo de Avanço

Os primeiros ensaios realizados visaram avaliar a adequação às tarefas de segmentação e classificação da dimensão da janela, para a divisão dos vectores de características em blocos, e o respectivo passo de avanço ou, de uma outra forma, a sobreposição entre janelas sucessivas. Os resultados obtidos resumem-se na tabela 5.3.

Hop	Janela (ms)			Hop	Janela (ms)		
	15	20	25		15	20	25
2.5	77.5/10.9	76.3/12.6	75.3/18.3	2.5	65.2	63.2	61.0
5.0	75.0/40.1	74.7/40.5	73.8/41.2	5.0	69.7	67.2	65.7
7.5	72.3/47.2	72.5/51.3	70.0/50.6	7.5	64.9	64.5	63.9
10.0	68.4/52.8	69.3/53.5	68.6/52.8	10.0	62.9	62.7	61.1

(a) Classificação

(b) Segmentação

Tabela 5.3: Resultados da classificação (PC%/EX%) e segmentação (TA%) variando a dimensão da janela e o passo de avanço.

Na tarefa de classificação, considerando apenas as concordâncias (PC), ficou evidente que os resultados melhores se obtêm nas situações em que a janela e o passo têm menor dimensão. No entanto, para a exactidão esta não é a situação mais benéfica pois o seu valor reduz-se em grande parte devido ao elevado número de erros por inserção que surgem.

Em relação à taxa de acerto para a tarefa de segmentação os melhores resultados são encontrados para um passo de 5ms e para uma janela de 15ms. Em valores mais pequenos surgem problemas análogos ao referidos na classificação (a inserção de um fone numa sequência provoca também a inserção de uma fronteira).

Refira-se que em situações de decisão se optou por preferir escolhas que levam a erros por inserção em vez das que levam a erros por eliminação. Isto deve-se ao facto de os erros por inserção serem mais facilmente resolvidos pois a probabilidade associada ao fonema é geralmente baixa. Assim, definindo um valor de *threshold* para a probabilidade associada é possível *limpar* os fonemas menos fiáveis. À medida que se sobe o *threshold* desaparecem os erros de inserção e, em contra partida, vão surgindo outros por eliminação. O valor a escolher para o nível de *threshold* deve resultar de um compromisso entre ambos os erros que permita maximizar a métrica da exactidão. Geralmente, para intervalos e passos pequenos, esta operação é vantajosa e permite melhorar a exactidão em cerca de 8%.

Número de Coeficientes Mel-Cepstrais

Foi também alvo de análise a influência do número de coeficientes mel-cepstrais nas tarefas da anotação. Mantiveram-se então os parâmetros apresentados e variou-se o número de coeficientes entre 10 e 18 utilizando sempre os respectivos coeficientes delta e aceleração. Os resultados obtidos apresentam-se na tabela 5.4.

De acordo com o esperado e como os resultados confirmam o aumento do número de coeficientes é benéfico para o desempenho do sistema. Quanto maior for o número de coeficientes melhor será descrito o sinal de áudio e consequentemente, após o treino, mais fiel será o modelo resultante. A utilização de apenas 10 coeficientes é nitidamente insuficiente para o efeito e resulta numa diminuição de PC e TA. A exactidão (EX) apesar de sair beneficiada pela redução do número de erros por inserção não foi considerada relevante pois nos restantes ensaios utilizando uma correcção por nível de *threshold*, como já foi descrito,

#Coef	Classificação		Segmentação
	PC (%)	EX (%)	TA (%)
10	64.59	55.89	58.96
12	68.60	52.75	61.14
14	71.12	56.89	62.46
16	72.35	56.82	63.92
18	73.01	57.01	64.11

Tabela 5.4: Resultados da classificação e segmentação variando o número de coeficientes mel-cepstrais.

Características	Dim. Vector	Classificação		Segmentação
		PC (%)	EX (%)	TA (%)
MFCC	12	49.64	35.78	40.54
MFCC_E	13	57.71	47.01	47.95
MFCC_0	13	56.78	45.10	46.13
MFCC_ED	26	65.77	56.68	56.37
MFCC_DA	36	66.12	53.52	59.03
MFCC_EDA	39	68.60	52.75	61.14

Tabela 5.5: Resultados da classificação e segmentação variando o vector de características.

conseguem-se melhorias significativas. O recurso a 14 coeficientes, 2 mais do que a referência, provoca o aumento da dimensão do vector de características de 39 para 45 pois foram também utilizados os coeficientes delta e aceleração bem como a respectiva energia. Situação análoga acontece com os outros casos. O aumento da dimensão do vector de características reflecte-se no tempo de treino dos modelos e no número de iterações que aumentam consideravelmente. Para um maior número de coeficientes os valores dos vários indicadores vão aumentando ligeiramente sem que no entanto se notem melhorias drásticas. Tendo em conta a globalidade das métricas e o esforço de treino dos modelos (tempo e iterações) considerou-se que as melhores opções são para a classificação 14 coeficientes e para a segmentação 16 coeficientes.

Descrição do Sinal

Analizado o número de coeficientes a considerar decidiu-se também analisar o papel das características dinâmicas dos MFCCs e o de outras características complementares. Assim sendo e tendo por referência os parâmetros referidos efectuaram-se ensaios e obtiveram-se resultados para a introdução das características delta (assinaladas com o carácter "D" na tabela 5.5), aceleração (carácter "A"), energia (carácter "E") e coeficiente de ordem nula (carácter "0") tendo sempre por base 12 coeficientes MFCC. Os valores obtidos reúnem-se na tabela 5.5.

A análise da tabela mostra que os coeficientes dinâmicos delta e aceleração têm um papel de substancial importância na tarefa de segmentação, cerca de menos 5% em relação à referência, mas por outro lado são menos relevantes para

Partindo do conjunto de MFCCs simples conclui-se também que a utilização da energia proporciona melhores resultados do que a utilização do coeficiente de ordem nula seja qual for a métrica ou tarefa em questão.

Depois de definidas e extraídas as características do sinal que melhor o descrevem para os efeitos desejados é necessário especificar a topologia dos modelos markovianos a utilizar. Na versão mais simplificada estes modelos são definidos um a um, em ficheiros separados. Um exemplo do conteúdo de um protótipo de HMM apresenta-se abaixo:

No início da definição do HMM são indicadas a dimensão e tipo dos vetores de características e o nome do modelo. No exemplo dado, coerente com o ficheiro de configurações apresentado antes, o vector tem dimensão 39 pois resulta de 12 coeficientes MFCC mais energia a que se juntam os respectivos parâmetros dinâmicos delta e delta-delta (ou aceleração). Seguidamente especificam-se as médias e variâncias relativas aos estados do modelo. No HTK o primeiro e último estado têm um valor simbólico e apenas são utilizados internamente pelo *software* para efeitos de ligações entre HMMs. Assim, o modelo apresentado,

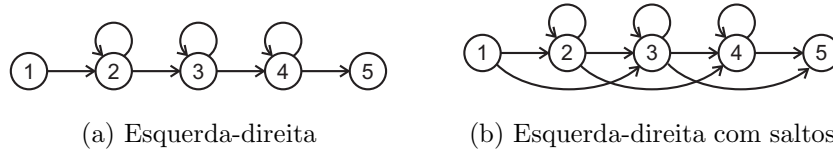


Figura 5.10: Topologias HMM.

Topologia	Classificação		Segmentação TA (%)
	PC (%)	EX (%)	
Esq-Dir	68.60	52.75	61.14
Esq-Dir com saltos	75.31	48.39	64.95

Tabela 5.6: Resultados da classificação e segmentação variando a topologia dos modelos.

onde se declaram 5 estados pelo identificador *NumStates*, possui efectivamente 3 estados. Na figura 5.10 mostram-se dois modelos de HMM possíveis correspondendo o da esquerda à descrição do protótipo apresentada.

Por estado deverão existir 39 valores para a média e 39 valores para a variância (no caso apenas surgem alguns por uma questão de apresentação). Finalmente, para terminar a definição do modelo, indicam-se as probabilidades de transição entre estados que, quando diferentes de zero, implicitamente irão definir a topologia. A soma dos valores ao longo de uma linha tem de ser obrigatoriamente um com excepção da última linha que corresponde a um estado terminador (designado por *non-emitting state*). Os valores indicados no modelo apresentado são aleatórios, respeitando a gama de valores admissíveis, e destinam-se apenas à definição inicial da topologia do modelo. Estes valores serão depois alterados na fase de treino pelas ferramentas adequadas. A topologia apresentada corresponde a um HMM simples, da esquerda para a direita, com uma única distribuição gaussiana e representativa de um fonema isolado (sem considerar vizinhanças). Na figura 5.10 tem-se, à esquerda, um HMM simples da esquerda para a direita, e à direita, um HMM esquerda-direita com saltos. Estas são as topologias mais comuns na descrição de voz mas são possíveis outras.

Com o objectivo de encontrar a melhor topologia realizaram-se dois ensaios cujos resultados se afixam na tabela 5.6. Utilizaram-se os parâmetros referência e todos os fonemas têm, em cada ensaio, modelos com igual topologia.

Os modelos com saltos proporcionam melhores resultados no geral sendo a sua utilização particularmente benéfica para a tarefa de segmentação. Analisando os resultados modelo a modelo ou fone a fone verifica-se que, com alguma regularidade (mas com excepções), que a topologia com saltos oferece maiores vantagens nos fonemas com menor duração.

Número de Estados

Ainda sobre a topologia do modelo é possível variar o número de estados. Assim, partindo mais uma vez do modelo de referência, com três estados, investigaram-se as possibilidades oferecidas pela alteração das topologias neste sentido. Os

#Estados	Classificação		Segmentação
	PC (%)	EX (%)	TA (%)
3	68.60	52.75	61.14
4	65.83	56.45	55.62
5	63.98	54.10	56.39
6	60.40	52.75	53.52
7	60.41	52.75	53.54

Tabela 5.7: Resultados da classificação e segmentação variando o número de estados dos modelos.

#Misturas	Classificação		Segmentação
	PC (%)	EX (%)	TA (%)
1	68.60	52.75	61.14
3	73.41	60.76	64.17
5	75.88	62.99	65.83
7	76.68	64.84	69.02
9	77.54	65.33	67.96

Tabela 5.8: Resultados da classificação e segmentação variando o número de misturas gaussianas por estado.

resultados encontram-se na tabela 5.7.

O aumento do número de estados não constitui uma boa opção para a maioria dos modelos. A utilização de 4 estados poderá, do ponto de vista de classificação, ser considerada positiva não se podendo afirmar o mesmo para os outros casos. Analisando os resultados modelo a modelo conclui-se que os resultados são melhores nos modelos cujo sinal acústico é menos cíclico e modelável mas estes são apenas uma pequena minoria.

Foram ainda realizados outros ensaios em visaram apenas alguns fones em especial. Por exemplo verificou-se que as oclusivas orais propriamente ditas, sem considerar a zona de oclusão, são melhor modelizadas por HMMs (esquerda-direita sem saltos) de 2 estados. Na zona de oclusão os três estados representam o número ideal. Nas nasais conseguem-se melhorias em modelos de 4 estados. Outras situações foram testadas sem grande expressão nos resultados finais. O ajuste particular de modelos foi relevando então para segundo plano passando-se ao ensaio de outras possibilidades mais promissoras.

Número de Misturas

Um outro parâmetro importante é o número de gaussianas utilizadas em cada estado. O modelo de referência baseia-se numa única distribuição gaussiana não existindo portanto as chamadas *misturas*.

A utilização de várias distribuições por estado contribui em muito para o correcto ajuste dos modelos HMM conforme se pode analisar na tabela 5.8. O maior incremento é notado logo com a introdução das primeiras misturas, entre a referência e as três misturas, onde surgem aumentos absolutos de cerca de 5% para PC, 8% para EX e mais de 3% em TA. A partir daqui os incrementos

são sendo sucessivamente menores mas sempre bastante consistentes. Na segmentação obteve-se o melhor valor de TA para as 7 misturas ao passo que para a classificação as 9 misturas foram as mais abonatórias.

O aumento do número de misturas para lá das 9 ensaiadas com certeza contribuirá ainda mais para a melhoria da fidelidade dos modelos. Estes benefícios nos resultados fazem todo o sentido pois, sendo os HMMS comprovadamente adequados à tarefa, ao aumentar as misturas aumenta-se também o número de valores que efectivamente constituem o modelo e isso só poderá melhorar a sua fidelidade. O tempo de treino e o número de iterações aumenta também consideravelmente sendo necessário cerca de 6 vezes mais tempo entre o primeiro e o último e mais exigente ensaio. O benefício para a introdução de misturas terá sempre um limite, ainda que não se atinja a perfeição dos 100% nos resultados, pois a partir de certo momento surgirão situações de redundância de valores ou casos de sobre-adaptação (estes mais facilmente resolúveis).

Os parâmetros estudados que constituem um HMM são armazenados num ficheiro tipo texto. Para a definição de um conjunto de HMMS podem criar-se vários ficheiros independentes com as várias descrições dos modelos ou, em alternativa, um único ficheiro designado por *Master Macro File* (tipicamente com extensão MMF), cujo conteúdo corresponderá à concatenação dos conteúdos dos vários ficheiros independentes. Esta última alternativa oferece vantagens em termos de organização e rapidez de processamento.

Número de Fones

Depois de analisadas as características que melhor descrevem o sinal acústico e avaliados os parâmetros dos HMM que melhor descrevem sequências temporais de vectores de características passar-se-á à análise de outras situações mais relacionadas com a língua para que o sistema de anotação se desenvolva. Aqui incluem-se o conjunto de fonemas utilizado para a classificação e a gramática com regras de construção de palavras e orientadora das sequências fonéticas. O sistema foi também testado com a base de dados TIMIT em inglês e pode ser facilmente adaptado para outra qualquer língua se forem utilizados conjuntos de fones mais universais e gramáticas adequadas.

Para a base de dados FEUP/IPB foi utilizado até ao momento e em todos os ensaios apresentados um conjunto de 41 modelos de fones que correspondem aos 38 fonemas apresentados na tabela 2.3 acrescidos de modelos para a oclusão, a pausa e a aspiração.

Decidiu-se efectuar alguns testes tendo por objectivo encontrar o conjunto de fonemas mais adequado a cada situação. Os resultados obtidos encontram-se na tabela 5.9 e devem ser interpretados tendo em mente que são valores globais e que a influência do acréscimo/eliminação de um fone está dependente da sua frequência na sequência de teste. Nos ensaios observaram-se, para além dos apresentados, outros indicadores centrados nas alterações específicas.

Devido às grandes diferenças existentes nas durações das pausas, incluiu-se um modelo novo, designado por *short-pause* (SP), para a modelização de silêncios ou pausas com dimensão inferior a 180ms. Esta opção levou a uma melhoria apreciável na classificação, pois os silêncios eram frequentemente confundidos com oclusões, mas não se mostrou tão importante para a segmentação.

As oclusivas orais foram também alvo de ensaio. O símbolo único e comum foi eliminado e criaram-se 6 novos símbolos para a representação da oclusão de

#Descrição	Num. Fones	Classificação		Segmentação TA (%)
		PC (%)	EX (%)	
Referência	41	68.60	52.75	61.14
Com <i>short-pause</i>	42	69.02	55.37	62.19
Oclusões específicas	46	69.41	53.12	64.24
Pares de oclusivas	38	65.43	52.09	60.46
Com ditongos	49	67.99	54.18	63.22

Tabela 5.9: Resultados da classificação e segmentação variando o léxico.

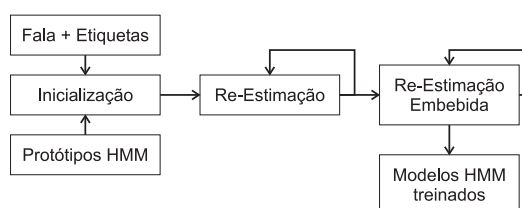


Figura 5.11: Procedimento para treino dos HMMs

cada fonema com esta classificação. Os resultados obtidos para a segmentação foram animadores não encontrando porém paralelo na classificação. Num outro ensaio, mais orientado para a segmentação, decidiu-se agrupar as oclusivas sonoras e surdas em pares por ponto de articulação. O novo conjunto de 38 modelos não se revelou interessante.

Devido à dificuldade que por vezes surge na definição da fronteira interior dos ditongos realizou-se também um ensaio onde se acrescentaram 8 novos modelos para a representação destes pares sonoros. Os resultados foram um pouco inferiores ao esperado tendo-se registado valores semelhantes à referência no caso da classificação e para a segmentação benefícios ligeiros.

5.5.3 Treino

Para a criação de uma primeira estimativa dos parâmetros dos modelos existem duas metodologias que podiam ter sido seguidas: *flatstart*, não considerando os ficheiros de etiquetas, ou *bootstrap*, utilizando os ficheiros com etiquetas. Optou-se por esta última opção por levar mais rapidamente a melhores resultados e por não existir o perigo de dificuldades de convergência iniciais. Os vários passos do procedimento utilizado para o treino estão esquematizados na figura 5.11.

Assim, por um processo de *bootstrap*, utilizaram-se as informações da fala, sobre a forma de vectores de características, e as informações de etiquetagem, já no formato HTK, para o cálculo de uma primeira aproximação aos parâmetros dos HMMs respeitando a topologia especificada nos protótipos.

Nesta primeira aproximação cada HMM é visto como um gerador de vectores de características. Cada ocorrência acústica de um determinado fonema é vista como a saída de um HMM cujos parâmetros é necessário estimar. Assim, se o estado que gerou cada vector de características for conhecido, então as médias e as variâncias podem ser estimadas através do cálculo da média de todos os

vectores associados a esse estado. Do mesmo modo, a matriz de transição pode ser estimada pela contagem do número de janelas que cada estado ocupa.

O processo apresentado é implementado através de uma sequência recursiva. Para isto é necessário obter estimativas iniciais que serão depois refinadas. Para isto é realizada uma segmentação uniforme dos dados e associam-se segmentos sucessivos a estados sucessivos. No caso de existirem várias misturas os vectores de características são associados às misturas sendo o número de vectores associada a cada mistura utilizado para estimar o peso de cada mistura. Possuindo a primeira estimativa para os modelos utiliza-se recursivamente um alinhamento Viterbi para procurar a sequência mais provável correspondente a cada exemplo de treino e que levará a um re-ajuste dos parâmetros.

Depois da inicialização dos modelos efectua-se uma re-estimação independente de cada modelo. O procedimento a seguir é análogo ao da inicialização mas dispensa o cálculo da primeira estimativa e substitui o alinhamento Viterbi por uma re-estimação Baum-Welch. Assim, procura-se a probabilidade de estar em cada estado em cada segmento de fala utilizando o algoritmo *Forward-Backward*. Quando os modelos possuem misturas pode surgir a situação de algumas componentes terem um número muito reduzido de informação associado o que levará a que os valores das variâncias ou do peso da mistura em causa sejam muito pequenos. Nesta situação é eliminada a mistura, se existir mais do que uma para o estado respectivo, de modo a não comprometer a validade do modelo. É também mantida a sequência iterativa onde se procura a convergência dos novos parâmetros.

O processo de re-estimação independente pode ser repetido algumas vezes até que o número de iterações que procuram a convergência seja globalmente reduzido. O número de iterações é bastante variável de fonema para fonema e é dependente da topologia do modelo onde se incluem todos os parâmetros já referidos. Para os modelos mais simples verificou-se que 3 repetições da re-estimação independente eram suficientes para garantir uma boa aproximação, nos modelos mais complexos podem ser necessárias cerca de 7 a 9 repetições.

Finalmente, para terminar a sequência de treino, efectua-se uma re-estimação embebida. Aqui, ao contrário dos processos que foram até agora descritos, o treino embebido actualiza simultaneamente todos os HMMS do sistema utilizando todos os dados de treino. No início são carregados todos os modelos envolvidos, reunidos propositadamente num único ficheiro, e à vez, são também carregados cada um dos ficheiros de treino e respectiva anotação. A sequência de etiquetas fonéticas da anotação é utilizada para a construção de um HMM composto que é alargado aos dados existentes. O HMM composto surge da concatenação dos modelos existentes com correspondência nas etiquetas que surgem na anotação. É depois utilizado o algoritmo *Forward-Backward* e vão sendo acumulados, com o peso apropriado, os resultados do cálculo das médias relativas a cada ficheiro de treino. Após o processamento de todos os ficheiros de treino são calculadas as novas estimativas dos parâmetros dos modelos a partir das somas pesadas. Com este passo termina o processo de treino dos modelos que constituem um sistema baseado em HMMS.

Na figura 5.12 pode observar-se o desempenho dos modelos ao longo de sucessivas iterações para o caso de uma única distribuição gaussiana e para o caso de 5 distribuições gaussianas.

O processo apresentado pode ainda ser modificado quando se verificar que os modelos estão sobre-treinados, ou seja, quando os modelos estão demasiado

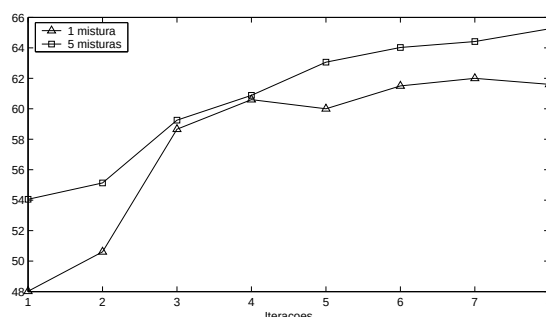


Figura 5.12: Taxa de acerto (TA) na tarefa de segmentação ao longo de várias iterações

ajustados aos dados de treino. Do ponto de vista dos parâmetros gaussianos dos modelos significa que as variâncias ficam bastante diminuídas (curva gaussiana mais reduzida). Esta situação só pode ser confirmada pelo aumento dos erros nos resultados finais ao longo de sucessivas iterações (tarefa seguinte de avaliação de resultados após a segmentação e classificação). Quando isto acontece os ficheiros de etiquetas resultantes da anotação automática substituem as etiquetas originais produzidas manualmente. Desta forma os modelos são re-treinados com informação ligeiramente distorcida que lhes induzirá alguma liberdade (operação frequentemente designada por relaxamento dos modelos). Para além disso torna-se possível analisar outras sequências fonéticas válidas mas que diferem das criadas na anotação manual. Um exemplo seria a identificação da sequência fonética correspondente à palavra "Porto" como /puortu/ quando o anotador decidiu por /portu/. Estas sequências resultam de transcrições alternativas que podem ser na verdade as que melhor correspondem ao sinal acústico. Sempre que estas situações surjam frequentemente, por exemplo na criação de um *corpus* baseado numa variante dialectal do Português mais longínqua do Português padrão, a realização de um pequeno relaxamento dos modelos não deve ser esquecida sob pena de as taxas de erro relativas serem bastante elevadas.

O treino pode ser concluído neste ponto pois os modelos já possuem os seus parâmetros otimizados para a informação que lhe foi fornecida. No entanto, cada fone, do ponto de vista acústico, é altamente influenciado pela sua vizinhança sendo por isso de grande interesse ter este aspecto em consideração.

Contexto

Na análise do contexto podem referir-se várias orientações que consideram vizinhanças apenas à esquerda, vizinhanças apenas à direita ou ambas. Em relação à profundidade da análise podem considerar-se apenas os fones imediatamente contíguos ou abranger uma maior quantidade que se considere relevante (menos útil do ponto de vista da análise de influência acústica). As situações mais comuns têm apenas em conta a vizinhança à esquerda (fone anterior) ou, em alternativa, utilizam trifones que reúnem os fones à esquerda e à direita. Numa outra perspectiva existe ainda a distinção entre n-fones intra-palavra, apenas no interior de palavras delimitadas por silêncios ou pausas, ou inter-palavra,

Contexto	Classificação		Segmentação
	PC (%)	EX (%)	TA (%)
Sem contexto	68.60	52.75	61.14
À esquerda e à direita	75.53	62.94	65.88

Tabela 5.10: Resultados da classificação e segmentação considerando ou não o contexto de cada fone.

que consideram todas as possibilidades incluindo silêncios e pausas. A título de curiosidade apresenta-se no anexo D um tabela com as ocorrências de bi-fones na base de dados FEUP/IPB.

O sistema a construir passa agora a ser baseado em trifones no lugar dos monofones que até ao momento têm sido falados. Cada trifone é representado por um novo HMM resultante da conjugação dos HMMs de cada um dos monofones constituintes. No lugar dos cerca de 40 monofones utilizadas passar-se-á a ter um conjunto muito mais vasto de trifones correspondentes às possibilidades permitidas pela língua. No caso utilizaram-se apenas aqueles que surgiam na base de dados. Incluíram-se ainda alguns difones que surgem no início e fim de palavras.

Existem trifones bastante frequentes, como os pares oclusiva-líquida seguidos de vogal, e que permitem reunir um conjunto de ocorrências suficientemente diverso para a realização de um bom ajuste do modelo representativo. Porém, outros trifones surgem muito raramente, com 2 ou menos ocorrências, o que é insuficiente para a realização de um treino adequado. Assim, para resolver este problema criaram-se novos modelos que partilham parâmetros, essencialmente a matriz de transições e que são designados por *tied-state triphones*. Estes modelos conduzem a estimativas melhores pois reúnem mais informação de treino. A escolha dos parâmetros a partilhar deve ser cuidadosa e cingir-se apenas aos factores que não influenciam a capacidade de discriminação dos modelos.

Para o treino é também necessário efectuar a conversão das anotações com monofones para trifones. Existirá assim todo um novo conjunto de informação que leva a novos resultados que se apresentam na tabela 5.10.

As melhorias são extremamente significativas para a classificação conseguindo-se um ganho absoluto de cerca de 10% na exactidão. Na tarefa de segmentação as melhorias são também animadoras mas um pouco mais modestas.

Terminadas as tarefas de treino, os parâmetros estimados para os modelos serão, de acordo com os algoritmos utilizados, os que melhor permitirão descrever a sequência temporal de características acústicas representativas do sinal de fala.

5.5.4 Segmentação e Classificação

Nesta fase o sistema de anotação tem já os modelos treinados e está pronto a ser ensaiado. Para este efeito é necessário, para além do conjunto de HMM, um conjunto de ficheiros de teste com informação distinta da utilizada para o treino. Apesar de não serem obrigatórias neste momento, os ficheiros de teste deverão possuir também informações de anotação de modo a que os resultados das tarefas de segmentação e classificação possam ser posteriormente avaliados.

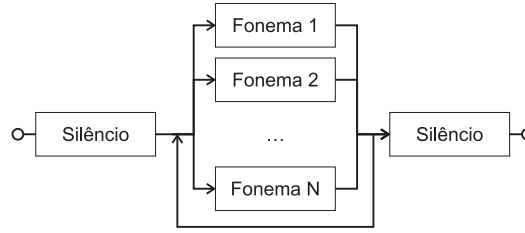


Figura 5.13: Exemplo de um modelo de linguagem

Se, para além dos recursos referidos, for ainda fornecida à ferramenta de anotação a sequência de símbolos fonéticos dos dados a processar, a tarefa fica bastante simplificada. Será assim simplesmente realizado o alinhamento forçado, restando depois apenas a identificação das fronteiras (limítrofes), ou seja, a segmentação.

Se, por outro lado, se considerar que apenas são fornecidos os ficheiros de fala, sem qualquer informação de anotação, o alinhamento forçado já não será possível e realizar-se-ão as tarefas de segmentação e classificação. A classificação pode ser facilitada ou os resultados melhorados se for fornecida uma gramática que contenha regras ou probabilidades para as sequências fonéticas admissíveis. Assim, criou-se uma rede fonética onde se indicam todas as possibilidades de sequência para o Português Europeu e, a partir da informação contida nos ficheiros de anotação de toda a base de dados (FEUP/IPB), obtiveram-se as probabilidades de cada sequência. Os valores aqui encontrados foram depois comparados com uma amostra de menor dimensão construída a partir dos recursos de corpora fonológico de imprensa do sítio web Linguateca[48]. Para os textos recolhidos obteve-se a transcrição fonética correspondente utilizando um sistema de conversão grafema-fonema automático e seguidamente utilizou-se um procedimento estatístico análogo. Não foram encontradas discrepâncias significativas.

A gramática referida, ou modelo de linguagem, pode ser inicialmente definida como uma simples rede cíclica de fonemas com algumas bifurcações. Um exemplo é dado na figura 5.13. As regras definidas para as situações possíveis conseguem limitar bastante as escolhas e assim reduzir a perplexidade da pesquisa Viterbi. No entanto, numa língua, o número de sequências fonéticas admissíveis é bastante grande e é difícil criar regras que possuem uma rigidez implícita para limitar as possibilidades.

Mais útil e mais realista é a utilização de modelos de linguagem onde se reúnam as regras possíveis e que estas sejam complementadas por informação estatística de ocorrências. A este tipo de gramática chamar-se-á *bi-gram*. A sua construção baseia-se nas probabilidades:

$$p(i, j) = \begin{cases} \frac{N(i, j) - D}{N(i)} & N(i, j) > t \\ b(i)p(j) & \text{outros valores} \end{cases} \quad (5.26)$$

Sendo $N(i, j)$ o número de vezes que o fonema j precede o fonema i , $N(i)$ é o número de vezes que o fonema i surge e D é um valor de desconto. Este desconto é feito no final às sequências mais frequentes e o seu valor é distribuído

Situação	Classificação		Segmentação
	PC (%)	EX (%)	TA (%)
Sem gramática	68.60	52.75	61.14
Com gramática	76.92	66.84	67.08

Tabela 5.11: Resultados da classificação e segmentação utilizando ou não uma gramática.

pelas sequências muito pouco frequentes. A variável t é um valor de *threshold* abaixo do qual a criação de uma sequência de dois fonemas (*bi-gram*) é anulada e utiliza-se apenas um fonema. Neste caso tem-se:

$$p(i) = \begin{cases} \frac{N(i)}{N} & N(i) > u \\ u/N & \text{outros valores} \end{cases} \quad (5.27)$$

O valor de u é novamente um valor mínimo para o número de ocorrências de um fonemas simples e N é dado por:

$$N = \sum_{i=1}^L \max[N(i), u] \quad (5.28)$$

Ainda em relação à expressão 5.26 resta esclarecer:

$$b(i) = \frac{1 - \sum_{j \in B} p(i, j)}{1 - \sum_{j \in B} p(j)} \quad (5.29)$$

Aqui, B é o conjunto dos fonemas para os quais existe a probabilidade de um *bi-gram*, $p(i, j)$.

As frequências absolutas de pares fonéticos encontrados na base de dados FEUP/IPB estão disponíveis para consulta no anexo D.

Foi confirmado em vários ensaios que a utilização de probabilidades associadas às redes fonéticas dá uma contribuição importante para a melhoria do desempenho do sub-sistema de classificação. Estes resultados estão apresentados na tabela 5.11.

Os resultados obtidos apesar de serem orientados essencialmente para a classificação acabam por influenciar também a segmentação, um fone extra traz consigo uma fronteira extra. Os benefícios da introdução de uma gramática com algumas regras foram consideráveis para ambas as tarefas da anotação. As regras criadas foram bastante simples e seria interessante conseguir um modelo de linguagem mais apurado.

Regressando à apresentação do procedimento executado onde se referia a prontidão dos modelos markovianos treinados para os testes, recorda-se que foram previstas duas possibilidades de prosseguir. Uma primeira, onde existe apenas a sequência fonética dos novos ficheiros de fala a anotar, e outra, em que nada é fornecido para além do áudio e se sugere a utilização de um modelo de linguagem. A seguir e independentemente das situações é calculado o melhor caminho (o caminho mais provável) ao longo do áudio de fala.

No caso do trabalho aqui apresentado optou-se pela não utilização de qualquer informação para além da voz com o objectivo de simplificar ao máximo a

criação de um corpus anotado. Com o sistema a funcionar deste modo é apenas necessário realizar a recolha do sinal de voz sendo eliminado todo o trabalho de organização dos textos e de transcrição fonética. O opção alternativa, onde se fornece a sequência fonética, é também aceite pelo sistema mas deixa-se para a parte final deste capítulo. Nesta situação, devido à informação extra, deixa de fazer sentido calcular erros de classificação.

A procura do melhor caminho é realizada com a ajuda do algoritmo de Viterbi já apresentado. Partindo da informação dos modelos de linguagem são criadas redes com estados ligados por arcos. A cada estado corresponde um modelo HMM que é também propriamente uma rede com estados ligados por arcos. No final possui-se uma rede global de grande complexidade que possui vários níveis. Assim, para uma sequência de n características acústicas analisar-se-ão n estados aos quais corresponderá uma determinada probabilidade. Cada caminho ao longo da rede possuirá também uma probabilidade que resulta da soma sucessiva das probabilidades ao longo dos sucessivos arcos. A procura do melhor caminho é realizada com a ajuda de uma técnica designada por passagem de testemunho (*token*). O testemunho neste caso contém a probabilidade acumulada até determinado estado. Inicialmente é colocado um testemunho nulo em cada um dos estados iniciais possíveis de acordo com a rede. Os testemunhos vão percorrendo a rede e sempre que se situam num estado que possui vários arcos de saída para outras sub-redes são replicados e vão avançando paralelamente. Cada testemunho armazena ainda uma referência temporal que permite ter uma ideia da sua localização e avanço dentro do sinal de áudio. No final de cada passo temporal os valores acumulados para cada testemunho são comparados e apenas é propagado o testemunho que possuir uma maior probabilidade acumulada. Como a rede de pesquisa é normalmente bastante complexa, o número de testemunhos acaba por ser também bastante elevado. Para otimizar a busca define-se um valor de *threshold* para a diferença, num determinado instante, entre a melhor probabilidade e as restantes. Sempre que um testemunho estiver abaixo desse valor é eliminado precocemente. Se este valor for demasiado reduzido corre-se o risco de eliminar o testemunho que viria a conduzir ao melhor caminho sendo por isso necessária alguma sensibilidade na definição do valor de *threshold*. No final obtem-se finalmente o caminho que permitiu acumular ao longo dos vários arcos a maior probabilidade.

O ajuste de cada estado de um determinado HMM dentro de uma sequência de características acústicas oferece grandes vantagens. Assim os HMMs, devido à sua natureza, e os algoritmos utilizados, possibilitam a obtenção simultânea de informação de classificação e segmentação bem como uma probabilidade que caracteriza o erro de avaliação cometido.

5.6 Resultados

Os vários ensaios que foram sendo apresentados permitem ter uma noção bastante clara de quais as soluções que resultam ou não resultam para cada uma das tarefas que se pretende desempenhar. Com esta informação em mãos iniciou-se a construção um sistema completo constituído por dois subsistemas, um para a classificação e outro para a segmentação.

Corpus	Classificação		Segmentação TA (%)	
	PC (%)	EX (%)	10ms	20ms
FEUP/IPB (com gram.)	82.02	76.43	77.55	84.03
FEUP/IPB (sem gram.)	77.31	71.89	72.14	79.97
TIMIT (sem gram.)	74.11	68.38	70.05	78.81

Tabela 5.12: Resultados da classificação e segmentação para os corpora utilizados.

5.6.1 Ensaios gerais

Reunindo as características que se consideraram mais adequadas construíram-se os subsistemas de acordo com as configurações abaixo:

- **Classificação.**

Janelas de análise com 25ms com passo de avanço de 7.5ms, 14 MFCCs mais energia com coeficientes delta, 41 modelos de três estados, esquerda-direita com saltos e 5 componentes gaussianas por estado.

- **Segmentação.**

Janelas de análise com 15ms com passo de avanço de 5ms, 16 MFCCs mais energia com coeficientes delta e aceleração, 47 modelos de três estados, esquerda-direita com saltos e 7 componentes gaussianas por estado.

O subsistema de classificação é bastante menos exigente e treina-se mais rapidamente que o subsistema de segmentação. Em funcionamento efectivo, após o treino, e apesar de a rapidez de um sistema de anotação não ser uma característica fundamental, ambos os subsistemas possuem um desempenho bastante célere.

Os primeiros resultados obtidos resumem-se na tabela 5.12 e dizem respeito a um intervalo de erro de 10ms. As métricas mostram valores interessantes para a base de dados FEUP/IPB tanto para a classificação como para a segmentação. Os resultados são obviamente influenciados pelo facto da base de dados conter apenas uma voz o que facilita a resolução do problema.

Para avaliar o sistema desenvolvido com vários falantes e para comparar também os resultados obtidos com os publicados por outros autores efectuou-se um ensaio com a base de dados TIMIT sem qualquer gramática e efectuou-se o mesmo para a base de dados FEUP/IPB. O peso da não utilização de uma gramática fica novamente bem evidenciado.

Os valores conseguidos para a base de dados TIMIT são bons mas estão um pouco abaixo dos melhores encontrados na literatura (que se encontram entre 79.8% e 92.6%). Tal facto deve-se provavelmente à optimização dos sistemas desenvolvidos para a base de dados FEUP/IPB que é constituída apenas por uma voz em língua portuguesa. A orientação para a língua portuguesa foi considerada prioritária devido ao menor número de recursos disponíveis para esta.

Importante dizer que os resultados que estão a ser utilizados para termo de comparação se referem a tarefas de alinhamento onde a sequência fonética é já conhecida. Esta situação, apesar de não ter sido considerada, deverá permitir a introdução de um benefício considerável nos resultados.

Corpus	Classificação		Segmentação TA (%)	
	PC (%)	EX (%)	10ms	20ms
FEUP/IPB (com gram.)	82.11	78.67	77.97	84.33
FEUP/IPB (sem gram.)	76.94	73.20	72.83	79.29

Tabela 5.13: Resultados da classificação e segmentação otimizados para a base de dados FEUP/IPB.

5.6.2 Refinamento

Tendo por objectivo melhorar ainda mais os resultados realizaram-se algumas tarefas adicionais de optimização. Como já se referiu, a utilização de HMMs permite obter implicitamente uma probabilidade de ajuste do modelo aos vectores acústicos representativos do sinal. Assim, é possível eliminar os modelos que possuem uma probabilidade de ajuste inferior a um determinado valor e assim limpar alguns erros por inserção. Esta tarefa foi realizada aumentando sucessivamente o nível de *threshold* até que a métrica exactidão se deteriorasse. A métrica PC sai sempre prejudicada pois são eliminados, em muito menor número, alguns fones que estão efectivamente correctos.

Na eliminação dos fones com durações abaixo de um nível considerado mínimo surge um problema. Um fone possui duas fronteiras, uma à esquerda e outra à direita. Assumindo que a sequência fonética é constituída por elementos contíguos então quando se procede à eliminação de um fone terá de ser feito um reajuste de uma das fronteiras adjacentes. Este reajuste pode ser feito alterando a marcação de fim do fone anterior ou alterando a marcação de início do fone seguinte. Numa tentativa de determinar a situação que mais favorecesse os resultados ensaiaram-se os dois casos. O anulação da primeira fronteira, ou seja, a absorção do fone a eliminar pelo fone anterior mostrou-se um pouco mais favorável e foi esta a utilizada nos resultados apresentados. Em princípio esta opção não poderá ser generalizada.

Uma outra tarefa que se realizou foi a colocação das fronteiras fonéticas no local mais adequado possível tendo por referência o sinal acústico. Assim, as fronteiras foram reposicionadas o mais próximo possível de uma passagem por zero. Sempre que o fone identificado possui vozeamento realizou-se também uma estimativa dos ciclos de f_0 e colocou-se a fronteira junto à passagem por zero mais próxima do início de um período. Esta operação é indicada quando os resultados da anotação se destinarem à constituição de uma base de dados para síntese de fala. No reconhecimento não se revela em geral de tanta importância. Os ciclos de f_0 foram estimados utilizando o Praat tendo no entanto sido experimentados alguns algoritmos desenvolvidos pelo autor que não ofereciam resultados com a exactidão desejada.

Para terminar, utilizaram-se algumas das informações obtidas nos vários ensaios que visaram apenas alguns fonemas em particular e incluíram-se estes pormenores no sistema final. Os resultados finais, na tabela 5.13, melhoraram ligeiramente.

Estas últimas operações centraram-se essencialmente nas unidades representativas do silêncio, silêncio curto, aspiração e momentos de oclusão por registarem um número elevado de erros. Estas e outras situações de erro onde se notou

l	@	m
l		m
(a)		
sil		asp
sil	ocl	d asp
(b)		
ocl		k
sil	ocl	k
(c)		
l	e	m
l		m
(d)		
ocl	k	o~
ocl	k	o~ w
(e)		
6~		w
6		w~
(f)		

Figura 5.14: Situações de erro encontradas com frequência nas anotações geradas automaticamente (em cima a sequência original e em baixo a hipotética sequência gerada automaticamente).

alguma recorrência e que se consideraram comuns surgem ilustradas na figura 5.14. O tratamento automático de situações como as representadas em (a), (b) e (c) serão alvo de trabalho futuro. Mais adiante realiza-se uma experiência com resultados significativos de forma já semi-automatizada que fundamenta as referidas intenções desenvolvimento do trabalho.

Em (a), (b) e (c) da mesma figura estão representadas situações de erro que apesar de poderem ser correctas do ponto de vista da correspondência acústica não o são do ponto de vista linguístico. Estas situações podem ser resolvidas com uma bom modelo de linguagem. Em (d), (e) e (f) surgem casos onde a sequência fonética faz sentido e pode perfeitamente estar correctamente identificada. No entanto são assinalados erros pois surgem discrepâncias quando se efectua a comparação com a referência. Estes problemas são devidos a fenómenos linguísticos, conforme se fez notar em 2.7.7, e cuja resolução pode passar pelo desenvolvimento de um sistema de análise mais complexo que permita a geração de transcrições fonéticas alternativas que contemplem os fenómenos da oralidade. A figura 5.14(f) tem por objectivo apresentar um padrão de erro comum mas não em especial o apresentado, ou seja, pretende-se apenas representar a situação frequente de substituição de uma vogal pela sua congénere nasalada e o recíproco.

Após uma análise pormenorizada dos erros concluiu-se que as fronteiras que registam uma maior disparidade em relação à referência produzida manualmente são as que dividem os difones vogal-vogal, vogal-glide, glide-vogal e vogal-nasal líquida. Esta observação é também feita por outros autores [50, 16]. Apesar do verificado não existe a evidência de que os resultados sejam necessariamente menos exactos. Analisando a evolução das formantes e do espectrograma ao longo da transição entre os sons referidos observa-se uma variação muito suave sendo bastante difícil identificar o ponto onde termina um som e se inicia o seguinte. A anotação manual destas situações é por isso também dúbia existindo grandes variações quando se compara o trabalho de vários anotadores humanos. Os erros nestes casos poderão então ser de algum modo considerados menores (este julgamento qualitativo não foi considerado em nenhum dos resultados numéricos apresentados pois não foi quantificado).

Para terminar foi testada uma forma de pós-processamento com alguma intervenção manual. O sequência fonética gerada automaticamente foi mani-

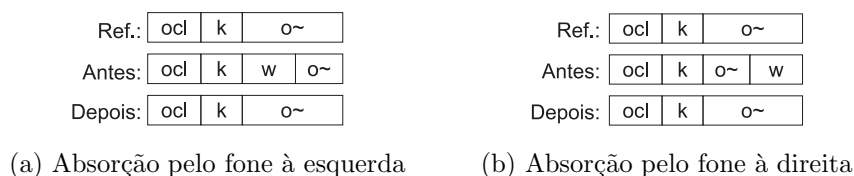


Figura 5.15: Procedimentos para a eliminação de fones (em cima a sequência original e em baixo a hipotética sequência gerada automaticamente).

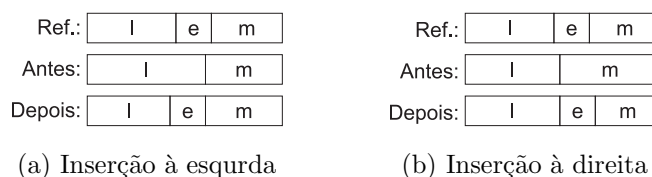


Figura 5.16: Procedimentos para a inserção de fones (em cima a sequência original e em baixo a hipotética sequência gerada automaticamente).

pulada, efectuando eliminações, inserções e substituições, tendo por objectivo conseguir uma sequência igual à utilizada para referência. A sequência gerada automaticamente possuía inicialmente um número de fones um pouco superior à referência. Sempre que se eliminou um fone foi escolhida individualmente e por intervenção manual a melhor fronteira. As duas situações possíveis nesta operação ilustram-se na figura 5.15 e são absorção pelo fone à esquerda ou absorção pelo fone à direita. Na figura 5.15(a) o fone $[w]$ está a mais, erro de inserção, e será absorvido pela fone imediatamente anterior. No final, após a absorção, o fone $[o \sim]$ manterá a sua fronteira inicial e passará a ter como nova fronteira final a do fone $[w]$ já desaparecido. Para a decisão entre uma ou outra situação teve-se em conta a duração do fone e a sua probabilidade (que resulta do ajustamento ao modelo). Para a inserção de fones na sequência gerada automaticamente, situação que ocorre muito raramente por se terem privilegiado os erros por inserção, usa-se como referência a fronteira mais próxima à esquerda ou à direita. Na figura 5.16(a) o fone $[e]$ estava em falta na sequência gerada automaticamente e é inserido alinhado pela fronteira mais próxima. A duração utilizada é a média das durações para o fonema em causa encontrada ao longo da sequência automática.

Com este processo obtiveram-se duas sequências fonéticas idênticas com diferenças apenas ao nível das marcas temporais de fronteira entre fones. Analisando as diferenças por fronteira conclui-se que 73.5% estão dentro de um intervalo de $\pm 10ms$, 88.4% estão dentro de um intervalo de $\pm 15ms$ e 92.6% estão dentro de um intervalo de $\pm 20ms$. Estes últimos resultados são bastante interessantes apesar de terem sido obtidos com alguma intervenção manual. No entanto o procedimento é de fácil automatização e será alvo de implementação num desenvolvimento futuro.

Capítulo 6

Conclusões e Trabalho Futuro

Ao longo deste trabalho estudou-se a composição de um sistema de segmentação automática e os muitos e diversos tipos de parâmetros que influenciam o seu desempenho. Após a investigação de outros trabalhos na área, levou-se a cabo, inicialmente, um estudo das características que se podem extrair dos sinais de fala no sentido da sua identificação.

Os modelos ARMA foram estudados nas várias configurações bem como os algoritmos que levam à obtenção dos parâmetros constituintes. Analisou-se separadamente e em conjunto a influência do número de pólos e de zeros na capacidade descritiva dos modelos. Concluiu-se que os pólos têm uma importância maior na descrição do sinal e que o seu número deverá ser superior a 12. O número de pólos pode ser sempre aumentado existindo porém limites para a beneficiação dos modelos. Depois de esgotada a capacidade de descrição de sinal a adição de novos pólos pode mesmo levar a resultados menos interessantes. Tendo em conta o poder computacional actualmente existente e os cada vez mais elevados requisitos de qualidade sugere-se que 16 seja o número mais adequado e óptimo considerando a relação esforço computacional/qualidade.

Os modelos exclusivamente baseados em pólos mostraram-se no entanto insuficientes para a caracterização de alguns sons, nomeadamente e de forma mais nítida nos sons não-vozeados e em particular nas consoantes nasalizadas. Nestas situações, em consequência de uma diferente configuração do tracto vocal, é benéfica a introdução de zeros. Nos vários ensaios realizados com sons destas categorias mostra-se que os modelos se revelaram pouco sensíveis à adição de novos pólos enquanto que, com novos zeros, se obtiveram melhorias significativas. A quantidade de zeros a utilizar merece os mesmos comentários feitos para os pólos sendo todavia necessários em menor número. Um modelo com 4 zeros e 16 pólos mostrou-se bastante adequado à maioria das situações.

As técnicas de cálculo dos parâmetros dos modelos, zeros e pólos ou coeficientes dos polinómios do filtro de predição, foram também analisadas. Para os pólos os métodos da autocorrelação e covariância, amplamente conhecidos, foram sucintamente abordados. Para os zeros apresentaram-se os métodos de Prony e Steiglitz-McBride.

A grande sensibilidade dos coeficientes de predição linear dos modelos ARMA

levou a que se estuda-se adicionalmente a representação por *Line Spectral Frequencies* (LSF). Esta técnica, baseada na predição linear, gera um conjunto de parâmetros mais robusto que, devido à sua natureza particular, se encontram distribuídos ao longo do círculo unitário. Explorando esta característica é possível fazer representações bastante claras da evolução destes coeficientes ao longo do tempo e analisar facilmente variações no sinal. Em conjunto com espectrogramas e pequenos trechos anotados foram realizados ensaios que validam as capacidades discriminativas das LSFs.

Outras características como a taxa de passagens por zero, a energia média deslizante ou os coeficientes Mel-cepstrais, que são utilizadas ao longo deste trabalho, foram também brevemente expostas.

Após este estudo preparatório sobre características e modelos conceberam-se dois sistemas de anotação com abordagens distintas.

Na primeira abordagem realiza-se numa fase preparatória a extracção de características do sinal. Utilizaram-se, em várias combinações, a taxa de passagens por zero, a energia e os coeficientes LSFs acompanhados dos respectivos parâmetros dinâmicos, delta e aceleração. As características representativas do sinal acústico são depois comparadas de dois modos: comparações sucessivas e comparações início-fim. Estes dois métodos combinados permitem obter resultados razoáveis para a segmentação. Quando o sistema é utilizado em modo semi-automático, isto é, com alguma intervenção humana, os resultados atingem valores próximos dos 90%.

Quando a sequência fonética é desconhecida e não é possível efectuar apenas o alinhamento fonético utilizam-se redes neuronais para a classificação. Utilizam-se duas redes, uma para segmentos vozeados e outra para segmentos não-vozeados, sendo a escolha realizada essencialmente a partir da análise da energia e da taxa de passagens por zero. Os resultados obtidos para a classificação atingiram os 80% para os sons vozeados e os 69% para os sons não vozeados. Neste ponto, por ser já um pouco desviado da segmentação que motiva este trabalho, deixou-se algum espaço para melhoria. A alteração da topologia das redes, número de camadas e número de neurónios por camada, levará provavelmente a resultados melhores.

As técnicas baseadas em HMMs constituem a segunda abordagem seguida. Aqui analisaram-se em pormenor vários parâmetros e a sua contribuição para o bom funcionamento do sistema. Estudaram-se a dimensão da janela de análise e respectivo avanço, a dimensão dos vectores de características e a sua constituição, as topologias dos HMMs, o número de estados por HMM, o número de misturas gaussianas por estado, o contexto ou enquadramento acústico dos fones, a dimensão do léxico fonético e finalmente a inclusão de modelos de linguagem. Todas as análises foram efectuadas partindo de uma configuração que se definiu como padrão permitindo assim compreender com clareza a intervenção de cada variável no desempenho do sistema. Deste estudo chegou-se a duas configurações distintas, uma mais adaptada à segmentação e outra ajustada à classificação. Os resultados obtidos foram muito bons quando o sistema foi aplicado à base de dados FEUP/IPB. Algumas das situações onde foi assinalado um erro, tanto na segmentação como na classificação, poderiam estar efectivamente correctas. O problema é explicado e resulta de fenómenos da oralidade onde por vezes se omitem, juntam ou criam sons. Um módulo ainda extra que permitisse considerar estes fenómenos poderia ter conduzido a um melhor desempenho. O facto da base de dados FEUP/IPB possuir apenas um falante reduz a dificul-

dade das tarefas de segmentação e classificação. Para a base de dados TIMIT os resultados para a segmentação são igualmente bons mas um pouco inferiores ao anunciado por outros autores. Na classificação verifica-se uma situação idêntica. Este *corpus* está em língua inglesa e devido a uma menor profundidade de conhecimento desta algumas possibilidades, como por exemplo a gramática, não foram convenientemente exploradas. Os restantes ensaios e resultados mantêm toda a sua validade.

O sistema baseado em HMMs utiliza um conjunto de ferramentas de *software* de distribuição gratuita que funciona na linha de comandos. Para facilitar a sua utilização desenvolveu-se uma aplicação denominada HTKFace que cria um ambiente de desenvolvimento integrado, com *interface* gráfica, que permite criar e manipular os ficheiros necessários ao sistema de anotação, executar treino e testes e avaliar o desempenho.

Observando globalmente os resultados das duas abordagens pode chegar-se a algumas conclusões mais sustentadas.

As janelas de menor duração para a análise do sinal e o passo de avanço das mesmas mais curto favorecem a tarefa de segmentação. Isto faz sentido pois uma fronteira é um fenómeno sem duração e deve ser procurado em intervalos de menor duração. Já a classificação, devido à dimensão dos fones, pode ser realizada a partir de janelas com durações um pouco maiores.

A dimensão e diversidade de características a utilizar devem ser ponderadas caso a caso. A quantidade pode ser aumentada até ao limite de descrição possível e analisando sempre o esforço computacional requerido. A diversidade é também favorável enquanto as perspectivas que se vão adicionando forem diferentes, ou seja, se os parâmetros extra não oferecerem um visão do sinal de um outro prisma não trarão qualquer valor acrescentado sendo por isso redundantes. Nas tarefas analisadas utilizaram-se MFCCs e energia que revelaram ter uma boa capacidade descritiva. Para a segmentação foram ainda considerados os parâmetros dinâmicos delta e delta-delta respectivos. Na classificação os parâmetros de aceleração não se mostraram tão interessantes sendo suficientes para a distinção entre dois sons fonéticos os MFCCs acompanhados dos parâmetros dinâmicos de primeira ordem. Já na segmentação, uma vez que se procuram variações no sinal, é importante considerar uma ordem superior.

Em relação à dimensão do inventário fonético concluiu-se que à medida que esta aumenta os resultados da segmentação melhoram. O inverso se passa com os resultados da classificação. Um maior número de hipóteses de escolha dificulta a tarefa de classificação e diminui a probabilidade global de acerto. Na segmentação, como apenas se procura distinguir a transição de um determinado fone para outro, é favorável a utilização de um léxico mais alargado.

Por último, as gramáticas e a análise do contexto têm um papel importante na melhoria dos resultados de ambas as tarefas.

Com o trabalho desenvolvido, onde se propõem duas soluções distintas para a etiquetagem de sinais de fala, e com a qualidade do desempenho obtido, bastante bom para a base de dados FEUP/IPB e ligeiramente inferior aos resultados de outros autores na base de dados TIMIT, considera-se que os objectivos propostos inicialmente foram cumpridos. O trabalho desenvolvido revelou-se um pouco mais extenso do que o previsto mas devido ao interesse que certos assuntos despertaram decidiu-se não prescindir da sua exploração. O percurso um pouco menos linear que foi seguido permitiu que outros conhecimentos fossem adquiridos revelando-se compensador no final.

6.1 Trabalho Futuro

O trabalho realizado deixou espaço por melhorias e despertou o desejo de experimentar outras soluções. Assim, num prazo mais curto pretende-se:

- Explorar e desenvolver modelos de linguagem mais apurados.
- Analisar e apurar os modelos HMM fone a fone procurando uma configuração optimizada e mais adaptada às características acústicas particulares de cada um.
- Explorar outras possibilidades oferecidas pelos HMMs.
- Melhorar a topologia das redes neuronais utilizadas e ensaiar outras possibilidades.
- Construção de um sistema híbrido baseado em HMMs e redes neuronais que permita conjugar o melhor de cada técnica.
- Desenvolver um módulo de pós-processamento para melhorar os resultados da anotação por HMMs.

A mais longo prazo, pretende-se estudar técnicas alternativas e construir um novo sistema de etiquetagem utilizando uma abordagem diferente. A utilização de *Bayesian Networks* tem suscitado algum interesse e os resultados apresentados por alguns autores são indicadores de um desempenho superior. Uma outra opção baseada numa abordagem DTW modificada revela-se também interessante.

A aplicação das actividades de segmentação e anotação na área de síntese da fala aponta também para a preparação de ferramentas adicionais de caracterização dos segmentos que permita alimentar algoritmos de diversos tipos para a síntese, tais como os de selecção de unidades.

Apêndice A

Divisão e Multiplicação de Polinómios por um Binómio

A divisão e multiplicação de polinómios de coeficientes reais por uma ou mais raízes é utilizada frequentemente como ferramenta de cálculo em vários domínios da área do processamento da fala. Faz-se aqui uma pequena revisão dos procedimentos elementares que permitem o seu cálculo através de métodos adaptáveis à implementação em computadores.

Num caso genérico, para um polinómio de grau N com coeficientes reais a_i e uma raiz r , a expressão é:

$$\sum_{i=0}^N a_i x^i = (x - r) \left[a'_1 + \sum_{i=2}^{N-1} a_i + r.a_{i-1} \right] \quad (\text{A.1})$$

Para clarificar, considere-se um polinómio genérico de grau 3:

$$Ax^3 + Bx^2 + Cx + D \quad (\text{A.2})$$

Os coeficientes são valores reais. Procura-se atingir um polinómio equivalente com a forma:

$$(x - r)(ax^2 + bx + c) \quad (\text{A.3})$$

Tem-se uma raiz r num primeiro factor e um polinómio de grau imediatamente inferior.

Fazendo a expansão dos termos:

$$\begin{aligned} (x - r)(ax^2 + bx + c) &= ax^3 + bx^2 + cx - r.ax^2 - r.bx - rc \\ &= ax^3 + (b - r.a)x^2 + (c - r.b)x - r.c \end{aligned} \quad (\text{A.4})$$

Comparando agora os coeficientes das expressões A.2 e A.4:

$$\begin{aligned} a &= A \\ b &= B + r.a \\ c &= C + r.b \end{aligned}$$

Figura A.1: Esquema exemplificativo para a divisão de polinômios

Ainda um último exemplo, frequentemente utilizado para o cálculo expedito e que recorre a uma pequena tabela. o polinômio será $x^3 - x^2 + 2x - 2$ com uma raiz em $x = 1$. No topo da tabela surgem os coeficientes do polinômio na forma canônica e mais à esquerda a raiz. O valor do primeiro coeficiente é copiado para a zona inferior ao passo que os restantes correspondem à soma entre o coeficiente que se encontra acima com a multiplicação da raiz pelo coeficiente encontrado anteriormente.

A multiplicação de polinômios baseia-se nas mesmas regras bastando inverter as igualdades apresentadas.

Apêndice B

Determinação dos Parâmetros AR

Os modelos de predição linear baseados apenas em pólos são extensivamente utilizados em múltiplas aplicações. Nesta secção serão apresentados os métodos de cálculo típicos para a determinação dos parâmetros destes modelos.

Considere-se então a equação genérica da predição linear aplicada ao modelo auto-regressivo.

$$H(z) = \frac{S(z)}{U(z)} = \frac{G}{1 - \sum_{k=1}^p a_k z^{-k}} = \frac{1}{A(z)} \quad (\text{B.1})$$

Neste caso o modelo fica inteiramente determinado pelos coeficientes a_k e o ganho G . O numerador $A(z)$ é designado como filtro de predição inverso. Centre-se a discussão apenas na determinação dos coeficientes.

De acordo com o exposto, e considerando os coeficientes b_l nulos ($b_0 = 1$), uma qualquer amostra de um sinal de fala pode ser determinada a partir de uma combinação linear de amostras anteriores e um termo adicional para considerar a excitação. Assim:

$$s(n) = a_1 s(n-1) + a_2 s(n-2) + \dots + a_p s(n-p) + G.u(n) \quad (\text{B.2})$$

Num formato simplificado:

$$s(n) = \sum_{k=1}^p a_k s(n-k) + G.u(n) \quad (\text{B.3})$$

Ou, no domínio Z :

$$S(z) = \sum_{k=1}^p a_k s(z) z^{-k} + G.U(z) \quad (\text{B.4})$$

Admitindo que apenas é possível considerar um número limitado de amostras anteriores e que se desconhece por completo o sinal de excitação do sistema, tem-se como aproximação ao valor real de uma amostra:

$$\tilde{s}(n) = \sum_{k=1}^p \alpha_k s(n-k) \quad (\text{B.5})$$

A diferença entre o valor real e o estimado será o erro de predição, resultando na expressão:

$$e(n) = s(n) - \tilde{s}(n) = s(n) - \sum_{k=1}^p \alpha_k s(n-k) \quad (\text{B.6})$$

Para a determinação dos parâmetros do modelo o objectivo será então encontrar os valores de α_k que minimizem o erro $e(n)$. Uma vez que à partida se desconhece qualquer tipo particular de distribuição de probabilidade para os parâmetros procede-se tradicionalmente ao cálculo do erro médio quadrático considerando um intervalo limitado de amostras próximas de n . Assim, sendo $s_n(m)$ um sinal constituído por amostras na vizinhança de n , o erro de predição total E_n (designado por erro de predição de curta duração ou *short-term prediction error*), será:

$$E_n = \sum_m e_n^2(m) = \sum_m \left[s_n(m) - \sum_{k=1}^p \alpha_k s_n(m-k) \right]^2 \quad (\text{B.7})$$

Através de derivadas parciais em relação a cada α_k e igualando a zero tem-se:

$$\frac{\partial E_n}{\partial \alpha_k} = 0, \quad k = 1, 2, \dots, p \quad (\text{B.8})$$

Chegando-se à expressão:

$$\sum_m s_n(m-i) s_n(m) = \sum_{k=1}^p \hat{\alpha}_k \sum_m s_n(m-i) s_n(m-k) \quad (\text{B.9})$$

Observando a equação B.9 surgem óbvias semelhanças entre os termos e a função de correlação $\Phi_n(i, k)$ que se apresenta como referência:

$$\Phi_n(i, k) = \sum_m s_n(m-i) s_n(m-k) \quad (\text{B.10})$$

Assim, efectuando a substituição de B.10 em B.9 chega-se à expressão final:

$$\sum_{k=1}^p \alpha_k \Phi(i, k) = \Phi(i, 0), \quad i = 1, 2, \dots, p \quad (\text{B.11})$$

A expressão B.11 representa um sistema de p equações com p variáveis, no formato de Yule-Walker, que pode ser facilmente resolvido para a determinação dos coeficientes α_k .

Para a resolução deste sistema de equações são sugeridos vários métodos com abordagens mais ou menos eficientes e com múltiplos graus de complexidade matemática.

B.1 Método da Autocorrelação

O método da auto-correlação é tido como a abordagem clássica e é o que resulta de se considerarem as sequências de sinais s_n como obtidas de sinais infinitos por truncatura.

A equação B.9 pode ser re-escrita efectuando uma substituição pela função de auto-correlação avaliada em $(i - k)$.

$$R_n(k) = \sum_{m=0}^{N-1-k} s_m(m)s_n(m+k) \quad (\text{B.12})$$

com

$$\Phi_n(i, k) = R_n(i - k) \quad (\text{B.13})$$

Assim, é possível escrever na forma matricial:

$$\begin{bmatrix} R(0) & R(1) & \dots & R(p-1) \\ R(1) & R(0) & \dots & R(p-2) \\ \vdots & \vdots & & \vdots \\ R(p-1) & R(p-2) & \dots & R(0) \end{bmatrix} * \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \vdots \\ R(p) \end{bmatrix} \quad (\text{B.14})$$

A matriz principal é uma matriz de auto-correlações dando portanto o nome a este método.

O erro de predição vem também como:

$$E_n = R_n(0) - \sum_{k=1}^p \alpha_k R_n(k) \quad (\text{B.15})$$

O problema pode ser resolvido pelo tradicional método de Gauss-Jordan mas, tendo em conta a simetria da matriz e a igualdade dos elementos da diagonal, matriz Toeplitz, é aconselhado o algoritmo de Levinson-Durbin, computacionalmente eficiente e optimizado para este caso particular. Este algoritmo inicialmente desenvolvido e apresentado por Levinson em 1947, foi posteriormente modificado e optimizado por Durbin em 1959 [30], é um algoritmo recursivo que começa com o cálculo de um modelo de ordem unitária que será sucessivamente aumentada sendo os novos coeficientes calculados a partir dos do modelo de menor ordem.

Resumidamente as equações que traduzem o algoritmo apresentam-se:

$$\begin{aligned} E^{(0)} &= R[0] \\ k_i &= \frac{\left\{ R[i] - \sum_{j=1}^{(i-1)} a_j R[i-j] \right\}}{E^{(i-j)}}, 1 \leq i \leq p \\ a_i^{(i)} &= k_i \\ a_j &= a_j^{(i-1)} - k_i a_{i-j}^{(i-1)}, i \leq j \leq i-1 \\ E^{(i)} &= (1 - k_i^2) E^{(i-1)} \end{aligned}$$

A segunda e última equações são resolvidas de modo iterativo com $i = 1, 2, \dots, p$ e o resultado final é dado por:

$$a_j = a_j^{(p)}$$

O algoritmo de Schur (1917) pode também ser utilizado em alternativa ao de Durbin mas com algumas adaptações. Através de uma técnica de processamento paralelo os tempo de cálculo são quadraticamente reduzidos. Mais pormenores podem ser encontrados em [64].

B.2 Método da Covariância

O método da covariância baseia-se em princípios semelhante mas deixa de considerar o sinal limitado a um intervalo. As equações conduzem neste caso a uma sistema idêntico mas baseado numa matriz de covariâncias, também simétrica mas não Toeplitz.

$$\begin{bmatrix} C(1,1) & C(1,2) & \dots & C(1,p) \\ C(2,1) & C(2,2) & \dots & C(2,p) \\ \vdots & \vdots & & \vdots \\ C(p,1) & C(p,2) & \dots & C(p,p) \end{bmatrix} * \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_p \end{bmatrix} = \begin{bmatrix} C(1,0) \\ C(2,0) \\ \vdots \\ C(p,0) \end{bmatrix} \quad (\text{B.16})$$

Mais uma vez, para o cálculo das soluções, pode ser aplicado um algoritmo optimizado designado por decomposição de Cholesky e que conduz a resultados mais eficientemente. A utilização do método da covariância pode resultar num número de equações mais simples, com resultados um pouco diferentes do método da autocorrelação, mas que, de acordo com [15], conduz a modelos com melhor desempenho. Existe porém a possibilidade de o filtro obtido não ser estável, situação que não ocorre optando pela autocorrelação. A dificuldade em ajustar o modelo à realidade que ocorre deriva dos problemas de condicionamento da matriz do sistema que pode facilmente degradar-se. O esforço computacional dos métodos apresentados é, de acordo com [68], sensivelmente o mesmo.

Apêndice C

Tabelas IPA e SAMPA para o Português Europeu

O actual alfabeto fonético IPA (*International Phonetic Alphabet*), evolução de sucessivas propostas de Rousselot, Giliéron e Straka, e desenvolvido com o objectivo de ser uma norma internacional possui um conjunto vasto de símbolos para a representação dos sons na maioria das línguas conhecidas e estudadas. Os símbolos IPA de difícil representação num computador levaram ao desenvolvimento do SAMPA. Este utiliza caracteres internacionais e disponíveis de forma quase universal nos teclados dos computadores.

Os alfabetos podem ser consultados na sua totalidade nos endereços internet respectivos.

- IPA: <http://www.arts.gla.ac.uk/IPA/index.html>
- SAMPA: <http://www.phon.ucl.ac.uk/home/sampa/portug.htm>

Na tabela apresentam-se apenas os símbolos principais respeitantes ao português europeu, 20 consoantes, 14 vogais e 4 semi-vogais.

#	Classificação	SAMPA	IPA	Exemplo
01	Silêncio	-		
02	Oclusivas	p	p	pai
03		b	b	bar
04		t	t	tia
05		d	d	data
06		k	k	casa
07		g	g	gato
08	Fricativas	f	f	férias
09		v	v	vaca
10		s	s	selo
11		z	z	azul
12		S	ʃ	chave
13		Z	ʒ	agir
14	Nasais	m	m	meta
15		n	n	neta
16		J	ɲ	senha
17	Líquidas	l	l	lado
18		l~	ɫ	sal
19		L	ʎ	folha
20	Vibrantes	r	r	caro
21		R	R	carro
22	Vogais	i	i	fita
23		e	e	pêra
24		E	ɜ	seta
25		a	a	caro
26		6	ɐ	cama
27		O	ɔ	corda
28		o	o	sopa
29		u	u	muda
30		@	ə	deste
31		i~	ĩ	pinta
32		e~	ẽ	menta
33		6~	õ	manta
34		o~	õ	ponta
35		u~	ũ	mundial
36	Semi-vogais	w	w	pau
37		j	j	pai
38		w~	ẽ	pão
39		j~	ĩ	mãe

Tabela C.1: Símbolos IPA e SAMPA para o português europeu

Apêndice D

Características da Base de Dados FEUP/IPB

A base de dados FEUP/IPB, desenvolvida pelo Laboratório de Sinais e Sistemas (LSS) da Universidade do Porto, surge de uma colaboração entre a Faculdade de Engenharia da Universidade do Porto e a Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Bragança. A principal motivação para a sua criação foi a inexistência de uma base de dados de fala de elevada qualidade que pudesse ser utilizada para o desenvolvimento de sistemas de fala em português europeu.

Algumas das características desta base de dados foram já apresentadas ao longo do texto principal deste trabalho, as restantes, ou por serem adjacentes ao tema ou por motivos de organização foram reservadas para este anexo. Nos quadros seguintes apresentam algumas particularidades deste corpus que em muitas situações, considerando as devidas reservas, podem ser extrapoladas para a língua Portuguesa.

	1	2	3	4	5	6	7
p	77	91	50	63	57	74	35
b	38	27	18	30	7	38	8
t	141	117	90	102	89	81	64
d	136	111	84	85	73	128	56
k	89	120	57	79	66	87	31
g	23	42	15	54	10	7	10
f	25	27	15	27	15	12	15
v	30	33	36	34	21	13	20
s	102	116	64	86	69	106	34
z	46	26	27	36	30	17	10
S	113	108	64	50	71	121	40
Z	35	45	43	43	31	15	17
m	62	84	42	47	44	70	20
n	45	56	31	48	30	38	8
J	10	9	11	6	7	4	1
l	52	38	40	38	22	43	38
l*	19	32	22	28	7	0	0
L	13	11	5	6	0	6	1
r	182	141	93	132	123	132	78
R	19	16	11	15	20	11	4
i	141	109	78	112	93	129	79
e	26	77	36	35	21	106	30
E	35	34	35	27	30	31	24
a	87	97	50	99	66	90	54
6	249	201	174	201	195	181	127
O	25	35	18	44	26	5	20
o	43	44	19	39	29	77	15
u	120	124	76	109	72	142	82
@	45	36	33	24	32	68	24
i~	21	10	2	22	13	6	11
e~	37	29	25	22	12	28	29
6~	73	86	44	46	46	38	18
o~	16	26	14	19	13	49	11
u~	15	24	8	9	5	12	7
w~	3	0	0	0	0	23	10
j~	6	0	0	0	0	20	5
w	73	73	37	41	51	15	19
j	61	89	47	59	41	43	26
Σ	2.333	2.344	1.514	1.917	1.537	2.066	1.081

Tabela D.1: Ocorrências de fonemas por cada track da base de dados FEUP/IPB (Tracks de 1 a 7 de 13).

X	8	9	10	11	12	13	Total
p	44	98	502	209	11	16	1327
b	22	27	134	95	0	3	447
t	85	204	838	490	24	46	2371
d	93	203	914	554	27	20	2484
k	76	150	691	324	11	52	1833
g	12	27	182	101	9	7	499
f	17	39	165	92	11	5	465
v	24	26	104	97	10	9	457
s	65	135	620	291	23	10	1721
z	36	32	139	78	5	1	483
S	110	162	680	532	14	34	2099
Z	48	18	107	63	1	3	469
m	53	131	591	301	12	21	1478
n	45	85	420	311	18	11	1146
J	7	8	34	24	2	2	125
l	20	101	336	275	14	12	1029
l*	15	0	120	62	0	3	308
L	10	22	26	20	0	1	121
r	154	259	1000	553	29	25	2901
R	6	28	95	76	4	2	307
i	108	266	981	662	31	22	2811
e	33	151	917	441	35	20	1928
E	51	26	192	81	8	4	578
a	111	143	719	383	25	19	1943
6	253	340	1283	950	38	53	4245
O	24	4	46	29	0	0	276
o	36	143	601	347	14	39	1446
u	128	304	1135	647	24	64	3027
@	52	113	638	321	26	16	1428
i~	13	20	84	39	5	4	250
e~	24	38	233	149	8	5	639
6~	68	77	452	196	10	19	1173
o~	20	48	186	106	1	10	519
u~	11	12	93	50	2	1	249
w~	0	36	161	81	3	3	320
j~	0	30	168	77	6	4	316
w	59	36	204	74	3	3	688
j	73	44	234	117	6	10	850
Σ	2.006	3.586	16.025	9.298	470	579	44.756

Tabela D.2: Ocorrências de fonemas por cada track da base de dados FEUP/IPB (Tracks de 8 a 13 de 13 e Totais).

	p	b	t	d	k	g	f	v	s	z	S	Z	m	n	J
p	0	0	0	0	0	0	0	0	0	3	0	0	2	0	0
b	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
t	0	0	0	1	0	0	1	1	4	1	12	0	5	0	0
d	0	0	0	1	0	0	6	1	4	2	4	2	1	0	1
k	0	0	0	0	0	0	1	0	9	0	1	1	2	7	1
g	0	0	0	0	0	0	0	0	1	0	1	0	0	4	0
f	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0
v	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
s	1	0	0	0	0	0	2	1	1	1	1	1	1	3	0
z	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
S	0	0	1	0	0	0	2	0	10	0	1	1	2	1	0
Z	0	0	0	1	0	0	0	3	0	0	2	0	9	5	0
m	0	0	0	0	0	0	0	0	1	0	2	0	0	1	0
n	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
J	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
l	0	0	0	0	0	0	0	2	0	0	0	1	0	0	0
l*	0	0	0	0	0	0	2	2	1	0	6	0	3	0	0
L	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
r	0	0	0	0	0	0	1	1	16	1	3	1	1	3	0
R	0	0	0	0	0	0	0	2	1	2	1	0	0	0	0
i	0	0	0	0	2	0	1	9	19	9	8	4	21	6	1
e	0	0	0	0	0	0	0	2	5	1	10	4	10	5	0
E	0	0	0	0	0	0	0	0	14	0	6	1	1	2	0
a	0	0	0	0	0	0	0	6	7	7	4	1	3	3	0
6	0	0	1	0	0	0	6	9	24	5	31	17	9	7	3
O	0	0	0	0	0	0	2	3	3	1	1	1	6	1	0
o	0	0	0	0	0	0	0	2	1	1	0	1	0	0	1
u	0	0	0	0	0	0	7	3	8	2	21	7	25	4	2
@	0	0	0	0	0	0	0	5	3	0	2	2	1	4	0
i~	0	0	0	0	0	0	0	2	1	0	0	2	0	0	0
e~	0	0	1	0	0	0	4	3	6	0	1	0	0	0	0
6~	0	0	1	0	0	0	0	0	1	0	0	2	0	0	0
o~	1	0	0	0	0	0	1	0	0	0	1	1	1	0	0
u~	0	0	0	0	0	0	1	0	1	2	2	0	1	2	0
w~	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
j~	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
w	0	0	0	0	0	0	2	0	4	0	8	0	3	6	0
j	0	0	0	0	0	0	1	0	3	3	17	9	2	0	0
X	2	0	2	0	2	0	2	0	9	0	1	0	2	2	0
XX	8	0	5	0	13	0	4	1	8	0	1	1	4	4	0
!	137	43	189	171	160	66	0	0	0	0	0	0	0	0	0

Tabela D.3: Ocorrências de alguns difones na base de dados FEUP/IPB para os *tracks* 1 e 2.

Apêndice E

Símbolos Utilizados

Para a anotação do corpus FEUP/IPB utilizou-se um conjunto de 41 fonemas a que correspondeu o mesmo número de modelos HMM no sistema baseado nesta técnica. Este conjunto apresenta-se na tabela E.1.

Para considerar outras possibilidades este conjunto foi alargado para 47 fonemas. Este novo conjunto apresenta-se na tabela E.2.

Para a anotação da base de dados TIMIT utilizaram-se 48 fonemas da totalidade dos 61 disponíveis. A redução obedeceu ao procedimento utilizado por Lee [46] e os dois conjuntos mostram-se na tabela E.3.

#	Classificação	SAMPA	FEUP/IPB
01	Oclusivas	p	p
02		t	t
03		t	t
04		b	b
05		d	d
06		g	g
07	Fricativas	f	f
08		v	v
09		s	s
10		z	z
11		S	sm
12		Z	zm
13	Nasais	m	m
14		n	n
15		J	jm
16	Líquidas	l	l
17		l~	l_
18		L	lm
19	Vibrantes	r	r
20		R	rm
21	Vogais	i	i
22		e	e
23		E	em
24		a	a
25		6	sx
26		O	om
27		o	o
28		u	u
29		@	at
30		i~	i_
31		e~	e_
32		6~	sx_
33		o~	o_
34		u~	u_
35	Semi-vogais	w	w
36		j	j
37		w~	w_
38		j~	j_
39	Silêncio	-	sil
40	Aspiração		asp
41	Oclusão		ocl

Tabela E.1: Conjunto base de HMMs para a base de dados FEUP/IPB
 Conjunto base de modelos HMM considerados para a base de dados
 FEUP/IPB e sons SAMPA correspondentes.

#	Classificação	SAMPA	FEUP/IPB
01	Oclusivas	p	p
02		t	t
03		t	t
04		b	b
05		d	d
06		g	g
07	Fricativas	f	f
08		v	v
09		s	s
10		z	z
11		S	sm
12		Z	zm
13	Nasais	m	m
14		n	n
15		J	jm
16	Líquidas	l	l
17		l~	l_
18		L	lm
19	Vibrantes	r	r
20		R	rm
21	Vogais	i	i
22		e	e
23		E	em
24		a	a
25		6	sx
26		O	om
27		o	o
28		u	u
29		@	at
30		i~	i_
31		e~	e_
32		6~	sx_
33		o~	o_
34		u~	u_
35	Semi-vogais	w	w
36		j	j
37		w~	w_
38		j~	j_
39	Silêncio	-	sil
40	Aspiração		asp
41..46	Oclusão		oclp, oclt, oclb, oclg
47	Pausa		sp

Tabela E.2: Conjunto alargado de HMMs para a base de dados FEUP/IPB
Conjunto alargado de modelos HMM considerados para a base de dados
FEUP/IPB e sons SAMPA correspondentes

#	TIMIT	Lee	#	TIMIT	Lee
01	p	p	32	b	b
02	t	t	33	d	d
03	k	k	34	g	g
04	pcl	cl	35	bcl	vcl
05	tcl	cl	36	dcl	vcl
06	kcl	cl	37	gcl	vcl
07	dx	dx	38	q	
08	m	m	39	em	m
09	n	n	40	en	en
10	ng	ng	41	eng	ng
11	nx	n	42	sh	sh
12	s	s	43	zh	zh
13	z	z	44	jh	jh
14	ch	ch	45	dh	dh
15	th	th	46	v	v
16	f	f	47	el	el
17	l	l	48	w	w
18	r	r	49	h#	sil
19	y	y	50	epi	epi
20	pau	sil	51	hv	hh
21	hh	hh	52	ih	ih
22	eh	eh	53	ae	ae
23	ao	ao	54	ah	ah
24	aa	aa	55	uh	uh
25	uw	uw	56	ux	uw
26	er	er	57	oy	oy
27	ay	ay	58	iy	iy
28	ey	ey	59	ow	ow
29	aw	aw	60	axr	er
30	ax	ax	61	ax-h	ah
31	ix	ix			

Tabela E.3: Conjunto de HMMs para a base de dados TIMIT
Conjunto de modelos HMM considerados para a segmentação da base de dados TIMIT.

Apêndice F

Acrónimos e Abreviaturas

ANN - Artificial Neural Networks
AR - Auto-Regressive
ARMA - Auto-Regressive Moving Average
CD - Compact Disc
DARPA - Defense Advanced Research Projects
DCT - Discrete Cosine Transform
DTW - Dynamic Time Warping
FEUP - Faculdade de Engenharia da Universidade do Porto
FFT - Fast Fourier Transform
FIR - Finite Impulse Response
HMM - Hidden Markov Model
IIR - Infinite Impulse Response
IPB - Instituto Politécnico de Bragança
KLT - Karhunen-Loeve Transformation
LPC - Linear Predictive Coding
LSF - Line Spectral Frequencies
LSP - Line Spectrum Pairs
MA - Moving Average
MAP - Maximum A posteriori Probability
MFCC - Mel Frequency Cepstral Coefficients
MIT - Massachusetts Institute of Technology
MLP - Multi Layer Perceptron
MLR - Maximum Likelihood Ratio
NIST - National Institute of Standards and Technology
OGI - Oregon Graduate Institute
PCA - Principal Component Analysis
PLP - Perceptual Linear Prediction
PSOLA - Pitch Synchronous Overlap and Add
RMS - Root Mean Square
SRI - Stanford Research Institute
TI - Texas Instruments
ToBI - Tone and Break Indices
WER - Word Error Rate

Bibliografia

- [1] AKANSU, A. N., AND HADDAD, R. A. *Multiresolution Signal Decomposition - Transforms, Subbands, Wavelets*, 2nd ed. Academic Press, 2, 2001.
- [2] AMARAL, R., CARVALHO, P., CASEIRO, D., TRANCOSO, I., AND OLIVEIRA, L. Anotação fonética automática de corpora de fala transcritos ortograficamente. *Actas do PROPOR'99 - IV Encontro para o Processamento Computacional da Língua Portuguesa Escrita e Falada* (Setembro 1999).
- [3] AMARAL, R., PERDIGÃO, F., MARTA, E. S., AND DE SÁ, L. V. Automatic segmentation and labelling of a portuguese telephone-speech digits database. *RecPad'97* (1997).
- [4] BARRAS, C., GEOFFROIS, E., WU, Z., AND LIBERMAN, M. Transcriber: A free tool for segmentation, labelling and transcribing speech. In *Proc. LREC '98* (1998), pp. 1373–1376.
- [5] BARROS, M. J. Estudo comparativo e técnicas de geração de sinal para a síntese da fala. Master's thesis, Faculdade de Engenharia da Universidade do Porto, 2001.
- [6] BARROSO, H. *Forma e Substância da Expressão da Língua Portuguesa*. Almedina, Coimbra, Maio 1999.
- [7] BAUM, L. E., AND EAGON, J. A. An inequality with applications to statistical estimation for probabilistic functions of markov processes and to a model for ecology. *Bulletin of American Mathematical Society* 73 (1967), 360–363.
- [8] BOERSMAN, P., AND WEENICK, D. <http://www.fon.hum.uva.nl/praat>. *PRAAT: Doing Phonetics by Computer* (1999).
- [9] BRAGA, D., AND FREITAS, D. Towards an intonation module for a portuguese tts system. In *Proc. ICSLP '02* (Denver-Colorado, USA, 2002).
- [10] BRAGA, D., FREITAS, D., TEIXEIRA, J. P., BARROS, M. J., AND LATSH, V. Back close non-syllabic vowel [u] behaviour in european portuguese: Reduction or suppression. In *Proc. ICSP '01* (Korea, 2001).
- [11] BRAGA, D., FREITAS, D., TEIXEIRA, J. P., BARROS, M. J., AND LATSH, V. Correlation between phonetic factors and linguistic events

- regarding a prosodic pattern of european portuguese: a practical proposal. In *Proc. ICSP '01* (Korea, 2001).
- [12] BRAGA, D., FREITAS, D., TEIXEIRA, J. P., MARQUES, M. A., AND COELHO, L. On the use of prosodic labelling in corpus-based linguistic studies of spontaneous speech. In *Proc. TSD '03* (Ceske Budejovice, Tchech Republic, 2003).
 - [13] BRUGNARA, F., FALAVIGNA, D., AND OMOLOGO, M. A hmm-bases system for automatic segmentation and labelling of speech. *Proceedings of ICSLP'92* (October 1992), 803–806.
 - [14] BRUGNARA, F., FALAVIGNA, D., AND OMOLOGO, M. A hmm-bases system for automatic segmentation and labelling of speech basend on hidden markov models. *Speech Communication* (1993), 357–370.
 - [15] BURRUS, C. S., AND MCCLELLAN, J. H. *Computer-Based Exercises for Signal Processing*. Prentice-Hall, 1994.
 - [16] CARVALHO, P. *Determinação Automática de Segmentos para Síntese de Fala por Concatenação*. PhD thesis, Universidade Técnica de Lisboa - Instituto Superior Técnico, 2004.
 - [17] CASEIRO, D., TRANCOSO, I., OLIVEIRA, L., AND VIANA, C. Grapheme-to-phone using finite-state transducers. *IEEE Workshop on Speech Synthesis (TTS'02)* (September 2002).
 - [18] CHAPPELL, D. T., AND HANSEN, J. Articulation and phonetics: a new approach. *Speech Communications* 36, 3-4 (March 2002), 343–373.
 - [19] CHOMSKY, N., AND HALLE, M. *The Sound Pattern of English*. Harper and Row, New York, 1968.
 - [20] COLE, R., OSHIKA, B. T., NOEL, M., AND LANDER, T. Labeler agreement in phonetic labeling of continuous speech. In *Proc. ICSLP '94* (Yokohama, Japan, 1994), pp. 2131–2134.
 - [21] COSI, P., FALAVIGNA, D., AND OMOLOGO, M. A preliminary statiscal evaluation of manual and automatic segmentation discrepancies. *Proceedings of Eurospeech '91* (1991), 693–696.
 - [22] DAVIS, S., AND MERMELSTEIN, P. Comparison of parametric representations for monosyllable word recognition in countinuosly spoken sentences. *IEEE Trans. on Acoustics, Speech and Signal Procesing* 28, 4 (1980), 357–336.
 - [23] DE LURDES MOUTINHO, M. *Uma Introdução ao Estudo da Fonética e Fonologia do Português*, 1st ed. Plátano Edições Técnicas, Lisboa, August 2000.
 - [24] DELGADO-MARTINS, M. R. Vogais e consoantes do português: Estatística de ocorrências, duração e intensidade. *Boletim de Filologia XXIV*, 1/4 (1975), 1–11.

- [25] DELGADO-MARTINS, M. R. *Ouvir falar: Introdução à Fonética do Português*. Editorial Caminho, 1988.
- [26] DELGADO-MARTINS, M. R. *Fonética do Português: Trinta Anos de Investigação*. Editorial Caminho, 2002.
- [27] DELLER, J., HANSEN, J., AND PROAKIS, J. *Discrete Time Processing of Speech Signals*. IEEE Press, Wiley-Interscience, New York, 2000.
- [28] DEMUYNCK, K., AND LAUREYS, T. A comparison of different approaches to automatic speech segmentation. *Speech Communication* (2002), 357–370.
- [29] DONOVAN, R. E. A new distance measure for costing spectral discontinuities in concatenative speech syntheses. In *Proc. SSWC '01* (2001).
- [30] DURBIN, J. Efficient estimation of parameter in moving average models. *Biometrika* 46 (1959), 306–316.
- [31] EUROPEAN UNION EXPERT ADVISORY GROUP ON LANGUAGE ENGINEERING STANDARDS. *Handbook of Standards and Resources for Spoken Language Systems*. Dafydd Gibbon and Roger More and Richard Winski, 1997.
- [32] GALLISON, R., AND COSTE, D. *Dicionário de Didáctica das Línguas*. Livraria Almedina, 1983.
- [33] GHOLAMPOUR, I., AND NAYEBI, K. A new fast algorithm for automatic segmentation of continuous speech. *Proceedings of ICSLP '98* (1998).
- [34] GOLDBERG, R., AND RIEK, L. *A Practical Handbook of Speech Coders*. CRC Press, 2000.
- [35] GOUVEIA, P., TEIXEIRA, J. P., AND FREITAS, D. Divisão silábica automática do texto escrito e falado. *Propor 2000* (2000). S. Paulo.
- [36] GUT, U., AND BAYER, P. S. Measuring the reliability of manual annotations of speech corpora. In *Proc. Speech Prosody 2004* (Japan, 2004).
- [37] HOSOM, J. *Automatic Time Alignment of Phonemes using Acoustic-Phonetic Information*. PhD thesis, Oregon Graduate Institute of Science and Technology, May 2000.
- [38] HUANG, X., ACERO, A., AND HON, H.-W. *Spoken Language Processing, A Guide to Theory, Algorithm, and System Development*. Prentice-Hall PTR, NJ, 2001.
- [39] HUCKVALE, M. <http://www.phon.ucl.ac.uk/resource/sfs>. *Speech Filling System Tools for Speech Research* (2002).
- [40] ITAKURA, F. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Signal Processing* 23, 1 (Feb 1975), 67–72.
- [41] JAKOBSON, R., FANT, G., AND HALLE, M. *Preliminaries to Speech Analysis*. MIT Press, Cambridge, 1952.

- [42] KABAL, P. The computation of line spectral frequencies using chebyshev polynomials. *IEEE Transactions on Acoustics, Speech and Signal Processing* 34, 6 (December 1986), 1419–1426.
- [43] KELLER, E. Neural network motivation for segmental distribution. In *Proc. ICSLP '98* (Sydney, Australia, 1998).
- [44] KIPP, A., WESENICK, M., AND SCHIEL, F. Automatic detection and segmentation of pronunciation variants in german speech corpora. *Proceedings of ICSLP'96* (October 1996), 106–109.
- [45] LANDER, T. *The CSLU Labeling Guide*. Oregon Graduate Institute, May 1997.
- [46] LEE, K. F., AND HON, H. W. Speaker independent phone recognition using hidden markov models. *IEEE Transactions on Acoustics, Speech and Audio* (1989).
- [47] LEUNG, H., AND ZUE, V. A procedure for automatic alignment of phonetic transcriptions with continuous speech. *Proceeding of ICASSP '84* (1984), 2.7.1–2.7.4.
- [48] LINGUATECA. <http://www.linguateca.pt/acdc/>.
- [49] LIPPMANN, R., AND SINGER, E. Hybrid neural networks/hmm approaches to wordspotting. In *Proceedings of ICASSP'93* (1993).
- [50] LJOLJE, A., HIRSCHBERG, J., AND SANTEN, J. V. *Automatic Speech Segmentation for Concatenative Inventory Selection*. Springer-Verlag, New York, 1997.
- [51] LJOLJE, A., AND RILEY, M. Automatic segmentation and labelling of speech. *Proceeding of ICASSP '91* (1991), 473–476.
- [52] LJUNG, L. *System Identification: Theory for the User*. Prentice-Hall, Englewood Cliffs, NJ, 1987.
- [53] MAKHOUL, J. Linear prediction: A tutorial review. *Proceedings of IEEE* 63, 4 (April 1975), 561–580.
- [54] MALFRERE, F., DEROO, O., AND DUTOIT, T. Phonetic alignment: Speech synthesis vs. hybrid hmm/ann. *Proceedings of ICSLP '98 4* (December 1998), 1571–1574.
- [55] MARTINS, C., MASCARENHAS, M. I., MEINADO, H., AND NETO, J. P. Spoken language corpora for speech recognition and synthesis in european portuguese. *Proceedings of RECPAD '98* (March 1998).
- [56] MATEUS, M. H. M. Aspectos da fonologia do portugues. *Revista do Centro de Linguística da Universidade de Lisboa* (1972).
- [57] MATEUS, M. H. M. Aspectos da fonologia portuguesa. *Revista do Centro de Estudos Filológicos* (1974).
- [58] MICROSOFT. *Encarta*. Microsoft, 2003.

- [59] MOREIRA, L. F. S. Desenvolvimento de um sistema de reconhecimento de fala. Master's thesis, Faculdade de Engenharia da Universidade do Porto, Setembro 2004.
- [60] MUTHUSAMY, Y. K., COLE, R. A., AND OSHIKA, B. T. The ogi multi-language telephone speech corpus. *Proceedings of ICSLP '92* (October 1992), 895–898.
- [61] PARKS, T., AND BURRUS, C. *Digital Filter Design*. John Wiley and Sons, 1987.
- [62] PORTO EDITORA. *Diciopédia 2005*. Porto Editora, Porto, 2004.
- [63] PÊRA, V. C. *Reconhecimento de Fala Contínua com Processamento Simultâneo de Diferentes Características do Sinal*. PhD thesis, Faculdade de Engenharia da Universidade do Porto, 2001.
- [64] PROAKIS, J. G., RADER, C. M., LING, F., NIKIAS, C. L., MOONEN, M., AND PROUDLER, I. K. *Algorithms for Statistical Signal Processing*. Prentice-Hall, 2002.
- [65] QIAN, S. *Time-Frequency and Wavelets Transforms*. Prentice-Hall, 2002.
- [66] RABINER, L. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77, 2 (February 1989), 257–286.
- [67] RABINER, L. R., AND JUANG, B.-H. *Fundamentals of Speech Recognition*. Prentice-Hall, NJ, 1993.
- [68] RABINER, L. R., AND SCHAFER, R. W. *Digital Processing of Speech Signals*. Prentice-Hall, NJ, 1978.
- [69] RANCHHOD, E. M. *Tratamento das Línguas por Computador*. Caminho, 2001.
- [70] RAPP, S. Automatic phonemic transcription and linguistic annotation from known text with hidden markov models. *Proceedings of ELSNET Goes East and IMACS Workshop* (1995).
- [71] ROBINSON, T., AND FALLSIDE, F. Phoneme recognition from the timit database using recurrent error propagation networks. *CEUD/F-INFENG/TR.42* (1990).
- [72] ROTHWEILER, J. A rootfinding algorithm for line spectral frequencies. *Proceedings of ICASSP '99* (1999).
- [73] ROWDEN, C. *Speech Processing*. McGraw-Hill Book Company-The Essex Series in Telecommunications and Information Systems, 1992.
- [74] SETHY, A., AND NARAYANAN, S. Refined speech segmentation for concatenative speech synthesis. *Proceedings of ICSLP 2002* (2002).

- [75] SILVERMAN, K., BECKMAN, M., PITRELLI, J., OSTENDORF, M., WHIGHTMAN, C., PRICE, P., PIERREHUMBERT, J., AND HIRSCHBERG, J. Tobi: a standard for labeling english prosody. In *Proc. ICSLP '92* (Banff, Alberta, 1992), pp. 867–870.
- [76] SOONG, F. K., AND JUANG, B. H. Line spectrum pair (lsp) and speech data compression. *Proceedings of ICASSP '84* (1984), 1.10.1–1.10.4.
- [77] STEIGLITZ, K., AND MCBRIDE, L. E. A technique for the identification of linear systems. *IEEE Transactions on Automatic Control AC-10* (1965), 461–464.
- [78] STROM, N. A tonotopic artificial neural network architecture for phoneme probability estimation. In *Proc. ASRU '97* (1997).
- [79] SVENDSEN, T., AND KVALE, K. Automatic alignment of phonemic labels with continuous speech. *Proceedings of ICSLP'90* (1990), 997–1000.
- [80] SVENDSEN, T., AND SOONG, F. On the automatic segmentation of speech signals. *Proceedings of ICASSP '87* (April 1987), 77–88.
- [81] TEIXEIRA, J. P. Modelização paramétrica de sinais para aplicações em sistemas de conversão texto-fala. Master's thesis, Faculdade de Engenharia da Universidade do Porto, Outubro 1995.
- [82] TEIXEIRA, J. P. *Prosody Models for Speech Synthesis*. PhD thesis, Faculdade de Engenharia da Universidade do Porto, 2004.
- [83] TEIXEIRA, J. P., FREITAS, D., BRAGA, D., BARROS, M. J., AND LATSCH, V. Phonetic events from the labeling of the european portuguese database for speech synthesis, feup/ipb-db. *Proceedings of EuroSpeech '01* (2001).
- [84] TORKKOLA, K. Automatic alignment of speech with phonetic transcriptions in real time. *Proceedings of ICASSP'88* (1988), 611–614.
- [85] TRUBETZKOY, N. *Principes de Phonologie*. Royal Academic Publishing, Paris, 1939.
- [86] VAN ERP, A., HOUBEN, C., BARRY, B., GRICE, M., BOE, L. J., BRAUN, G., COSI, P., DYHR, N., PERENON, G., VIGOROUS, N., AND AUTESERRE, D. A unified approach to labelling of speech: First multi-lingual results. *Proceedings of Eurospeech '89 2* (1989), 88–91.
- [87] VAN SANTEN, J., AND SPROAT, R. Highaccuracy automatic segmentation, 1999.
- [88] VINTSYUK, T. Element-wise recognition of continuous speech composed of words from a specified dictionary. *Kibernetika 1*, 7 (March-April 1971), 133–143.
- [89] VITERBI, A. J. Error bounds for convolutional codes and a asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory 13*, 2 (1967), 260–269.

- [90] VORSTERMANS, A., MARTENS, J. P., AND VAN COILE, B. Automatic segmentation and labelling of multi-lingual speech data. *Speech Communications* 19, 4 (1996), 271–293.
- [91] WAGNER, M. Automatic labelling of continuous speech with a given phonetic transcription using dynamic time warping programming algorithms. *Proceeding of ICASSP '81* (1981), 1156–1159.
- [92] WANG, D., LU, L., AND ZHANG, H.-J. Speech segmentation without speech recognition. In *Proc. ICASSP 2003* (2003).
- [93] WELLS, J. Sampa computer readable phonetic alphabet. <http://www.phon.ucl.ac.uk/home/sampa/home.htm> (2000).
- [94] WESENICK, M., AND KIPP, A. Estimating the quality of phonetic transcriptions and segmentations of speech signals. *Proceeding of ICSLP '96* (1996), 129–132.
- [95] WIGHTMAN, C., AND TALKIN, D. *The Aligner: Text-to-Speech Alignment Using Hidden Markov Models*. Springer-Verlag, New York, 1997.
- [96] WOODLAND, P., LEGETTER, C., ODELL, J., VALTCHEV, V., AND YOUNG, S. The 1994 htk large vocabulary speech recognition system. *Proceedings of ICASSP '95* (May 1995), 73–76.
- [97] WOOTERS, C. H. *Lexical Modelling in a Speaker Independent Speech Understanding System*. PhD thesis, U. C. Berkley, 1993.
- [98] YOUNG, S., EVERMANN, G., HAIN, T., KERSHAW, D., MOORE, G., ODELL, J., OLLASON, D., POVEY, D., VALTCHEV, V., AND WOODLAND, P. *The HTK Book*, 3.2 ed. Microsoft Corporation and Cambridge University Engineering Department, December 2002.
- [99] YOUNG, S. J., ODELL, J. J., AND WOODLAND, P. Tree-based state tying for high accuracy acoustic modelling. In *Proc. ARPA Human Language Technology Conference '94* (1994).
- [100] ZEMLIN, W. R. *Speech and Hearing Science. Anatomy and Physiology*, 3rd ed. Prentice Hall, New York, 1988.