



UNIVERSIDADE TÉCNICA DE LISBOA

Instituto Superior Técnico

Identificação Automática da Língua em Fala Contínua

Diamantino António Caseiro
(Licenciado)

Dissertação para a obtenção do grau de Mestre em
Engenharia Electrotécnica e de Computadores

Tese Realizada sob a orientação de
Isabel Maria Martins Trancoso
Professora Associada do Instituto Superior Técnico
da Universidade Técnica de Lisboa

Constituição do Júri:

Doutora Isabel Maria Martins Trancoso
Doutor Luís António Serralva Vieira de Sá
Doutor António Joaquim dos Santos Romão Serralheiro
Doutor Luís Miguel Veiga Vaz Caldas de Oliveira

Lisboa, Março de 1998

ÍNDICE

1. Introdução	1
1.1 O Problema	1
1.2 Motivação	2
1.3 Estrutura da tese	4
2. Enquadramento da Identificação Automática da Língua	
Falada	5
2.1 Relação da identificação automática da língua falada com outras áreas	5
2.1.1 Tarefas em identificação da língua	6
2.1.2 Identificação da língua falada vs. identificação da língua escrita	7
2.2 Informação utilizada em identificação automática da língua falada	7
2.2.1 Acústica/Fonética	8
2.2.2 Fonotáctica	8
2.2.3 Prosódia	9
2.2.4 Vocabulário	9
2.2.5 Gramática	9
2.3 Identificação humana da língua falada	10
3. Abordagens à Identificação Automática da Língua	
Falada	14
3.1 As primeiras abordagens	14
3.2 Abordagens recentes	25

3.2.1	Utilização de informação acústica	25
3.2.2	Utilização de informação prosódica	26
3.2.3	Identificação da língua falada por identificação de orador	28
3.2.4	Utilização de informação fonotáctica	29
3.2.4.1	<i>Tipos de informação segmental utilizados</i>	30
3.2.4.2	<i>Reconhecimento paralelo de fones</i>	32
3.2.4.3	<i>Phoneme-Recognition followed by Language Modeling (PRLM)</i>	33
3.2.4.3.1	<i>PRLM – Paralelo (PRLM-P)</i>	34
3.2.4.3.2	<i>Double Bigram Decoding (DBD)</i>	35
3.2.5	Utilização de vocabulário	36
3.2.5.1	<i>Pseudo-palavras</i>	37
3.2.5.2	<i>Ocorrência de palavras comuns</i>	37
3.2.6	Identificação da língua falada por reconhecimento de fala continua de largo vocabulário	38
4.	Corpus	40
4.1	O <i>corpus</i> SpeechDat	40
4.2	Seleção de material de treino e teste	41
5.	Um Sistema de Base para a Identificação Automática da Língua Falada	44
5.1	Arquitectura	44
5.2	Extracção de parâmetros acústicos	45
5.3	Reconhecedor de fones	47
5.3.1	Unidades fonéticas	48
5.3.2	Treino	48
5.3.3	Reconhecimento de fones	49

5.3.4	Desempenho do reconhecedor	50
5.4	Modelos de língua	51
5.5	Avaliação do sistema	52
6.	Extensões ao Sistema de Base	55
6.1	Modelos de língua	55
6.1.1	Trigramas	55
6.1.2	Bigramas Esquerdo e Direito	56
6.1.3	Mapeamento do contexto fonético	57
6.1.4	Generalização do mapeamento do contexto fonético	59
6.2	Arquitectura	60
6.2.1	<i>Bootstrapped</i> DBD	60
6.3	Avaliação das extensões ao sistema base	61
6.3.1	Modelos de língua	61
6.3.2	Arquitectura DBD	63
6.3.3	Efeito da duração da locução de teste	64
7.	Conclusões e Trabalho Futuro	67
7.1	Análise de resultados	67
7.2	Desenvolvimentos futuros	68
A.	Identificação Automática da Língua Escrita	70
A.1	O que é a identificação automática da língua escrita	70
A.2	Aplicações da identificação da língua escrita	71
A.3	Metodologias utilizadas em identificação da língua escrita	72
A.3.1	Fontes de informação para a identificação da língua escrita	73
A.3.1.1	<i>O alfabeto</i>	74

A.3.1.2	<i>Sequências de caracteres</i>	74
A.3.1.3	<i>Palavras comuns</i>	75
A.4	Experiências em identificação da língua escrita	76
A.4.1	<i>Corpus</i> e preparação dos dados	77
A.4.2	Resultados	78
	Bibliografia	80