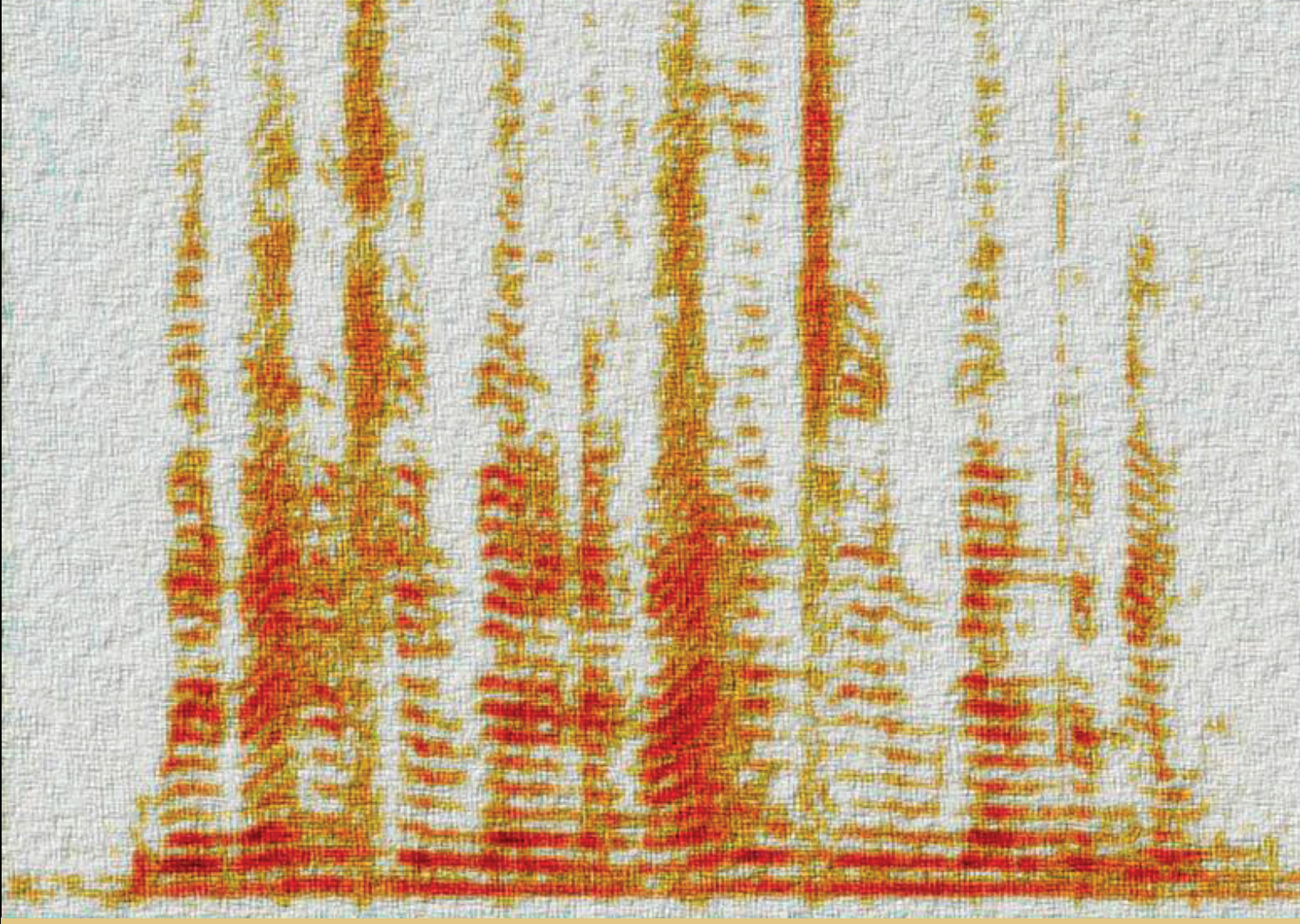

Arlindo Oliveira da Veiga

TREINO NÃO SUPERVISIONADO DE MODELOS ACÚSTICOS PARA RECONHECIMENTO DE FALA

Tese de Doutoramento em Engenharia Eletrotécnica e de Computadores, orientada pelo Doutor Fernando Manuel dos Santos Perdigão e pelo Professor Doutor Luís António Serralva Vieira de Sá e apresentada ao Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Setembro de 2013

· U C ·

Universidade de Coimbra

• U C •

FCTUC FACULDADE DE CIÊNCIAS
E TECNOLOGIA
UNIVERSIDADE DE COIMBRA

Arlindo Oliveira da Veiga

TREINO NÃO SUPERVISIONADO DE MODELOS ACÚSTICOS PARA RECONHECIMENTO DE FALA

Tese de Doutoramento em Engenharia Eletrotécnica e de Computadores apresentada ao Departamento de Engenharia Eletrotécnica e de Computadores da Faculdade de Ciências e Tecnologia da Universidade de Coimbra para obtenção do grau de Doutor

Orientadores: Doutor Fernando Manuel dos Santos Perdigão e Professor Doutor Luís António Serralva Vieira de Sá

Setembro de 2013

Dedicada à minha mãe

Agradecimentos

Este trabalho é a continuidade da linha de investigação que tenho seguido desde a conclusão da Licenciatura. Durante este percurso tive o privilégio de trabalhar com várias pessoas que, de uma forma ou outra, contribuíram para a elaboração desta tese e que aqui quero deixar os meus agradecimentos.

Antes de mais, aos meus orientadores, o Doutor Fernando Manuel dos Santos Perdigão e o Professor Doutor Luís António Serralva Vieira de Sá, pela oportunidade, disponibilidade e ajuda prestada.

Aos companheiros de laboratório com os quais partilhamos esforços e entreajudas. Destaco deste grupo a Carla Alexandra Lopes, a Sara Candeias, o Cláudio Neves, o José Nunes, o José David, o Jorge Proença e Dirce Celerico.

Agradeço aos meus amigos de infância e aos meus familiares, em especial os meus avós Augusto Martins Gonçalves e Domingas Mendes Oliveira, a minha mãe Fernanda Martins Oliveira, a minha esposa Sónia Maria Vaz Semedo e o meu filho Alexandre Semedo da Veiga. Aos meus colegas das residências universitárias por onde passei e a todos os colegas de curso que me proporcionaram uma boa integração na universidade e na sociedade portuguesa.

Agradeço a todos os professores que constituem o meu curriculum académico desde a Escola Primária de Pilão Cão até agora a Universidade de Coimbra. Neste percurso, destaco os meus primeiros professores, Belmiro Mendes Furtado e Gregório Sanches Cardoso. Ao meu orientador de Licenciatura, Doutor Vítor Silva, que me abriu as portas do Instituto de Telecomunicações e continua a proporcionar-me boas conversas sobre a minha terra natal.

Também quero agradecer as várias instituições que foram decisivas no meu percurso. À Universidade de Coimbra e ao Instituto de Telecomunicações pelo acolhimento, pelas condições que criaram e financiamento da minha investigação e à Fundação Calouste Gulbenkian pelo financiamento da minha Licenciatura.

A falar é que nos entendemos

Resumo

Esta tese resume os trabalhos desenvolvidos na área de processamento automático de fala com o objetivo de incrementar a quantidade de recursos linguísticos disponíveis para o português europeu. O estágio de desenvolvimento e a aplicação das tecnologias de fala para uma língua estão relacionados com a quantidade e a qualidade de recursos disponíveis para esta língua. Poucas línguas apresentam, no domínio público e livre, todos os recursos necessários para desenvolver as tecnologias de fala. A língua portuguesa, como muitas outras, tem escassez de recursos públicos e livres, o que pode dificultar o desenvolvimento e a aplicação de tecnologias de fala que incorporam esta língua. Os trabalhos descritos nesta tese apresentam uma abordagem para criar bases de dados de fala, recorrendo apenas aos recursos do domínio público e livres, partindo de sinais multimédia sem transcrições ortográficas ou fonéticas. É apresentada uma solução para aproveitar a grande disponibilidade de material multimédia existente no domínio público (*podcasts* por exemplo) e selecionar segmentos de fala adequados para treinar modelos acústicos. Para isso, foram desenvolvidos vários sistemas para segmentar e classificar automaticamente os noticiários. Estes sistemas podem ser combinados para criar bases de dados de fala com transcrição fonética sem a intervenção humana.

Foi desenvolvido um sistema de conversão automático de grafemas para fonemas que se apoia em regras fonológicas e modelos estatísticos. Esta abordagem híbrida é justificada pelos desenvolvimentos de algoritmos de aprendizagem automática aplicados a conversão de grafemas para fonemas e pelo fato do português apresentar uma razoável regularidade fonética e fonológica bem como uma ortografia de base fonológica. Com auxílio deste sistema, foi criado um dicionário de pronunciação com cerca de 40 mil entradas, que foram verificadas manualmente.

Foram implementados sistemas de segmentação e de diarização de locutor para segmentar sinais de áudio. Estes sistemas utilizam várias técnicas como a impressão digital acústica, modelos com misturas de gaussianas e critério de informação *bayesiana* que normalmente são aplicadas noutras tarefas de processamento de fala.

Para selecionar os segmentos adequados ou descartar os segmentos com fala não preparada que podem prejudicar o treino de modelos acústicos, foi desenvolvido um sistema de deteção de estilos de fala. A deteção de estilos de fala baseia-se na combinação de parâmetros acústicos e parâmetros prosódicos, na segmentação automática e em classificadores de máquinas de vetores de suporte. Ainda neste âmbito, fez-se um estudo com o intuito de caracterizar os eventos de hesitações presentes nos noticiários em português.

A transcrição fonética da base de dados de fala é indispensável no processo de treino de modelos acústicos. É frequente recorrer a sistemas de reconhecimento de fala de grande vocabulário para fazer transcrição automática quando a base de dados não apresenta nenhuma transcrição. Nesta tese, é proposto um sistema de *word-spotting* para fazer a transcrição fonética dos segmentos de fala. Fez-se uma implementação preliminar de um sistema de *word-spotting* baseado em modelos de fonemas. Foi proposta uma estratégia para diminuir o tempo de resposta do sistema, criando, *a priori*, uma espécie de "assinatura acústica" para cada sinal de áudio com os valores de todos os cálculos que não dependem da palavra a pesquisar, como a verosimilhanças de todos os estados dos modelos de fonemas. A deteção de uma palavra utiliza medidas de similaridade entre as verosimilhanças do modelo da palavra e do modelo de enchimento, um detetor de picos e um limiar definido por forma a minimizar os erros de deteção.

Foram publicados vários recursos para a língua portuguesa que resultaram da aplicação dos vários sistemas desenvolvidos ao longo da execução desta tese com especial destaque para o sistema de conversão de grafemas para fonemas a partir do qual se publicaram vários dicionários de pronunciação, dicionários com as palavras homógrafas heterofónicas, dicionário com estrangeirismos, modelos estatísticos para a conversão de grafemas para fonemas, o código fonte de todo sistema de treino e conversão e um demonstrador *online*.

Palavras-chave: reconhecimento automático de fala, treino de modelos acústicos, conversão de grafemas para fonemas, segmentação de áudio, *word-spotting*.

Abstract

This thesis summarizes the works done in the automatic speech processing field aiming to increase the amount of the linguistic resources available for European Portuguese language. The development stage and the application of speech technologies into a language are related to the quantity and quality of resources available for that given language. Few languages have all the required resources to implement speech technologies within free-access and public domain. Like many other language, the Portuguese language lacks public and free resources which may hinder the development and the application of speech technologies that incorporate the Portuguese language. The works described in this thesis present an approach to create speech databases, using only the public and free-access resources, starting from multimedia signals without orthographic or phonetic transcriptions. It this sense, a solution is presented to take advantage of the wide availability in the public domain of multimedia material (e.g. podcasts) and select appropriate speech segments to train acoustic models. To this end, several systems have been developed to automatically segment and classify broadcast news. These systems can be combined to build speech databases with phonetic transcription without human intervention.

A system was developed to automatically convert graphemes to phonemes based on phonological rules and statistical models. This hybrid approach is justified by the developments in machine learning algorithms applied to the conversion of graphemes into phonemes and by the fact that the Portuguese language presents a reasonable phonetic/phonologic regularity and an orthography that is roughly phonologically based. Using this system, a pronunciation dictionary was created including about 40 thousands entries that where manually confirmed.

They were implemented a system for segmentation into five predetermined acoustic classes (speech, music, noise, speech with music and speech with noise) and a system for speaker diarization. These systems use various techniques such as acoustic fingerprint,

Gaussian mixture model and Bayesian information criterion that normally are used in other speech processing tasks.

In order to select appropriate audio segments or discard non-prepared speech segments that may impair acoustic models training, it was developed a system to detect speaking styles. The detection of speaking styles is based on the combination of acoustic and prosodic parameters, on automatic segmentation and on support vector machine classifiers. Also in this scope, a study was made in order to characterize the hesitation events present in the Portuguese broadcast news.

The transcription of the audio databases is essential in the process of acoustic models training. The large-vocabulary continuous speech recognition system is usually used to do automatic transcription wen the database do not have any transcripts. In this thesis, it is proposed to use word-spotting system to provide phonetic transcriptions of speech segments. A preliminary implementation of a word-spotting system based on phoneme models was conducted. A strategy was proposed to decrease the system response time, creating, a priori, a sort of “acoustic signature” for each audio signal with the values of all calculations which do not depend on the searching word as for example the likelihood of all states of phoneme models. The detection of a word uses similarity measures based on likelihood of word model and likelihood of filler model, a peak detector and a threshold value defined as to minimize detection errors.

Several resources for the Portuguese language were published that resulted from the application of the various systems developed throughout the development of this thesis with particular emphasis on the graphemes to phonemes system from which it was published several dictionaries of pronunciation, dictionary with heterophonic homographs words, dictionary of foreign words, statistical models for converting graphemes to phonemes, the source code of the whole system of training as well as conversion and an online demo.

Keywords: automatic speech recognition, training acoustic models, grapheme-to-phoneme conversion, audio segmentation, word-spotting.

Lista de siglas e acrónimos

AO90	Acordo Ortográfico de 1990
BER	*Bit Error Rate*
BIC	*Bayesian Information Criterion*
CRF	*Conditional Random Fields*
DCT	*Discrete Cosine Transform*
DER	*Diarization Error Rate*
DFT	*Discrete Fourier transform*
DNN	*Deep Neural Network*
DTW	*Dynamic Time Warping*
EER	*Equal Error Rate*
EM	*Expectation Maximization*
FFT	*Fast Fourier Transform*
FST	*Finite-State Transducers*
G2P	*Grapheme to Phoneme*
GMM	*Gaussian Mixture Model*
HMM	*Hidden Markov Model*
HNR	*Harmonics to Noise Ratio*
HTK	*Hidden Markov Model Toolkit*
IPA	*International Phonetic Alphabet*

L2S *Letter to Sound*

LLR *Log-Likelihood Ratio*

LP *Linear Prediction*

LPCC *Linear Predictive Cepstral Coefficients*

LRT *Likelihood Ratio Test*

MAP *Maximum A Posteriori*

MCE *Minimum Classification Error*

MFCC *Mel-Frequency Cepstral Coefficients*

ML *Maximum Likelihood*

MLLR *Maximum Likelihood Linear Regression*

MLP *Multi-Layer Perceptron*

MMI *Maximum Mutual Information*

MPE *Minimum Phone Error*

MWE *Minimum Word Error*

NIST *National Institute of Standards and Technology*

NMF *Non-negative Matrix Factorization*

PCM *Pulse-Code Modulation*

PDF *Probability Density Function*

PE Português Europeu

PER *Phoneme Error Rate*

PLP *Perceptual Linear Prediction*

PoS *Part of Speech*

ROC *Receiver Operation Characteristics*

SAMPA *Speech Assessment Methods Phonetic Alphabet*

SVM *Support Vector Machine*

UBM *Universal Background Model*

VOP Vocabulário Ortográfico do Português

WER *Word Error Rate*

ZCR *Zero Crossing Rate*

Índice de tabelas

Índice de figuras

Índice de conteúdo

Capítulo 1. Introdução

1.1. Introdução

A generalização de acesso aos equipamentos eletrónicos é uma realidade presente em muitas sociedades. A disponibilidade de computadores e equipamentos móveis com capacidades de executar aplicações cada vez mais complexas, impulsionam o desenvolvimento de novas formas de interface, além das formas tradicionais tais como teclados e ratos.

A fala pode tornar a interação com os equipamentos eletrónicos mais eficaz e cómoda em muitas tarefas, uma vez que é o meio privilegiado de interação entre humanos. Pode proporcionar a interação com os utilizadores que têm as mãos ocupadas ou portadores de deficiências que reduzem a capacidade de interação ou lhes impossibilitam a interação usando os meios tradicionais.

A área de processamento automático da fala debruça-se sobre os aspetos relacionados com a produção e reconhecimento dos sinais da fala. O estado da arte desta área está muito aquém de reproduzir o sistema de codificação e descodificação desenvolvido pelos humanos. O desempenho dos sistemas que simulam o sistema de audição humana (reconhecedores automático de fala) é fortemente influenciado pelas características do vocabulário utilizado, pelas características dos locutores e pelos ruídos de fundo que afetam o sinal da fala (ambientes acústicos). Mesmo sem atingir o desempenho humano, em ambientes controlados, existem aplicações com resultados satisfatórios, como é o caso de reconhecedores de sequências de dígitos que podem apresentar taxas de acertos acima de 99% (Falavigna et al., 2009). Existem também sistemas de transcrição automática de noticiários com desempenhos aceitáveis para o pivô em estúdios e com modelos previamente adaptados (Woodland, 2002; Zdansky and David, 2004; Neto et al., 2008; Wessel and Ney, 2005; Batista et al., 2009; Kaufmann et al., 2009). Fora destes ambientes, a fiabilidade dos resultados de um sistema de reconhecimento automático de fala de grande vocabulário independente do locutor sofre uma forte redução,

inviabilizando por vezes a sua aplicação prática. Este decréscimo é fortemente explicado pela impossibilidade de treinar modelos acústicos que modelam toda a variabilidade que o sinal de fala possa apresentar. Para reduzir este problema, é necessário usar uma grande variedade e quantidade de amostras de sinal de fala por forma a modelar corretamente a maioria das variações acústicas.

O treino de modelos acústicos para um sistema de reconhecimento de fala contínua de grande vocabulário requer uma base de dados de fala muito grande, com transcrições fonéticas ou, pelo menos, transcrições ortográficas que se podem converter numa transcrição fonética larga.

A obtenção de transcrições fonéticas de qualidade é um processo moroso e requer o emprego de especialistas com experiência e conhecimentos fonéticos e linguísticos. É um processo crucial no desenvolvimento de um sistema de reconhecimento de fala mas consome grande parte de recursos e esforços despendidos neste processo. A transcrição fonética manual é fiável e garante a qualidade exigida mas pode acarretar custos impraticáveis. Por exemplo, em (Kawai and Toda, 2004) é indicado que a transcrição fonética manual pode demorar 130 vezes o tempo real. Para minimizar o impacto da geração de transcrição fonética no desenvolvimento do sistema de reconhecimento, surgiram propostas de implementação de sistemas automáticos ou semiautomáticos de transcrição que, apesar de apresentarem resultados de menor qualidade, conseguem produzir grande quantidade de transcrições num curto espaço de tempo, tentando compensar a qualidade com a quantidade.

Uma vez que as transcrições manuais são mais coerentes que as transcrições automáticas, é expectável que o conjunto de modelos acústicos treinados com as transcrições manuais apresente desempenho melhor que o conjunto de modelos treinados com as transcrições automáticas. Contudo, mostra-se que, usando iterativamente um transcritor automático para segmentar e treinar os modelos acústicos, a discrepância de desempenho entre os dois conjuntos de modelos pode ser minimizada (Wessel and Ney, 2005).

1.2. Trabalhos desenvolvidos

Os trabalhos descritos nesta tese apresentam soluções para minimizar os problemas que advêm da falta de recursos linguísticos para o português europeu no treino de modelos acústicos. Neste sentido foram implementados vários sistemas com a finalidade de segmentar, classificar e transcrever automaticamente segmentos de fala de noticiários ou outro tipo de sinal de fala que não apresenta transcrição fonética.

Foi desenvolvido um sistema de conversão automático de grafemas para fonemas que utiliza uma abordagem híbrida, combinando regras fonológicas com modelos estatísticos, (Veiga et al., 2013a, 2011a, 2011b, 2013b; Candeias et al., 2012a).

Foi apresentado um sistema de segmentação e diarização de locutor que combina várias técnicas como a impressão digital acústica, modelos com misturas de gaussianas e critério de informação *bayesiana*, (Veiga et al., 2010a; Lopes et al., 2010).

Foi desenvolvido um sistema de deteção de estilos de fala, recorrendo a combinações dos parâmetros acústicos e parâmetros prosódicos. A deteção de estilos de fala pode ser usada para rejeitar segmentos que apresentem caraterísticas de fala não preparada. Ainda sobre este assunto, fez-se um estudo para caracterizar hesitações presentes em noticiários em português, (Veiga et al., 2012a, 2012b, 2011c; Candeias et al., 2012b, 2013a, 2013b; Proença et al., 2013; Veiga et al., 2012c)

Por fim, fez-se um teste preliminar de um sistema de *word-spotting* baseado em modelos de fonemas. É apresentada uma proposta para diminuir o tempo de resposta do descodificador de *word-spotting*, calculando, *a priori*, as verosimilhanças de todos os estados dos modelos de fonemas, e para criar uma espécie de "assinaturas acústicas" para cada sinal de áudio. Ao combinar os vários trabalhos, é possível criar uma base de dados de fala, a partir de gravações de noticiários das rádios e das televisões, que pode ser usada para treinar modelos acústicos independentes do locutor, (Veiga et al., 2012d).

1.3. Estado da arte de treino não supervisionado de modelos acústicos

Um sistema de aprendizagem automática pode ser classificado como sendo supervisionado ou não supervisionado. Na aprendizagem supervisionada todas as classes são conhecidas de antemão e os parâmetros acústicos são acompanhados de transcrições que identificam o seu conteúdo linguístico. Dado um protótipo inicial dos modelos, o algoritmo de aprendizagem procura, em cada iteração, estimar os parâmetros do modelo por forma a minimizar uma função de distância entre a resposta do sistema e a transcrição providenciada.

Quando não existem descrições suficientes do material acústico, nomeadamente a ausência de transcrições, é preciso utilizar a aprendizagem não supervisionada. Todas as informações são inferidas a partir dos parâmetros acústicos recorrendo à procura de padrões que se repetem e agrupando parâmetros similares. A ausência de qualquer informação ou recurso linguístico inicial limita o detalhe de informações que pode ser aprendido quando é usando uma abordagem puramente não supervisionada. Na prática, existe sempre algum recurso inicial que pode ser utilizado para limitar o número de classes, inicializar os parâmetros dos modelos e fazer uma transcrição automática inicial. Por vezes, esta abordagem é identificada como semissupervisionada (Yu et al., 2010; Zhang and Rudnicky, 2006), por utilizar modelos previamente inicializados e/ou uma fração da base de dados de áudio previamente transcrita.

O desenvolvimento de um sistema de reconhecimento de grande vocabulário para uma determinada língua é fortemente influenciado pela disponibilidade de recursos linguísticos nesta língua. Treinar modelos acústicos que produzam desempenhos satisfatórios pode requerer centenas ou milhares de horas de dados acústicos transcritos (Evermann et al., 2005; Wessel and Ney, 2005). A transcrição manual da base de dados é uma opção que é inviável em muitos casos. A utilização de técnicas não supervisionadas ou semissupervisionadas é uma alternativa rápida e de baixo custo para o treino de modelos acústicos. Existem várias abordagens de implementação de sistemas de treino não supervisionado ou semissupervisionado. A maioria dos trabalhos utiliza um sistema

de reconhecimento de grande vocabulário, desenvolvido previamente, para transcrever automaticamente os materiais acústicos e utilizam medidas de confiança do reconhecedor para selecionar os dados de treino. Esta abordagem é normalmente utilizada para incrementar o desempenho dos sistemas previamente desenvolvidos (Zavaliagkos et al., 1998).

Existem muitos trabalhos que abordam o problema de treinar modelos acústicos com poucos recursos iniciais ou até mesmo sem nenhum recurso inicial (exceto os sinais de áudio) e são capazes de aprender entidades lexicais tais como palavras, sub-palavras e sílabas. Em (Riccardi and Hakkani-Tur, 2003), é apresentada uma abordagem que utiliza um sistema de reconhecimento de fala para transcrever os dados (não transcritos) e calcula uma medida de confiança para as transcrições geradas. Os segmentos com baixo valor de confiança são transcritos manualmente. Por vezes, a descrição do material acústico não apresenta todos os detalhes presentes no sinal, por exemplo as legendas de dados audiovisuais que, normalmente, não detalham todos os eventos acústicos presentes tais como hesitações, correções, pausas de preenchimento e eventos não linguísticos. Mesmo assim, em (Chan and Woodland, 2004), é aproveitada a disponibilidade da legendagem de noticiários para treinar modelos de linguagem e para auxiliar a seleção dos segmentos usados no treino de modelos acústicos. É comparado o desempenho de método de treino de máxima verosimilhança (ML – *Maximum Likelihood*) com os métodos de treino discriminativo: máxima informação mútua (MMI – *Maximum Mutual Information*) e mínimo erro de fone ou palavra (MPE – *Minimum Phone Error* ou MWE – *Minimum Word Error*). São testados vários critérios de seleção de dados de treino como a medida de confiança, uma medida de similaridade entre o resultado de reconhecimento e a legendagem e é também testada a escolha aleatória dos segmentos. O treino discriminativo apresentou melhores resultados mas nenhum dos critérios de seleção de dados resultou num incremento significativo do desempenho dos modelos acústicos quando comparados com o desempenho sem nenhum critério de seleção, ou seja, utilizar todos os segmentos de áudio disponível.

Em (Wessel and Ney, 2005), é estudada a influência da quantidade dos dados transcritos inicialmente disponíveis e os limiares das medidas de confiança no desempenho dos modelos acústicos. Também, em (Ma et al., 2006) é avaliado o treino não supervisionado inicializado com modelos treinados com quantidades variáveis de dados. São também testados vários limiares de medidas de confiança, vários modelos de linguagens (genéricos e contextual) e diferentes métodos de treino de modelos acústicos (ML, MMI e treino com adaptação do orador). É apresentado ainda o problema de utilizar um limiar de medida de confiança muito restritivo. A utilização de um limiar muito restrito diminui a possibilidade de utilizar dados incorretamente transcritos no treino, mas também impede a seleção de muitos segmentos corretamente transcritos, diminuindo assim o número de amostras para o treino. É preciso encontrar um compromisso entre o limiar de medida de confiança e a quantidade de dados selecionados para o treino.

Utilizar apenas medidas de confiança para selecionar os dados de treino pode gerar amostras desequilibradas entre as classes. O trabalho descrito em (Zhang and Rudnicky, 2006) apresenta uma nova estratégia de seleção de dados de treino que não se baseia apenas nas medidas de confiança. A abordagem apresentada garante a escolha de exemplares de todas as classes e é testada em tarefas de segmentação de imagem, de reconhecimento de carateres e de reconhecimento de fala. O agrupamento de observações é feito sem supervisão usando o algoritmo *k-means*. É estimada a probabilidade de cada grupo e a probabilidade de cada classe pertencer a um determinado grupo. A seleção de dados de treino utiliza estas probabilidades e as medidas de confiança dos resultados reconhecidos para fazer uma escolha equilibrada das classes.

Em (Stouten et al., 2008) é introduzida a técnica não supervisionada denominada de fatorização de matriz não-negativa (NMF – *Non-negative Matrix Factorization*) para descobrir padrões de fones presentes numa locução de fala que possam representar palavras. A aprendizagem de padrões de fones é feita a partir de um conjunto de grelhas de hipóteses (*lattices*) de fones resultante da descodificação de locuções de fala e não é preciso nenhum conhecimento prévio das palavras presentes nos sinais de áudio. No

entanto, foi aplicada apenas numa base de dados de pequeno vocabulário, TI-Digits (Leonard, 1984), que contem apenas dígitos, por isso, muito pobre em termos de variedade linguística e fonética.

Existem vários trabalhos que propõem utilizar apenas os parâmetros acústicos como entrada sem nenhuma descrição do conteúdo. Em (Varadarajan et al., 2008) é apresentada uma abordagem não supervisionada para aprender os parâmetros dos modelos acústicos usando HMM (*Hidden Markov Model*) sem uma topologia inicial definida. Partindo de um modelo HMM com um estado emissor e, em cada iteração, é feita a estimação dos parâmetros e o incremento de número de estados até atingir um limite máximo de estados. A partir da sequência ótima de estados, os autores inferiram a sequência ótima de fonemas e mostraram que é possível identificar vários fonemas recorrendo ao padrão de sequência de estados.

Em (Park and Glass, 2008) é apresentado um trabalho que procura descobrir segmentos de fala que se repetem, sem auxílio de um sistema de reconhecimento automático de fala e sem nenhum conhecimento prévio, utilizando técnica de programação dinâmica e aplicação de grafos. Ao acumular padrões semelhantes, verificou-se que podiam ser agrupados em sequências acústicas similares, formando entidades lexicais (palavras ou frases). Entretanto, foi utilizado um reconhecedor de fonemas para auxiliar na deteção de silêncio a partir do qual fez-se a segmentação de frases. Os autores admitem que esta proposta funciona bem apenas para fala produzida por um único locutor e num ambiente adequado, não sendo propícia para fala de noticiários.

Em (Muscariello et al., 2009) os autores procuram padrões de fala que se repetem numa locução, identificados aqui como “motivos” em analogia com a aplicação da mesma técnica em biologia. A procura de motivos é efetuada dividindo a locução em pequenos segmentos e utilizando a técnica de programação dinâmica.

Os modelos acústicos normalmente usados em reconhecimento de fala são modelos de Markov não observáveis (HMM – *Hidden Markov Model*). Um HMM é definido através de uma cadeia de Markov e uma matriz de probabilidades de transição entre estados para modelar a variação temporal. Os estados contêm parâmetros que representam a

distribuição dos dados acústicos que é, normalmente, definida por mistura (soma) de gaussianas (GMM – *Gaussian Mixture Model*). Os GMM associados aos estados são usados para calcular a probabilidade de uma trama de áudio pertencer a um determinado estado (probabilidade *a priori*). Têm surgido abordagens que utilizam, com sucesso, redes neuronais para substituir os GMM nos HMM. Estas abordagens utilizam redes neuronais para calcular probabilidades *a posteriori* dos estados e HMM para modelar as variações temporais dos estados. São consideradas abordagens híbridas e o estado da arte dos sistemas de reconhecimento de fala utiliza as redes neuronais profundas (DNN – *Deep Neural Network*) para substituir os GMM (Dahl et al., 2012; Hinton et al., 2012). As DNN são redes neuronais artificiais *feed-forward* com mais de uma camada escondida entre a camada de entrada e camada de saída e com uma grande camada de saída (dimensão igual ao número de estados dos HMMs). A introdução de DNN nos sistemas de reconhecimento de fala motivou o aparecimento de abordagens de aprendizagem não supervisionada direcionadas para DNN, como a apresentada em (Thomas et al., 2013).

A maioria das abordagens utiliza um sistema de reconhecimento para auxiliar a tarefa de transcrever automaticamente os sinais de áudio (não transcritos). As abordagens que não utilizam um sistema de reconhecimento apenas conseguem aprender algumas formas lexicais em condições muito restritas. Aplicar técnicas de aprendizagem não supervisionada aos sinais de áudio de noticiários sem auxílio de um sistema de reconhecimento de fala de grande vocabulário é um desafio que pode abarcar a área da linguística e vários tópicos da área do processamento automático de fala.

1.4. Motivação e desafios

Os primórdios desta tese situam-se nos desafios encontrados durante e depois do desenvolvimento de um sistema de comandos independente do locutor para reconhecer cerca de duas centenas de comandos (Lopes et al., 2008a). Para desenvolver este sistema foi criada uma base de dados de fala com transcrição ortográfica, que foi verificada manualmente. O treino dos modelos de fonemas com contexto à direita e à esquerda

(modelos de trifones) com esta base de dados resultou em aproximadamente 900 modelos de trifones, correspondendo a cerca de 2000 estados de HMM, o que é insuficiente para desenvolver um sistema de grande vocabulário. Um estudo preliminar identificou mais de 30 mil trifones diferentes num dicionário de pronunciação para o português europeu com cerca de 40 mil palavras. Um sistema de reconhecimento de fala contínua de grande vocabulário independente de locutor precisa, normalmente, de centenas de horas de fala para treinar modelos acústicos, tem um vocabulário com 20 mil a 65 mil palavras e apresenta entre 4 mil a 12 mil unidades acústicas ou estados dos HMMs (Zweig and Picheny, 2004). Por exemplo, em (Dahl et al., 2012) é apresentado um sistema com um vocabulário de 65 mil palavras e com 2 mil trifones físicos (53 mil trifones lógicos) que correspondem a cerca de 6 mil estados (por norma, os trifones têm 3 estados) e é indicado que existem 761 estados partilhados pelos trifones.

Todos os trabalhos desenvolvidos nesta tese têm como principal objetivo proporcionar amostras de fala que possam ser usadas para incrementar o número de trifones e o número de amostras de trifones, sem auxílio de um sistema de reconhecimento de fala contínua de grande vocabulário. Com um limitado número de modelos iniciais e não tendo disponível um sistema de reconhecimento de grande vocabulário, é necessário explorar outras técnicas para segmentar, classificar e transcrever os segmentos de áudio que apresentam uma boa qualidade em termos de ruído de fundo e dicção do locutor.

A seleção de segmentos de fala é feita com base em restrições definidas *a priori* de forma a garantir que os segmentos selecionados tenham alguma qualidade (por exemplo, aceitar apenas segmentos sem muito ruído de fundo e que evidenciam fala preparada). Estas restrições são necessárias uma vez que a grande quantidade do material de áudio utilizado nesta tese provém de gravações de noticiários da rádio ou de televisão, onde é possível encontrar segmentos áudio que não correspondem a fala (como por exemplo *jingles*, músicas, palmas, ou *spots* publicitários). Num segmento de fala, é possível encontrar uma grande variabilidade em termos linguísticos e de locutores bem como uma grande variedade de ambientes acústicos. Exemplos desta variabilidade são as gravações de fala em diversos ambientes (estúdio, rua ou salas com reverberação), com diferentes

canais (microfones, telefones, telemóveis), com vários estilos de fala (lida, formal ou espontânea) e com vários tipos de locutores (profissionais do audiovisual, falantes comuns, dirigentes políticos, intervenientes em debates, etc., além da divisão em género masculino e feminino). É possível restringir a seleção de material de fala para apenas segmentos que tenham indícios de conter fala preparada ou fala proferida por um determinado locutor ou um conjunto predefinido de locutores, como, por exemplo, os *pivôs* de noticiários. Por norma, os *pivôs* têm uma boa dicção, estão em estúdios durante a emissão de noticiários e seguem um guião, podendo as suas falas serem consideradas falas lidas. As locuções produzidas durante a leitura de textos em estúdios apresentam poucos problemas, ao contrário do que acontece, por exemplo, com as locuções de fala espontânea, que são propícias à presença de eventos de hesitações ou disfluências, como pausas preenchidas, prolongamentos vocálicos, repetições e correções. Se for possível detetar e rejeitar os segmentos problemáticos, os restantes podem integrar uma base de dados de treino de modelos acústicos.

Para aplicar as restrições de escolha de segmentos de fala, foi necessário implementar vários algoritmos que lidam com os seguintes problemas:

- deteção de mudança de locutores;
- identificação de determinadas classes acústicas;
- identificação de segmentos produzidos pelo mesmo locutor (diarização de locutores);
- deteção de músicas e *jingles*;
- deteção de estilos de fala;
- transcrição fonética dos segmentos.

Após a seleção dos segmentos de áudio que cumprem os requisitos predefinidos, é preciso gerar as suas transcrições fonéticas. Nesta tese, propõe-se o uso da técnica de *word-spotting* para colmatar a indisponibilidade de um sistema de reconhecimento de fala de grande vocabulário. O *word-spotting* é uma técnica para detetar a presença e a localização de uma palavra ou uma sequência de palavras numa locução (Amir et al., 2001). É necessário dispor de um dicionário de pronunciação extenso ou um conversor de

grafemas para fonemas. O conversor de grafemas para fonemas possibilita pesquisas sem restrições de vocabulário e auxilia no incremento de número de trifones ou seleção de palavras com um determinado padrão de fonemas. Foi implementado um sistema de conversão de grafemas para fonemas, foi gerado um dicionário de pronunciação com cerca de 40 mil palavras e foi também implementado um sistema preliminar de *word-spotting*.

Outro assunto investigado prende-se com o problema da coarticulação, mais propriamente, com a geração da transcrição fonética com a possibilidade de existência de coarticulação interpalavras e intrapalavra, diretamente a partir da transcrição ortográfica.

A Figura 1 ilustra o esquema que resume o encadeamento dos vários trabalhos desenvolvidos com o objetivo de criar uma base de dados de fala com transcrição fonética. Os segmentos de fala com a transcrição fonética podem ser utilizados para treinar os modelos acústicos, aumentando progressivamente o número de amostras de trifones e o seu número. Todo este processo pode ser executado sem intervenção humana e sem o auxílio de um sistema de reconhecimento de fala de grande vocabulário para obter transcrições fonéticas, como é comum nos sistemas de treino não supervisionados descritos na literatura (Gales et al., 2006; Huijbregts, 2008; Matsoukas et al., 2006; Wang et al., 2007).

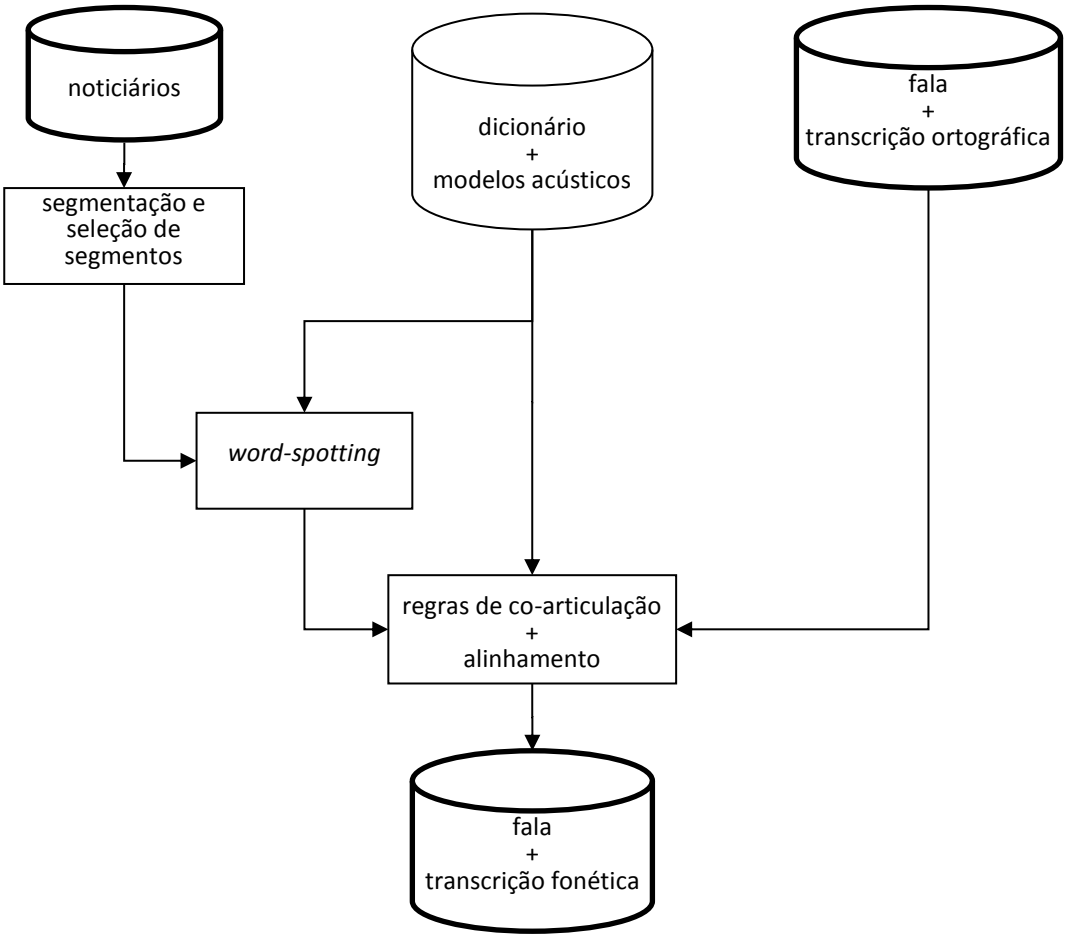

Figura 1 – Geração de transcrição fonética de segmentos de fala.

Capítulo 2. Sistemas de reconhecimento automático de fala

2.1. Introdução

Um sistema de reconhecimento automático de fala processa o sinal acústico da fala e produz uma saída que corresponde ao texto da mensagem falada. O texto produzido é composto pela sequência de palavras que foram identificadas no sinal de fala. A Figura 2 ilustra os módulos principais de um sistema de reconhecimento automático de fala.

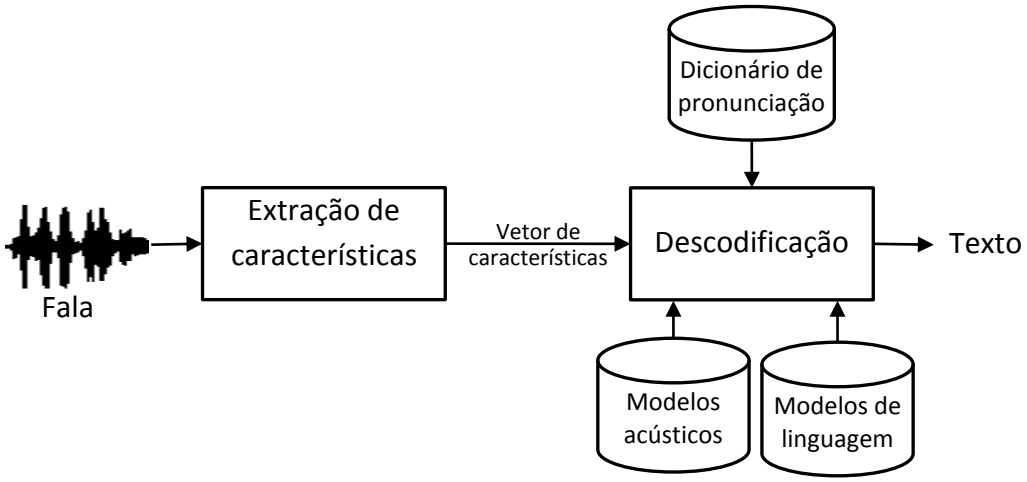

Figura 2 – Módulos de reconhecimento automático de fala.

O módulo da extração de características converte o sinal de áudio numa representação compacta e robusta à variabilidade das condições acústicas, mas sensível ao conteúdo linguístico do sinal de áudio. O módulo de descodificação procura a melhor sequência de palavra num conjunto de hipóteses possíveis dada a representação de características (observações). Esse módulo precisa, além das características do sinal, de dicionários de pronunciação de todas as palavras que é possível reconhecer (ou um sistema de conversão de grafemas para fonemas), de modelos acústicos (de fonemas, palavras ou subpalavras) e de modelos de linguagem ou gramáticas, que definem as hipóteses possíveis.

O desempenho de um sistema de reconhecimento de fala depende da precisão dos modelos acústicos, da complexidade da tarefa definida pelo modelo de linguagem (ou gramática) e da qualidade do sinal de áudio adquirido, a qual pode variar substancialmente com o ambiente acústico. A Tabela 1 apresenta valores nominais de desempenho dos sistemas de reconhecimento automático de fala em termos da percentagem da taxa de erro das palavras (WER – *Word Error Rate*) para várias tarefas (Falavigna et al., 2009).

Tabela 1 – Desempenho dos sistemas de reconhecimento automático de fala.

Tarefa de reconhecimento	**WER**
Dígitos ligados	≤ 0.5 %
Ditado	≤ 5 %
Noticiário em estúdio	≤ 10 %
Reportagem telefónica	≤ 20 %
Conversa telefónica	≤ 30 %
Reunião com microfone de lapela	≤ 30 %
Reunião com microfone distante	≤ 50 %

A Figura 3 ilustra a evolução histórica do desempenho dos sistemas de reconhecimento de fala nos desafios propostos pelo grupo do NIST (*National Institute of Standards and Technology*) (NIST, 2009).

Estes dados ilustram a discrepância que existe entre o sistema de reconhecimento humano e os sistemas de reconhecimento automático de fala.

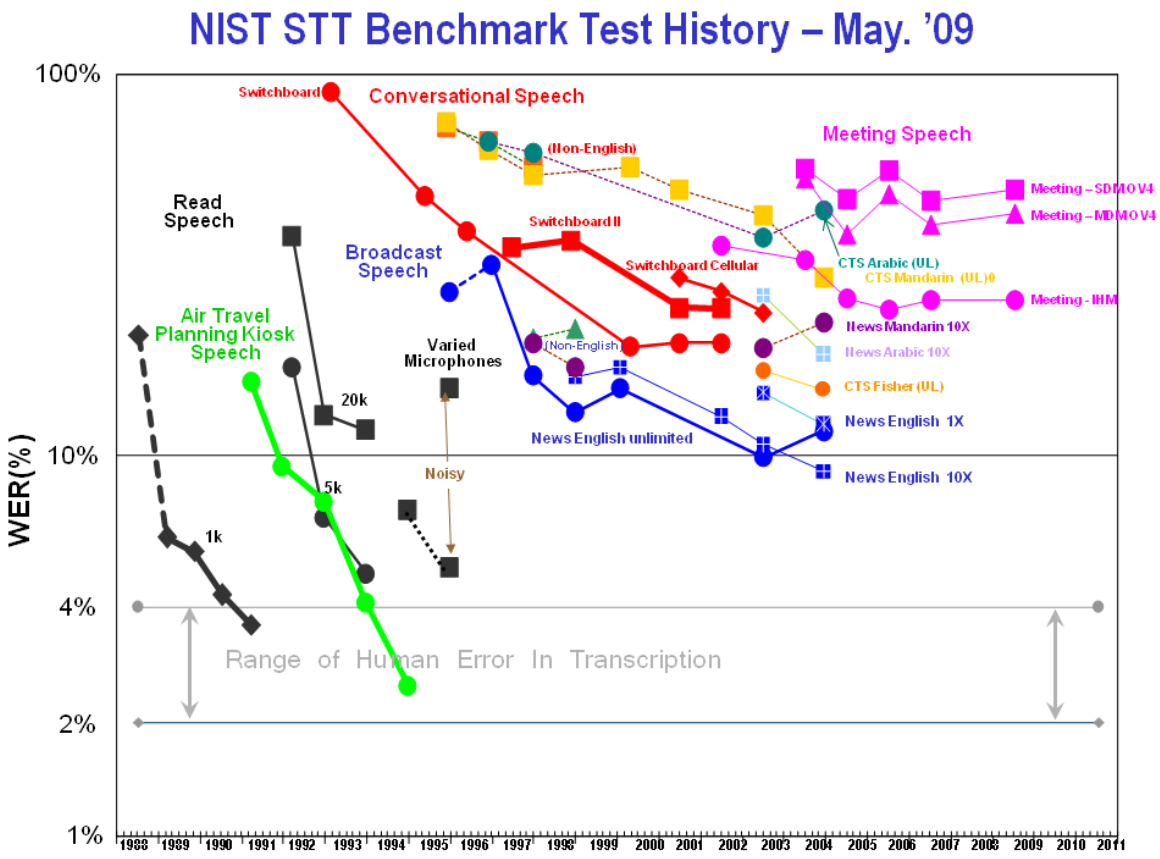

Figura 3 – Evolução do desempenho dos sistemas de reconhecimento **(NIST, 2009)** (com permissão de reprodução).

2.2. Extração de características

A extração de características, ou parametrização, é um processo recorrente em todos os sistemas que envolvem reconhecimento de padrões. Tem como principal objetivo compactar o sinal de forma a reter apenas a informação útil para reconhecimento e desta forma permitir um processamento eficaz e robusto.

No processamento automático de fala, os parâmetros mais utilizados são os coeficientes cepstrais na escala *Mel* (MFCC – *Mel-Frequency Cepstral Coefficients*) (Davis and Mermelstein, 1980). Outros parâmetros usuais são os baseados na análise de predição linear (LP – *Linear Prediction*), como os coeficientes cepstrais de predição linear (LPCC – *Linear Predictive Cepstral Coefficients*) ou os derivados da análise da predição linear percetual (PLP – *Perceptual Linear Prediction*) (Hermansky, 1990). É comum adicionar

parâmetros baseados em energia da trama e também os coeficientes dinâmicos (parâmetros *delta* e *delta-delta*) que correspondem a uma aproximação discreta da primeira e segunda derivada dos coeficientes base (coeficientes estáticos) em ordem ao tempo.

Toda a parametrização utilizada neste trabalho é baseada em MFCC, com as configurações típicas usadas em muitos trabalhos relacionados com o reconhecimento automático de fala. O sinal de áudio é segmentado em tramas de 25 milissegundos, com avanço de 10 milissegundos, usando uma janela de *Hamming*, o que perfaz um ritmo de análise de 100 tramas por segundos. São usados 12 coeficientes cepstrais (c_1 até c_{12}) e o logaritmo da energia como parâmetros básicos e são adicionados os parâmetros *delta* e *delta-delta*, perfazendo um total de 39 parâmetros por cada trama. A Figura 4 apresenta um esquema com os passos para calcular os MFCCs para uma trama de *N* amostras.

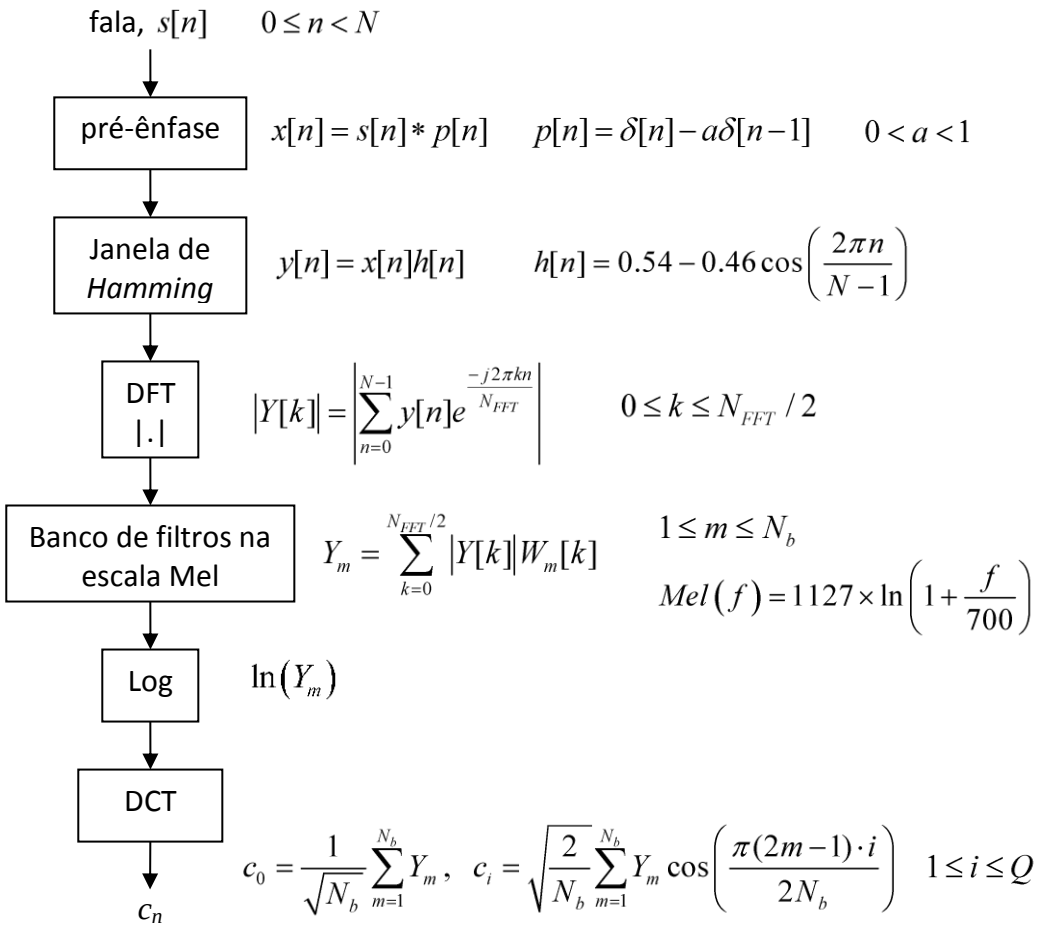

Figura 4 – Diagrama do cálculo dos parâmetros MFCCs.

A pré-ênfase é usada para aumentar a resolução das altas frequências, compensando a inclinação (*tilt*) espetral (o espetro do sinal da fala tem mais energia em baixas frequências) causada pela natureza dos tecidos moles que formam o aparelho fonador. É descrita, tipicamente, por um sistema de resposta a impulso $p[n]=\delta[n]-a\delta[n-1]$ com a definido no intervalo entre 0 e 1, tipicamente 0.97. Após a pré-ênfase, o sinal é decomposto em tramas de *N* amostras e avanço de *M* amostras como é ilustrada na Figura 5.

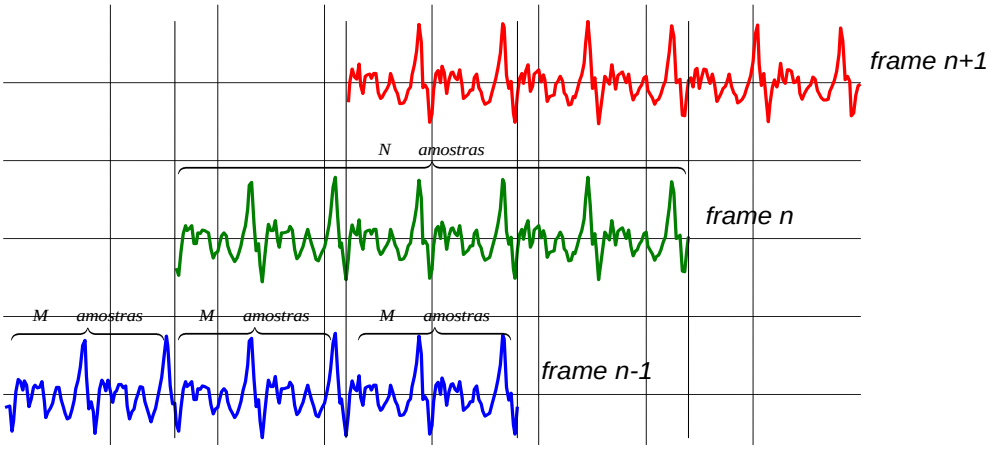

Figura 5 – Decomposição do sinal em tramas.

Os N_b filtros na escala *Mel*, $W_m(k)$, têm respostas triangulares, reais (fase zero) com frequências centrais igualmente espaçadas na escala *Mel*. Definido o intervalo de frequências que varia entre o valor mínimo (f_{Low}) e o valor máximo (f_{High}), são calculados N_b+2 pontos na escala *Mel* entre $Mel\left(f_{Low}\right)$ e $Mel\left(f_{High}\right)$ que formam os vértices dos triângulos (resposta dos filtros). Com exceção dos pontos extremos, os outros N_b pontos definem as frequências características dos filtros. Nesta tese, foram usados 32 filtros ($N_b=32$) entre 0 Hz ($f_{Low}=0$) e *fs*/2 Hz ($f_{High}=fs/2$) (fs - frequência de amostragem do sinal) e foram utilizados sinais com frequência de amostragem de 8 kHz e de 16 kHz.

A Figura 6 ilustra um exemplo de um banco com 8 filtros. Pode-se verificar que a resposta em escala Hertz é aproximadamente triangular.

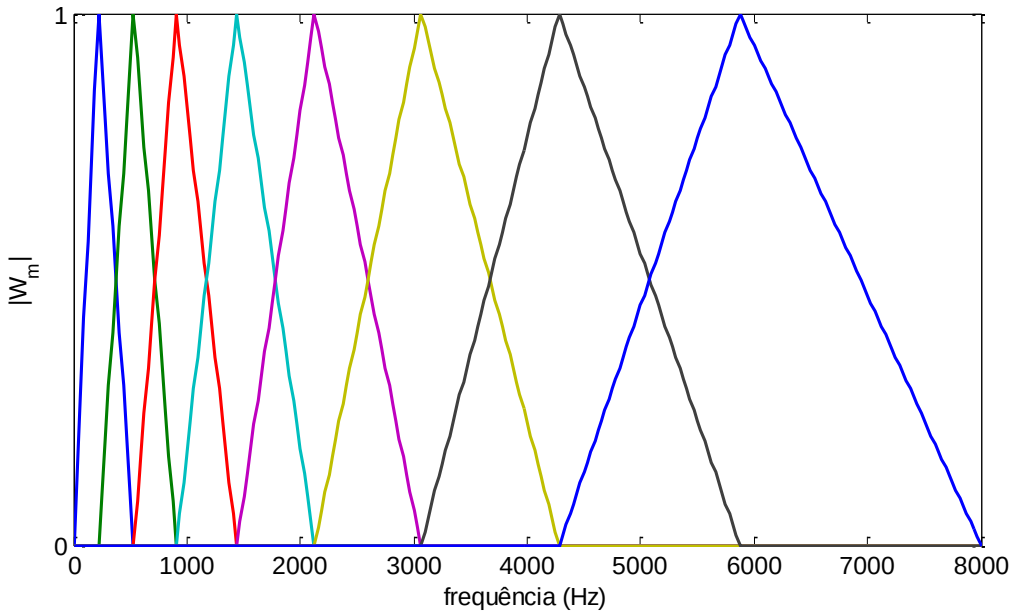

Figura 6 – Exemplo das respostas de um banco com 8 filtros.

A resposta global do banco de filtros, $W(k)$, onde k é o índice da DFT, é a soma das respostas parciais dos canais:

$$W(k) = \sum_{m=1}^{N_b} W_m(k). \qquad \textbf{(1)}$$

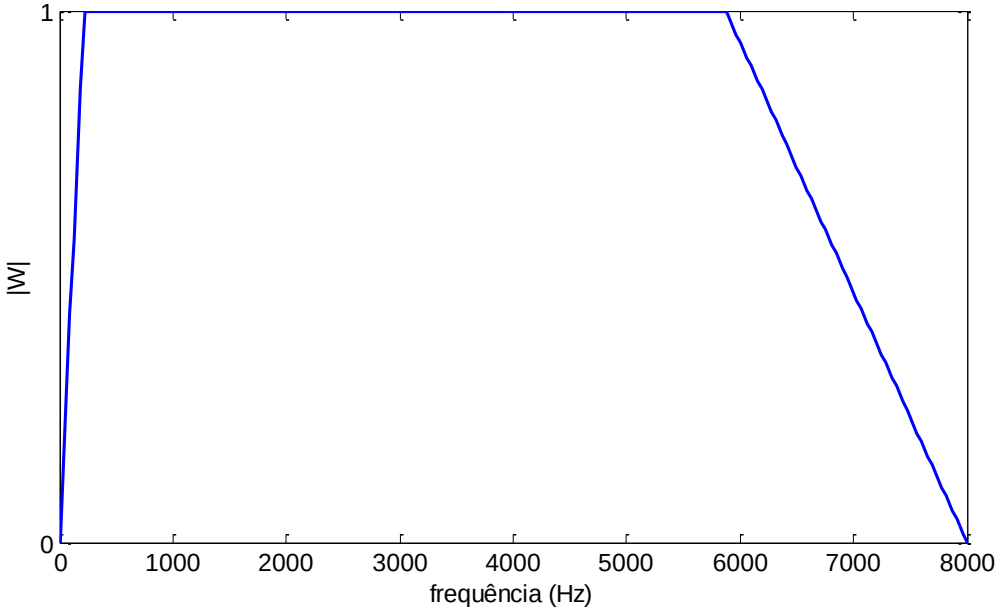

Figura 7 – Resposta global do banco de filtros do exemplo anterior.

A resposta é aproximadamente um trapézio, plana desde $f_{CF}(1)$ a $f_{CF}(N_b)$ (a primeira e última frequências características) e aproximadamente linear nos extremos.

Os MFCC são calculados a partir da aplicação da transformada discreta de cosseno (DCT – *Discrete Cosine Transform*, mais propriamente a DCT-II (Young et al., 2006)) ao logaritmo das energias à saída dos filtros do banco de filtros. O resultado da DCT é truncado para *Q* coeficientes (normalmente 13). A utilização do banco de filtros permite dois efeitos: uma suavização espetral por integração em bandas e uma resolução em frequência que diminui com a frequência, tal como acontece no ouvido humano. A truncagem dos coeficientes corresponde a uma suavização das energias do banco de filtros, perdendo-se algum detalhe que pode ser considerado informação redundante.

2.3. Descodificação

O problema da descodificação consiste em obter a melhor sequência de palavras que descreve uma dada sequência de observações acústicas. As palavras são representadas por sequências de modelos acústicos e a escolha das palavras obedece a um modelo dito de linguagem (ou gramática da tarefa de reconhecimento). A associação entre palavras e modelos acústicos é normalmente auxiliada por um dicionário que impõe um limite máximo ao vocabulário de palavras diferentes que podem ser reconhecidas. Os modelos de linguagem ou gramáticas definem o universo de hipóteses das sequências de palavra do descodificador.

Definindo $O = X_1^T = \{o_1, o_2, \cdots, o_T\}$ como sendo a sequência de observações acústicas e W como uma sequência de palavras qualquer, a determinação da sequência ótima de palavras, W^*, é descrita, usando o critério de *Bayes*, da seguinte forma:

$$W^* = \arg\max_W P(W \mid O) = \arg\max_W \frac{P(O \mid W) \cdot P(W)}{P(O)}. \qquad \textbf{(2)}$$

A probabilidade $P(O \mid W)$ é calculada usando os modelos acústicos; a probabilidade $P(W)$ é definida pelo modelo de linguagem e a probabilidade das observações acústicas,

$P(O)$, pode ser ignorada na pesquisa da sequência ótima de palavras, uma vez que é um denominador comum para todas as sequências de palavras.

2.3.1. Modelos acústicos

Os modelos acústicos são normalmente HMM de fonemas, com 3 estados emissores. Podem ser modelos sem contexto, monofones, ou modelos com contexto: difones (contexto à direita ou à esquerda) e trifones (contexto à direita e à esquerda). Em modelos de síntese de fala usam-se contextos ainda mais alargados, a 5 fonemas: dois à esquerda e dois à direita).

É possível encontrar HMM a representar outras entidades lexicais como palavras (em sistemas de reconhecimento de pequeno vocabulário) ou mesmo sílabas.

Um HMM é definido pelos seguintes parâmetros: número de estados, matriz de probabilidades de transição entre estados e, associada a cada estado, uma função de densidade de probabilidade (PDF - *Probability Density Function*) que caracteriza os parâmetros acústicos observados nesse estado. Nesta tese, as PDF associadas aos estados são contínuas e são definidas como uma soma pesada de gaussianas (mistura de gaussianas com densidades contínuas). Dada uma observação acústica, o_t, a PDF para um estado *j* é definida como:

$$b_j(o_t) = \sum_{m=1}^{G} c_{jm} \mathcal{N}\left(o_t, \mu_{jm}, \Sigma_{jm}\right), \quad \textbf{(3)}$$

onde G é o número de componentes gaussianas, c_{jm} é o peso da gaussiana *m* do estado *j* e $\mathcal{N}(\cdot;\mu;\Sigma)$ representa uma função gaussiana multivariável com o vetor de médias μ e a matriz de covariâncias Σ definida da seguinte forma:

$$\mathcal{N}(o,\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^Q |\Sigma|}} \exp\left(-\frac{1}{2}(o-\mu)'\Sigma^{-1}(o-\mu)\right), \quad \textbf{(4)}$$

onde Q é a dimensão da observação o (39 no presente caso). Para um determinado estado a soma dos pesos das misturas é unitária, respeitando as chamadas restrições estocásticas.

As matrizes de covariâncias podem ser diagonais quando são usados coeficientes MFCC uma vez que os MFCCs geralmente não estão muito correlacionados entre si (Young et al., 2006). Esta simplificação reduz enormemente o número de parâmetros a estimar e uma matriz de covariância reduz-se a um vetor de variâncias na diagonal da matriz. Além disso, numa mistura de gaussianas definidas com matrizes de covariância diagonais, a matriz global da PDF não é diagonal.

A topologia dos HMM é escolhida por forma a restringir a sequências de estados possíveis e, em reconhecimento de fala, é comum a utilização da topologia "esquerda-direita" ou Bakis, como a ilustrada na Figura 8.

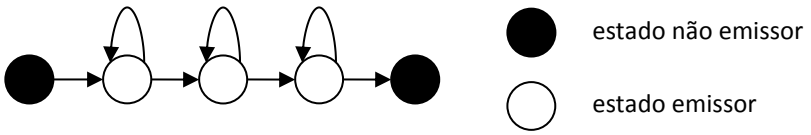

Figura 8 – Modelo HMM com topologia "esquerda-direita".

Com esta topologia, um estado só pode transitar para ele próprio ou para o estado seguinte. Os estados não emissores são estados de entrada e de saída dos modelos, não têm PDF associadas e servem apenas para auxiliar a concatenação de modelos HMM e facilitar o processo de descodificação.

O número de estados e a topologia dos modelos HMM são definidos antes de iniciar o processo de treino. O processo de treino estima, a partir de amostras de observações acústicas, os pesos das misturas, os vetores das médias, as matrizes de covariâncias e as matrizes de probabilidades de transição entre estados. O método de treino de máxima verosimilhança (ML – *Maximum Likelihood*) é muito utilizado para estimar os parâmetros dos modelos recorrendo ao algoritmo de *Viterbi* ou *Baum-Welch* (Young et al., 2006). Este método treina um determinado HMM apenas com as observações acústicas etiquetadas

para este HMM (classes observadas), e por isso não é discriminativo na medida em que não usa estas observações para treinar todos os outros modelos como contraexemplos. Existem técnicas de treino discriminativo para HMM tais como MMI, MCE (*Minimum Classification Error*) e MPE/MWE que complementam o treino ML e que possibilitam a estimação de parâmetros de um HMM com amostras de observações certas e amostras de observações erradas (Chan and Woodland, 2004; Silva et al., 2009; Young et al., 2006).

2.3.2. Adaptação de modelos

A adaptação dos modelos é aplicada quando se pretende melhorar ou refinar os parâmetros dos modelos para uma determinada condição acústica. Normalmente utiliza-se em situações onde não existem amostras de áudio suficientes para uma nova condição acústica, como por exemplo, para criar um conjunto de modelo dependente do orador a partir de um conjunto de modelos independentes do orador por forma a melhorar o desempenho da descodificação da fala do orador para o qual o modelo foi adaptado.

Destacam-se dois métodos de adaptação de modelos: regressão linear de máxima verosimilhança (MLLR – *Maximum Likelihood Linear Regression*) e probabilidade *a posteriori* máxima (MAP – *Maximum A Posteriori*) (Young et al., 2006), também conhecida como adaptação *bayesiana*.

Uma das abordagens de MLLR passa pelo cálculo de uma matriz de transformação a partir de resolução do problema de maximização de verosimilhança dos dados utilizados na adaptação. A matriz de transformação é aplicada aos parâmetros dos HMM (as médias e as variâncias) iniciais segundo uma transformação linear.

A utilização do critério da máxima probabilidade *a posteriori* implica o conhecimento da probabilidade *a priori* dos parâmetros, $P(\theta)$. A probabilidade *a priori* dos parâmetros reduz o problema de "sobre-adaptação" dos parâmetros dos modelos.

Considerando que θ representa os parâmetros de um modelo a adaptar e O observações com a nova condição acústica, os novos parâmetros do modelo, θ^*, são estimados com MLLR e com MAP da seguinte forma:

$$MLLR: \theta^* = \arg\max_{\theta} P(O \mid \theta), \quad \textbf{(5)}$$

$$MAP: \theta^* = \arg\max_{\theta} P(O \mid \theta) \cdot P(\theta). \quad \textbf{(6)}$$

A escolha de um ou outro método de adaptação depende da quantidade de dados disponível para adaptar os modelos. O MAP adapta apenas as componentes gaussianas que foram observadas nos dados e o MLLR adapta todas as componentes gaussianas do modelo. O MLLR é mais apropriado quando existe uma quantidade reduzida de dados para adaptar os modelos. Se houver dados representativos de todas as componentes gaussianas, o desempenho do MAP supera o desempenho do MLLR (Wang et al., 2003).

2.3.3. Modelos de linguagem

A estimação dos modelos de linguagem consiste em estimar as probabilidades de sequências de palavras. É feita a partir de uma grande quantidade de textos e tem um limite máximo para o comprimento das sequências das palavras (*n-grama*). Usando *n-gramas*, a probabilidade de uma palavra é condicionada no máximo por n-1 palavras precedentes (história). Definindo $W = W_1^K = \{w_1, w_2, \cdots, w_K\}$ como sendo uma sequência de *K* palavras, a probabilidade de toda a sequência pode ser estimada usando a probabilidade condicionada às n-1 palavras precedentes:

$$\hat{P}(w_1, w_2, \cdots, w_K) = \prod_{i=1}^{K} P(w_i \mid w_{i-n+1} \ldots w_{i-1}). \quad \textbf{(7)}$$

Assume-se que esta probabilidade é muito semelhante à probabilidade da sequência completa. As probabilidades condicionais podem ser facilmente estimadas a partir da

contagem de ocorrências de *n-gramas* encontradas nos textos. Assim, $P(w_i \mid w_{i-n+1}...w_{i-1})$ pode ser descrita como sendo:

$$P(w_i \mid w_{i-n+1}...w_{i-1}) = \frac{C(w_{i-n+1}...w_i)}{C(w_{i-n+1}...w_{i-1})}, \quad \textbf{(8)}$$

onde *C*(.) é a função da frequência do *n-grama*. É de notar que o denominador pode ser visto como a soma das frequências de todos os *n-gramas* que partilham a mesma história:

$$C(w_{i-n+1}...w_{i-1}) = \sum_j C(w_{i-n+1}...w_{i-1}w_j). \quad \textbf{(9)}$$

A estimação robusta dos modelos de linguagem não se baseia apenas na contagem de ocorrências de *n-gramas*. Na prática também são aplicados algoritmos de "descontos" e "suavizações" (ver secção 3.2.3) para minimizar os efeitos da falta de amostras de sequências de palavras ou até mesmo ausência de amostras na estimação de probabilidades de *n-gramas*.

A perplexidade é utilizada como uma medida de desempenho dos modelos de linguagem. Usando uma sequência de *K* palavras para teste, $W = W_1^K = \{w_1, w_2, \cdots, w_K\}$, a perplexidade é definida (Young et al., 2006):

$$PP = \hat{P}(w_1, w_2, ..., w_K)^{-\frac{1}{K}}. \quad \textbf{(10)}$$

Quanto menor é o valor de perplexidade, melhor é um modelo de linguagem (Young et al., 2006).

2.3.4. Algoritmo de Viterbi

O cálculo direto da probabilidade de uma sequência de observações acústicas para um dado modelo HMM é praticamente impossível uma vez que, para a maioria dos casos, envolve um número elevado de sequências de estados possíveis. Dado um modelo HMM, *M*, se se considerar $X = \{x(1), x(2), ..., x(T)\}$ como o conjunto de todas as sequências de

estados possíveis do HMM para uma sequência de T observações acústicas, a probabilidade desta sequência de observações é expressa da seguinte forma:

$$P(O \mid M) = \sum_{X} a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) \cdot a_{x(t)x(t+1)} \,, \tag{11}$$

em que $a_{x(0)x(1)}$ é a probabilidade de transição do estado inicial para o estado $x(1)$,

$a_{x(t)x(t+1)}$ é a probabilidade de transição do estado x(t) para o estado $x(t+1)$ para t maior que 1, e $b_{x(t)}(o_t)$ é a probabilidade da observação o_t pertencer ao estado $x(t)$, dada pela PDF do estado.

Esta probabilidade é muitas vezes substituída pela probabilidade máxima que resulta de considerar a sequência mais provável (também conhecida como "caminho ótimo"), em vez da soma das probabilidades de todas as sequências:

$$\hat{P}(O \mid M) = \max_{X} \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) \cdot a_{x(t)x(t+1)} \right\}. \tag{12}$$

Existe um método recursivo e simples que estima eficientemente esta probabilidade: o algoritmo de Viterbi (Viterbi, 1967). Considerando uma grelha de pesquisa com a dimensão vertical igual ao número de estados do HMM (N_e) e a dimensão horizontal igual ao número de observações acústicas (T), o algoritmo de Viterbi procura o melhor caminho desde o estado de entrada até ao estado de saída do HMM (ver Figura 9). Para isso, em cada ponto da grelha é calculada e guardada a informação do melhor caminho que leva desde o estado inicial até este ponto. Na última observação acústica é identificado o melhor caminho ao consultar as informações que estão no estado de saída e, através de *backtracking* (pesquisa do fim para o início), descobre os pontos da grelha que pertencem ao caminho ótimo.

A verosimilhança parcial máxima de uma observação no instante t pertencer a um estado j, $\psi_j(t)$, é calculada a partir das máximas verosimilhanças parciais da observação no instante t-1:

$$\psi_j(t) = \max_i \left\{ \psi_i(t-1) \cdot a_{ij} \right\} \cdot b_j(o_t), \tag{13}$$

em que a_{ij} é a probabilidade de transitar do estado *i* para o estado *j*. Na primeira trama, considera-se que $\psi_1(1) = 1$ (estado de entrada) e que $\psi_j(1) = a_{1j} \cdot b_j(o_1)$ para $1 < j < N_e$ (estados emissores). A máxima verosimilhança final é calculada na última trama no estado de saída:

$$\hat{P}(O \mid M) = \max_i \left\{ \psi_i(T) \cdot a_{iN_e} \right\}. \tag{14}$$

A Figura 9 ilustra um exemplo de uma grelha de pesquisa em que cada nodo representa a probabilidade de uma trama pertencer a um estado num determinado instante e cada arco representa a probabilidade de transição entre estados.

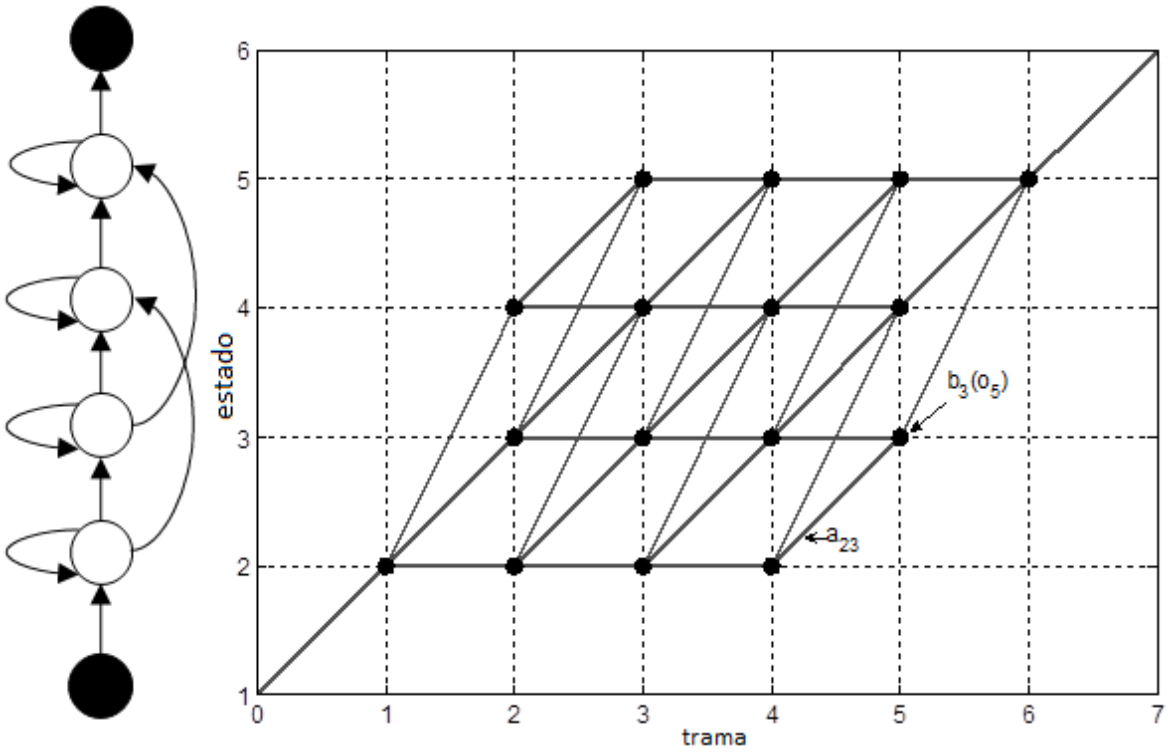

Figura 9 – Grelha de pesquisa do algoritmo de Viterbi.

Para evitar possíveis problemas de precisão numérica por causa da multiplicação de probabilidades (valores entre 0 e 1) e aumentar a gama dinâmica dos valores, é frequente formalizar o algoritmo de Viterbi no domínio logarítmico:

$$\psi_j(t) = \max_i \left\{ \psi_i(t-1) + \log\left(a_{ij}\right) \right\} + \log\left(b_j(o_t)\right). \quad \textbf{(15)}$$

O logaritmo da probabilidade de um caminho é calculado através da soma do logaritmo das probabilidades de transição com o logaritmo das probabilidades das saídas dos estados ao longo deste caminho.

O algoritmo de Viterbi pode ser generalizado para uma rede de palavras definida pelo modelo de linguagem ou gramática. Auxiliado pelo dicionário, converte cada palavra numa sequência de HMM dos fonemas da palavra e a probabilidade de transição entre HMM é definida pelo modelo de linguagem.

O resultado da descodificação pode ser o caminho ótimo ou um conjunto de candidatos condensados numa grelha de hipóteses (*lattice*).

2.4. Ferramentas para o desenvolvimento de sistemas de reconhecimento automático de fala

Existem vários pacotes de programas disponíveis em código aberto que dispõem de ferramentas que possibilitam o desenvolvimento de sistemas de reconhecimento automático de fala. Oferecem opção de treino de modelos acústicos, treino de modelos de linguagem e descodificação da fala. Os pacotes mais utilizados são o HTK (Young et al., 2006) e o CMU Sphinx (Lamere et al., 2003).

O HTK é um conjunto de ferramentas que foi desenvolvido especificamente para manipular modelos HMM para sistemas de reconhecimento de fala, embora possa ser utilizado para modelar quaisquer dados que tenham variação temporal que se possa considerar markoviana. A versão 3.41 apresenta cerca de 35 ferramentas desenvolvidas

em linguagem C e implementa muitos algoritmos de treino de modelos acústicos e treino de modelos de linguagem.

O CMU Sphinx é constituído por um pacote de ferramentas para o treino de modelos HMM (SphinxTrain), um pacote de ferramentas para o treino de modelos de linguagem (CMUclmtk), três descodificadores (Sphinx 2, Sphinx 3 e Sphinx 4), uma biblioteca com as funcionalidades de um reconhecedor (Pocketsphinx) e uma biblioteca com as funcionalidades básicas (Sphinxbase).

Outros pacotes de ferramentas disponibilizam soluções completas para implementar um sistema de reconhecimento como é o caso do RWTH ASR (Rybach et al., 2009), do SHoUT (Huijbregts, 2008) e do Kaldi (Povey et al., 2011). O Kaldi difere dos restantes pacotes por desenvolver reconhecedores baseados em máquinas de estados finitos.

O motor de reconhecimento Julius (Lee et al., 2001) implementa um descodificador eficiente para reconhecedores de grande vocabulário. Utiliza modelos acústicos treinados com HTK.

O Simon (Simon, 2013) é um sistema de reconhecimento de fala que foi desenvolvido para pessoas portadoras de deficiência motora. É auxiliado por outras ferramentas para implementar os vários módulos de um sistema de reconhecimento de fala (HTK, Julius e CMU Sphinx).

O SRILM (Stolcke, 2002) é um conjunto de ferramentas muito utilizado para manipular modelos de linguagem.

Nos trabalhos desta tese foram desenvolvidas várias ferramentas e foram utilizadas as ferramentas do HTK principalmente para a parametrização de sinal de áudio e treino de modelos acústicos. O Praat (Boersma and Weenink, 2001) foi utilizado para calcular e visualizar parâmetros prosódicos como a frequência fundamental dos segmentos. Também é de referir a utilização do Transcriber (Barras et al., 2001) nas tarefas de transcrição manual e visualização dos resultados de segmentação, classificação e deteção de eventos de hesitações e estilos de fala. A ferramenta WEKA (Hall et al., 2009) foi utilizada para treinar e testar os classificadores SVM.

Capítulo 3. Conversão de grafemas para fonemas

3.1. Introdução

A conversão de grafema para fonema diz respeito à tarefa de encontrar a pronúncia (representação fonológica ou fonética) de um vocábulo dado na sua forma escrita (representação grafemática). Um sistema automático de mapeamento inequívoco entre grafemas e fonemas tem várias aplicações, nomeadamente no ensino da língua, nos estudos fonético/fonológicos e, principalmente, na área de processamento de fala, uma vez que, tradicionalmente, a unidade básica utilizada na modelação da fala é o fonema. Assim, é necessário, tanto na síntese como no reconhecimento automático de fala, converter símbolos grafemáticos, que representam palavras, em símbolos fonológicos que representam os modelos. Um sistema de conversão de grafemas para fonemas é muitas vezes designado por G2P (do inglês *Grapheme to Phoneme*; por vezes também designado L2S: *Letter to Sound*) e a sua implementação é fortemente dependente da língua. O português europeu (PE) tem uma sólida e madura investigação nesta área, mas o problema do G2P ainda não se encontra totalmente resolvido, como se pode comprovar quer pelas taxas de erro publicadas nos artigos da área, quer pelos erros de conversão que persistem nos atuais sistemas existentes no mercado. Por outro lado, ainda que a maior causa para os problemas encontrados tenha já sido diagnosticada, e envolve questões de natureza essencialmente morfológica e sintática (cf., a título de exemplo, (Braga and Marques, 2007)), as soluções até ao momento apresentadas, muitas vezes acompanhadas por um dicionário extenso de exceções, não estão claramente publicadas nem se encontram acessíveis para possível melhoria e extensão.

As primeiras abordagens propostas para solucionar o problema de conversão automática de grafemas para fonemas são baseadas em regras linguísticas – regras definidas a partir do conhecimento teórico da língua normalmente definidas por especialista na linguística (Ainsworth, 1973). Mais tarde apareceram abordagens que aplicam algoritmos de aprendizagem, como redes neuronais (Hain, 1999), modelos de *Markov* não observáveis (HMM – *Hidden Markov Model*) (Taylor, 2005), campos aleatórios condicionais (CRF –

Conditional Random Fields) (Wang and King, 2011), modelos de sequências conjuntas de grafemas e fonemas (*joint-sequence models*) (Bisani and Ney, 2008) para inferir as regras ou para fazer o mapeamento baseado em modelos probabilísticos.

Considerando a língua portuguesa, encontram-se na literatura várias abordagens para o problema de conversão automática de grafemas para fonemas para o português europeu, destacando-se as seguintes:

- regras linguísticas (Braga et al., 2006; Candeias and Perdigão, 2008; Oliveira et al., 1992; Teixeira, 2004);
- regras inferidas a partir dos dados (Teixeira et al., 2006);
- máquinas de estados finitos (FST – *Finite-State Transducers*) (Caseiro et al., 2002; Oliveira et al., 2004);
- máxima entropia (Barros and Weiss, 2006);
- redes neuronais (Trancoso et al., 1994);
- árvores de decisão (Oliveira et al., 2001).

A disponibilidade de grandes quantidades de recursos linguísticos sob a forma digital faz com que as abordagens baseadas em modelos estatísticos apresentem resultados que formam, presentemente, o estado da arte da conversão de grafemas para fonemas. Uma das técnicas a referenciar, e que tem sido aplicada essencialmente em línguas que não apresentam uma clara correspondência entre grafema e fonema, como é o caso do inglês, é a que se baseia numa abordagem por modelos probabilísticos, apresentada por Bisani e Ney (Bisani and Ney, 2008). Contrastando com as abordagens baseadas em regras, as quais são suportadas por um conhecimento da estrutura linguística da língua, que se pretende exaustivo, a abordagem estatística baseia-se no pressuposto de que a pronunciação de um vocábulo é possível de ser inferida, por analogia, a partir de exemplos de sequências suficientes de grafonemas – unidades identificativas da associação entre grafema e respetivo fonema (cf. (Bisani and Ney, 2008)). Uma das vantagens apontada aos modelos probabilísticos é não implicarem uma verificação constante da interdependência das regras, em especial quando surge uma sequência de grafemas que sai fora das regularidades até então admitidas. Por outro lado, tem-se

verificado que a conversão de grafemas para fonemas proveniente de modelos probabilísticos não capta um contexto suficientemente amplo de forma a impedir que a estrutura fonológica da língua seja violada. A título de exemplo, é difícil treinar um modelo estatístico capaz de converter corretamente a palavra <amigavelmente> devido à presença da acentuação secundária na sílaba <-ga->, a par da acentuação primária na sílaba <-men-> . A língua portuguesa na sua vertente europeia, assim como as línguas românicas na sua generalidade, apresenta uma razoável regularidade fonética e fonológica bem como uma ortografia de base fonológica (Mateus and Andrade, 2002). Estas características justificam o sucesso da inclusão de regras linguísticas na conversão de grafemas para fonemas, como é o caso da marcação da tónica (descrito em (Candeias and Perdigão, 2008; Braga et al., 2006; Almeida and Simões, 2001; Teixeira et al., 1998; Caseiro et al., 2002).

O presente trabalho apresenta uma abordagem híbrida para a conversão automática de grafemas para fonemas para o português europeu, utilizando modelos de sequências conjuntas com um módulo de pré-processamento para a inclusão de regras linguísticas inequívocas. Além disso, todos os recursos utilizados e produzidos encontram-se disponíveis em (LABFALA, 2011), tais como o dicionários de pronunciação, modelos de sequências conjuntas, bem como o programa para treinar e testar o sistema de conversão de grafemas para fonemas.

No presente estudo, considerou-se trabalhar ao nível do fonema, uma vez que o procedimento de conversão adotado admite valências do contexto mais ou menos alargado no âmbito da unidade acentual (palavra), considerando a unidade para a qual o grafema é convertido como uma escolha significativa por entre todas as outras unidades que o sistema de língua coloca ao dispor. Assim, aceitou-se a unidade fonológica, ou fonema, como uma classe à qual pode corresponder um fone ou um feixe de realizações alofónicas disponíveis no português europeu (acolhendo-se, assim, a possível inserção de pronúncias alternativas).

3.2. Modelo probabilístico de sequências conjuntas

A tarefa de conversão de grafemas para fonemas pode ser formulada em termos da determinação da sequência ótima de fonemas dada a sequência de grafemas, usando uma abordagem probabilística. Definindo $G = G_1^N = \{g_1, g_2, \cdots, g_N\}$ como sendo a sequência de grafemas de entrada e F como uma sequência de fonemas qualquer, a determinação da sequência ótima de fonemas, F^*, é descrita da seguinte forma:

$$F^* = \arg\max_F P(F \mid G). \tag{16}$$

É praticamente impossível determinar F^* calculando diretamente a probabilidade *a posteriori* $P(F \mid G)$ para todas as sequências F possíveis. É mais fácil treinar modelos de fonemas e usar os grafemas como observações. Assim, usando o teorema de *Bayes* pode-se rescrever-se o problema como:

$$F^* = \arg\max_F \frac{P(G \mid F)}{P(G)} \cdot P(F). \tag{17}$$

O fator $1/P(G)$ pode ser eliminado uma vez que é comum para todas as sequências F. Assim, o seu valor não influencia a escolha de F^*, pelo que o problema pode ser simplificado de acordo com a seguinte equação:

$$F^* = \arg\max_F P(G \mid F) \cdot P(F). \tag{18}$$

A estimação de $P(F)$ é feita usualmente recorrendo aos modelos *n-grama*. Quanto à determinação de $P(G \mid F)$, as abordagens que utilizam modelos de *Markov* simplificam o problema assumindo a independência entre os grafemas que constituem uma sequência. Neste caso, o cálculo de $P(G \mid F)$ pode ser decomposto da seguinte forma:

$$P(G \mid F) = \prod_{n=1}^{N} P(g_n \mid F). \tag{19}$$

Esta simplificação parte do princípio de que a dependência entre fonemas é suficiente para modelar o problema e que os contextos de fonemas replicam os contextos de grafemas (Taylor, 2005; Jiampojamarn and Kondrak, 2009).

Existe outra abordagem que propõe a utilização de modelos de probabilidade conjunta, $P(G,F)$, para determinar a sequência ótima de fonemas (Bisani and Ney, 2002; Galescu and Allen, 2001), rescrevendo a Equação (18) da seguinte forma:

$$F^* = \arg\max_F P\left(G \mid F\right) \cdot P(F) = \arg\max_F P\left(G,F\right). \quad \textbf{(20)}$$

Esta abordagem possibilita a modelação da dependência entre grafemas, a dependência entre fonemas e a dependência entre grafemas e fonemas. Para simplificar a aprendizagem, esta abordagem define uma nova entidade, o *grafonema* que significa a junção de grafemas e fonemas e que está descrita na secção 3.2.2. O presente trabalho segue esta abordagem, dado que apresentou melhores desempenhos comparado com as abordagens que utilizam modelos de *Markov*.

Qualquer abordagem estatística adotada na tarefa de conversão de grafemas em fonemas requer a existência de um dicionário fonológico, necessário para estimar as probabilidades dos padrões encontrados, e a maioria das abordagens requer ainda um algoritmo que permita o alinhamento entre grafemas e fonemas.

3.2.1. Alinhamento entre grafemas e fonemas

Muitos grafemas do português têm uma correspondência unívoca com os fonemas, situação da qual resulta uma conversão direta de grafemas para fonemas. É o caso de muitas das consoantes, como o <p> e o <t>, presentes por exemplo em <português>, as quais são diretamente convertidas nos fonemas /p/ e /t/, respetivamente[1]. Contudo, existem grafemas em que a correspondência com os fonemas é dependente de vários fatores, nomeadamente o contexto grafemático (caso dos grafemas <r> e <u> em <português>) e o estatuto morfológico (em inglês PoS – *Part of Speech*), algumas das vezes com interdependência sintática (caso dos grafemas <e> e <o>, os quais, dependendo da sua condição morfológica, podem ser convertidos nos fonemas /e/ ou /E/ e /o/ ou /O/, respetivamente (notados aqui segundo o alfabeto SAMPA – *Speech*

[1] Os grafemas anotam-se geralmente entre os símbolos '<' e '>' e os fonemas entre barras '/'.

Assessment Methods Phonetic Alphabet (Wells, 1997), ver Anexo A). Seguem alguns exemplos para esta última situação: <selo> (nome) → /selu/, <selo> (verbo) → /sElu/; <olho> (nome) → /oLu/, <olho> (verbo) → /OLu/). Existem também situações em que um único grafema pode originar vários fonemas, assim como existem outras situações em que vários grafemas podem originar um único fonema.

Todas as abordagens estatísticas se deparam com problemas como os aqui descritos, sendo necessário, durante o processo de treino, segmentar e alinhar a sequência de grafemas e a sequência de fonemas por forma a ficarem com igual número de segmentos. A solução nem sempre é trivial ou única e depende da forma como os algoritmos de alinhamento associam os grafemas aos fonemas de um dado vocábulo.

De acordo com (Jiampojamarn et al., 2007), os alinhadores podem ser classificados, basicamente, em dois tipos: "um-para-um" e "muitos-para-muitos".

3.2.1.1. Alinhamento "um-para-um"

No alinhamento "um-para-um" cada grafema é associado a apenas um único fonema, originando segmentos com apenas um símbolo. Ainda assim, é necessário utilizar um símbolo nulo (símbolo '_') para lidar com casos em que um grafema pode originar vários fonemas (inserção de fonemas) ou casos em que vários grafemas originam apenas um fonema (apagamento de fonemas). As inserções de fonemas podem ser evitadas, no caso do português europeu, uma vez que ocorrem em pouquíssimos contextos e facilmente identificados, como é o caso do iode que ocorre em algumas das estruturas, tais como em <extra> → /6**j**Str6/ (SAMPA) ou /ɐ**j**ʃtrɐ/ (IPA).

Este tipo de alinhador é de fácil implementação (por exemplo, através do algoritmo de Levenshtein (Gusfield, 1997)), mas necessita do conhecimento prévio do mapeamento entre grafemas e fonemas. Existem duas vertentes dentro do alinhamento "um-para-um": "01-01", quando inserções e apagamentos de fonemas são permitidos e "1-01", quando apenas apagamentos de fonemas são permitidos. No presente estudo, o alinhador usado é o de "um-para-um", na vertente de "1-01", isto é, um grafema para

zero ou um fonema. Neste caso, fonemas múltiplos associados a um único grafema (como <e> → /ɐj/ no exemplo anterior) têm de ser representados por um único símbolo.

3.2.1.2. Alinhamento "muitos-para-muitos"

No alinhamento "muitos-para-muitos", os segmentos podem ser compostos por vários símbolos, o que possibilita a associação de vários grafemas a vários fonemas. Este alinhador é mais genérico, pode ser utilizado sem nenhum conhecimento prévio do mapeamento entre grafemas e fonemas, e lida com os casos de inserções e de apagamentos de fonemas sem necessidade de se recorrer a símbolos especiais. No entanto, os modelos resultantes são mais difíceis de estimar e o desempenho é geralmente inferior ao dos modelos com alinhamento "um-para-um" (Bisani and Ney, 2008). Este tipo de associação é também conhecido como alinhamento "m-n".

3.2.2. Modelos com grafonemas

Depois de efetuado o alinhamento entre grafemas e fonemas, as sequências de grafemas e de fonemas apresentam o mesmo número de segmentos. É proposta na literatura uma nova entidade, composta pela associação de um segmento de grafemas a um segmento de fonemas, denominada de *grafonema* (Bisani and Ney, 2002). Mostra-se um exemplo com o vocábulo <compõem>, no qual os *grafonemas* surgem entre parênteses retos. Neste exemplo considerou-se, tanto quanto possível, um alinhamento de "1-para-1", mas onde se admitem também casos de alinhamento de "2-para-1" e de "1-para-2" (que são resolvidos mais tarde com a abordagem "1-01").

$$\begin{matrix} Grafemas \\ Fonemas \end{matrix}\begin{bmatrix} c \\ k \end{bmatrix}\begin{bmatrix} om \\ o\sim \end{bmatrix}\begin{bmatrix} p \\ p \end{bmatrix}\begin{bmatrix} \tilde{o} \\ o\sim\, j\sim \end{bmatrix}\begin{bmatrix} e \\ 6\sim \end{bmatrix}\begin{bmatrix} m \\ j\sim \end{bmatrix}$$

Uma sequência de K *grafonemas* é anotada como $Q(G,F) = \{q_1, q_2, \cdots, q_K\}$ e o problema de conversão de grafemas para fonemas é escrito, tal como em (21), como:

$$F^* = \arg\max_F P\big(Q(G,F)\big). \tag{21}$$

Sem admitir independência entre símbolos, o cálculo de $P(Q(G,F))$ pode ser decomposto da seguinte forma:

$$P\big(Q(G,F)\big) = P\big(q_1\big)\cdot P\big(q_2 \mid q_1\big)\cdot P\big(q_3 \mid q_1 q_2\big)\cdots P\big(q_K \mid q_1 q_2 ... q_{K-1}\big). \tag{22}$$

No modelo estatístico é frequente limitar-se o contexto (ou história) dos *grafonemas* utilizando os chamados modelos *n-grama*, que correspondem a sequências limitadas a um comprimento até *n* símbolos. Deste modo, a Equação (22) pode ser aproximada a:

$$P\big(Q(G,F)\big) \approx \prod_{i=1}^{K} P\big(q_i \mid q_{i-n+1} ... q_{i-1}\big). \tag{23}$$

3.2.3. Estimação do modelo

Os modelos *n-grama* são utilizados para estimar a probabilidade de um símbolo, neste caso *grafonema*, numa sequência conhecendo os *n*-1 símbolos anteriores da sequência (história). A estimação da probabilidade de um *n-grama* é baseada em contagens de ocorrências de sequências de símbolos num dado conjunto de treino. Definindo a frequência de um *n-grama* por $C(.)$, a sua probabilidade pode ser estimada através de:

$$P\big(q_i \mid q_{i-n+1} ... q_{i-1}\big) = \frac{C\big(q_{i-n+1} ... q_i\big)}{C\big(q_{i-n+1} ... q_{i-1}\big)}, \tag{24}$$

onde o denominador pode ser calculado como:

$$C\big(q_{i-n+1} ... q_{i-1}\big) = \sum_j C\big(q_{i-n+1} ... q_{i-1} q_j\big). \tag{25}$$

A estimação desta probabilidade, baseada apenas na contagem de ocorrências de padrões de sequências de *n* símbolos, acarreta o problema de atribuir probabilidade zero aos *n-gramas* que não estão presentes no dicionário de treino. Além disso, podem existir *n-gramas* que estão presentes no dicionário mas em número sem significado estatístico. Para evitar estes constrangimentos, é preciso precaver a existência de sequências que nunca foram encontradas no dicionário de treino (usando os chamados "descontos" ou

discounts), ou que são pouco frequentes (a "suavização" ou *smoothing*). Assim, uma pequena massa de probabilidade é retirada dos *n-gramas* mais frequentes e é reservada para os *n-gramas* ausentes ou pouco frequentes no dicionário de treino. Existem vários algoritmos propostos para redistribuir a massa de probabilidade dos *n-gramas*, destacando-se os descontos de Good-Turing (Good, 1953), de Witten-Bell (Witten and Bell, 1991), de Kneser-Ney (Kneser and Ney, 1995), o desconto absoluto de Ney (Ney et al., 1994) e a suavização de Katz (Katz, 1987). Utilizam-se também as probabilidades de *n-gramas* de ordens inferiores (*backoff*) na estimação de probabilidade de uma sequência que não foi treinada.

O algoritmo implementado neste trabalho faz a suavização por interpolação e utiliza a versão modificada do algoritmo de Kneser-Ney (Chen and Goodman, 1999). Este algoritmo utiliza a probabilidade de *n-grama* de ordem inferior (*backoff*) na estimação da probabilidade de um *n-grama*. A estimação da probabilidade de um *n-grama* é reescrita da seguinte forma:

$$P\left(q_i \mid q_{i-n+1}...q_{i-1}\right) = \frac{C\left(q_{i-n+1}...q_i\right) - D\left(C\left(q_{i-n+1}...q_i\right)\right)}{C\left(q_{i-n+1}...q_{i-1}\right)} + \\ +\gamma\left(q_{i-n+1}...q_{i-1}\right) \cdot \frac{C\left(q_{i-n+2}...q_i\right)}{C\left(q_{i-n+2}...q_i\right)}, \quad \textbf{(26)}$$

em que $D(c)$ é a função de desconto, definida:

$$D(c) = \begin{cases} 0, & se \quad c = 0 \\ D_1, & se \quad c = 1 \\ D_2, & se \quad c = 2 \\ D_{3+}, & se \quad c \geq 3 \end{cases}. \quad \textbf{(27)}$$

A função $\gamma(\cdot)$ é o peso do *backoff* e é definida de forma a que a soma das probabilidades resulta em 1:

$$\gamma\left(q_{i-n+1}...q_{i-1}\right) = \\ \frac{D_1 \cdot N_1\left(q_{i-n+1}...q_{i-1}\right) + D_2 \cdot N_2\left(q_{i-n+1}...q_{i-1}\right) D_{3+} \cdot N_{3+}\left(q_{i-n+1}...q_{i-1}\right)}{C\left(q_{i-n+1}...q_{i-1}\right)}, \quad \textbf{(28)}$$

em que em que D_1, D_2 e D_{3+} são aplicados aos n-gramas com uma, duas e 3 ou mais contagens, respetivamente e:

$$\begin{aligned} N_1(q_{i-n+1}...q_{i-1}) &= \left|\left\{i : C(q_{i-n+1}...q_{i-1}) = 1\right\}\right| \\ N_2(q_{i-n+1}...q_{i-1}) &= \left|\left\{i : C(q_{i-n+1}...q_{i-1}) = 2\right\}\right| \\ N_{3+}(q_{i-n+1}...q_{i-1}) &= \left|\left\{i : C(q_{i-n+1}...q_{i-1}) \geq 3\right\}\right| \end{aligned} . \quad \textbf{(29)}$$

O valor ótimo dos descontos D_1, D_2 e D_{3+} é calculado:

$$\begin{aligned} D_1 &= 1 - 2 \cdot Y \cdot \frac{n_2}{n_1} \\ D_2 &= 2 - 3 \cdot Y \cdot \frac{n_3}{n_2} \\ D_{3+} &= 3 - 4 \cdot Y \cdot \frac{n_4}{n_3} \end{aligned} , \quad \textbf{(30)}$$

onde Y é definido como:

$$Y = \frac{n_1}{n_1 + 2n_2} . \quad \textbf{(31)}$$

Os valores de n_1, n_2 e n_3 são números de *n-grama* com frequência de ocorrência igual a 1, 2 e 3 respetivamente.

Este algoritmo demonstrou ter um desempenho superior quando comparado com outros também utilizados na suavização de modelos de *n-grama* com a suavização de Katz ou de Witten-Bell (Chen and Goodman, 1999). Uma outra vantagem é que dispensa a estimação de qualquer parâmetro empírico.

3.3. Criação do modelo híbrido

O modelo híbrido é criado a partir da transformação da sequência de grafemas, introduzindo novos símbolos com significado fonológico preciso, proporcionando desta forma a integração de regras fonológicas no modelo estatístico. É o caso dos dígrafos <nh>, <lh>, <ch>, <ss>, <rr>, etc. Os procedimentos para a criação do modelo estatístico

não são alterados; apenas passam a existir mais símbolos, precisando mais claramente a associação entre grafema e fonema.

Foi necessário criar um dicionário de pronunciação para treinar os modelos *n-grama*. A seguir, descrevem-se em pormenor os passos que foram dados até à criação dos modelos.

3.3.1. Vocabulário

Foi necessário, numa primeira fase, definir uma listagem de vocábulos atuais e representativos do português europeu. O material utilizado para esse fim foi o corpus *CETEMPúblico* (Santos and Rocha, 2001), uma coleção de extratos do jornal *Público*, publicado entre 1991 e 1998 com aproximadamente 180 milhões de palavras[2]. Foram filtradas as palavras da listagem inicial do *CETEMPúblico*, aceitando apenas as palavras que obedecem simultaneamente aos seguintes critérios:

- começar com um grafema do alfabeto português (a-z, A-Z, á-ú, Á-Ú);
- não conter dígitos;
- não apresentar todos os grafemas em maiúscula (caso de siglas);
- não conter o caráter '.' (caso de *URLs*);
- terminar com um grafema do alfabeto português ou com '-';
- o lema correspondente não conter o caráter '=' (caso de nomes compostos no *CETEMPúblico*).

A partir do resultado obtido, formou-se uma lista de cerca de 50.000 vocábulos (excluindo nomes próprios), os quais correspondem a uma contagem de ocorrências no corpus de mais do que 70 vezes. Sendo arbitrária, a consideração desta medida para a configuração do vocabulário de base deveu-se ao facto de anular a possibilidade de se estarem a incluir erros tipográficos e de se obter uma primeira listagem representativa do PE extensível até

[2] Por palavras entendem-se, aqui, todos os átomos do corpus que contêm, pelo menos, um grafema ou dígito. Esta significação é extensível a todas a vezes que o termo "palavra" surge no contexto do *CETEMPúblico*. No âmbito deste trabalho, palavra é sinónimo de vocábulo, podendo coexistir os dois termos.

cerca de 50.000 vocábulos. Por fim, foram retirados quer vocábulos estrangeiros quer estrangeirismos, usando, em primeiro lugar, critérios automáticos e, seguidamente, uma verificação manual. A pesquisa automática excluiu todos os vocábulos que apresentavam grafemas ou sequências grafemáticas que não fazem parte do sistema do português europeu, tais como <k>, <w> e <y>; <sh> e <pp>; e <b>, <d> ou <p> em posição final de vocábulo. Alguns destes vocábulos foram usados como base de constituição de um dicionário de pronunciação de estrangeirismos (cf. 3.3.2). Como resultado final deste processo, construiu-se uma lista de cerca de 40.000 vocábulos, os quais correspondem ao vocabulário de referência tomado para este trabalho, referenciado por "*voc_CETEMP_40k*". Na medida em que as palavras que constituem o *CETEMPúblico* apresentam uma grafia de acordo com as normas anteriores ao Acordo Ortográfico de 1990 (AO90), houve necessidade de se constituir uma listagem adicional com vocábulos grafados de acordo com o AO90. Usou-se a ferramenta *Lince* (Lince, 2010) para converter os vocábulos para a nova grafia. Das 41.589 palavras do vocabulário sem aplicação do AO90 (pré-AO90), 915 sofreram alterações de grafia, nomeadamente a eliminação das consoantes mudas (<c> e <p>), a eliminação dos hífenes e a alteração da acentuação gráfica. De acordo com a possibilidade de coexistirem duas grafias, este novo vocabulário apresenta pares de vocábulos ditos "parónimos", tais como <conceptual> e <concetual> ou <desconectar> e <desconetar>. O vocabulário de acordo com o AO90 (pós-AO90) é constituído por 41.602 vocábulos, sendo referenciado como "*voc_CETEMP_40k_ao*".

Nas secções seguintes não será feita a distinção entre estes dois vocabulários, a não ser quando pertinente, como é o caso da secção dos resultados.

3.3.2. Transcrição fonológica

A transcrição fonológica do vocabulário de referência foi efetuada por um processo iterativo. Em primeiro lugar, foi feito um modelo estatístico, conforme descrito acima em 3.2.2, tendo por base o dicionário de pronunciação da base de dados *SpeechDat* (SpeechDat, 1998), com cerca de 15.000 vocábulos. Para a constituição do dicionário foram retirados os estrangeirismos e foram efetuadas algumas correções e normalizações

de pronunciação. Não se distinguiu a lateral velarizada da lateral, ainda que sistemas reconhecidos de anotação para o português, como o usado na *SpeechDat*, admitam a presença de /l~/ (/5/ em X-SAMPA) e de /l/. Convencionou-se que os símbolos representativos das glides /j/ e /w/ (semivogais) fossem notados, inicialmente, como as vogais correspondentes. A inclusão das semivogais é feita num módulo de pós-processamento que utiliza um conjunto de regras para o efeito. Assim, as semivogais foram suprimidas durante o processo de treino, eliminando alguma ambiguidade que poderia suscitar a sua utilização. Optou-se, igualmente, pela inclusão do iode, como foi já observado em 3.2.1, com o objetivo de uma maior aproximação à pronúncia padronizada do português europeu.

Os símbolos SAMPA adotados (cf. Tabela 2 e Tabela 3) resultaram de uma ponderação cuidada sobre representatividade do português europeu falado. A observação atenta dos alfabetos fonéticos SAMPA para o Português (SAMPA-PT) e X-SAMPA, deu conta de alguma indefinição de regularidade, exemplificada na atribuição de mais do que um símbolo para o mesmo som. Na verdade, o símbolo /r/ no SAMPA-PT parece ter como correspondente no X-SAMPA o símbolo /4/ (IPA: /ɾ/), simbolizando o /r/ no X-SAMPA a vibrante alveolar múltipla (IPA: /r/). Para facilitar o alinhamento entre grafemas e fonemas, definiu-se um novo conjunto de símbolos (SAMPA unicaráter) para codificar os símbolos SAMPA que são representados por mais de um caráter.

Usando o dicionário de pronunciações do *SpeechDat* normalizado, criou-se uma primeira versão do modelo estatístico que foi utilizado para transcrever o vocabulário *CETEMPúblico.* De forma informal, verificou-se que o resultado da aplicação do modelo estatístico ao vocabulário *CETEMPúblico* (“voc_CETEMP_40k”) era já bastante preciso, apresentando, pontualmente, algumas incorreções.

Seguiu-se então um processo moroso de verificação e correção manual das transcrições obtidas automaticamente. O passo seguinte consistiu em comparar as transcrições do dicionário com as transcrições geradas por um sintetizador de fala comercial. Esta comparação permitiu-nos confiar no nosso resultado já que, maioritariamente, as transcrições coincidiram. As transcrições que diferiram foram analisadas individualmente

e corrigidas quando necessário, no sentido da maior representatividade do português europeu. Deste processo resultou o dicionário de transcrição fonológica que será referenciado como "dic_CETEMP_40k". Com este dicionário foi feito um novo modelo estatístico. O teste do modelo com o próprio dicionário de treino permitiu ainda corrigir alguns erros remanescentes, bem como uniformizar algumas transcrições. Por exemplo, os vocábulos iniciados por <ex-> são transcritos como /6iS/ (observando-se a inserção do iode) assim como em <extra> /6iStr6/, mas não em <extenso> /@Ste~su/, não se tendo transcrito a sequência <ex> como /6iS/ em certos contextos de atonicidade.

Tabela 2 – Símbolos SAMPA, símbolos *unicarácter* (uc) e IPA[3] das vogais com os grafemas possíveis e exemplos.

SAMPA	uc	IPA	Grafemas	Exemplos
6		ɐ	a, â, e, ê	d**a**, cr**â**nio, v**e**nho, **a**m**ê**ijo**a**
a		a	a, á, à	p**a**la, p**á**, **à**
@		ə	e	d**e**
e		e	e, ê	d**e**do, v**ê**
E		ɛ	e, é	p**e**le, p**é**
i		i	i, í, e, y	v**i**, a**í**, r**e**al, henr**y**
o		o	o, ô, ou	**o**co, av**ô**, l**ou**co
O		ɔ	o, ó	p**o**te, p**ó**
u		u	u, ú, o, w	t**u**, ba**ú**, at**o,** ki**w**i
6~	ã	ɐ̃	ã, an, ân, am, âm, e, é, a	v**ã**, **an**ca, **ân**sia, **am**pla, **âm**bar, t**e**m, al**é**m, i**a**m
e~	ë	ẽ	ên, en, êm, em	ag**ên**cia, p**en**te, **êm**bolo, **em**pate
i~	ï	ĩ	i, in, im, ím, ín, m, n, e	mu**i**ta, **in**ca, s**im**, **ím**pio, **ín**dio, alé**m**, be**n**s, põ**e**
o~	õ	õ	õ, ôn, ôm, on, om	i**õ**es, c**ôn**sul, t**ôm**bola, p**on**te, p**om**bo
u~	ü	ũ	u, ún, un, um, úm, o, m	m**u**ito, n**ún**cio, **un**s, at**um,** c**úm**plice, vã**o**, ia**m**

[3] IPA – *International Phonetic Alphabet*

Tabela 3 – Símbolos SAMPA e símbolos IPA das consoantes com os grafemas possíveis e exemplos.

SAMPA	IPA	Grafemas	Exemplos
b	b	b	**b**e**b**er
d	d	d	**d**a**d**o
g	g	g	**g**ato
p	p	p	**p**ato
t	t	t	**t**oca
k	k	q, c, k	**q**uando, **c**asa, **k**iwi
f	f	f	**f**é
s	s	s, ç, x, c, ss	**s**ol, ca**ç**a, au**x**ílio, **c**ima, a**ss**im
S	ʃ	ch, s, z, x	**ch**ave, pá**s**, pa**z**, **x**á
v	v	v	**v**ida
z	z	z, s, x	ca**s**a, **z**ebra, e**x**emplo
Z	ʒ	j, g, s, z, x	**j**á, **g**ira, me**s**mo, e**x**-líder
l	l	l	**l**ua
L	ʎ	lh	ve**lh**o
r	ɾ	r	ca**r**o
R	ʀ	r, rr	ca**rr**o, **r**ato
m	m	m	**m**ão
n	n	n	**n**ada
J	ɲ	nh	se**nh**a

3.3.3. Alinhador de grafemas com fonemas

O alinhamento entre a sequência de grafemas e a sequência de fonemas é um passo crucial para a criação do modelo *n-grama*. Optou-se, como já foi referido antes, pelo alinhamento "um-para-um" na vertente "1-01", em que um grafema pode corresponder a zero ou a um fonema. Esta opção foi baseada no facto de se ter encontrado apenas 7 contextos em que um grafema pode corresponder a mais do que um fonema. O problema criado por estes casos foi solucionado com a criação de símbolos unicaráter que

correspondem a mais do que um fonema. Foi utilizado o símbolo unicaráter '_' para indicar os casos em que um grafema corresponde a zero fonemas. Estes casos estão ilustrados na Tabela 4.

Tabela 4 – Símbolos SAMPA, símbolos *unicaráter* (uc) e IPA dos casos especiais para permitir alinhamento "1-01" com os grafemas possíveis e exemplos.

SAMPA	uc	IPA	Grafemas	Exemplos
	_		c, h, p, u, z	se**c**tor, **h**ábil, rece**p**ção, toq**u**e, jaz**z**
ks	K	ks	x, cc, cç	tele**x**, fi**cc**ional, fi**ç**ão
o~i~	¤	õĩ	õ	p**õ**em
Oi	®	ɔi	o	constr**o**em
ai	Å	ai	a	s**a**em
6~i~6~	Ê	ɐ̃ĩɐ̃	ê	t**ê**m
6i	æ	ɐi	e, ê	**e**x-líder, t**ê**xtil

O alinhamento entre grafemas e fonemas é então obtido usando o conhecido algoritmo de alinhamento entre cadeias de carateres (*edit distance* ou algoritmo de *Levenshtein*) (Gusfield, 1997). Para tal, foi necessário definir uma matriz de distância ou uma matriz de custo de associação entre grafemas e fonemas baseada na probabilidade condicional, $P(f \mid g)$, em que *f* representa um fonema e *g* representa um grafema. Esta probabilidade é estimada a partir de um dicionário de transcrições inicial alinhado. Definiu-se também um valor máximo para essa distância, $d_{\max}$, para os casos onde não existe qualquer associação entre grafema e fonema. Existe um apagamento de um grafema sempre que esse grafema não dá origem a um fonema e, para que isso aconteça, o apagamento tem de ter um custo menor que $d_{\max}$, admitindo ser preferível apagar um grafema a fazer uma associação errada.

As sequências (de grafemas e de fonemas) alinhadas têm o mesmo número de símbolos, tornando assim trivial a criação de grafonemas (junção de um símbolo de grafema com um símbolo de fonema).

A aplicação de regras linguísticas, nomeadamente os dígrafos, afeta o processo de alinhamento uma vez que são definidos novos símbolos grafemáticos. A correspondência dos grafemas <c> e <p> ao fonema /_/ só se verifica no vocabulário pré-AO90.

3.3.4. Regras fonológicas

O português europeu apresenta uma razoável regularidade fonética e fonológica bem como uma ortografia de base fonológica que justifica a inclusão de constrições linguísticas que auxiliam a tarefa de conversão de grafemas em fonemas (Mateus and Andrade, 2002). Assim, foram adicionadas regras fonológicas para a acentuação vocálica, reconhecendo o núcleo de sílaba tónica de cada vocábulo, e para a identificação da correspondência exata entre um grafema e respetivo fonema, de acordo com o contexto.

As regras resultam na definição de símbolos grafemáticos que as exprimem e que são introduzidos no modelo estatístico. Foram, assim, criados símbolos para dígrafos, vogais tónicas e grafemas em certos contextos fonológicos.

3.3.4.1. Dígrafos

Um dígrafo ocorre quando dois grafemas correspondem a um único fonema. Propõe-se alterar a representação dessas sequências de dois grafemas de forma a permitir uma associação ótima entre o símbolo grafado e o símbolo sonoro. Neste estudo foram consideradas como dígrafos sequências vocálicas e sequências consonânticas. No âmbito das sequências vocálicas, considerou-se a sequência oral <ou>, a qual, seguindo a pronúncia padronizada do português europeu, corresponde ao fonema singular /o/, e as sequências nasais que seguem os padrões <V + m + C> e <V + n + C> (‘V’ – vogal; ‘C’ – consoante; ‘+’ – concatenação) em que o <m> e o <n> indicam a nasalidade da vogal precedente. As sequências consonânticas consideradas são <ch>, <lh>, <nh>, <rr>, <ss>. Não foram consideradas as sequências <qu> e <gu> porque apresentam ambiguidades nas suas correspondências com os fonemas (existem casos em que a vogal <u> é

pronunciada e casos em que não é pronunciada: <quente>, <frequente>). Foram consideradas também as sequências consonânticas <cc> e <cç> quando é utilizado o vocabulário pré-AO90.

Cada dígrafo é reduzido a um único símbolo grafemático antes do alinhamento conforme indicado na Tabela 5.

Tabela 5 – Dígrafos e símbolos *unicaráter* (uc) dos casos especiais para permitir alinhamento "1-01" com os grafemas possíveis e exemplos.

Dígrafos	Novos símbolos	Fonemas	Exemplos
ch	S	S	**ch**ave → **S**ave
lh	L	L	a**lh**o → a**L**o
nh	J	J	ni**nh**o → ni**J**o
rr	R	R	ca**rr**o → ca**R**o
ss	§	s	ma**ss**a → ma**§**a
cc	C	s, ks	a**cc**iona → a**C**iona
cç	Ç	s, ks	a**cç**ão → a**Ç**ão
ou	º	o	p**ou**co → p**º**co
an, am	Ä	6~	c**an**tina → c**Ä**tina
ân, âm	Â	"6~	**âm**bito → **Â**bito
en, em	Ë	e~	s**en**tido → s**Ë**tido
ên, êm	Ê	"e~	c**ên**timos → c**Ê**timos
in, im	Ï	i~	l**im**peza → l**Ï**peza
ín, ím	Í	"i~	**ín**dio → **Í**dio
on, om	Ö	o~	c**on**tar → c**Ö**tar
ôn, ôm	Ô	"o~	c**ôn**cavo → c**Ô**cava
un, um	Ü	u~	c**um**prir → c**Ü**prir
ún, úm	Ú	"u~	c**úm**plice → c**Ú**plice

Quando são aplicadas as regras de marcação da tonicidade, os dígrafos vocálicos que ocorrem em posições tónicas são codificados com novos símbolos.

3.3.4.2. Marcação de tonicidade

Seguindo os pressupostos teóricos discutidos em (Mateus and Andrade, 2002), admitiu-se tratar de uma tarefa de maior importância a marcação das vogais acentuadas, núcleos de sílaba, no âmbito de um vocábulo enquanto unidade acentual. A informação sobre a vogal tónica tem sido reconhecida em trabalhos prévios de conversão de grafema para fonema, quer para a implementação de regras de transmutação do grafema em fonema, quer para a modulação de índices prosódicos (em especial se a informação for alargada à sílaba tónica). Sendo o contexto do *n-grama* fixo, curto e sem informação silábica, o conhecimento da vogal tónica traduziu-se num melhoramento substancial do modelo estatístico, uma vez que permitiu definir vários grafonemas de forma unívoca. Assim como em (Andrade and Viana, 1985), a proposta atual considerou ser pertinente a marcação da vogal tónica (identificada com o símbolo SAMPA ") e não a respetiva unidade silábica.

O processo de identificação de vogal tónica foi conseguido de uma forma original, sem o recurso à informação silábica, o que não é usual noutros trabalhos (pelo menos nos trabalhos pesquisados sobre este assunto). A transformação da grafia para a marcação da vogal tónica é efetuada quando a palavra não apresenta nenhum diacrítico para sinalizar a vogal tónica. Excetuando os dígrafos nasais (vogais nasais) e o dígrafo <ou>, esta marcação é indicada pela transformação das vogais na posição tónica em maiúsculas. Os dígrafos nasais na posição tónica são transformados em dígrafos nasais com a acentuação gráfica correspondente. Foi também considerada a marcação da vogal tónica secundária para as palavras terminadas em <mente>.

Os passos para a marcação da vogal tónica numa palavra são os seguintes:

I. Separa as palavras delimitadas por hífen. Existem, nos vocabulários utilizados, cerca de 6% das palavras que contêm hífen. Exemplos: <abaixo-assinado> (<abaixo> e <assinado>), <aceita-se> (<aceita> e <se>) e <água-de-colónia> (<água>, <de> e <colónia>);

II. Verifica se a palavra termina em <mente> (ou <mËte> quando são utilizados os dígrafos nasais), separa a raiz da terminação <mente> e marca a penúltima vogal (<e>) como tónica (<m**E**nte> ou <m**Ê**te>). Se assim for, verifica se a palavra é um advérbio derivado de um adjetivo cuja acentuação gráfica foi suprimida. Neste caso, tenta recuperar a acentuação gráfica, auxiliado por um conjunto de regras aplicadas à raiz. Exemplos: <timidamente> (<t**í**mida> e <m**E**nte>), <solidamente> (<s**ó**lida> e <m**E**nte>), <somente> (<s**ó**> e <m**E**nte>). A partir deste ponto, o processamento incide apenas sobre a raiz (palavra sem o sufixo <mente>). Existem cerca de 1.6% das palavras a terminar em <mente> nos vocabulários utilizados;

III. Se a palavra apresentar algum diacrítico que sinaliza a vogal tónica, termina a marcação da tónica. Cerca de 20% das palavras dos vocabulários utilizados têm diacríticos que sinalizam a vogal tónica. Exemplo: <g**é**lida>;

IV. Ignora o <s> no final das palavras. Abrange cerca de 27% das palavras presentes nos vocabulários utilizados. Exemplos: <delfin**s**> (<delfin>), <casa**s**> (<casa>), <atrize**s**> (<atrize>);

V. Se a palavra terminar em <u>, <in>, <im>, <un>, <um> (também <º>, <Ï> ou <Ü> quando são utilizados os dígrafos), ou em uma consoante que não seja <m> nem <n>, a posição da última vogal é inicialmente indicada como tónica (aproximadamente 19% das palavras). Exemplos: <men**U**>, <delf**I**n(s)>, <delf**I**m>, <at**U**n(s)>, <at**U**m>, <ca**I**r>, <can**A**l> e <in**O**x>. Exemplos em que a marcação inicial não se confirma (ver o passo IX): <ma**U**>, <te**U**>, <dormi**U**>, <atra**I**>, <sue**I**>, <fo**I**> e <polu**I**>;

VI. Se a palavra terminar em <i> e a vogal <i> for não for precedida de <qu> ou <gu>, a posição da última vogal (<i>) é inicialmente indicada como tónica. Exemplos: <un**I**>, <ag**I**>, <perf**I**(s)>. Se terminar em <gui> ou <qui> em que a vogal <u> não é pronunciada, a última vogal (<i>) é indicada como tónica. Exemplos: <aqu**I**> e <segu**I**>. Se a vogal <u> é pronunciada, a penúltima vogal (<u>) é marcada como tónica. Verificou-se que existem apenas cinco verbos que em algumas flexões têm este padrão final (terminar em <qui(s)> ou <gui(s)> e a vogal <u> é pronunciada). São os verbos: <arguir> (<arg**U**is>, <arg**U**i>), <redarguir> (<redarg**U**is>,

<redarg**U**i>), <relinquir> (<relinq**U**is>, <relinq**U**i>), <delinquir> (<delinq**U**is>, <delinq**U**i>) e <codelinquir> (<codelinq**U**is>, <codelinq**U**i>). Esta regra é aplicada apenas a 1.5% das palavras dos vocabulários utilizados;

VII. Se a palavra terminar em <que>, <gue>, <quem> ou <guem> em que a vogal <u> não é pronunciada, a posição da antepenúltima vogal é inicialmente indicada como tónica. Exemplos: <d**E**ngue>, l**I**gue, gr**O**gue, s**A**ngue. Se a vogal <u> é pronunciada, esta é marcada como tónica. Exemplos: <ag**U**e> (<aguar>), <adeq**U**e> (<adequar>), <apazig**U**e> (<apaziguar>), <relinq**U**em> (<relinquir>) e <arg**U**em> (<arguir>). Existe um número residual de palavras (0,3%) nos vocabulários utilizados onde esta regra é aplicada;

VIII. Se o padrão final da palavra não for nenhum dos casos anteriores (cerca de 59% das palavras dos vocabulários utilizados), a posição da penúltima vogal é inicialmente indicada como tónica (caso a palavra tenha pelo menos duas vogais). Exemplos: <c**A**sa>, <can**E**la>, <incl**I**na>, <cart**O**la> e <cost**U**me>;

IX. se a posição indicada inicialmente como tónica corresponder a vogal <i> ou <u> e se for precedida de uma vogal diferente daquela que for indicada como tónica, a vogal precedente é marcada como tónica. Exemplos: <m**A**u>, <pol**U**i>, <vel**E**iro> e <p**A**usa>. Existem, no entanto, alguns contextos em que a tónica não é passada para a vogal precedente, quando:

   a. a posição da tónica corresponde a um vogal nasalada. Exemplos: <transe**U**nte>, <a**I**nda> e <ca**I**ndo>;
   b. a posição da tónica é precedida de <i> e não corresponde ao último grafema. Exemplos: <esmi**U**ça> e <multi**U**so>;
   c. a posição da tónica corresponde ao penúltimo grafema e é seguida do grafema <r>, <l> ou <z>. Exemplos: <polu**I**r>, <ra**U**l> e <ra**I**z>;
   d. a posição da tónica é seguida do padrão <r+C>, <l+C> ou <z+C> (C – consoante). O padrão <rr>, <ll> ou <zz> não é válido para reter a tónica. Exemplos: <di**U**rno>, <ca**I**rmos> e <atra**I**rdes>;
   e. a posição da tónica é precedida do padrão <au> ou <ao> (ditongos). Exemplos: <tau**I**smo> <mao**I**sta>, <alau**I**ta> e <cau**I**la>;

f. a posição da tónica é seguida do padrão <nh> (ou <J> quando são utilizados os dígrafos). <ra**l**nha>, <gra**U**nha> e <remo**l**nho>;
g. a posição da tónica é precedida do padrão <qu> ou <gu>. Exemplos: <segu**l**da> e <traqu**l**na>

Existem palavras em latim que não foram eliminadas dos vocabulários utilizados. Foram identificadas 7 destas palavras que não obedecem às regras da determinação da vogal tónica (<c**a**mpus>, <cl**a**usus>, <c**o**rpus>, <v**e**rsus>, <gen**e**ris>, <hon**o**ris> e <posteri**o**ri>). Estes casos são identificados e acentuados antes do módulo de marcação da vogal tónica.

Foi criado mais um conjunto de símbolos unicaráter (SAMPA_uc) para incorporar a informação da vogal tónica num único símbolo, conforme listado na Tabela 6.

Tabela 6 – Símbolos SAMPA e símbolos *unicaráter* (uc) das vogais tónicas.

SAMPA	SAMPA_uc
"6	â
"a	á
"e	ê
"E	É
"i	í
"o	ô
"O	Ó
"u	ú
"6~	Ã
"e~	Ë
"i~	Ï
"o~	Õ
"u~	Ü

Além destes símbolos unicaráter e os casos especiais apresentados na Tabela 4 que ocorrem sempre na posição tónica (/¤/, /Ê/, /®/ e /Å/), foi definido o símbolo /Æ/ para indicar o símbolo /æ/ na posição tónica.

3.3.4.3. Regras para contextos frequentes

A descodificação da transmutação de grafema em fonema sem ambiguidade foi também auxiliada pela indicação de regras simples que atendem ao contexto grafemático. A título de exemplo, a determinação da sequência grafemática "<al+C>" resulta na notação de <a> em /a/ (ex.: <**a**lmoçar> → /**a**lmusar/); a definição de "<V+s+V" resulta na notação de <s> em /z/ (ex.: <ca**s**a> → /ka**z**6/). Foram ainda definidas outras regras para o <s> e para os grafemas <r>, <z>, <c>, <g> e <x>, inseridos em contextos mais restritos. Considerando um contexto mais alargado, na sequência grafemática <muit>, as vogais orais <u> e <i> passam a vogais nasais (<muit> → /mu~i~t/).

3.4. Multipronunciação e palavras homógrafas heterofónicas

Muitas palavras apresentam pronúncias que variam com a região (regionalismos), com os contextos onde aparecem nas frases, e com a velocidade da pronúncia (coloquial ou pausada). Neste trabalho fez-se um esforço para utilizar a pronúncia pausada (ou silabada) e normalizar a pronúncia das palavras ao eixo Coimbra-Lisboa, embora se reconheça que a normalização da pronúncia apresenta uma subjetividade maior quando comparada com a normalização da grafia.

A multipronunciação de palavras homógrafas heterofónicas é um problema adicional para um sistema de conversão de grafemas para fonemas que utiliza a apenas a grafia (não utiliza o contexto nem outras informações sobre a palavra).

As palavras homógrafas heterofónicas apresentam um problema adicional para os sistemas de conversão de grafemas para fonemas, uma vez que têm mais de que uma pronúncia e as pronúncias corretas dependem do contexto onde aparecem nas frases. Muitas pronúncias desambiguam com a determinação da classe gramatical, como por exemplo a palavra <dobro> pronunciada /dobru/ (substantivo) ou /dObru/ (verbo) e a palavra <esmero> pronunciada /@Smeru/ (substantivo) ou /@SmEru/ (verbo). No entanto, existem ainda outras em que não é possível desambiguar a pronúncia através da

determinação da classe gramatical, como por exemplo a palavra <travesso> pronunciada /tr6vesu/ (adjetivo) ou /tr6vEsu/ (adjetivo), a palavra <apegar> pronunciada /6p@gar/ (verbo) ou /6pEgar/ (verbo) e a palavra <cesto> pronunciada /seStu/ (substantivo) ou /sEStu/ (substantivo). Para lidar com este problema, propôs-se a criação de uma lista exaustiva de palavras homógrafas heterofónicas que será usada como uma das listas de exceções num sistema de conversão de grafemas para fonemas. A lista foi criada a partir das bases de dados CETEMPúblico (Santos and Rocha, 2001), o Vocabulário Ortográfico do Português (VOP) (VOP, 2010) e a consulta dos dicionários do português da Priberam e da Porto Editora que podem ser consultados livremente na internet (DLPO, 2013; Infopédia, 2013). Foram transcritas manualmente e adicionadas outras informações como a classe gramatical, a identificação da vogal cuja pronúncia é alternada, a pronúncia por defeito e uma indicação sobre a frequência de utilização da palavra no uso corrente. Esta lista foi tornada pública (LABFALA, 2011) e contem cerca de 500 homógrafas heterofónicas. As ambiguidades encontradas recaem todas nas pronúncias das vogais <e> (/e/, /E/ ou /@/) e <o> (/o/ ou /O/). As alternâncias da pronúncia da vogal <e> em /e/ ou /E/ e da pronúncia da vogal <o> em /o/ ou /O/ são as mais comuns e acontecem quando estas vogais são tónica (<este> - /”eSt@/ ou /”ESt@/, <seco> - /s”eku/ ou /s”Eku/, <abono> - /6b”onu/ ou /6b”Onu/, <molho> - /m”oLu/ ou /m”OLu/). A alternância da pronunciação da vogal <e> em /@/ ou /E/ foi identifica em posições não tónicas em várias flexões dos verbos <pregar> (/pr@g”ar/ ou /prEg”ar/) e <apegar> (/6p@g”ar/ ou /6pEg”ar/) além da palavra <colherão> pronunciada como /kuLEr”6~w~/ ou /kuL@r”6~w~/.

Os vocabulários utilizados para criar os modelos estatísticos apresentam algumas homógrafas heterofónicas mas apenas com as pronúncias mais frequentes.

3.5. Resultados

Todas as experiências foram baseadas no dicionário de pronunciação descrito na secção anterior. Internamente, o sistema utiliza os símbolos SAMPA_uc por forma a representar todos os fonemas com um único símbolo. O dicionário base tem 41589 palavras, perfazendo um total de 367933 grafemas (média de 8,85 grafemas por palavras) sendo que 2797 são hífenes. A transcrição fonológica resultou em 341859 fonemas, uma média de 8,22 fonemas por palavra.

Para determinar o impacto de cada conjunto de regras fonológicas, foram criados vários dicionários a partir da aplicação de diferentes formas de pré-processamento ao dicionário base, "dic_CETEMP_40k". Assim, foram criados 4 dicionários conforme a Tabela 7.

Tabela 7 – Identificação dos dicionários conforme as regras fonológicas.

Regras fonológicas		**Nome do dicionário**
Tónica	Dígrafos	
		"dic_CETEMP_40k_alinhado"
	X	"dic_CETEMP_40k_alinhado_digrafos"
X		"dic_CETEMP_40k_alinhado_tonica"
X	X	"dic_CETEMP_40k_alinhado_digrafos_tonica"

O "dic_CETEMP_40k_alinhado" é o resultado do alinhamento do dicionário base, sem a aplicação de nenhuma das regras fonológicas. Foram identificados 380 casos especiais indicados na Tabela 4, o que reduz o número de fonemas para 341465, e foi necessária a inserção de 26468 fonemas nulos (/_/) para garantir o alinhamento entre grafemas e fonemas.

Foram identificadas 21274 situações onde é possível utilizar os símbolos especiais para os dígrafos. Os dicionários que utilizam estes símbolos têm menos grafemas e utilizam menos o fonema nulo (/_/) para garantir o alinhamento entre grafemas e fonemas. A indicação da vogal tónica nos dicionários é sinalizada apenas nos fonemas, uma vez a transformação da grafia é operada internamente no sistema de conversão de grafemas

para fonemas. Também a aplicação da regra para contextos frequentes, que implica apenas a transformação na grafia, é feita internamente quando for necessária.

As experiências foram feitas usando a estratégia de validação cruzada, particionando cada dicionário de forma aleatória em 5 partes disjuntas. Usando uma parte como conjunto de teste e as restantes quatro para treinar o modelo estatístico, pode-se obter 5 resultados para cada condição de teste, rodando os conjuntos de testes. O resultado final pode ser expresso como a média dos 5 resultados para cada condição de teste.

O desempenho do sistema de conversão de grafemas para fonemas é expresso usando duas taxas de erros: PER (*Phoneme Error Rate*) referente a taxa de erro de fonemas e WER (*Word Error Rate*) referente a taxa de erro de palavras. A Tabela 8 sumariza os resultados obtidos sem a aplicação de regras fonológicas e a Tabela 9 apresenta os resultados com a aplicação de todas as regras fonológicas, variando o *n-grama* entre 2 e 10.

Tabela 8 – Resultados sobre o dicionário "dic_CETEMP_40k_alinhado" sem regras fonológicas.

n-grama	**WER (%)** média (± desvio padrão)	**PER (%)** média (± desvio padrão)
2	35.23 ± 0.69	4.71 ± 0.10
3	15.28 ± 0.30	1.84 ± 0.03
4	7.81 ± 0.36	0.93 ± 0.04
5	5.60 ± 0.33	0.67 ± 0.04
6	5.50 ± 0.28	0.66 ± 0.03
7	5.60 ± 0.24	0.67 ± 0.03
8	5.59 ± 0.26	0.67 ± 0.03
9	5.60 ± 0.26	0.67 ± 0.03
10	5.59 ± 0.26	0.67 ± 0.03

Tabela 9 – Resultados sobre o dicionário "dic_CETEMP_40k_alinhado_digrafos_tonica" com todas as regras fonológicas.

n-grama	**WER (%)** média (± desvio padrão)	**PER (%)** média (± desvio padrão)
2	10.05 ± 0.56	1.31 ± 0.06
3	4.77 ± 0.18	0.61 ± 0.03
4	2.46 ± 0.09	0.32 ± 0.01
5	2.15 ± 0.15	0.28 ± 0.02
6	2.20 ± 0.15	0.28 ± 0.02
7	2.27 ± 0.20	0.29 ± 0.02
8	2.27 ± 0.19	0.29 ± 0.02
9	2.26 ± 0.18	0.29 ± 0.02
10	2.27 ± 0.17	0.29 ± 0.02

A utilização de regras fonológicas resultou numa diminuição significativa das taxas de erros (a redução relativa das taxas de erros varia entre 57% e 76%). Os gráficos apresentados na Figura 10 e na Figura 11 ilustram o desempenho individual de cada conjunto de regras fonológicas. É de notar que o incremento de *n-grama* a partir de 6 não traduz numa melhoria do desempenho, como se poderia esperar. Isto pode ser explicado pela falta de amostras suficientes para estimar convenientemente *n-gramas* com grandes contextos, como é demonstrado no gráfico da Figura 12. Pode-se verificar que a marcação da vogal tónica é que dá maior contributo para a melhoria do desempenho do sistema global. As outras regras têm uma influência residual para *n-gramas* igual ou superior a 5. De facto, a utilização de símbolos especiais para dígrafos só é eficaz para os *n-grama*s menores que 5, uma vez que acentua o problema da falta de amostras para treinar grandes contextos. O melhor resultado obtido foi com 5-*grama*, com as regras dos contextos frequentes e com a marcação da vogal tónica (WER de 2.03% e PER de 0.25%).

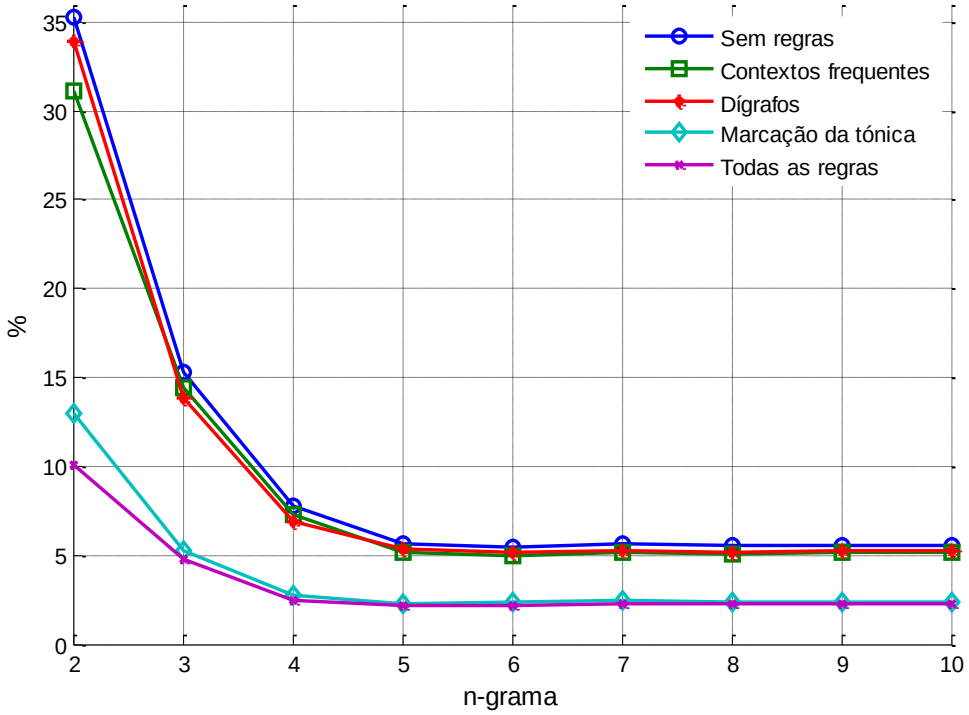

Figura 10 – WER das regras fonológicas em função do *n-grama*.

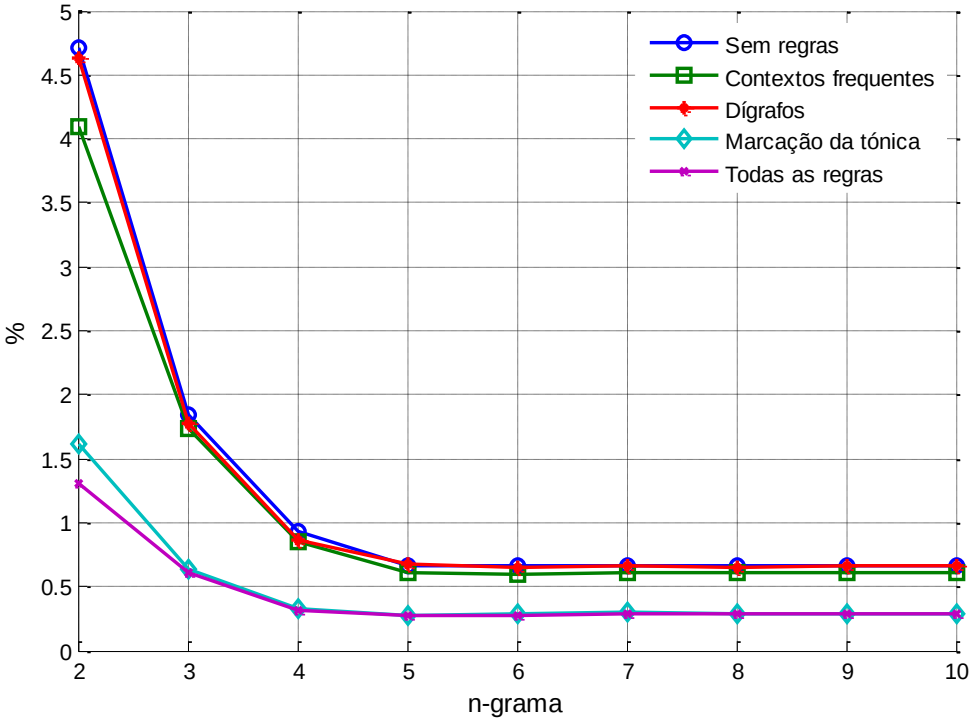

Figura 11 – PER das regras fonológicas em função do *n-grama*.

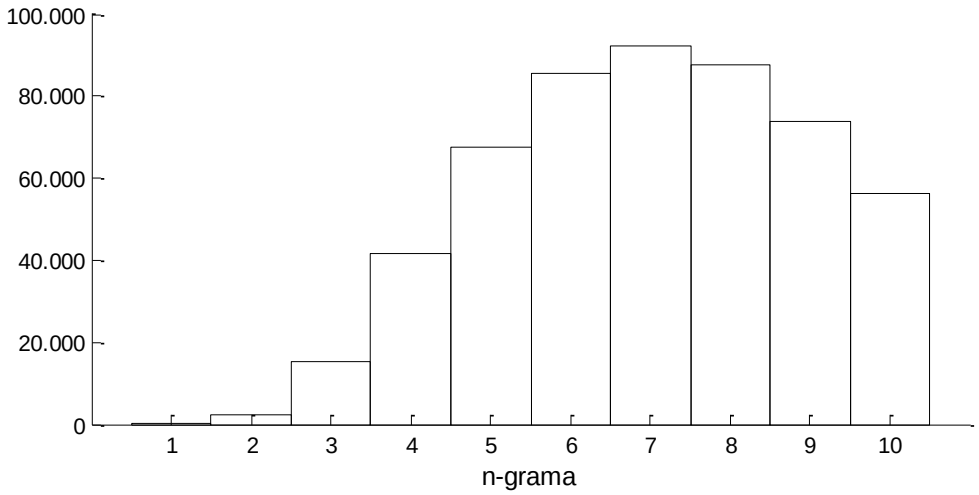

Figura 12 – Histograma de número de *n-grama* presentes no dicionário "dic_CETEMP_40k_alinhado_digrafos_tonica".

Foram feitas experiências com dicionários feitos a partir do vocabulário "*voc_CETEMP_40k_ao*", que contém palavras grafadas de acordo com o Acordo Ortográfico de 1990. A Figura 13 e a Figura 14 ilustram os melhores desempenhos obtidos com os vocabulários pré-AO90 e pós-AO90. O desempenho com o vocabulário pós-AO90 é ligeiramente inferior ao desempenho com o vocabulário pré-AO90. Isto verificou-se com todas as combinações de aplicação das regras fonológicas e para todos os valores de *n-grama* testados.

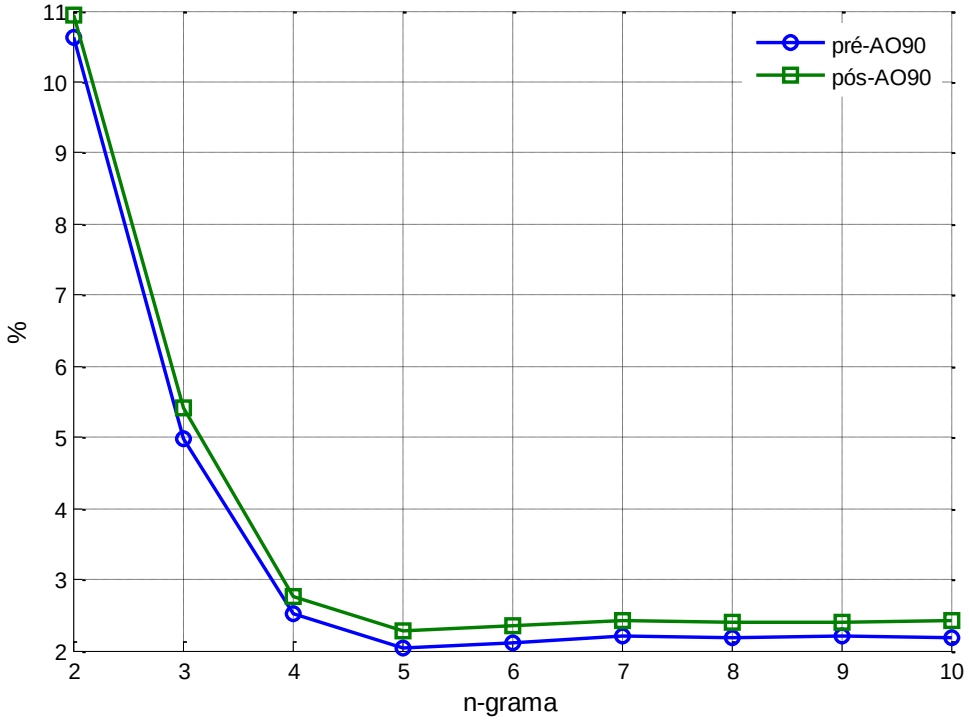

Figura 13 – WER dos vocabulários pré-AO90 e pós-AO90.

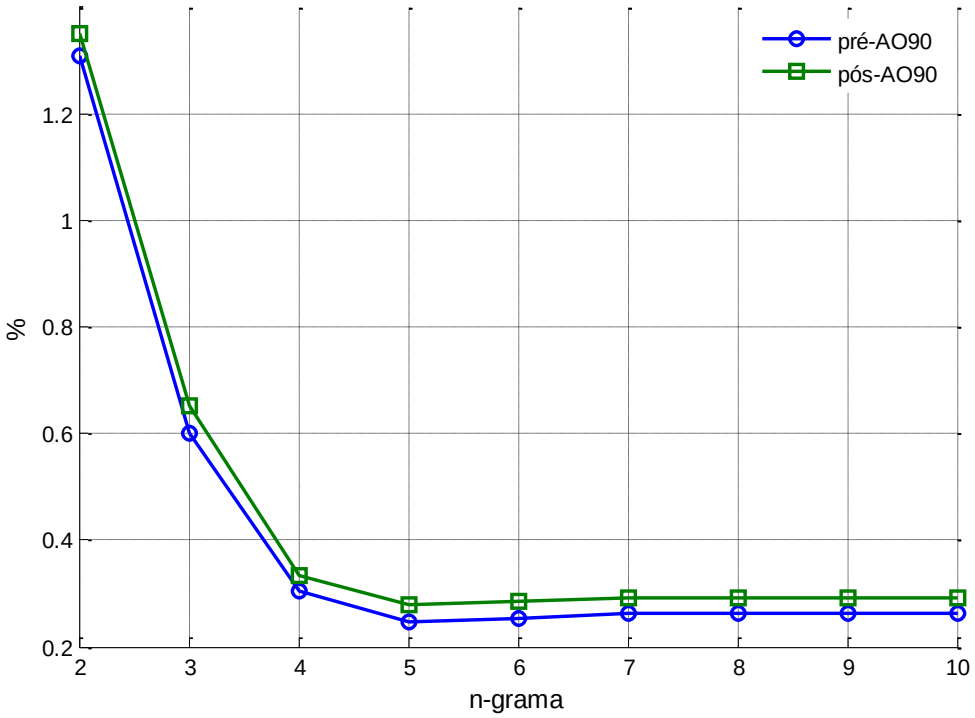

Figura 14 – PER dos vocabulários pré-AO90 e pós-AO90.

Numa análise aos erros por grafemas nos conjuntos dos melhores resultados (5-*grama*, marcação da vogal tónica e regras para contextos frequentes) obteve-se os números de erros apresentados na Tabela 10.

Tabela 10 – Número de erros com vocabulários pré-AO90 e pós-AO90 por grafemas.

Grafema	# erros pré-AO90	# erros pós-AO90
o	413	465
e	298	334
a	112	150
x	30	34
c	17	0
u	14	20
p	8	0
outros	13	21
Total	905	1024

Mais de 90% dos erros são provocados pela confusão gerada pelo grau de abertura dos fonemas que devem corresponder aos grafemas <o>, <e> e <a>, os quais podem ser pronunciados como /o/ ou /O/, /e/ ou /E/ e /6/ ou /a/. Nos modelos pós-AO90, não existem erros provocados pelas consoantes mudas <c> e <p>, tendo-se verificado um acréscimo expressivo de erros provocados pela confusão de abertura das vogais. Tal facto pode justificar o decréscimo do desempenho dos modelos pós-AO evidenciando o papel das consoantes mudas, indicativo, muitas das vezes, da abertura da vogal a elas antecedente e auxiliador, em muitos dos casos, da desambiguação entre a pronúncia aberta e a pronúncia fechada. A supressão da acentuação gráfica em alguns vocábulos, como é exemplo <boia>, (<bóia>, pré-AO), contribui também para aumentar a ambiguidade na aprendizagem e na determinação da pronunciação. A supressão do hífen usado no processo de prefixação veio igualmente dificultar a tarefa da determinação da vogal tónica.

Um outro resultado, obtido com unigramas (1-*grama*, ou sem informação da história), demonstra a regularidade fonética e fonológica do português europeu. Neste caso, a descodificação não utiliza a informação do contexto, e é escolhido sempre o melhor grafonema possível (com maior probabilidade) para cada grafema, resultando numa taxa PER de 24.82% sem as regras fonológicas e de 5.86% com as regras. No entanto, conversores com estas taxas de PER são impraticáveis para as tarefas de processamento automático de fala, uma vez que correspondem a uma taxa WER de 92.01% sem as regras fonológicas e a uma taxa WER de 38.86% com as regras.

Em termos de implementação e recursos computacionais, o sistema de conversão de grafemas para fonemas usa uma tabela de exceções que é carregada para uma tabela *hash* (*hastable*). A conversão com o modelo estatístico será invocada se um vocábulo não constar na tabela de exceção. Carregar uma tabela de exceção e o modelo estatístico requer mais recursos de memória, mas, por outro lado, aumenta a precisão e a rapidez da resposta. Um exemplo prático: usando uma tabela *hash* com 100000 entradas para um dicionário de 40000 vocábulos, é necessário cerca de 2MB de memória e apresenta cerca de 7500 colisões. Usando apenas a tabela *hash*, o ritmo de conversão é de cerca 1 milhão de vocábulos por segundo. Usando um conversor com apenas o modelo estatístico (2-*grama*), o ritmo de conversão passa para 20 mil vocábulos por segundo (num computador "*quad core*" a 2.8GHz, 4GB de memória).

Não é possível comparar, inequivocamente, o desempenho deste sistema com os desempenhos dos outros sistemas de conversão de grafemas para fonemas para o português europeu, uma vez que, até agora, não se encontram disponíveis publicamente e de acesso livre nenhum sistema de conversão ou a base de dados onde foram testados. No entanto, os resultados apresentados neste trabalho são os melhores reportados, até agora, em tarefas semelhantes. Por exemplo, (Braga et al., 2006) apresentam um resultado que é equivalente ao PER de 1.19%. Usaram o vocabulário de 1802 palavras (10884 grafemas) proveniente de 6 artigos do jornal Público (Público, 2013). O trabalho de (Caseiro et al., 2002) indica uma taxa WER de 3.94% e uma taxa PER de 0.59% conseguido com 7-*grama* e com informação de tónica. Este trabalho já tinha enfatizado a

importância da indicação da tónica no desempenho de um conversor e o resultado foi obtido num dicionário com mais de 200 mil palavras transcritas automaticamente. Em (Barros and Weiss, 2006) é apresentado um valor de 89% de palavras corretamente transcrita o que corresponde a um WER de 11%, com uma base de dados de treino com 7352 palavras e o teste foi feito com 550 palavras (3430 fonemas).

3.6. Conclusão

Este trabalho foi desenvolvido com o intuito de produzir um dicionário de pronunciação do português europeu. Experimentou-se uma abordagem que integra conhecimentos linguísticos com um modelo de base estatístico. Sequências de grafonemas foram modeladas através de um algoritmo de alinhamento entre grafemas e fonemas, nas quais foram também consideradas informações provenientes do contexto fonológico da língua portuguesa, tais como dígrafos, a acentuação tónica e a vizinhança fonético-fonológica. Todas estas informações foram testadas individualmente, tendo-se verificado que a inclusão de informação sobre a tonicidade da vogal foi decisiva para o aumento do desempenho do conversor. Contrariamente, a inclusão de informação sobre dígrafos não trouxe benefícios acentuados.

Os modelos de *n-gramas* foram treinados e testados usando a grafia pré-AO90 e pós-AO90, tendo-se verificado um ligeiro, mas consistente, decréscimo de desempenho dos modelos pós-AO90.

Decorrente da tarefa de conversão, foi gerado um dicionário de pronunciação com mais de 40 mil vocábulos oriundos do corpus CETEMPúblico, do qual derivaram outros dicionários, com informação de alinhamento, de acentuação e de dígrafos. Um dicionário de múltipla pronunciação com cerca de 500 homógrafos heterofónicas foi criado e tornado público junto com os restantes recursos produzido em (LABFALA, 2011).

O sistema de conversão de grafemas para fonemas desenvolvido pode ser facilmente adaptado para tarefas específicas adicionando módulos de pré-processamento como, por

exemplo, deteção de formas flexionadas e deteção de prefixos e sufixos. Será sempre necessário um dicionário de estrangeirismos para uma utilização mais abrangente deste conversor (também criado manualmente e disponibilizado junto com os outros recursos). Uma possível melhoria, mas que carece de um estudo cuidado, é a introdução de outras variantes do português e regionalismos bem como meta informação das palavras com as classes gramaticais.

Capítulo 4. Segmentação e diarização de locutor

A segmentação e a diarização de locutor são tarefas essenciais na geração de meta informação de sinais de áudio de noticiários. A segmentação identifica os limites dos segmentos homogéneos tornando possível a fragmentação de um longo ficheiro de áudio em segmentos mais pequenos de fácil tratamento. A diarização de locutor identifica os segmentos que foram proferidos pelo mesmo locutor (ou locutores com características muito semelhantes), possibilitando o agrupamento de segmentos por locutor.

4.1. Base de dados

Os sistemas de segmentação e de diarização do locutor foram desenvolvidos no âmbito de um desafio sobre este tema proposto na conferência FALA2010 (Butko et al., 2010; Zelenák et al., 2010). Para este desafio foi fornecida uma base de dados de áudio referente à estação de televisão catalã 3/24 TV. Além do catalão, a base de dados apresenta cerca de 17% de fala castelhana. Não tem uma distribuição equilibrada em termos de género, uma vez que apresenta cerca de 63% de fala masculina contra 37% de fala feminina. O sinal de áudio foi disponibilizado no formato PCM, mono, com 16 bits de resolução e com a frequência de amostragem de 16 kHz.

Para a segmentação foram disponibilizadas anotações correspondentes a aproximadamente 87 horas de áudio, distribuídos em 24 sessões, com 5 eventos acústicos:

- fala limpa (SP) –segmentos de fala proferidos em estúdio;
- música (MU) – segmentos com apenas música;
- fala com ruído de fundo (SN) – segmentos de fala gravados fora do estúdio ou fala com algum tipo de ruído sobreposto (por exemplo aplausos e ruído de tráfego) ou fala com tradução simultânea (muito frequente em noticiários);

- fala com música de fundo (SM) – segmentos onde há sobreposição da classe SP ou da classe SN com a classe MU;
- outros (OT) – segmentos de ruído ou segmentos com eventos acústicos que não correspondem a nenhuma das 4 classes anteriores.

Foram disponibilizadas 16 sessões para treino e 8 para teste. A distribuição das classes é apresentada na Tabela 11.

Tabela 11 – Distribuição das classes acústicas na base de dados.

Classe acústica	Percentagem na base de dados
SP	37 %
UM	5 %
SN	40 %
SM	15 %
OT	3 %

Para reduzir o tamanho das sessões de áudio, foi utilizado um detetor de silêncio simples, baseado nas energias das tramas de áudio a cada 100 milissegundos utilizando janelas de 200 milissegundos. A duração mínima de um segundo foi definida como um limiar de segmentação, o que corresponde a um bom compromisso para uma segmentação inicial dos ficheiros de áudio.

Para a tarefa de diarização, a mesma base de dados de áudio é usada, alterando apenas as etiquetas, que correspondem agora a uma enumeração de locutores.

Esta base de dados será referida como FALA2010, daqui para frente.

4.2. Deteção de segmentos repetidos

É frequente as emissões de rádio e televisão repetirem reportagens, *jingles* e *spots* publicitários durante uma sessão ou entre sessões. A deteção de segmentos repetidos

ajuda a tarefa de segmentação, uma vez que não é necessário segmentar mais do que uma vez os segmentos repetidos. As emissões normalmente apresentam *jingles* e genéricos que indicam o início ou o fim de um programa ou intervalos de publicidade que podem ser aproveitados para fazer uma segmentação das locuções em termos de programas contínuos.

A deteção de segmentos repetidos foi conseguida a partir da adaptação de uma técnica que inicialmente foi desenvolvida para detetar *jingles* nas emissões de rádio e televisão, denominada de impressão digital acústica.

A impressão digital acústica refere-se a uma forma condensada de representar um sinal de áudio por forma a identificar uma amostra desse sinal ou para localizar segmentos semelhantes num *stream* ou base de dados de áudio. Já existem várias aplicações que utilizam esta técnica, nomeadamente na monitorização de canais multimédia (Batlle et al., 2003), na identificação de músicas (pesquisa por amostra) (Wang, 2006) e na televisão interativa (Fink et al., 2006). Neste trabalho, esta técnica tem duas aplicações: a primeira consiste em, a partir de uma amostra de um *jingle* ou segmento áudio, identificar todos os segmentos iguais existentes na BD de áudio; a segunda aplicação consiste em identificar segmentos semelhantes, mas sem nenhuma amostra prévia de referência.

A impressão digital acústica é bastante precisa e rápida e exige apenas que os segmentos de áudio repetidos apresentem pouca distorção espetral e temporal. Deve reter as principais características percetuais e ser robusto a ruído. Para cumprir estes requisitos, foram propostas várias estratégias para criar a impressão digital acústica e foram usados vários métodos de pesquisa de ocorrência de uma impressão chave (assinatura) numa impressão de um sinal de áudio maior. Em (Pinquier and André-Obrecht, 2004) é utilizada a distância euclidiana para detetar a presença de uma assinatura. Em (Betser et al., 2007) é apresentada uma solução baseada na modelação sinusoidal em que as impressões digitais acústicas são caracterizadas por um conjunto pequeno de componentes espetrais fortes. Em (Johnson and Woodland, 2000) é usado um modelo estatístico baseado na parametrização cepstral e uma métrica baseada na covariância das impressões. Em (Ogle and Ellis, 2007) é apresentada uma solução bastante rápida que pode ser aplicada a sinais

de áudio longos. Utiliza os picos de energia no espaço tempo-frequência e tabelas "*hash*" para acelerar a pesquisa de assinaturas. Em (Haitsma and Kalker, 2002) é apresentada uma forma eficiente de representar impressões para indexação de excertos de músicas que pode ser implementada em dispositivos com poucos recursos computacionais. Em (Meinedo and Neto, 2004) é apresentada uma outra solução, baseada em redes neuronais para detetar padrões acústicos em noticiários.

Neste trabalho explorou-se a utilização de uma solução baseada em modelos HMM, tal como é proposto em (Johnson and Woodland, 2000; Batlle et al., 2003). No entanto, verificou-se que os parâmetros que normalmente são aplicados aos sistemas de reconhecimento de fala não são suficientemente robustos para esta tarefa. Optou-se então por implementar uma solução inspirada na proposta de (Haitsma and Kalker, 2002). Esta solução apresenta uma forma de criar as assinaturas bastante eficiente e um algoritmo de pesquisa que recorre a operações simples de comparação binária. As assinaturas são representações binárias dos padrões espetrais. A proposta de (Haitsma and Kalker, 2002) foi testada com impressões de sinais distorcidos e verificou-se que esta solução é pouco robusta à compressão ou expansão temporal mas que é muito robusta à corrupção com ruído. Neste trabalho foram feitos vários testes para determinar a influência de análise espetral, nomeadamente o tamanho das janelas e o ritmo de análise espetral, a definição da máscara e o tamanho do padrão de bit no desempenho do algoritmo.

4.2.1. Criação de padrões de impressão digital acústica

O processo de criação de assinatura digital acústica inicia-se com a conversão da frequência de amostragem do sinal para 8 kHz seguida de análise espetral. Foram testadas janelas retangulares que variam de 80 a 240 ms e com uma taxa de tramas (*frame-rate*) de 25 a 100 janelas por segundo. É usado um banco com 33 filtros na escala de Mel (tal como os usados na parametrização MFCC). No final é feita a convolução 2D do espetrograma com uma máscara (Figura 15), sendo o resultado binarizado (utilizando

zero como limiar). Foram testadas 4 máscaras ilustradas da Figura 15, (Neves et al., 2009) onde os pesos a negrito indicam a origem (coordenada 0,0) na convolução 2D.

Máscara 1

+1	-1
-1	**+1**

Máscara 2

-1	-1
+1	**+1**

Máscara 3

-1	-1
+2	**+2**
-1	-1

Máscara 4

-1	-1	-1
+2	+2	**+2**
-1	-1	-1

Figura 15 – Máscaras para criar padrões binários.

A primeira máscara (identificada como máscara 1 na Figura 15) foi proposta em (Haitsma and Kalker, 2002). Além desta máscara, foram testadas mais 3 máscaras em que os pesos foram definidos por forma a reter informações específicas do espetro do sinal. Assim, a máscara 2 é usada para detetar declives negativos no espetrograma em duas tramas consecutivas, a máscara 3 é usada para identificar picos no sonograma que podem evidenciar presença de componentes do tom e a máscara 4 que é idêntica a máscara 3 mas requer a presença de componentes do tom mais longo. A Figura 16 mostra os sonogramas binarizados com as máscaras 1, 2 e 3 e pode-se verificar que as riscas horizontais presentes no sonograma original são mantidas nos sonogramas binarizados com as máscaras 2 e 3, o que não acontece com o sonograma binarizado com a máscara 1.

Em cada trama de áudio, aplica-se uma máscara às 33 energias à saída do banco de filtros e obtém-se um padrão de 32 bits que é convenientemente representado por uma palavra de 32 bits. Os padrões são concatenados para formar a impressão digital acústica de um segmento.

No final, obtemos um número de 32 bits com um dado *frame-rate*, que representa a base de impressão digital acústica do sinal de áudio, com o objetivo de procurar ou detetar nessa base, impressões digitais ou assinaturas de objetos de áudio bem definidos. Nesta descrição usamos os termos “base de impressão digital acústica” para a sequência de números de 32 bits que representam os padrões binários de uma peça de áudio completa. À sequência de padrões de 32 bits relativa a um (pequeno) objeto de áudio chamamos “impressão digital” ou “assinatura”.

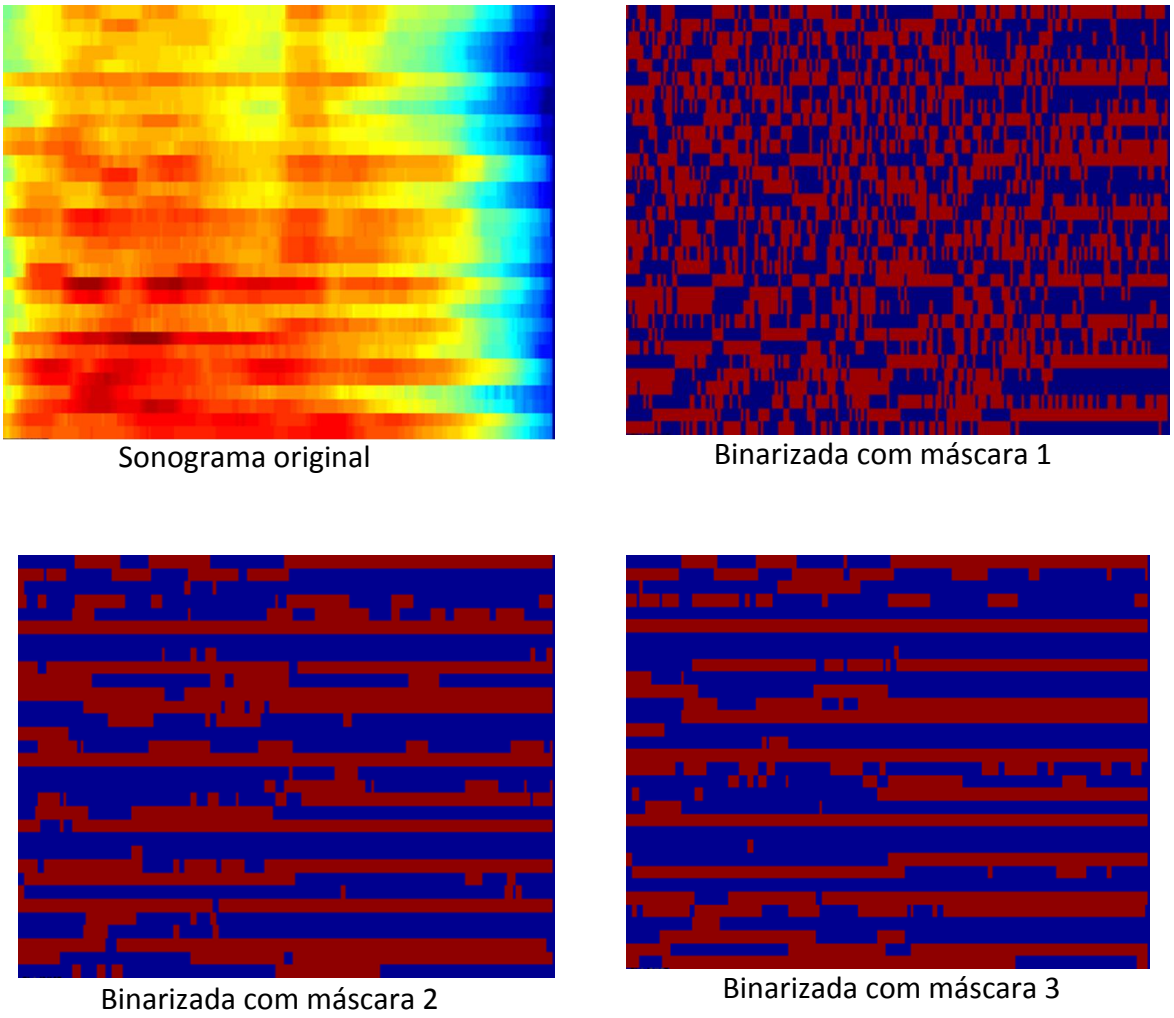

Figura 16 – Padrões binários gerados com as máscaras 1, 2 e 3.

4.2.2. Pesquisa de assinaturas

O processo de pesquisa de uma assinatura numa base de impressão digital acústica baseia-se na contagem dos bits que o padrão de assinatura tem em comum com a impressão digital acústica do sinal de áudio onde é procurada a assinatura. O processo é repetido a cada trama de áudio e, utilizando um limiar, é possível identificar os instantes onde ocorre a assinatura no sinal de áudio. Neste trabalho optou-se por fazer um processo inverso, ou seja, contar o número de bits diferentes em vez da correspondência

de bits. É aplicada a operação "ou exclusivo" (XOR) entre cada padrão de assinatura e cada padrão da base de impressão digital acústica onde é procurada a assinatura. Trata-se de uma convolução entre a base e a assinatura, o que resulta na contagem de bits a 1 e que indica o número de bits diferentes ou errados, num dado instante. A razão entre o número de bits errados (não coincidentes) e o número de bits total da assinatura define uma taxa de erro de bit (BER – *Bit Error Rate*) que é usada para detetar a ocorrência de uma assinatura, independente do seu tamanho. O valor médio do BER é de 0.5 em segmentos onde a assinatura não ocorre (bit 1 com a mesma probabilidade que bit 0). Fixando um limiar menor que 0.5 para o BER, as tramas que apresentarem BER abaixo deste limiar são indicadas como tramas onde ocorre a assinatura. Na realidade, é preciso usar uma pequena histerese porque o BER pode apresentar flutuações no seu decréscimo. Figura 17 apresenta um exemplo com valores do BER calculados durante uma pesquisa de um *jingle* de poucos segundos numa peça de noticiário de mais de uma hora.

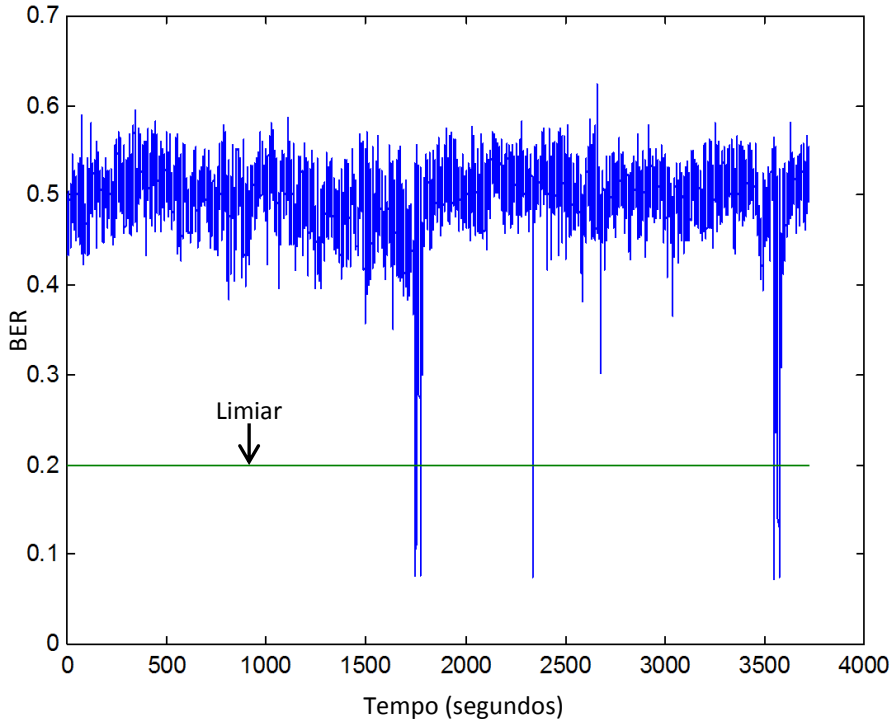

Figura 17 – Exemplo do BER de uma assinatura.

O processo de pesquisa é bastante rápido. Num computador comum (Pentium 4 a 3GHz e 2 GB de memória) pode-se pesquisar uma assinatura de 10 segundos num sinal de áudio de 1 hora em menos de meio segundo com uma análise espetral a um ritmo de 50 tramas por segundos.

O maior desafio do processo de pesquisa é a implementação de um algoritmo eficiente de contagem de bits. Usando uma *lookup table* pode-se implementar contagens de bits bastante eficientes e mais rápidas quando comparadas com as soluções que usam deslocamento de bits. Os processadores mais recentes têm instruções específicas para contagem de bits (instrução POPCNT, "*population count*"), aumentando ainda mais a velocidade do processo de pesquisa.

4.2.3. Resultados da pesquisa de assinaturas

O algoritmo de assinatura digital acústica foi testado numa outra base de dados, constituída por sessões de áudio de 5 canais de televisão portuguesa, nomeadamente RTP1, RTP2, SIC, SIC Notícias e TVI. Foram gravadas 25 sessões, cada uma com uma hora de duração. Foram marcadas manualmente as ocorrências de 9 *jingles* e 7 *spots* publicitários perfazendo um total de 16 assinaturas diferentes. Existem no total 103 ocorrências de assinaturas com a distribuição por canal indicada na Tabela 12 (Neves et al., 2009).

O prefixo "jng" no nome das assinaturas identifica os *jingles* e o prefixo "pub" identifica os *spots* publicitários. Muitas assinaturas contêm música, mas existem algumas que têm segmentos apenas com fala. Foram identificadas 4 situações em que a ocorrência de assinaturas tinha algum problema (*jingle* cortado, *jingle* sobreposto com fala ou *jingle* com música em *fade-out*).

Tabela 12 – Assinaturas para testes.

Nome de assinatura	Canal	Tamanho (s)	# Ocorrências
jng1RTP1	RTP1	4	23
jng2RTP1	RTP1	3	11
pubcutRTP1	RTP1	6	3
pubZonTVCaRTP1	RTP1	30	4
pubModBombRTP1	RTP1	30	4
jng1RTP2	RTP2	4	4
jng2RTP2	RTP2	2	2
jng3RTP2	RTP2	6	2
jng1SIC	SIC	4	14
pubcutSIC	SIC	3	2
pubVodADSLSIC	SIC Notícias	25	6
jng1SNOT	SIC Notícias	4	13
pubcutSNOT	SIC Notícias	1	2
jng1TVI	TVI	2	6
jng2TVI	TVI	3	2
pubcutTVI	TVI	3	5

Foram testadas as quatro máscaras, a análise espetral com vários ritmos e janelas com diferentes tamanhos. A Tabela 13 resume os resultados obtidos com diferentes combinações desses parâmetros. São apresentados os números de assinaturas corretamente detetados. Estes resultados foram obtidos usado um limiar do BER igual a 0.23 para as máscaras 2 a 4 mas de 0.35 para a máscara 1, conforme é sugerido em (Haitsma and Kalker, 2002). De facto, o padrão aleatório que o sonograma binarizado com a máscara 1 produz (ver exemplo na Figura 16), resulta num BER com menor variância em torno de 0.5, o que justifica a utilização de um limiar maior. Testes com máscara 1 e com o limiar de 0.23 mostram uma pequena degradação do desempenho face à utilização das outras máscaras.

Não foi detetado nenhum dos 4 casos em que a ocorrência de assinaturas apresentava problema (o número máximo de segmentos detetados foi 99 e havia 103 segmentos anotados). Os testes que foram efetuados mostram que, com os limiares corretamente estabelecidos, é pouco provável que aconteça um falso alarme, ou seja, o BER descer muito abaixo de 0.5 em tramas onde não ocorre a assinatura. Com o limiar de 0.23 apenas a máscara 4 produziu sempre um falso alarme nos testes com janela de 240 milissegundos (assinalado com '*' na Tabela 13). Mesmo nestes casos, seria possível evitar falsos alarmes, usando um limiar personalizado para cada máscara.

Tabela 13 – Número de assinaturas corretamente identificadas.

Tamanho janela (ms)	Ritmo de análise (Hz)	Máscara 1	Máscara 2	Máscara 3	Máscara 4
240	25	98	98	99	99*
240	50	95	**99**	99	99*
240	100	91	99	99	99*
120	25	95	98	99	99
120	50	95	98	99	99
120	100	86	98	99	99
80	25	95	98	99	99
80	50	93	97	99	99
80	100	81	98	99	99

O desempenho das máscaras é afetado pelos parâmetros da análise espetral mas as máscaras propostas (máscaras 2-4) mostram pouca sensibilidade à variação dos parâmetros da análise espetral.

4.2.4. Aplicação de assinatura digital acústica na segmentação

Após estes testes, foram fixados os parâmetros para a criação da assinatura digital acústica que foi utilizada nos trabalhos de segmentação descritos nesta tese (a negrito na

Tabela 13). Ou seja, optou-se pela utilização da máscara 2, análise espetral a um ritmo de 50 tramas por segundo, janela de 240 milissegundos (1920 amostras à frequência de amostragem de 8000 Hz com FFT (*Fast Fourier Transform*) de 2048 pontos).

Dada a base de dados de treino etiquetada em termos dos 5 eventos acústicos ou em termos do locutor, é criada uma assinatura para cada segmento. Depois é feita a procura das assinaturas na base de dados de teste. Nos testes com a base de dados descrita acima em 4.1 foram encontrados muitos segmentos repetidos, o que justifica a opção tomada de detetar repetições. Existiam algumas inconsistências na etiquetagem dos segmentos, o que se pode justificar pelo facto de este processo ter sido manual. Sempre que foi encontrada uma assinatura repetida na base de dados de teste correspondente a várias etiquetas na base de dados de treino, ou seja, segmentos iguais no treino mas com etiquetas diferentes, procedeu-se a um sistema de votação para atribuir uma etiqueta ao segmento de teste.

O resultado de classificação a partir da procura de assinaturas foi combinado com o resultado do detetor de silêncio, antes de se proceder a deteções de eventos acústicos nos restantes segmentos.

4.3. Deteção de classes acústicas

A deteção de classes acústicas é baseada num descodificador híbrido entre redes neuronais MLP (*Multi-Layer Perceptron*) e HMM como é proposto em (Lopes, 2011). A rede MLP é constituída por uma camada de entrada, uma camada escondida e uma camada de saída associada às classes definida na base de dados FALA2010. A camada escondida tem 200 nodos e a camada de saída tem 5 nodos, um para cada evento acústico considerado (SP, MU, SN, SM e OT). Usaram-se como parâmetros de entrada da rede 48 parâmetros descritos em 4.3.1 em baixo, com contexto de 10 tramas à direita e 10 tramas à esquerda. Foi usado o treino com retropropagação resiliente do erro proposto em (Riedmiller and Braun, 1993) para treinar os parâmetros da rede MLP. Na camada de saída foi utilizada a função “softmax” como função de ativação. Assim as

saídas podem ser interpretadas como probabilidade *a posteriori* dos eventos, que são associadas aos estados dos modelos HMM das classes. Os modelos HMM das classes foram treinados com as ferramentas do HTK (Young et al., 2006) e têm a topologia esquerda-direita com 10 estados. Os 10 estados de um HMM são todos iguais (partilham a mesma saída de rede MLP) e foram definidos assim para garantir uma duração mínima apropriada aos eventos acústicos reconhecidos.

4.3.1. Parâmetros

Para treinar a rede neuronal foram usados dois conjuntos de parâmetros: os tradicionais coeficientes MFCC e um conjunto de 8 medidas que são utilizadas em várias tarefas de processamento da fala: energia das tramas de áudio em decibel, taxa de cruzamentos por zero (ZCR - *Zero Crossing Rate*), centroide espetral, *roll-off* espetral (a 90%), correlação máxima normalizada dos MFCC, frequência correspondente a correlação máxima normalizada, harmonicidade (proporcional a duração dos harmónicos) e fluxo espetral. O vetor de parâmetros tem 48 parâmetros contendo 16 coeficientes cepstrais, incluindo o c_0, as 8 medidas adicionais e a primeira derivada destes 24 valores (deltas).

Os parâmetros são calculados a um ritmo de 10 tramas por segundo com janelas de *Hamming* de 200 milissegundos.

4.3.2. Resultados de segmentação

O desempenho de segmentação foi medido usando a métrica proposta na FALA2010 (Butko et al., 2010). É calculado o desempenho individual de todas as classes consideradas e o desempenho final é a média dos erros das classes. A média foi utilizada porque a distribuição das classes é desequilibrada na base de dados e, assim, todos as classes têm o mesmo peso no desempenho final. O desempenho da classe OT não foi considerado na avaliação final. O erro é contabilizado por trama (a um ritmo de 50 tramas por segundos). Uma trama é considerada errada se a sua classificação não corresponder a

etiqueta de referência. É dada uma tolerância de 1 segundo à volta dos limites dos segmentos (colar) dentro da qual não são consideradas tramas erradas (tolerância admitindo alguma subjetividade na marcação manual dos limites de segmentos). Para uma classe, podem acontecer dois tipos de erros: o apagamento ou a inserção desta classe. Um erro de substituição (segmento classificado com classe errada) contribui para penalizar duas classes, uma com erro de apagamento (a classe correta indicada pela anotação de referência) e outra com erro de inserção (a classe indicada pelo classificador).

Considerando $Dur(.)$ uma função de duração de um evento, a métrica é definida da seguinte forma:

$$Erro = \frac{1}{4}\sum_{i=1}^{4}\frac{Dur\left(Apagamento_i\right)+Dur\left(Inserção_i\right)}{Dur\left(Classe_i\right)}, \quad \textbf{(32)}$$

onde o índice *i* se refere às classes {SP, MU, SN e SM}.

A Tabela 14 e a Tabela 15 mostram os desempenhos por classe, obtidos nas bases de dados de treino e de teste respetivamente.

Tabela 14 – Desempenho na base de dados de treino.

Classe acústica	Erro (%)
SP	22.787
UM	13.62
SN	26.610
SM	20.10

Tabela 15 – Desempenho na base de dados de teste.

Classe acústica	Erro (%)
SP	48.03
UM	21.43
SN	48.49
SM	51.66

O desempenho final na base de dados de treino foi de 20.68% e na base de dados de teste passou para 42.40%. A duplicação do erro na base de dados de teste, quando comparada com o erro na base de dados de treino, pode ser explicada pela diferença que foi verificada entre os sinais de áudio do treino e os sinais de áudio de teste. A estratégia implementada baseia-se muito na deteção de repetições usando as impressões digitais acústicas. Nos sinais de áudio de treino foram encontrados 4427 segmentos que aparecem pelo menos duas vezes e, no total, as repetições correspondem a 65% do tempo total dos áudios de treino. No entanto, os áudios de teste apenas 12% do tempo correspondem a repetição de segmentos.

4.4. Diarização de locutor

A tarefa de diarização de locutor é a de segmentar um sinal de áudio em termos dos locutores presentes, isto é, identificando os segmentos de fala que foram produzidos pelo mesmo locutor. Nenhuma informação sobre os locutores é conhecida de antemão.

Tal como aconteceu com a segmentação de áudio em classes acústicas, apresentada anteriormente, o sistema de diarização aqui proposto foi desenvolvido no âmbito do desafio proposto na conferência FALA2010 (Butko et al., 2010; Zelenák et al., 2010). O sistema de diarização de locutor partilha muitos módulos com o sistema de segmentação já descrito. A própria base de dados de áudio é igual, alterando apenas as etiquetas (na diarização as etiquetas são enumerações de locutores em vez de eventos acústicos). O sistema de segmentação foi incorporado no sistema de diarização para identificar segmentos que podem ser descartados como música ou ruído. A diarização é feita apenas nos segmentos identificados como sendo de fala.

Além dos módulos já descritos na segmentação em classes acústicas, foram implementados mais dois módulos, ambos para fazer agrupamento (*clustering*) de segmentos. Após o módulo de segmentação, é aplicado um algoritmo de *clustering* aos segmentos de fala (SP, SN e SM), onde cada segmento é tomado como um locutor diferente. Embora possa acontecer que um segmento tenha presente mais que um

locutor, a probabilidade de isso acontecer é baixa porque qualquer evento de silêncio (OT) ou mesmo fala sobreposta ou com ruído (SN), obrigaria a dividir esse segmento de fala.

O algoritmo de *clustering* é baseado em misturas das componentes gaussianas (GMM), seguindo uma abordagem análoga à usada em reconhecimento de orador e é discutido na secção seguinte. Após este processo os segmentos de fala ficam etiquetados segundo um conjunto de locutores. Segue-se um processo de verificação em que se testa se existem segmentos repetidos mas com identificados com etiquetas diferentes. Nestes casos, utiliza-se a etiqueta mais frequente encontrada neste conjunto ou, em caso de empate, a mais utilizada. Este processo baseia-se no facto do sistema de deteção de repetições apresentar taxa de falso alarme praticamente nula.

No final deste processo de agrupamento e com o objetivo de reduzir o número de locutores, foi aplicado um outro algoritmo de agrupamento. Este é baseado num algoritmo utilizado sobretudo na deteção de mudanças de locutores, BIC (*Bayesian Information Criterion*) (Delacourt and Wellekens, 2000). Neste processo tomam-se todos os pares de segmentos para verificar se o locutor envolvido é ou não o mesmo. Se a decisão for positiva (mesmo locutor), então os dois segmentos são etiquetados com o mesmo índice, correspondente ao locutor do primeiro segmento do par. Um teste final verifica se existem segmentos sucessivos com a mesma etiqueta, sendo fundidos nesse caso. A Figura 18 sistematiza os módulos que compõem o sistema de diarização de locutor.

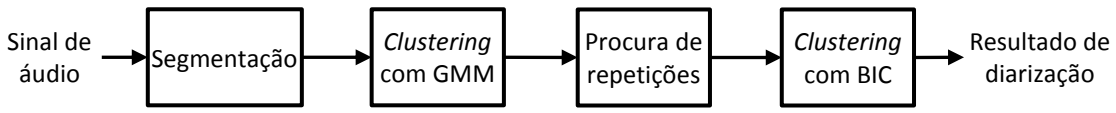

Figura 18 – Diagrama do sistema de diarização de locutor.

4.4.1. *Clustering* com GMM

A abordagem tradicional dos sistemas de identificação e verificação de locutor utiliza um modelo com muitas componentes gaussianas (modelo global) para representar todas as distribuições dos locutores e utiliza técnicas de adaptação para criar modelos de cada locutor. Este modelo global é denominado de modelo de fundo universal (UBM – *Universal Background Model*) (Reynolds and Rose, 1995) e é treinado com amostras de toda a base de dados contendo todos os locutores. Dado que no presente caso não existe informação sobre os locutores, procurou-se utilizar o UBM para fazer segmentação não supervisionada de locutor, considerando que cada componente gaussiana pode modelar características de um conjunto restrito de locutores. Assim, durante a descodificação são interpretadas as transições entre misturas como sendo mudanças de características que indiciam mudança de locutor. Foi utilizada uma penalização de transição elevada entre misturas por forma a garantir que um segmento homogéneo proferido por um locutor ocupa a mesma mistura da gaussiana durante o tempo do segmento. Assim, dados dois segmentos pode-se determinar se foram proferidos pelo mesmo locutor, ao descodificar um segmento constituído pela concatenação dos dois segmentos. A transição entre misturas não ocorre tipicamente na marca de transição entre locutor como é ilustrada na Figura 19. O método de *clustering* não altera as marcas previamente indicadas, apenas confirma ou não se os seguimentos pertencem ao mesmo locutor.

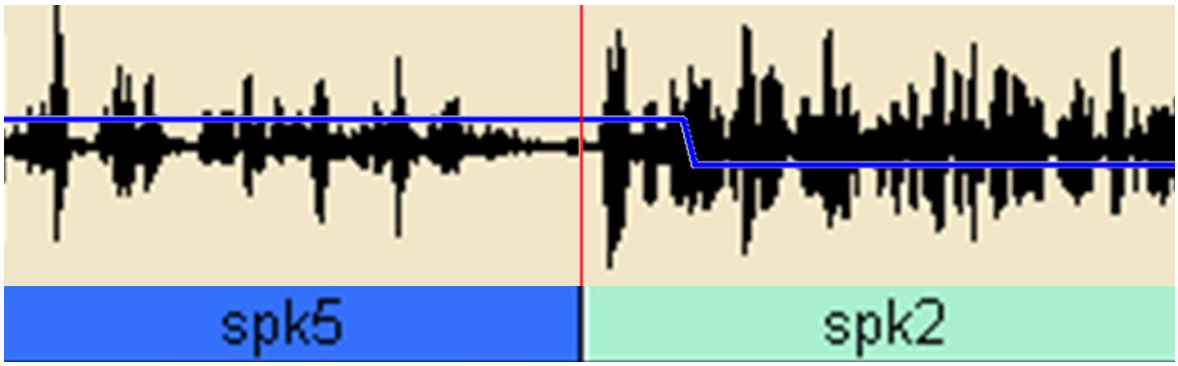

Figura 19 – *Clustering* com GMM - Exemplo onde a transição entre duas componentes de mistura (linha a azul) não ocorre na marca de separação de segmentos.

Neste trabalho foi criado um UBM por sessão, cada um com uma mistura de 256 componentes gaussianas. O número de misturas foi escolhido baseado no facto da

descrição da base de dados indicar que cada sessão pode ter entre 30 a 250 locutores diferentes. Os parâmetros para treinar os UBM são idênticos aos descritos em 4.3.1. O descodificador foi implementado em Matlab, onde as misturas são associadas a estados em paralelo de um HMM.

Como foi referido, após o *clustering* com GMM é utilizada a impressão digital acústica para procurar segmentos repetidos e verificar a consistência da classificação.

4.4.2. *Clustering* com BIC

Um outro método de *clustering*, baseado em BIC, foi implementado a fim de tentar reduzir o número de locutores diferentes presentes numa sessão de áudio. A vantagem do método BIC é que apenas é necessário estimar um parâmetro.

Considerando dois segmentos de fala, X_1 e X_2, ambos representados por uma única gaussiana multivariável, $X_1 \sim \mathcal{N}(x;\mu_{X1};\Sigma_{X1})$ e $X_2 \sim \mathcal{N}(x;\mu_{X2};\Sigma_{X2})$, a concatenação dos dois segmentos, $X=\{X_1,X_2\}$, continua a ser representado por uma mesma gaussiana se X_1 e X_2 forem segmentos similares. Assumimos neste caso que o segmento homogéneo X diz respeito a uma locução proferida pelo mesmo locutor. A razão de máxima verosimilhança logarítmica entre a hipótese dos segmentos serem similares (mesmo locutor) e a hipótese de serem diferentes (mudança de locutor) é dada pela seguinte expressão:

$$R=\frac{N_X}{2}\log\left(|\Sigma_X|\right)-\frac{N_{X1}}{2}\log\left(|\Sigma_{X1}|\right)-\frac{N_{X2}}{2}\log\left(|\Sigma_{X2}|\right), \quad \textbf{(33)}$$

onde N_X, N_{X1} e N_{X2} são os comprimentos dos segmentos X, X_1 e X_2 respetivamente. A variação do critério BIC (Delacourt and Wellekens, 2000) ou delta BIC (ΔBIC) é dada pela seguinte expressão:

$$\Delta BIC=-R+\lambda\cdot P, \quad \textbf{(34)}$$

onde λ é o fator de penalização (único parâmetro a estimar, com valor ideal 1) e *P* é a penalização da complexidade dos modelos dado em função da dimensão do vetor de parâmetros utilizado (*p*).

$$P = \frac{1}{2}\left(p + \frac{1}{2} p\left(p+1 \right) \right). \quad \textbf{(35)}$$

Um valor de ΔBIC negativo é indicativo de que os dois segmentos foram proferidos por locutores diferentes. Se for positivo o sistema identifica os dois segmentos com a mesma etiqueta.

Para fazer *clustering* com BIC foram utilizados os 16 MFCC indicados em 4.3.1, descartando os parâmetros adicionais. Daí, o valor de *p* ser 16, o que resulta numa penalização de 76. Determinou-se empiricamente o fator de penalização e chegou a um valor de 0.6 após testes com algumas sessões de áudio.

4.4.3. Resultados de diarização

O desempenho dos sistemas de diarização utiliza, por norma, uma medida definida pelo NIST: a taxa de erro de diarização (DER – *Diarization Error Rate*) (NIST, 2013). A taxa DER é a soma das taxas de erro de apagamento, de erro de inserção e de erro de substituição. É definida da seguinte forma:

$$DER = \frac{\sum_{\substack{todos \\ segs}} \left\{ Dur(seg) \cdot \left(\max\left\{ N_{ref}(seg), N_{rec}(seg) \right\} - N_{correcto}(seg) \right) \right\}}{\sum_{\substack{todos \\ segs}} \left\{ Dur(seg) \cdot N_{ref}(seg) \right\}}, \quad \textbf{(36)}$$

onde:

- *Dur(seg)* é a duração de um segmento;
- $N_{ref}(seg)$ é o número de etiquetas de referência num segmento;
- $N_{rec}(seg)$ é o número de etiquetas detetadas num segmento;
- $N_{correcto}(seg)$ é o número de etiquetas corretamente detetadas num segmento.

Nas sessões de teste, a taxa DER foi de 55.8%. Os erros de substituição (52.4%) é que mais pesaram para este valor de DER. Os erros de inserção (2.3%) e os erros de apagamento (1.1%) foram um dos mais baixos do desafio de FALA2010 e a segmentação fala/não fala produziu o menor erro de entre os sistemas presentes no desafio.

Uma possível melhoria que podia incrementar o desempenho do sistema seria a utilização de supervetores de médias de GMM/UBM na classificação de segmentos ou melhorar o algoritmo de *clustering* com GMM.

Capítulo 5. Deteção de estilos de fala

5.1. Introdução

A deteção de estilos de fala proporciona a segmentação de dados multimédia em partes consistentes. Neste trabalho é utilizada para identificar segmentos de fala que são apropriados para treinar modelos acústicos, descartando os segmentos com estilos de fala que poderiam causar algum problema no processo de treino.

A definição de estilo de um segmento de fala é muito abrangente e varia com os objetivos dos estudos e a área de aplicação em que é utilizada. Existem definições que são comuns na área de processamento de fala. Numa pesquisa de literatura sobre este matéria verificou-se a existência de estudos, como o de (Eskenazi, 1993), onde o autor define diferentes eixos de análise das características da fala que melhor capturam a natureza do estilo de fala. No entanto, a definição de estilo de fala mantém-se ainda imprecisa. Prova disso é o uso de variadas expressões sinonímicas sob as quais os estilos têm vindo a ser classificados. Termos como "fala lenta", "fala rápida", "fala monocórdica", "fala espontânea", "fala limpa", "fala planeada", "fala informal" e "fala formal", entre outros, têm sido usados e definidos em perspetivas tão diversas quanto o número de autores que a eles fazem referência (Llisterri, 1992). De igual forma verifica-se que não existe uma característica específica que defina um estilo de fala, por isso várias características têm sido apontadas como relevantes à caracterização de alterações de estilos de fala. As características acústicas foram usadas por (Nakamura et al., 2008) em ambiente de reconhecimento automático de fala para diferenciar "fala espontânea" de "fala lida". No entanto, estudos como (Deshmukh et al., 2009) destacam a importância de utilizar as características prosódicas em conjunto com as características fonéticas para melhorar a compreensão da estrutura da fala, bem como para aumentar a precisão de classificação de segmentos de fala em termos de estilo (Biadsy and Hirschberg, 2009; Sanchez et al., 2011). No âmbito da língua portuguesa, na sua vertente europeia, existem vários trabalhos que tentam demonstrar evidências da presença de eventos de hesitação (Moniz et al., 2009; Veiga et al., 2011c) ou do grau de articulação das formas de superfície (Barros

et al., 2001; Candeias et al., 2011) na fala em contínuo. Um outro estudo, (Barbosa et al., 2009), compara o português europeu com o português do Brasil utilizando o ritmo de fala.

Neste trabalho pretende-se distinguir os dois estilos de fala mais evidentes nos sinais de áudio de noticiários: "fala espontânea" ou "fala não preparada" e "fala lida" ou "fala preparada". Para isso, explorou-se a combinação de parâmetros prosódicos e parâmetros fonéticos. A distinção entre informação fonética e informação prosódica presente numa língua advém da perspetiva metodológica adotada na análise prática/teórica dessa língua. A prosódia é a ciência que estuda a natureza e funcionamento das variações de tom, intensidade e duração na cadeia falada (Nespor and Vogel, 1986) e a fonética, como é definida por (Fry, 1979), é a ciência que se ocupa das propriedades físicas dos sons de fala, do seu funcionamento enquanto gerador de sons, e das principais correspondências entre traços acústicos e elementos dos sistemas fonológicos das línguas.

Foi ainda objeto de estudo a caracterização e deteção de hesitação em sinais de áudio de noticiários. Este estudo pode ser integrado na deteção de estilos de fala, dado que existe forte correlação entre número de ocorrência de eventos de hesitações e certos estilos fala como é o caso de "fala espontânea" ou "fala não preparada".

5.2. Caracterização da base de dados

Para desenvolver e testar um sistema de deteção de estilos de fala foi necessário criar uma base de dados de noticiários. Para isso, recorreu-se aos *podcasts* de noticiários disponibilizados pelas estações de televisão. A Figura 20 ilustra o processo de criação da base de dados de noticiários.

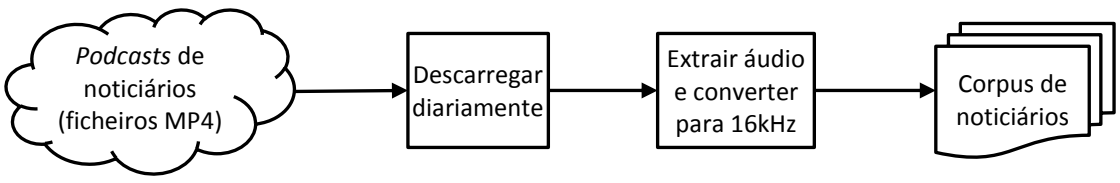

Figura 20 – Criação da base de dados de noticiários.

Como é ilustrada na Figura 20, os sinais de áudio foram extraídos dos respetivos sinais multimédia e convertidos de uma frequência de amostragem de 44.1 kHz para 16 kHz. Foram selecionados os sinais de áudio do mês de setembro de 2011 do programa diário "Telejornal" da empresa RTP (Rádio Televisão de Portugal). Estes áudios correspondem a cerca de 27 horas distribuídas por 30 "Telejornais", cada um dividido em duas partes (existem, no entanto, dois "Telejornais" que não foram divididos) perfazendo um total de 58 ficheiros de áudio.

Os sinais de áudio contêm diversos ambientes e condições acústicas, incluindo fala em estúdio e intervenção de oradores via telefone, fala de *pivôs*, oradores profissionais, comentadores, segmentos com reportagens e peças de entrevistas. Os segmentos de fala dita "preparada" (lida) são predominantes mas muitas vezes a fala está sobreposta com outros eventos acústicos (ruído de fundo) como fala, ruído ou música. Existem também segmentos de áudio que não contêm fala onde é possível encontrar música, *jingles*, risos, tosse ou palmas.

A anotação manual da base de dados em termos de estilos de fala foi auxiliada por um passo inicial de segmentação automática usando o ΔBIC (Delacourt and Wellekens, 2000), da mesma forma que é apresentado na secção 5.3, para indicar possíveis marcas onde ocorrem mudanças de locutores entre dois segmentos consecutivos. Usando a ferramenta Transcriber (Barras et al., 2001), toda a base de dados foi anotada em 4 níveis, cada um com um tipo de informação sobre os segmentos.

No primeiro nível, denominado de nível do sinal, é indicado tipo do sinal de áudio (fala, silêncio, música, *jingle*, tosse, suspiro, ruído, ...). Neste trabalho, este nível foi dividido em duas classes: "fala" referente aos segmentos originalmente marcados como "fala", e "não-fala" referente aos segmentos que originalmente não foram marcados como "fala". Este nível é preenchido em todos os segmentos, ao contrário dos outros níveis que apenas são utilizados para segmentos classificados como "fala".

No segundo nível é indicado se os segmentos de fala são "fala limpa", "fala telefónica" ou fala com algum ruído de fundo (música, fala, ruído de multidão, ...).

O estilo de fala é indicado no terceiro nível, distinguindo a fala preparada, a fala Lombard[4] e a fala espontânea. É indicado ainda o nível de espontaneidade para segmentos da fala espontânea que depende do número de disfluências (hesitações) que estão presentes nos segmentos. Foram anotados três níveis de espontaneidade: baixo, médio e alto. A determinação do nível de espontaneidade da fala é subjetiva e verificou-se que os segmentos classificados com o nível baixo de espontaneidade eram inconsistentes. Assim, neste trabalho, apenas foram considerados segmentos classificados com nível de espontaneidade médio ou alto. A fala Lombard também não foi considerada dado que o número de segmentos com este estilo de fala é muito reduzido (0.7%).

O quarto nível tem informações sobre o locutor e o detalhe de informação pode chegar até a identificação do mesmo quando isso é possível. Neste nível é também indicada a presença de outras línguas (que não sejam o português) no segmento. As informações apresentadas neste nível não foram tidas em conta na identificação de estilos de fala.

A Tabela 16 resume as contagens e as durações médias dos segmentos que foram considerados neste estudo.

Tabela 16 – Estatística dos segmentos na base de dados.

Tipo de segmento	# Segmento	Duração média (± desvio padrão) (s)
Fala	7971	11.0 (± 9.4)
Não fala	2529	4.1 (± 5.3)
Fala lida	4989	10.6 (± 8.5)
Fala espontânea	1738	12.0 (± 10.4)

5.3. Metodologia

O método de classificação e deteção de estilos de fala propõe a utilização de parâmetros fonéticos em conjunto com os parâmetros prosódicos. Convém diferenciar uma tarefa de classificação de uma tarefa de deteção. A classificação preocupa-se apenas em atribuir

[4] O efeito (ou reflexo) Lombard consiste na tendência involuntária dos locutores em aumentar o esforço vocálico quando falam num ambiente ruidoso.

uma classe aos segmentos que foram detetados previamente. Por outras palavras, esta tarefa não determina as marcas temporais dos segmentos, apenas lhe atribui uma classificação. Por outro lado, a tarefa de deteção inclui a segmentação e a classificação.

Neste trabalho a tarefa de classificação é feita por classificadores SVM (*Support Vector Machine*) e são usadas as marcas de referências obtidas na anotação manual da base de dados. A tarefa de deteção é feita em dois passos. No primeiro passo é feita a segmentação automática baseada em BIC (*Bayesian Information Criterion*) e no segundo passo são usados os mesmos classificadores que foram usados na tarefa de classificação.

5.4. Segmentação automática

O algoritmo de segmentação implementado é baseado no BIC (*Bayesian Information Criterion*), uma abordagem utilizada para detetar mudanças de locutores que já foi apresentada na secção 4.4.2. Usando a implementação do BIC proposta em (Delacourt and Wellekens, 2000), denominada de *distBIC*, fez-se a segmentação automática que é utilizada durante a deteção.

O *distBIC* utiliza uma medida de distância ou divergência entre as PDF no passo inicial e utiliza o ΔBIC para validar essas marcas numa segunda fase. As observações acústicas são constituídas por 16 coeficientes MFCC e o logaritmo da energia obtidos com janelas de 25 milissegundos e avanço de 10 milissegundos. Foi utilizada a distância de Kullback-Leibler (divergência) no passo inicial para calcular a distância entre gaussianas de segmentos com tamanho fixo de 2 segundos. Considerando dois segmentos de fala consecutivos, X_1 e X_2, ambos representados por uma gaussiana, $X_1 \sim \mathcal{N}(x;\mu_{X1};\Sigma_{X1})$ e $X_2 \sim \mathcal{N}(x;\mu_{X2};\Sigma_{X2})$, a divergência é definida como:

$$J_D(X1,X2) = \mathrm{tr}\left(\frac{\Sigma_{X1}\Sigma_{X2}^{-1} + \Sigma_{X2}\Sigma_{X1}^{-1}}{2}\right) + (\mu_{X1} - \mu_{X2})^T \left(\frac{\Sigma_{X1}^{-1} + \Sigma_{X2}^{-1}}{2}\right)(\mu_{X1} - \mu_{X2}) - p\,, \qquad \textbf{(37)}$$

onde *tr*(.) é a função do traço da matriz e o *p* é a dimensão das observações. Se o valor da divergência entre dois segmentos for muito maior que a média das divergências de todo

o sinal de áudio, é muito provável que haja alteração de condições acústicas. A deteção de marcas onde provavelmente houve alterações acústicas é feita usando um algoritmo de deteção de picos aplicado aos valores suavizados da divergência. Em (Delacourt and Wellekens, 2000) os máximos locais considerados são definidos pelos valores mínimos adjacentes e o desvio padrão pesado das divergências, como é indicado na Figura 21.

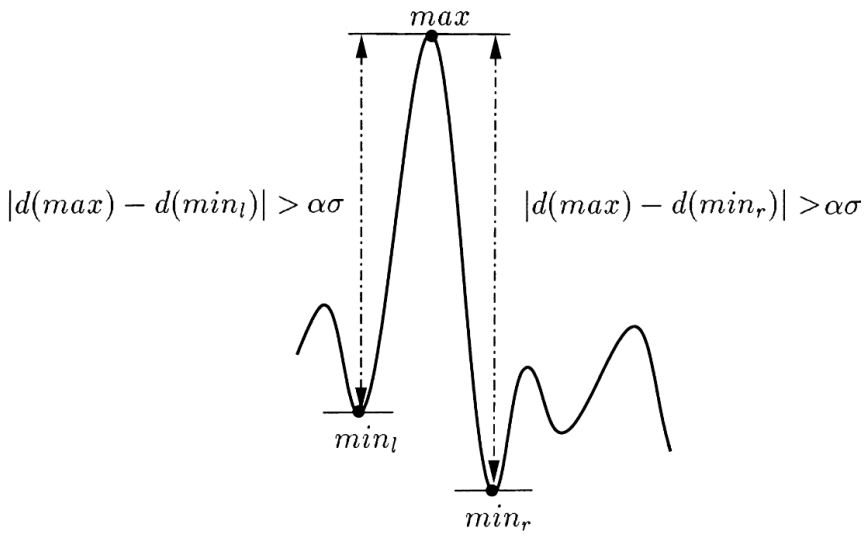

Figura 21 – Determinação máximos locais candidatos a alteração acústica **(Delacourt and Wellekens, 2000)**.

O peso do desvio padrão usado neste trabalho é de 0.6 e foi determinado de forma empírica para maximizar o desempenho de segmentação.

No último passo é calculado o valor de ΔBIC entre os segmentos definidos pelas marcas detetadas com a divergência. Uma marca é descartada se o valor de ΔBIC entre os dois segmentos adjacentes for positivo.

Antes da aplicação do *distBIC* são descartados os segmentos de silêncio que são detetados a partir de informação de energia. São considerados silêncios os segmentos com valor de energia abaixo de um limiar definido pelo desvio padrão de energias. A duração mínima de silêncio é de 0.5 segundos.

5.5. Classificação

Foram treinados dois classificadores SVM: um para classificar um segmento de áudio em "fala" ou "não-fala" e outro para classificar segmentos de fala em termos de estilos de fala "preparada" ou "não-preparada". As experiências foram feitas usando a estratégia de validação cruzada, particionando a base de dados de forma aleatória em 5 partes disjuntas. Usando uma parte como conjunto de teste e as restantes quatro para treinar os classificadores, obtêm-se 5 resultados, rodando os conjuntos de testes. O resultado final é expresso como a média dos 5 resultados.

Os classificadores SVM foram treinados com SMO (*Sequential Minimal Optimization*) (Platt, 1998) e o parâmetro de complexidade (C) do SMO foi determinado para cada classificador por forma a maximizar o desempenho de classificação. A ferramenta WEKA (Hall et al., 2009) foi utilizada para treinar e testar os classificadores.

O treino de dois classificadores para determinar estilos de fala foi requerido pela utilização de SVM, que apenas possibilita classificadores binários. Assim, a classificação de estilos de fala é feita em dois passos. No primeiro passo é feito a classificação "fala" / "não-fala" e só os segmentos classificados com "fala" são submetidos ao segundo classificador, que faz a classificação em termos de estilos de fala (lida / espontânea).

5.6. Parâmetros fonéticos e prosódicos

Os parâmetros fonéticos foram calculados a partir do resultado de um sistema automático de reconhecimento de fones. Foram criados 35 modelos HMM para fonemas do português e um modelo de silêncio que foram treinados com as frases da base de dados do TECNOVOZ (Veiga et al., 2010b) usando as ferramentas do HTK (Young et al., 2006) com as configurações tradicionais dos sistemas de reconhecimento de fala. Na descodificação foi usado um bigrama de fonemas calculado na base de dados TECNOVOZ. Medidas baseadas em duração dos fones e no valor da verosimilhança apresentado no

final da descodificação são usadas como parâmetros fonéticos. As medidas são calculadas por segmentos e compreendem as seguintes:

- quatro estatísticas com as durações dos fones (máximo, mediana, média e desvio padrão);
- a duração total de cada fone normalizada pela duração do segmento;
- a verosimilhança de descodificação de cada fone normalizada pela duração do fone;
- o número de fones normalizado pela duração do segmento;
- o número de silêncios normalizado pela duração do segmento;
- a duração dos fones normalizada pela duração do segmento;
- a duração dos silêncios normalizada pela duração do segmento;

Um vetor de parâmetros fonéticos é constituído por 214 parâmetros, sendo que 210 são referentes aos fones individuais (6 medidas vezes 35 fones) e 4 são referentes ao segmento no seu todo.

Os parâmetros prosódicos são baseados nos valores do tom (frequência fundamental ou F0) e nos valores da relação harmonicidade-ruído (HNR – *Harmonics to Noise Ratio*). Estes valores foram obtidos através da ferramenta Praat (Boersma and Weenink, 2001) usando passos de 10 milissegundos (100 tramas por segundo). O F0 foi limitado no intervalo entre 75 Hz e 300 Hz. Os parâmetros prosódicos são as estatísticas de primeira e segunda ordem das curvas de F0 e HNR calculados nas partes vozeadas dos segmentos. Além das estatísticas, foram adicionados os coeficientes dos polinómios resultantes da regressão da primeira e da segunda ordem das curvas de F0 e HNR e medidas relacionadas com a duração dos segmentos vozeados, duração do silêncio e número de vezes que F0 e HNR se anulam (*reset-rate*). No total, são usados 108 parâmetros prosódicos por cada segmento.

5.7. Resultados e análise

São apresentados três conjuntos de resultados, sendo os primeiros relativos à tarefa de segmentação, os segundos, relativos à tarefa de classificação e os terceiros, relativos à tarefa de deteção automática dos estilos de fala.

Foram usadas as seguintes medidas de desempenho:

- F1 e "Recall" para a tarefa de segmentação;
- "Accuracy" para a tarefa de classificação;
- Taxa de tempo de concordância (AT) para a tarefa de deteção (ou reconhecimento).

O *F1* é a média harmónica entre "Recall" e "Precision" definido da seguinte forma:

$$F_1 = 2 \cdot \frac{Precision \times Recall}{Precision + Recall}, \quad \textbf{(38)}$$

em que o "*Precision*" é definido:

$$Precision = \frac{\#\ marcas\ correctamente\ detetadas}{\#\ marcas\ de\ detectadas}, \quad \textbf{(39)}$$

e o "*Recall*" é definido:

$$Recall = \frac{\#\ marcas\ correctamente\ detetadas}{\#\ marcas\ de\ referência}. \quad \textbf{(40)}$$

Na segmentação, uma marca é considerada correta se houver uma marca de referência a menos de um determinado tempo de tolerância (colar). O desempenho de segmentação depende da duração do colar. A Figura 22 e a Figura 23 apresentam os desempenhos da tarefa de segmentação, apresentando os valores de *F1* e "*Recall*" em função do tamanho do colar, respetivamente.

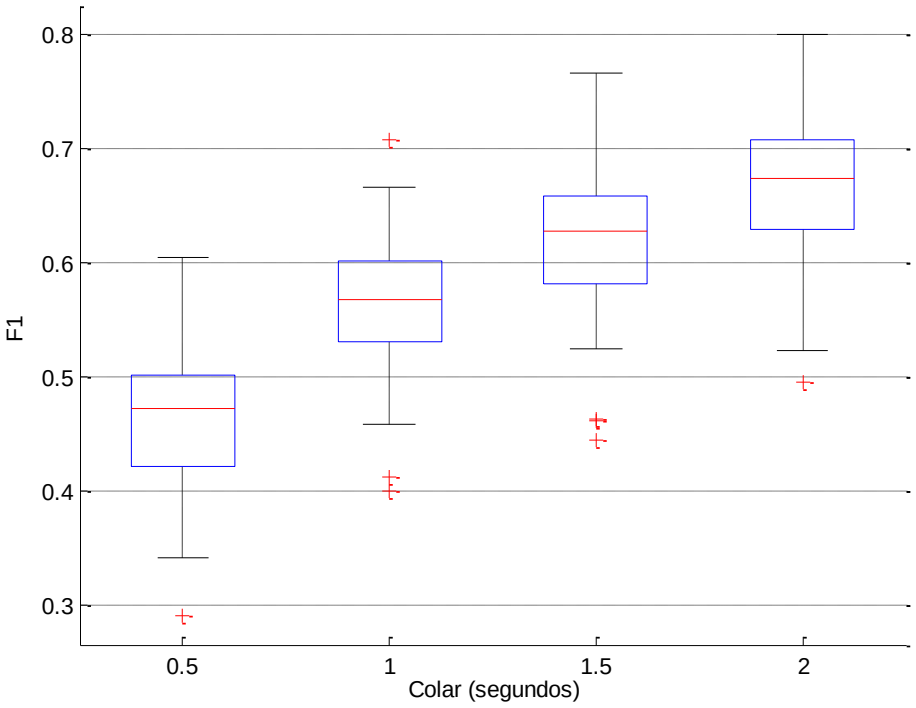

Figura 22 – F1 com colar entre 0.5 e 2 segundos.

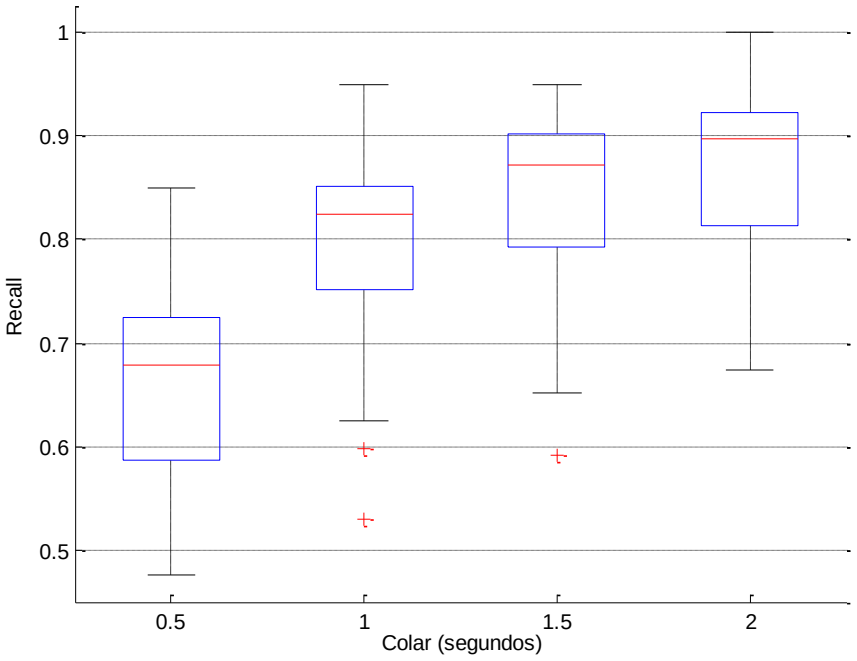

Figura 23 – Recall com colar entre 0.5 e 2 segundos.

O desempenho da classificação é expresso em *Accuracy* que é definido da seguinte forma:

$$Accuracy = \frac{\#\ segmentos\ correctamente\ classificados}{\#\ segmentos\ de\ referência}. \quad \textbf{(41)}$$

A Tabela 17 e a Tabela 18 apresentam os valores de *Accuracy* para os dois classificadores SVM.

Tabela 17 – *Accuracy* do classificador fala/não-fala.

Tipo de parâmetros	***Accuracy* de fala**	***Accuracy* de não-fala**	***Accuracy* global**
Fonético	96.7 %	82.0 %	93.8 %
Prosódico	97.5 %	81.9 %	93.8 %
Combinação	97.6 %	84.0 %	94.4 %

Tabela 18 – *Accuracy* do classificador fala lida/fala espontânea.

Tipo de parâmetros	***Accuracy* de fala lida**	***Accuracy* de fala espontânea**	***Accuracy* global**
Fonético	92.8 %	55.4 %	83.2 %
Prosódico	95.0 %	61.6 %	86.4 %
Combinação	93.7 %	69.5 %	87.4 %

Na Tabela 18 pode ser observado que a informação prosódica é mais pertinente, em relação à informação agrupada como fonética, no que concerne classificar os dois estilos de fala. Se se combinar a informação fonética e prosódica, a fala espontânea é classificada com um maior grau de acerto.

O desempenho da tarefa de deteção de estilos de fala é avaliado pela taxa de acordo temporal (AT) entre a anotação de referência (manual) e a classificação automática. O AT é definido:

$$AT = \frac{duração\ de\ segmentos\ corretamente\ classificados}{duração\ total\ dos\ segmentos}. \quad \textbf{(42)}$$

A Tabela 19 e a Tabela 20 apresentam os valores de AT para os dois classificadores SVM.

Tabela 19 – AT do classificador fala/não-fala.

Tipo de parâmetros	AT de fala	AT de não-fala	AT global
Fonético	94.9 %	62.2 %	91.5 %
Prosódico	97.0 %	61.0 %	93.2 %
Combinação	96.6 %	64.9 %	93.3 %

Tabela 20 – AT do classificador fala lida/fala espontânea.

Tipo de parâmetros	AT de fala lida	AT de fala espontânea	AT global
Fonético	91.9 %	38.6 %	76.7 %
Prosódico	93.0 %	51.2 %	81.1 %
Combinação	92.7 %	59.6 %	83.3 %

Os resultados mostram que a informação prosódica é mais pertinente na tarefa de distinguir fala lida. No entanto, a combinação de informação fonética e prosódica resulta num incremento significativo do desempenho de identificação de fala espontânea à custa de um pequeno decréscimo do desempenho de identificação de fala lida, o que leva a um incremento do desempenho global da combinação.

Os resultados atingidos, ainda que baseados em apenas algumas informações linguísticas retiradas do sinal acústico, são já reveladores da importância que podem ter na caracterização de estilos de fala lida e espontânea. Deles destaca-se os parâmetros de F0 e de duração de fones, os quais se mostraram consistentes na tarefa de diferenciação de segmentos.

5.8. Caracterização de eventos de hesitações

A fala espontânea apresenta muitos eventos de hesitações como as pausas preenchidas, o corte de palavras, as repetições e as extensões de segmentos. A identificação

automática de hesitações é um desafio que ainda carece de investigação e pode auxiliar a identificação de segmentos de fala espontânea.

O objetivo deste estudo prende-se com a identificação e classificação dos diferentes tipos de eventos de hesitação presentes no português. Como eventos de hesitação são consideradas as pausas preenchidas com segmentos não-lexicais (exs.:“uum”, “mm”, “amm”, “aa”), os prolongamentos vocálicos (extensões) no âmbito de palavras (ex.: “deeeee”), as palavras cortadas e as repetições (exs. «de de», «para a para a»). O estágio de desenvolvimento de estudos sobre hesitações difere de língua para língua. Estudos em diferentes línguas, como o inglês (Bell et al., 2003; Tree and Clark, 1997), o sueco (Eklund, 2004), o mandarim (Lee et al., 2004) e o francês (Candea, 2000), tentaram distinguir propriedades linguísticas entre pausas preenchidas e extensões vocálicas, principalmente em busca das razões linguísticas pelas quais as extensões não podem ser eliminadas no pré-processamento. Outros (Henry and Pallaud, 2003), apontam princípios lexicais e sintáticos, que podem fazer a ligação entre repetições e palavras cortadas. Na tarefa de deteção de repetições, têm sido frequentemente utilizadas características acústicas, incluindo duração (Shriberg, 1995) e algumas características sintáticas (Clark and Wasow, 1998).

Para o português europeu existem também vários estudos linguísticos relativos a eventos de hesitação que tentaram proporcionar conhecimentos significativos sobre o tema. Relativamente a pausas preenchidas, trabalhos como (Delgado-Martins and Freitas, 1991; Freitas, 1990; Viana, 1989) podem ser mencionados como pioneiros. Em (Mata, 1999) e em (Veiga et al., 2011c), a frequência fundamental e a duração de pausas preenchidas são apresentadas como características que contribuem para o planeamento de fala espontânea ou leitura oral. Outros trabalhos sobre o tema para o português europeu podem ser encontrados em (Moniz et al., 2009; Veiga et al., 2012a, 2012b). Embora a classificação de pausas preenchidas não seja o tema principal destas duas últimas, mostra que tais eventos de hesitação são responsáveis pela distinção entre o discurso planeado e o não planeado.

5.8.1. Base de dados de pausas preenchidas e extensões

Foram utilizadas 22 horas de sinais de áudio extraídos dos *podcast* de noticiários de televisão e foram utilizados os modelos acústicos treinados com a base de dados TECNOVOZ (os mesmos apresentados em 5.6). Foram utilizados reconhecedores de fones com algumas restrições para detetar vogais longas. De seguida foram feitas confirmações manuais dos segmentos marcados como hesitações. A Figura 24 mostra o diagrama do detetor de hesitações implementado.

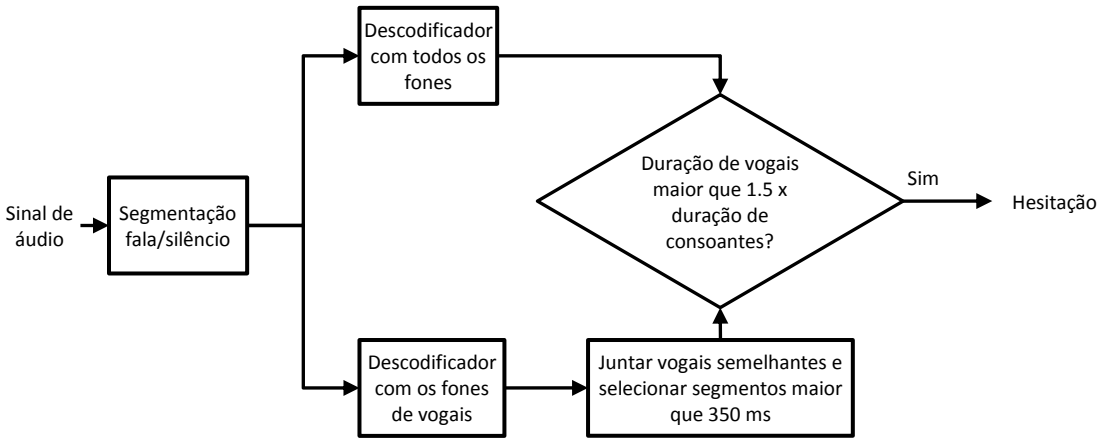

Figura 24 – Detetor de vogais longas.

O descodificador dos fones contém além dos fones, as consoantes nasais /m/ e /n/. O limiar de 350 milissegundos está de acordo com o proposto em (Bartkova, 2005). Este detetor foi pensado para auxiliar a anotação manual da base de dados, por isso os seus parâmetros foram definidos de forma a minimizar os erros de apagamentos à custa dos erros de inserção (falso alarme). Este processo semiautomático tornou a anotação pelo menos quatro vezes mais rápida do que a anotação manual.

Foram identificados 800 eventos de hesitações com este processo. As extensões são mais frequentes que as pausas preenchidas, totalizando 62 % dos eventos anotados. Foram marcadas 33 etiquetas diferentes para extensões e 15 para pausas preenchidas, mas

muitas delas ocorrem poucas vezes. A Tabela 21 apresenta o número de ocorrências das etiquetas mais frequentes.

Tabela 21 – Número de ocorrências de pausas preenchidas e extensões.

Tipo de hesitação	Fone (SAMPA)	# Ocorrências
Pausas preenchidas	6	198
	@	53
	6m	21
Extensões	@	70
	6	61
	i	58
	E	37
	u~	35
	a	25
	O	18
	6~w~	18
	o~	14
	E	13
	6~j~	8
	j6	6

Durante a anotação verificou-se que as extensões ocorrem maioritariamente em proposições e na última sílaba. Por vezes a distinção entre extensão e pausas preenchidas não é óbvia e só no contexto fonético é que se podem desambiguar.

Analisou-se alguns parâmetros acústicos dos eventos identificados com o intuito de confirmar as características já conhecidas para outras línguas. Assim, para cada segmento foram calculados as médias, os desvios padrão e os gradientes de F0 (tom) e de energia.

Os gradientes de F0 e de energia nos segmentos de hesitações são, na maioria das vezes, negativos, como se pode confirmar na Figura 25, o que significa que o F0 e a energia decaem durante um evento de hesitação.

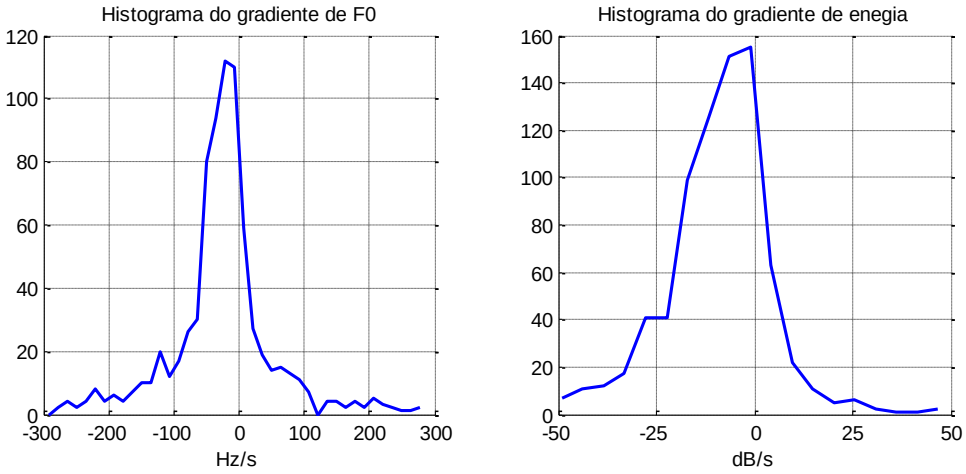

Figura 25 – Histogramas dos gradientes de F0 e de energia.

O desvio padrão de F0 é, em média, 15 Hz e o da energia é de 2.7 dB. Isto indica que, tanto F0 como energia variam pouco, ou seja, o decaimento é suave ao longo de um evento de hesitações. Verificou-se também que estas características não diferenciam entre pausas preenchidas e extensões, corroborando o facto de que perceptualmente a distinção entre esses dois eventos é ambígua se não for tido em conta o contexto.

Capítulo 6. Deteção de palavras

6.1. Introdução

O processo para obter segmentos de fala considerados adequados para serem usados no treino de modelos acústicos foi já descrito nos capítulos anteriores. Porém, é necessário transcrever estes segmentos em termos de uma sequência de palavras.

Normalmente, a transcrição em palavras é auxiliada por um sistema de reconhecimento de grande vocabulário. Não tendo disponível um sistema de reconhecimento de grande vocabulário, desenvolveu-se um sistema de deteção de palavras baseado na técnica de *word-spotting*. Esta técnica precisa apenas dos modelos acústicos de fonemas e da transcrição fonética das palavras a detetar.

Um sistema de *word-spotting* normalmente é muito mais rápido que um sistema de reconhecimento de fala e está limitado a um conjunto restrito de palavras a pesquisar. É apropriado para fazer monitorização em tempo real de transmissão de áudio, elaborar pesquisas de palavras numa grande base de dados de áudio ou indexação de conteúdos de áudio. A complexidade do algoritmo de *word-spotting* é proporcional ao desempenho de deteção das palavras e é necessário um compromisso entre o desempenho e a velocidade de execução de pesquisas.

O desenvolvimento das técnicas de *word-spotting* seguiu a evolução dos algoritmos dos sistemas de reconhecimento automático de fala. As primeiras abordagens utilizavam modelos baseados em DTW (*Dynamic Time Warping*) (Bridle, 1973; Higgins and Wohlford, 1985) e, tal como os sistemas de reconhecimento, evoluíram para modelos HMM (Rabiner, 1990; Rohlicek et al., 1989; Rose and Paul, 1990).

Existem várias estratégias de implementação de *word-spotting* usando modelos HMM. Os mais recorrentes utilizam as mesmas plataformas de treino de modelos acústicos utilizadas pelos sistemas de reconhecimento de fala. Além dos modelos acústicos, são treinados modelos de enchimento (*filler models*) por vezes designados de modelos de lixo (*garbage models*) e/ou os anti-modelos das palavras. Esses modelos são utilizados para

modelar a hipótese alternativa na terminologia estatística de testes de hipóteses. O *word-spotting* pode ser formulado como um problema estatístico de teste de hipóteses uma vez que em cada observação (uma sequência de vetores de parâmetros) tem que formular uma decisão binária:

- a observação pertence ao modelo da palavra – hipótese nula (*H0*);
- a observação não pertence ao modelo da palavra – hipótese alternativa (*H1*).

Para tomar uma decisão, é testada a hipótese nula contra a hipótese alternativa. Segundo o lema de Neyman-Pearson (Neyman and Pearson, 1933) em certas condições, o teste da razão de verosimilhança (LRT – *Likelihood Ratio Test*) é a solução ótima para teste de hipóteses. Dada uma observação e um limiar τ , o LRT é definido:

$$LRT = \frac{\hat{P}(O|H0)}{\hat{P}(O|H1)} \begin{matrix} H0 \\ \geq \\ \leq \\ H1 \end{matrix} \tau . \qquad \textbf{(43)}$$

A Figura 26 apresenta um sistema de *word-spotting* que utiliza LRT na deteção de palavras.

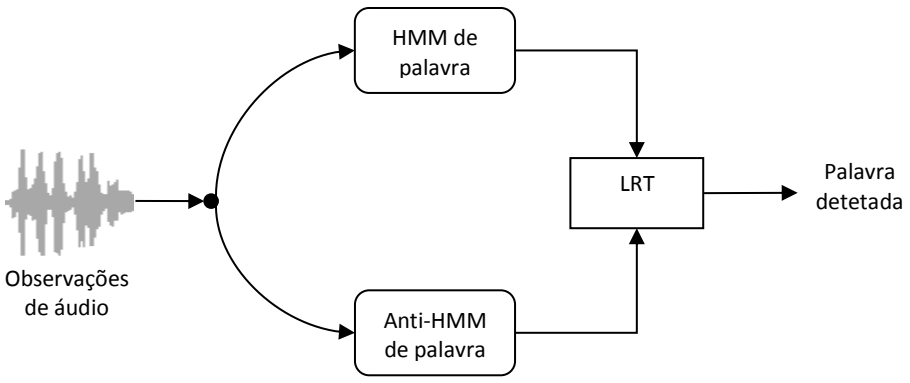

Figura 26 – Sistema de *word-spotting*.

A modelação da hipótese alternativa é o grande desafio dos sistemas de *word-spotting* que implementam esta abordagem. Em geral, não é possível determinar a distribuição da hipótese alternativa e por isso recorre-se a muitas experiências empíricas para melhorar a sua modelação.

A utilização de um modelo de preenchimento pode ser visto como um caso particular da utilização de anti-modelos em que é apresentado um anti-modelo genérico para todas as palavras. O modelo de enchimento mais simples é um modelo com um estado e com uma componente gaussiana onde os parâmetros foram treinados com todas as observações da base de dados. O treino de um anti-modelo para uma determinada palavra é ambíguo uma vez que é difícil determinar amostras adequadas para representar tudo que é diferente dessa palavra. Uma solução seria treinar o anti-modelo de uma palavra com locuções de outras palavras e excluindo locuções de palavras semelhantes. Existem outras soluções como a de (Weintraub, 1993) que utiliza, durante a pesquisa de uma palavra, os modelos de outras palavras como anti-modelo. Esta solução incrementa a dificuldade de descodificação, o que pode comprometer um dos princípios de *word-spotting*, a rapidez de pesquisa.

Uma solução interessante, como a ilustrada na Figura 27, utiliza um arranjo em paralelo de todos os modelos acústicos para formar um modelo de enchimento (Rohlicek et al., 1989; Rose and Paul, 1990).

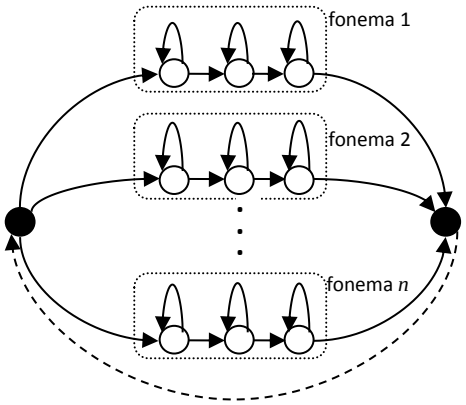

Figura 27 – Modelo de enchimento.

Este arranjo possibilita a modelação de qualquer sequência de fonemas incluindo a da palavra a pesquisar. Uma medida de distância entre a verosimilhança do modelo da palavra e a verosimilhança do modelo de enchimento pode ser utilizada para detetar ocorrências da palavra num vetor de observações. Se a palavra ocorre realmente no segmento em análise, as duas verosimilhanças são similares, correspondendo a uma pequena distância entre elas. Caso contrário a distância é maior.

6.2. *Word-spotting* com medidas de similaridade

Neste trabalho, implementou-se um sistema de *word-spotting* com um processo de descodificação inovador e que utiliza modelos de fonemas para criar o modelo de enchimento e os modelos das palavras. O modelo de uma palavra é feito a partir da concatenação de modelos de fonemas. O sistema de conversão de grafemas para fonemas descrito anteriormente serve para definir a sequência de fonemas da palavra.

Os sistemas de *word-spotting* baseados em modelos HMM utilizam a descodificação de Viterbi através de uma gramática que impõe a competição entre o modelo e o anti-modelo de uma palavra. A descodificação é feita em dois passos: o passo para frente (*forward*) onde se pode utilizar o paradigma de t*oken-passing* (Young et al., 1989) para propagar *tokens* com as verosimilhanças locais ótimas; e o passo de *backtracking* para identificar os segmentos onde o modelo da palavra "ganhou" ao anti-modelo de palavra. O processo de descodificação proposto neste trabalho é executado num único passo. O modelo de enchimento é semelhante ao ilustrado na Figura 27 mas sem o arco de realimentação (dentro do modelo). A gramática proposta (ver a Figura 28) possibilita que o melhor *token* à saída do modelo de enchimento propague para a entrada do modelo de palavra e para a entrada do modelo de enchimento (recuperando a realimentação da Figura 27).

A verosimilhança de um *token* à saída do modelo de enchimento é sempre maior do que a verosimilhança de um *token* à saída do modelo de palavra, exceto quando um *token* percorre no modelo de enchimento os mesmos fonemas que constituem a palavra (neste

caso as verosimilhanças são iguais). A configuração da gramática proposta alimenta o modelo de palavra sempre com o melhor *token*, pelo que o decaimento de verosimilhança verificado à saída do modelo de palavra (em comparação com a verosimilhança do modelo de enchimento) é justificado apenas pelas penalizações ocorridas dentro do modelo de palavra. A diferença entre a verosimilhança do *token* da palavra e a verosimilhança do *token* do enchimento pode ser explorada para implementar uma medida de similaridade.

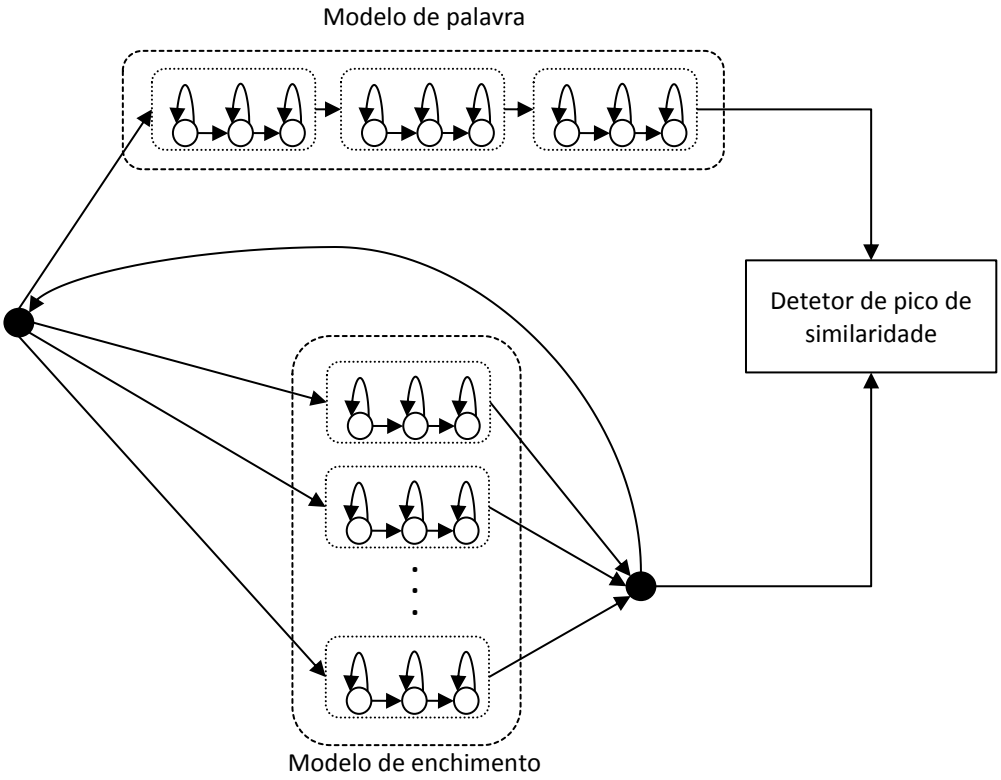

Figura 28 – *Word-spotting* com medida de similaridade.

As informações temporais dos *tokens* só são atualizadas à entrada do modelo de palavra com a indicação de possível trama de início da palavra. A propagação dessa informação permite determinar os limites temporais de uma palavra sem fazer o *backtracking* ou utilizar janelas deslizantes (*sliding window*) para este fim.

6.3. Medidas de similaridade

As medidas de similaridade utilizadas são baseadas no logaritmo das verosimilhanças à saída dos modelos. Um *token* é emitido à saída do modelo por cada observação acústica processada. Atendendo ao algoritmo de Viterbi, o *token* emitido é o *token* que sobreviveu (com a maior verosimilhança) ao percorrer os estados do HMM desde a entrada até à saída. O número médio de observações para propagar um *token* desde a entrada até à saída é imposto pela matriz de probabilidades de transição e no caso de um modelo HMM com topologia esquerda-direita, o número mínimo de observações é igual ao número de estados emissores do modelo.

Uma possível medida de similaridade pode ser o logaritmo da razão de verosimilhança (LLR – *Log-Likelihood Ratio*) entre os *tokens* que saem do modelo de palavra e os *tokens* que saem do modelo de enchimento. O valor de LLR num instante *t*, pela configuração da gramática, pode ser definido da seguinte forma:

$$LLR(t) = \sum_{i=t-N(t)+1}^{t} LW(i) - \sum_{i=t-N(t)+1}^{t} LF(i), \quad \textbf{(44)}$$

em que $N(t)$*)* é o número de observações que o *token* de palavra no instante *t* gastou para percorrer o modelo de palavra e $LW(i)$ é o logaritmo da verosimilhança parcial do *token* calculada no instante i. $LF(i)$ é o logaritmo da verosimilhança parcial do *token* do modelo de enchimento.

A verosimilhança de um *token* que sai do modelo de enchimento num instante *t* é maior ou igual à verosimilhança de um *token* que sai do modelo de palavra, dado que o modelo de enchimento pode modelar qualquer sequência de fonemas. Por isso, os valores de LLR são negativos ou zero. Verifica-se que o valor de LLR decresce com o número de observações de que os *tokens* precisam para percorrer o modelo, por isso não é uma boa medida de similaridade (que deve ser independente do $N(t)$) (ver a Figura 29).

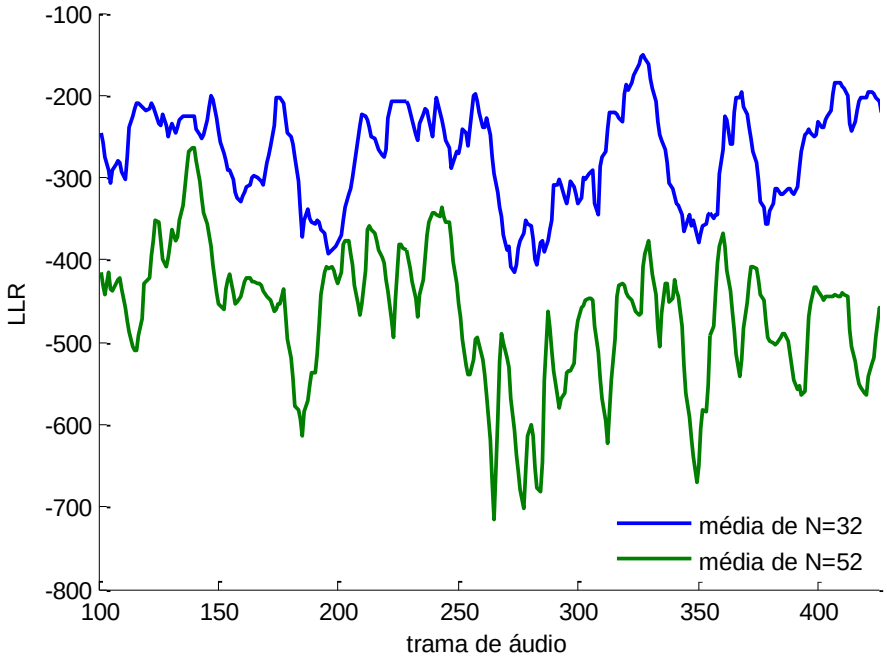

Figura 29 – Valores de LLR variam com N.

Normalizando o valor de LLR pelo número de observações que o *token* demorou a percorrer o modelo de palavra (N), obtém-se uma média geométrica da razão de verosimilhança que pode ser utilizada como medida de similaridade. Assim, definiu-se uma primeira medida de similaridade (SS_1) da seguinte forma:

$$SS_1(t) = \frac{LLR(t)}{N(t)}. \tag{45}$$

A vantagem desta normalização pode ser observada na Figura 30. Para determinar $N(t)$, é necessário que os *tokens* tenham a informação sobre o instante em que entraram no modelo de palavra. Uma outra normalização, baseada na duração média das palavras poderia ser utilizada. Com esta normalização é dispensado o cálculo do tempo que um *token* percorre o modelo de palavra à custa de um ligeiro decréscimo de desempenho. Isto motivou o desenvolvimento de uma outra medida de similaridade que leva em conta a duração média das palavras.

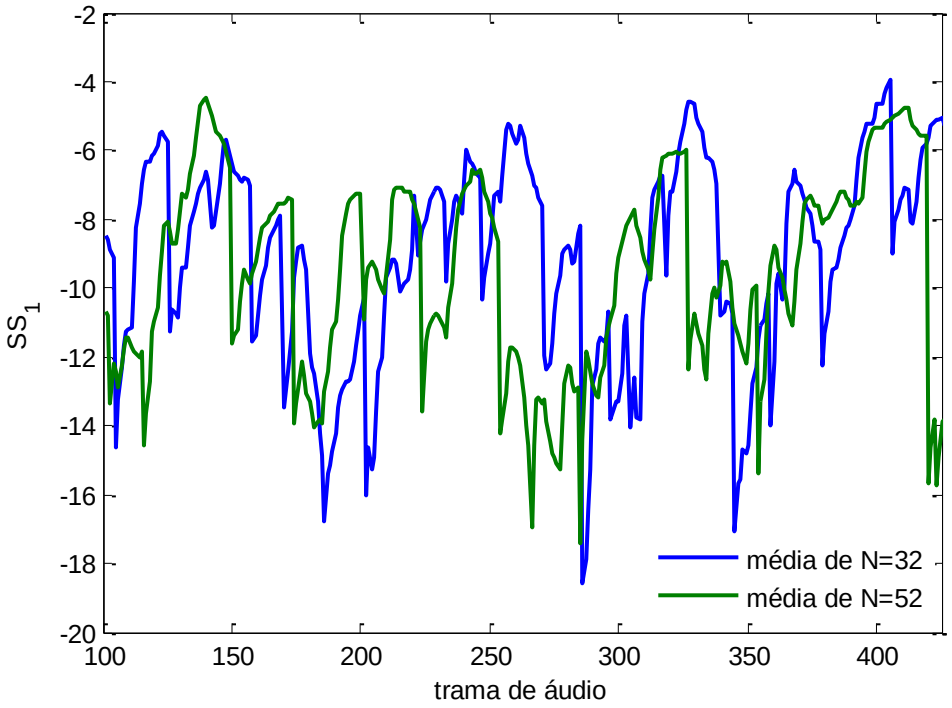

Figura 30 – Valores de SS_1 não variam com N.

A duração média de uma palavra pode ser calcula a partir dos valores da matriz de probabilidades de transição dos estados dos HMM. A duração média de um modelo HMM com a topologia esquerda-direita é simplesmente a soma da duração média de cada estado emissor (Papoulis, 1991). Considerando a_{jj} a probabilidade de um estado j transitar para ele próprio, a duração média deste estado é dada pela seguinte expressão:

$$\bar{d}_j = \frac{1}{1-a_{jj}}. \tag{46}$$

A duração média de um modelo HMM com 3 estados emissores (os modelos de fonemas utilizados neste trabalho tem 5 estados, sendo 3 emissores) é dada pela seguinte expressão:

$$\bar{d}_{HMM} = \sum_{j=2}^{4} \frac{1}{1-a_{jj}}. \tag{47}$$

Nesta expressão consideram-se apenas os 3 estados emissores de um HMM, tal como é indicado na Figura 8. E, por sua vez, a duração média de uma palavra é dada pela soma das durações médias dos modelos de fonemas que constituem a palavra.

Definindo $\bar{D}_W$ como a duração média de uma palavra, a nova medida de similaridade é expressa da seguinte forma:

$$SS_2(t) = \frac{LLR(t)}{\bar{D}_W}. \tag{48}$$

Experimentou-se outras formas de normalização dos valores de LLR e foi definida uma terceira medida de similaridade cuja normalização é baseada na integração dos valores LLR.

Definindo uma área de integração A, da seguinte forma:

$$A(t) = \sum_{i=1}^{N(t)} \big(LLR(t) - LLR(t-i) \big), \tag{49}$$

a nova medida de similaridade é definida como sendo:

$$SS_3(t) = \frac{LLR(t)}{A(t)}. \tag{50}$$

Esta medida leva em conta a duração dos *tokens* e a forma como o LLR evolui deste a entrada dos *tokens* na palavra até à sua saída.

Para tornar o processo de descodificação mais rápido, optou-se por criar um ficheiro de valores por cada ficheiro de áudio, com todos os cálculos que não dependem da palavra a pesquisar. Assim, é calculada, *a priori*, a probabilidade de todos os estados dos modelos HMM e a verosimilhança do modelo de enchimento para todas as observações. Com estes valores são criados uma espécie de parâmetros por cada sinal de áudio ("assinatura acústica") que dispensa o cálculo de gaussianas nas subsequentes pesquisas de palavras. O processo de descodificação limita-se a concatenar os modelos de fonemas das palavras, somar as probabilidades dos estados e somar os pesos da matriz de probabilidades de transição.

6.4. Base de dados

O sistema de *word-spotting* foi avaliado numa base de dados de fala em português, TECNOVOZ (Lopes et al., 2008a; Veiga et al., 2010b). Desta base de dados foram utilizadas 22627 locuções de frases que correspondem a aproximadamente 31.5 horas. Foi treinado um conjunto de 37 modelos de HMM utilizando as ferramentas do HTK (Young et al., 2006) (ver Capítulo 2). Destes 37 modelos, 35 correspondem aos modelos de fonemas da língua portuguesa, 1 corresponde ao modelo de silêncio (no início e no final das locuções) e 1 corresponde ao modelo de pausas curtas (normalmente entre palavras) que não foi considerado neste trabalho, resultando em 108 estados emissores (36 modelos vezes 3 estados). Durante o treino de modelos, foram incrementados o número de componentes gaussianas dos estados até atingir as 96 componentes. Este elevado número de componentes gaussianas justifica-se uma vez que é difícil a utilização dos modelos de trifones com a configuração da descodificação proposta. Os modelos de trifones existentes nesta base de dados correspondem a cerca de 2000 estados diferentes de HMM o que incrementa o número de valores que são guardados depois do cálculo das probabilidades dos estados (um incremento de 1752 % relativo ao número de estados dos monofones).

A base de dados tem 208 *prompts* (frases) diferentes, totalizando 1455 palavras diferentes. As *prompts* têm em média 14 palavras com um desvio padrão de 4.4. A pesquisa admite apenas uma ocorrência de uma dada palavra por locução e foram evitadas palavras no final da locução. As marcas temporais das palavras foram determinadas a partir do alinhamento forçado. Verificou-se que as locuções de frases do TECNOVOZ apresentavam inconsistências e eventos de hesitações em algumas locuções que podem corromper o alinhamento forçado. Para minimizar o efeito destes problemas, foram descartadas as pesquisas que não apresentaram nenhum máximo local nos segmentos definidos pelo alinhamento forçado. Com estas restrições, foram testadas 1170 palavras diferentes com o número de fonemas por palavra a variar ente 1 e 16.

6.5. Resultados

A deteção de uma palavra é determinada pela deteção de picos que ultrapassem um determinado limiar. Variando o limiar desde o valor mínimo até ao valor máximo de cada medida, pode-se contabilizar os erros (falsa aceitação ou falso alarme e falsa rejeição) e fazer a curva DET (*Detection Error Trade-off*) (Martin et al., 1997) ou a curva ROC (*Receiver Operation Characteristics*) (Egan, 1975). Qualquer destas curvas pode ser utilizada para comparar desempenhos dos sistemas de decisão binária. Para cada *token* à saída do modelo de palavra é calculada uma medida de similaridade e, dado um limiar de decisão, pode acontecer uma das quatro hipóteses seguintes:

- decidir que ocorreu uma palavra e a palavra existe – *token* corretamente aceite (CA);
- decidir que não ocorreu a palavra e a palavra não existe – *token* corretamente rejeitado (CR);
- decidir que ocorreu a palavra e a palavra não existe – *token* falsamente aceite (FA);
- decidir que não ocorreu a palavra e a palavra existe – *token* falsamente rejeitado (FR);

A decisão é considerada certa se realizar as hipóteses CA e CR. O decisor erra sempre que acontecem as hipóteses FR e FA. Na literatura designa-se a hipótese FR como erro do tipo I e FA como erro do tipo II. Uma curva DET é obtida pela projeção da probabilidade do erro do tipo I (FR) em função da probabilidade do erro do tipo II (FA) (ver Figura 31).

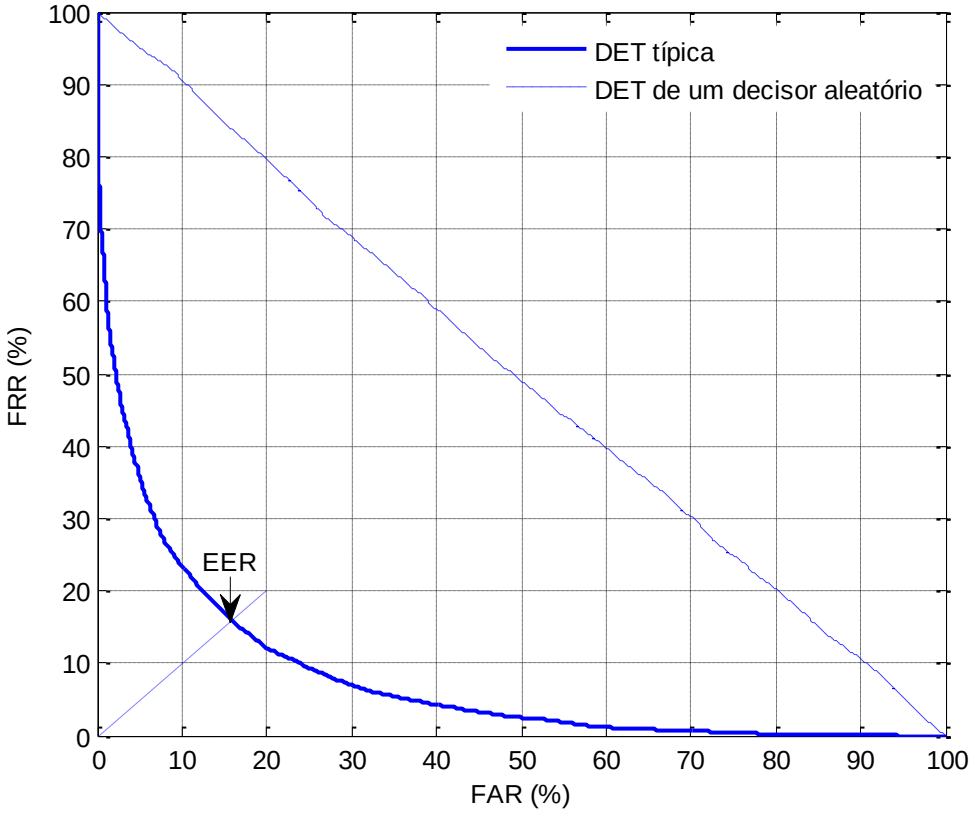

Figura 31 – Curva DET típica.

A curva DET ideal é aquela que passa pela origem. O ponto de operação ótimo depende da aplicação onde é utilizado o decisor. Há sistemas onde é preferível ter uma baixa taxa de falso alarme mesmo que isto seja conseguido à custa de um incremento de falsas rejeições. O peso de cada tipo de erro é que determina o ponto ótimo de operação. Um ponto de compromisso, em que ambos os erros têm o mesmo peso é o ponto onde as duas taxas de erros são iguais (EER – *Equal Error Rate*). Um outro ponto interessante é o ponto mais próximo da origem que muitas vezes não difere muito do EER. A curva DET pode ser utilizada para determinar o limiar de decisão ótimo por forma a minimizar os erros.

Para fazer o DET, para cada palavra, foram usadas amostras de segmentos que contém a palavra (amostras certas) e amostras de segmentos que não contém a palavra (amostras

erradas). Desta forma, obtêm-se os valores da medida quando há um acerto e valores da medida quando há um erro.

O desempenho do *word-spotting* é influenciado pelo número de fonemas da palavra. Uma sequência com poucos fonemas pode ser facilmente confundida ou incluída numa sequência maior. Para evitar isso, pode-se definir um número mínimo de fonemas que o *word-spotter* aceita. A Figura 32 mostra o histograma de número de fonemas por segmentos nas locuções de teste.

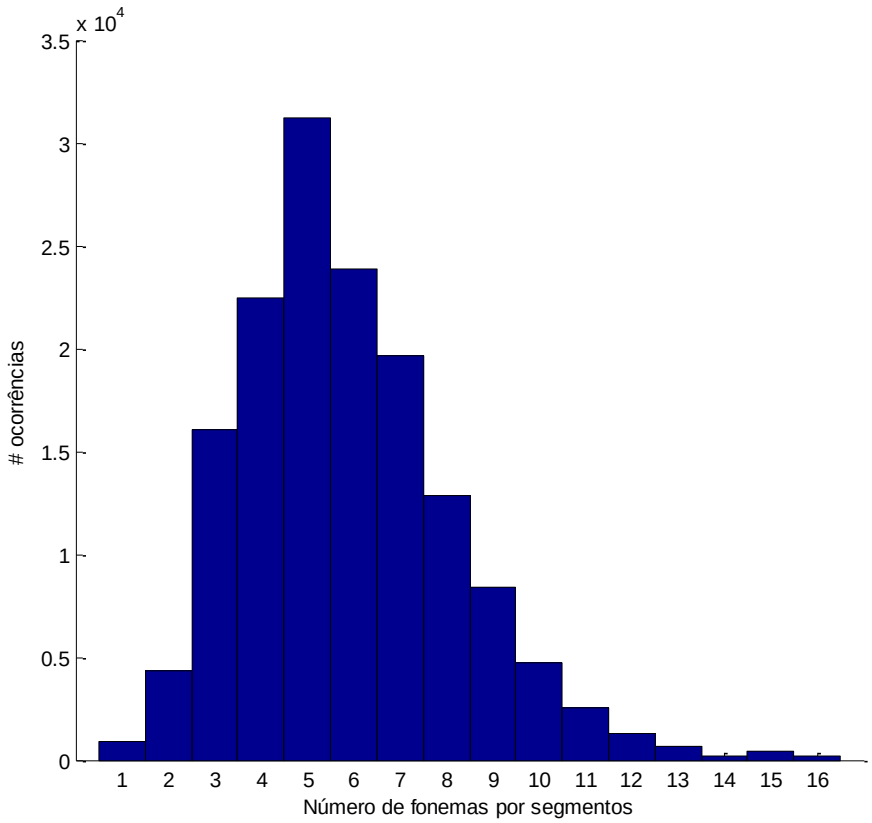

Figura 32 – Histograma de número de fonemas por segmentos.

A Figura 33 apresenta várias curvas DET para a medida SS_3, cada uma a utilizar um limite mínimo de fonemas por palavras.

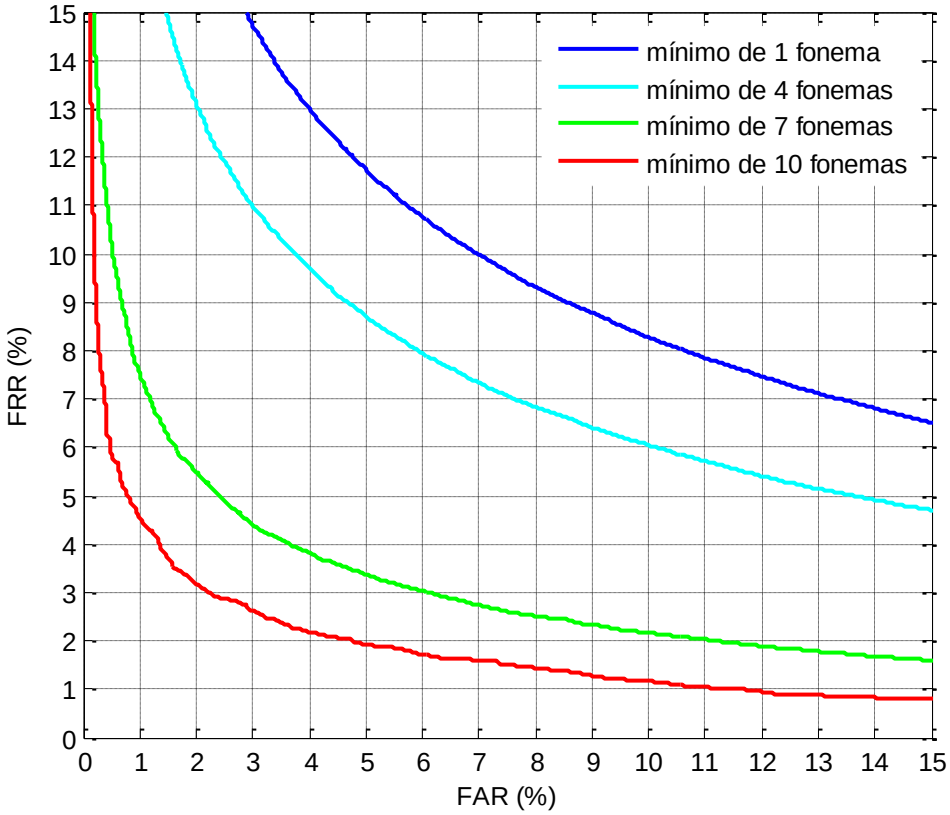

Figura 33 – DET com vários limites mínimos de fonemas por palavras.

Com um limite mínimo de 1 fonema, o EER é igual a 8.84 %, no entanto se utilizar 10 fonemas como limite mínimo o EER passa para 2.78 %.

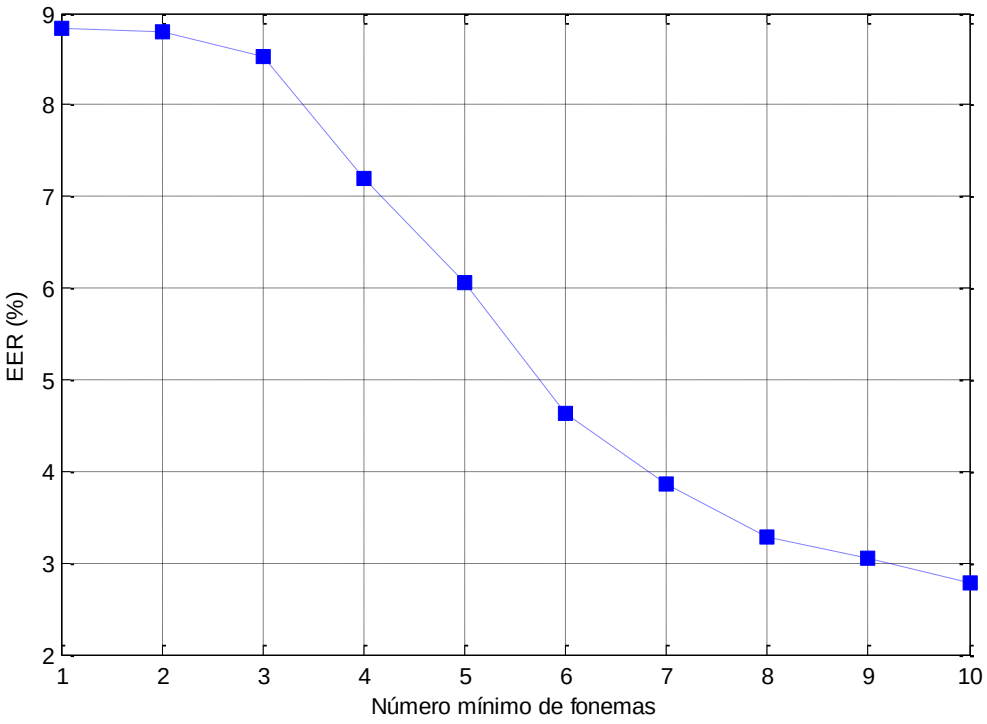

Figura 34 – EER em função do limite mínimo de fonemas por palavras.

Pode-se melhorar o desempenho de falso alarme (falsa aceitação) à custa de falsa rejeição. Por exemplo, se se impuser um limite máximo de falso alarme de 1 % e com um número mínimo de fonemas por palavras igual 10, obtém-se 4.54 % de falsas rejeições.

A Figura 35 apresenta os DET das várias medidas de similaridade para o caso de *word-spotting* com o mínimo de 10 fonemas por palavras.

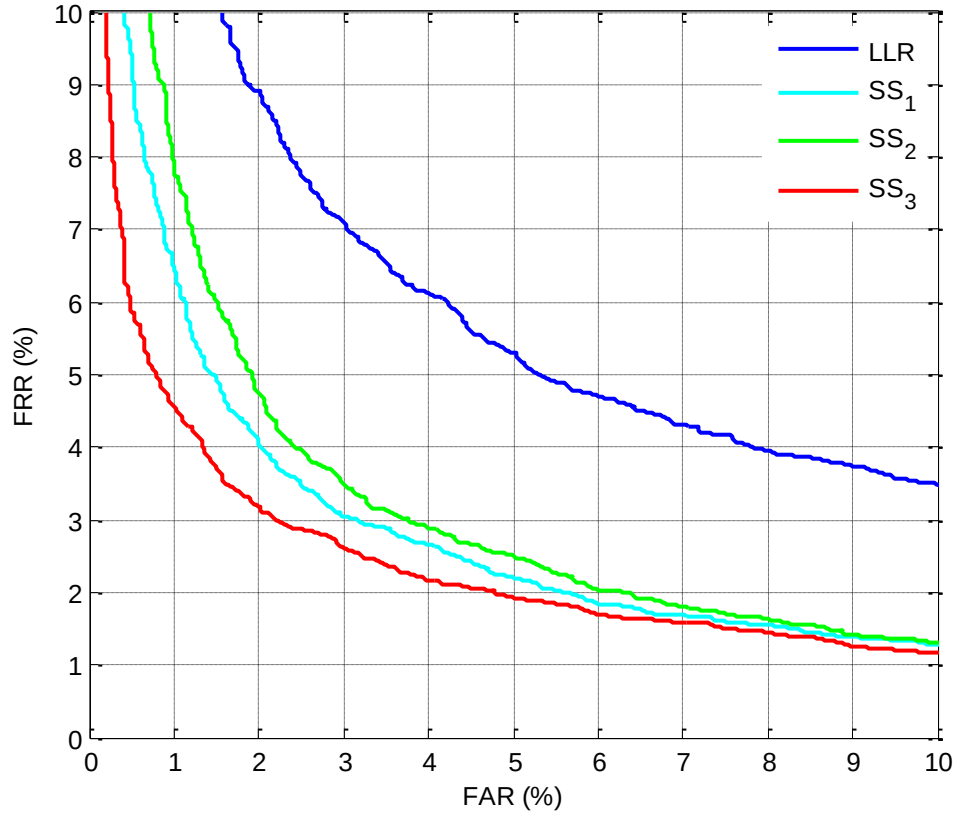

Figura 35 – DET das medidas de similaridade.

O DET confirma que é possível usar uma medida simples como LLR para detetar ocorrências de palavras. É possível melhorar o desempenho da deteção, normalizando o valor de LLR com informação da duração das palavras. Pode-se optar por uma normalização simples como a utilizada na medida SS_2, a duração média das palavras. Este valor é constante para cada palavra e por isso só precisa de ser calculado uma vez, no início de uma pesquisa, e o processo de deteção de ocorrências de uma palavra não precisa de gerir um *buffer* de LLR nem propagar, nos *tokens*, a informação temporal. A medida SS_2 tem um EER igual 3.27 % e uma diminuição de 36.26 % relativo ao EER do LLR (5.13 %).

Pode-se utilizar normalizações mais complexas que levam em conta o tempo que os *tokens* demoram a percorrer o modelo de palavra. As medidas SS_1 e SS_3 precisam de

informação temporal dos *tokens* e a medida SS_3 precisa de um gestor de *buffer* de LLR. O EER de SS_1 é de 3.05 % e o EER de SS_3 é de 2.78 % como já foi referido anteriormente.

Para aplicar este sistema de *word-spotting* às locuções de noticiários é necessário adaptar os modelos treinados na base de dados TECNOVOZ às características das locuções de noticiários e, para isso, era necessário transcrever alguns segmentos de noticiários. Testes preliminares, sem a devida adaptação, não produziram resultados satisfatórios. A falta de contexto dos modelos de monofones é mais um problema. Está demonstrado que os trifones apresentam melhores desempenhos que os monofones nos sistemas de reconhecimento de fala, como se pode observar em (Lopes et al., 2008b). O crescente interesse em redes neuronais profundas (DNN – *Deep Neural Network*) nos sistemas de reconhecimento de fala fez com que se equacionasse uma implementação de um sistema de *word-spotting* que modela os contextos usando as DNN. Para isso, é necessário treinar e testar várias configurações de redes e várias alternativas de inclusão de contexto. O treino de uma rede neuronal normalmente demora muito mais tempo que o treino de HMM com GMM e a implementação do *word-spotting* foi a última tarefa calendarizada no planeamento dos trabalhos. Assim, esta solução fica como recomendação para trabalho futuro.

Capítulo 7. Conclusão

O estágio de desenvolvimento e a aplicação das tecnologias de fala para uma língua estão relacionados com a quantidade e qualidade de recursos disponíveis para esta língua. A língua portuguesa carece de recursos livres, o que pode dificultar o desenvolvimento e a aplicação de tecnologias de fala que a incorporam. No decurso da execução desta tese foram implementadas várias tecnologias da área de processamento automático de fala que podem ser utilizadas para incrementar a disponibilidade de recursos linguísticos e acústicos da língua portuguesa.

Foram desenvolvidas várias técnicas para segmentar noticiários e gerar informações que possibilitam a escolha de segmentos adequados para treinar modelos acústicos. Desenvolveu-se um sistema de segmentação que utiliza a tecnologia de impressão digital acústica para detetar segmentos repetidos e desenvolveu-se um sistema de diarização de locutor que utiliza GMM e BIC para criar *cluster* de locutores. As abordagens propostas são originais e podem ser combinadas com outras técnicas para produzir melhores resultados. Ainda para a segmentação, desenvolveu-se um sistema capaz de identificar estilos de fala baseado em apenas algumas informações linguísticas (fonéticas e prosódicas) retiradas do sinal acústico. Fez-se um estudo sobre eventos de hesitações na língua portuguesa onde foram identificados e caracterizados os fonemas que são usados nos segmentos de hesitações.

Para a transcrição fonética dos segmentos fez-se uma proposta para um sistema de *word-spotting* baseado em medidas de similaridade. Propôs-se um descodificador que não precisa de calcular gaussianas nas pesquisas de palavras. Uma implementação desta abordagem foi testada numa base de dados de locuções de frases e resultou num bom desempenho. No entanto, é preciso adaptar o sistema *word-spotting* aos sinais de áudio dos noticiários e esta implementação carece de melhorias para incluir contextos dos fonemas e as novas tendências que se têm verificado na área de reconhecimento de fala.

Para utilizar o sistema de *word-spotting* é necessário utilizar um dicionário de pronunciação ou um sistema de conversão de grafemas para fonemas. Neste âmbito,

desenvolveu-se um sistema de conversão de grafemas para fonemas que recorre a regras fonológicas e modelos estatísticos e criou-se um dicionário de pronunciação com mais de 40000 entradas. O desempenho do conversor é um dos melhores publicados para a língua portuguesa e todos os recursos utilizados e criados foram publicados. Desenvolveu-se também um demonstrador deste sistema com interface web que foi denominado de grafone (LABFALA, 2011).

Apesar de não ter chegado ao estágio de integração de todas as tecnologias desenvolvidas para implementar um sistema de treino de modelos acústicos não supervisionado, todos os módulos propostos para desencadear o processo de treino partindo de sinais de áudio sem transcrição fonética foram implementados e testados. Os trabalhos desenvolvidos abrangeram uma vasta área de processamento automático de fala e requereram competências linguísticas. Esta abrangência e a multidisciplinaridade dos trabalhos desenvolvidos foram desafios superados e resultaram numa contribuição para a disponibilização de recursos linguísticos e acústicos para a língua portuguesa, como a publicação de uma base de dados de hesitações para o português, vários dicionários de pronunciação, um dicionário exaustivo de pronunciação de homógrafos heterofónicos, um dicionário de estrangeirismo e um sistema de conversão de grafemas para fonemas.

Bibliografia

Ainsworth, W. (1973). A System for Converting English Text Into Speech. IEEE Transactions on Audio and Electroacoustics *21*, 288–290.

Almeida, J.J., and Simões, A. (2001). Text to Speech, a Rewriting System Approach. Procesamiento Del Lenguaje Natural *27*, 247–254.

Amir, A., Efrat, A., and Srinivasan, S. (2001). Advances in Phonetic Word Spotting. In Proc. of the 10th International Conference on Information and Knowledge Management, (Atlanta, Georgia, USA), pp. 580–582.

Andrade, E., and Viana, M.C. (1985). CORSO I: Um Conversor de Texto Ortográfico em Código Fonético para o Português (Lisboa, Portugal: CLUL).

Barbosa, P.A., Viana, C., and Trancoso, I. (2009). Cross-Variety Rhythm Typology in Portuguese. In Proc. of the 10th Annual Conference of the International Speech Communication Association, (Brighton, United Kingdom: ISCA), pp. 1011–1014.

Barras, C., Geoffrois, E., Wu, Z., and Liberman, M. (2001). Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. Speech Communication *33*, 5–22.

Barros, M.J., and Weiss, C. (2006). Maximum Entropy Motivated Grapheme-to-Phoneme, Stress and Syllable Boundary Prediction for Portuguese Text-to-Speech. In Proc. of the IV Jornadas En Tecnologia Del Habla, (Zaragoza, Spain), pp. 177–182.

Barros, M.J., Braga, D., Freitas, D., Teixeira, J.P., and Latsch, V. (2001). Back Close Non-Syllabic Vowel [U] Behaviour in European Portuguese: Reduction or Suppression? In Proc. of the International Conference in Speech Processing, (Taejon, Korea: ISCA),.

Bartkova, K. (2005). Prosodic Cues of Spontaneous Speech in French. In Proc. of the 4th Workshop on Disfluency in Spontaneous Speech, (Aix-en-Provence, France), pp. 21–25.

Batista, F., Trancoso, I., and Mamede, N.J. (2009). Comparing Automatic Rich Transcription for Portuguese, Spanish and English Broadcast News. In Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding, (Merano, Itália), pp. 540–545.

Batlle, E., Masip, J., and Cano, P. (2003). System Analysis and Performance Tuning for Broadcast Audio Fingerprinting. In Proc. of the 6th International Conference on Digital Audio Effects, (London, United Kingdom),.

Bell, A., Jurafsky, D., Fosler-lussier, E., Girand, C., Gregory, M., and Gildea, D. (2003). Effects of Disfluencies, Predictability, and Utterance Position on Word form Variation in English Conversation. Journal of the Acoustical Society of America *113*, 1001–1024.

Betser, M., Collen, P., and Rault, J.-B. (2007). Audio Identification Using Sinusoidal Modeling and Application to Jingle Detection. In Proc. of the 8th International Conference on Music Information Retrieval, (Vienna, Austria), pp. 139–142.

Biadsy, F., and Hirschberg, J. (2009). Using Prosody and Phonotactics in Arabic Dialect Identification. In Proc. of the 10th Annual Conference of the International Speech Communication Association, (Brighton, United Kingdom: ISCA), pp. 208–211.

Bisani, M., and Ney, H. (2002). Investigations on Joint-Multigram Models for Grapheme-to-Phoneme Conversion. In Proc. of the 7th International Conference on Spoken Language Processing, (Denver, USA: ISCA), pp. 105–108.

Bisani, M., and Ney, H. (2008). Joint-Sequence Models for Grapheme-to-Phoneme Conversion. Speech Communication *50*, 434–451.

Boersma, P., and Weenink, D. (2001). Praat, a System for Doing Phonetics by Computer. Glot International *5*, 341–345.

Braga, D., and Marques, M.A. (2007). Desambiguador de Homógrafos Heterófonos para Sistemas de Conversão Texto-Fala em Português. Revista Diacrítica Série Ciências Da Linguagem *21*, 25–49.

Braga, D., Coelho, L., and Resende, F.G. (2006). A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese. In Proc. of the VI International Telecommunications Symposium, (Fortaleza, Brazil), pp. 328–333.

Bridle, J.S. (1973). An Efficient Elastic-Template Method for Detecting Given Words in Running Speech. In Proc. of the British Acoustical Society Meeting,.

Butko, T., Nadeu, C., and Schulz, H. (2010). Albayzin-2010 Audio Segmentation Evaluation: Evaluation Setup and Results. In Proc. of the GTM, RTTH and SIG-IL VI Jornadas En Tecnología Del Habla and II Iberian SLTech Workshop, (Vigo, Spain),.

Candea, M. (2000). Contribution a l'Etude des Pauses Silencieuses et des Phenomenes Dits «d'Hesitation» en Français Oral Spontane. PhD. Université Paris III.

Candeias, S., and Perdigão, F. (2008). Conversor de Grafemas para Fones Baseado em Regras para Português. In Perspectivas Sobre a Linguateca - Actas Do Encontro Linguateca: 10 Anos, L. Costa, D. Santos, and N. Cardoso, eds. (Linguateca), pp. 99–104.

Candeias, S., Veiga, A., Lopes, C., and Perdigão, F. (2011). A Realização do Schwa no Português Europeu. In Proc. of the 8th Brazilian Symposium in Information and Human Language Technology, (Cuiabá, Mato Grosso, Brazil), pp. 26–31.

Candeias, S., Veiga, A., and Perdigão, F. (2012a). Conversor de Grafemas em Fonemas: Um Recurso Linguístico para o Português Europeu. In Proc. of the XXVIII Encontro Nacional Da Associação Portuguesa de Linguística, (Faro, Portugal),.

Candeias, S., Veiga, A., Celorico, D., Proença, J., and Perdigão, F. (2012b). Eventos Linguísticos na Classificação Automática de Estilos de Fala. In Proc. of the XXVIII Encontro Nacional Da Associação Portuguesa de Linguística, (Faro, Portugal),.

Candeias, S., Celorico, D., Proença, J., Veiga, A., and Perdigão, F. (2013a). HESITA(tions) in Portuguese: a Database. In Proc. of the 6th Workshop on Disfluency in Spontaneous Speech, (Stockholm, Sweden: ISCA), pp. 13–16.

Candeias, S., Celorico, D., Proença, J., Veiga, A., and Perdigão, F. (2013b). Automatically Distinguishing Styles of Speech. In Proc. of the Conf. on Telecommunications, (Castelo Branco, Portugal), pp. 77–80.

Caseiro, D., Trancoso, I., Oliveira, L., and Viana, C. (2002). Grapheme-to-Phone Using Finite-State Transducers. In Proc. of the IEEE Workshop on Speech Synthesis, (California, USA), pp. 215–218.

Chan, H.Y., and Woodland, P. (2004). Improving Broadcast News Transcription by Lightly Supervised Discriminative Training. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (Montreal, Canada), pp. 737–740.

Chen, S.F., and Goodman, J. (1999). An Empirical Study of Smoothing Techniques for Language Modeling. Computer Speech & Language *13*, 359–393.

Clark, H.H., and Wasow, T. (1998). Repeating Words in Spontaneous Speech. Cognitive Psychology *37*, 201–242.

Dahl, G.E., Yu, D., Deng, L., and Acero, A. (2012). Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition. IEEE Transactions on Audio, Speech, and Language Processing *20*, 30–42.

Davis, S., and Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Transactions on Audio, Speech, and Language Processing *28*, 357–366.

Delacourt, P., and Wellekens, C.J. (2000). DISTBIC: A Speaker-based Segmentation for Audio Data Indexing. Speech Communication *32*, 111–126.

Delgado-Martins, M.R., and Freitas, M.J. (1991). Temporal Structures of Speech: “Reading News on TV.”In Proc. of the ESCA Workshop on Phonetics and Phonology of Speaking Styles, (Barcelona, Spain: ISCA), pp. 19.1–19.5.

Deshmukh, O., Kandhway, K., Verma, A., and Audhkhasi, K. (2009). Automatic Evaluation of Spoken English Fluency. In Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing, (Taipei, Taiwan: IEEE), pp. 4829–4832.

DLPO (2013). Dicionário Priberam da Língua Portuguesa, http://www.priberam.pt/DLPO.

Egan, J.P. (1975). Signal Detection Theory and ROC Analysis (New York, USA: Academic Press).

Eklund, R. (2004). Disfluency in Swedish Human-human and Human-machine Travel Booking Dialogues (Department of Computer and Information Science, Linköping Studies in Science and Technology).

Eskenazi, M. (1993). Trends in Speaking Styles Research. In Proc. of the 3th European Conference on Speech Communication and Technology, (Berlin, Germany: ISCA), pp. 501–509.

Evermann, G., Chan, H.Y., Gales, M.J.F., Jia, B., Mrva, D., Woodland, P.C., and Yu, K. (2005). Training LVCSR Systems on Thousands of Hours of Data. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, (Philadelphia, USA), pp. 209–212.

Falavigna, D., Giuliani, D., Gretter, R., Lööf, J., Gollan, C., Schluter, R., and Ney, H. (2009). Automatic Transcription of Courtroom Recordings in the JUMAS project. In Proceedings of the International Conference on ICT Solutions for Justice, (Skopje, Macedonia), pp. 65–72.

Fink, M., Covell, M., and Baluja, S. (2006). Social- and Interactive-Television Applications Based on Real-Time Ambient-Audio Identification. In Prc. of the 4th European Conference on Interactive Television, (Athens, Greece),.

Freitas, M.J. (1990). Estratégias de Organização Temporal do Discurso. MA. Universidade de Lisboa.

Fry, D.B. (1979). The Physics of Speech (Cambridge, United Kingdom: Cambridge University Press).

Gales, M.J.F., Kim, D.Y., Woodland, P.C., Chan, H.Y., Mrva, D., Sinha, R., and Tranter, S.E. (2006). Progress in the CU-HTK Broadcast News Transcription System. IEEE Transactions on Audio, Speech, and Language Processing *14*, 1513–1525.

Galescu, L., and Allen, J.F. (2001). Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model. In Proc. of the 4th ISCA Tutorial and Research Workshop on Speech Synthesis, (Perthshire, Scotland),.

Good, I.J. (1953). The Population Frequencies of Species and the Estimation of Population Parameters. Biometrika *40*, 237–264.

Gusfield, D. (1997). Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology (Cambridge University Press).

Hain, H.-U. (1999). Automation of the Training Procedure for Neural Networks Performing Multilingual Grapheme to Phoneme Conversion. In Proc. of the 6th European Conference on Speech Communication and Technology, (Budapest, Hungary), pp. 2087–2090.

Haitsma, J., and Kalker, T. (2002). A Highly Robust Audio Fingerprinting System. In Proc. of the 3rd International Conference on Music Information Retrieval, (Paris, France),.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA Data Mining Software: An Update. ACM SIGKDD Explorations Newsletter *11*, 10–18.

Henry, S., and Pallaud, S. (2003). Word Fragments and Repeats in Spontaneous Spoken French. In Proc. of the 3th Workshop on Disfluency in Spontaneous Speech, (Gothenburg, Sweden), pp. 77–80.

Hermansky, H. (1990). Perceptual Linear Predictive (PLP) Analysis of Speech. Journal of the Acoustical Society of America *87*, 1738–1752.

Higgins, A., and Wohlford, R. (1985). Keyword Recognition Using Template Concatenation. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, pp. 1233–1236.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T., et al. (2012). Deep Neural Networks for Acoustic Modeling in Speech Recognition: The Shared Views of Four Research Groups. IEEE Signal Processing Magazine *29*, 82–97.

Huijbregts, M. (2008). Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled. Tese de Doutoramento. University of Twente.

Infopédia (2013). Dicionário da Língua Portuguesa da Porto Editora, http://www.infopedia.pt/lingua-portuguesa.

Jiampojamarn, S., and Kondrak, G. (2009). Online Discriminative Training for Grapheme-to-Phoneme Conversion. In Proc. of the 10th Annual Conference of the International Speech Communication Association, (Brighton, United Kingdom), pp. 1303–1306.

Jiampojamarn, S., Kondrak, G., and Sherif, T. (2007). Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion. In Proc. of the Annual Conference of the North American Chapter of the Association for Computational Linguistics, (Rochester, New York, USA), pp. 372–379.

Johnson, S.E., and Woodland, P.C. (2000). A Method for Direct Audio Search with Applications to Indexing and Retrieval. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, (Istanbul, Turkey), pp. 1427–1430.

Katz, S.M. (1987). Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer. IEEE Transactions on Acoustics, Speech and Signal Processing *35*, 400–401.

Kaufmann, T., Ewender, T., and Pfister, B. (2009). Improving Broadcast News Transcription with a Precision Grammar and Discriminative Reranking. In Proceedings of the Conference

of the International Speech Communication Association, (Brighton, United Kingdom), pp. 356–359.

Kawai, H., and Toda, T. (2004). An Evaluation of Automatic Phone Segmentation for Concatenative Speech Synthesis. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (Montreal, Canada), pp. 677–680.

Kneser, R., and Ney, H. (1995). Improved Backing-off for M-gram Language Modeling. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, (Detroit, Michigan, U. S. A.), pp. 181–184.

LABFALA (2011). Conversor de Grafemas para Fonemas para Português Europeu, http://www.co.it.pt/~labfala/g2p.

Lamere, P., Kwok, P., Walker, W., Gouvêa, E.B., Singh, R., Raj, B., and Wolf, P. (2003). Design of the CMU Sphinx-4 Decoder. In Proc. of the 8th European Conference on Speech Communication and Technology, (Geneva, Switzerland: ISCA),.

Lee, A., Kawahara, T., and Shikano, K. (2001). Julius - an Open Source Real-Time Large Vocabulary Recognition Engine. In Proc. of the 7th European Conference on Speech Communication and Technology, (Aalborg, Denmark), pp. 1691–1694.

Lee, T.-L., He, Y.-F., Huang, Y.-J., Tseng, S.-C., and Eklund, R. (2004). Prolongation in Spontaneous Mandarin. In Proc. of the 8th International Conference on Spoken Language Processing, (Jeju Island, Korea: ISCA), pp. 2181–2184.

Leonard, R. (1984). A Database for Speaker-independent Digit Recognition. In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, (California, USA), pp. 328–331.

Lince (2010). conversor para a nova ortografia, http://www.portaldalinguaportuguesa.org/lince.php.

Llisterri, J. (1992). Speaking Styles in Speech Research. In Proc. of ELSNET/ESCA/SALT Workshop on Intergating Speech and Natural Language, (Dublin, Ireland), pp. 17–38.

Lopes, C. (2011). Classes Fonéticas Alargadas no Reconhecimento Automático de Fones. Tese de Doutoramento. Universidade de Coimbra.

Lopes, C., Veiga, A., and Perdigão, F. (2010). Using Fingerprinting to Aid Audio Segmentation. In Proc. of the GTM, RTTH and SIG-IL VI Jornadas En Tecnología Del Habla and II Iberian SLTech Workshop, (Vigo, Spain),.

Lopes, J.D., Neves, C., Veiga, A., Maciel, A., Lopes, C., Perdigão, F., and Sá, L. (2008a). Development of a Speech Recognizer with the Tecnovoz Database. In Computational Processing of the Portuguese Language, A. Teixeira, V.L.S. Lima, L.C. Oliveira, and P. Quaresma, eds. (Springer Berlin Heidelberg), pp. 260–263.

Lopes, J.D., Neves, C., Veiga, A., Maciel, A., Lopes, C., Sá, L., and Perdigão, F. (2008b). Training a Robust Command Recognizer with The Tecnovoz Database. In Proc. of the V Jornadas En Tecnología Del Habla, (Bilbao, Spain), pp. 19–22.

Ma, J., Matsoukas, S., Kimball, O., and Schwartz, R. (2006). Unsupervised Training on Large Amounts of Broadcast News Data. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (Toulouse, France), pp. 1056–1059.

Martin, A., Doddington, G., Kamm, T., Ordowski, M., and Przybocki, M. (1997). The DET Curve in Assessment of Detection Task Performance (DTIC Document).

Mata, A.I. (1999). Para o Estudo da Entoação em Fala Espontânea e Preparada no Português Europeu: Metodologia, Resultados e Implicações Didácticas. PhD. Universidade de Lisboa.

Mateus, M.H., and Andrade, E. (2002). The Phonology of Portuguese (Oxford, New York, U. S. A.: Oxford University Press).

Matsoukas, S., Gauvain, J.-L., Adda, G., Colthurst, T., Kao, C.-L., Kimball, O., Lamel, L., Lefevre, F., Ma, J.Z., Makhoul, J., et al. (2006). Advances in Transcription of Broadcast News and Conversational Telephone Speech within the Combined EARS BBN/LIMSI System. IEEE Transactions on Audio, Speech, and Language Processing *14*, 1541–1556.

Meinedo, H., and Neto, J.P. (2004). Detection of Acoustic Patterns in Broadcast News Using Neural Networks. In Proc. of the Acústica - Acoustics European Symposium, (Guimarães, Portugal),.

Moniz, H., Trancoso, I., and Mata, A.I. (2009). Classification of Disfluent Phenomena as Fluent Communicative Devices in Specific Prosodic Contexts. In Proc. of the 10th Annual Conference of the International Speech Communication Association, (Brighton, United Kingdom: ISCA), pp. 1719–1722.

Muscariello, A., Gravier, G., and Bimbot, F. (2009). Audio Keyword Extraction by Unsupervised Word Discovery. In Proceedings of the Conference of the International Speech Communication Association, (Brighton, United Kingdom), pp. 2843–2846.

Nakamura, M., Iwano, K., and Furui, S. (2008). Differences Between Acoustic Characteristics of Spontaneous and Read Speech and Their Effects on Speech Recognition Performance. Jornal of Computer Speech and Language *22*, 171–184.

Nespor, M., and Vogel, I. (1986). Prosodic Phonology (Foris Publications).

Neto, J.P., Meinedo, H., Viveiros, M., Cassaca, R., Martins, C., and Caseiro, D. (2008). Broadcast News Subtitling System in Portuguese. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (Las Vegas, USA), pp. 1561–1564.

Neves, C., Veiga, A., Sá, L., and Perdigão, F. (2009). Audio Fingerprinting System for Broadcast Streams. In Proc. of the Conf. on Telecommunications, (Santa Maria da Feira, Portugal), pp. 481–484.

Ney, H., Essen, U., and Kneser, R. (1994). On Structuring Probabilistic Dependences in Stochastic Language Modelling. Computer Speech & Language *8*, 1–38.

Neyman, J., and Pearson, E.S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences *231*, 289–337.

NIST (2009). The History of Automatic Speech Recognition Evaluations at NIST, http://www.itl.nist.gov/iad/mig/publications/ASRhistory.

NIST (2013). Rich Transcription Evaluation Project, http://www.itl.nist.gov/iad/mig/tests/rt.

Ogle, J.P., and Ellis, D.P. (2007). Fingerprinting to Identify Repeated Sound Events in Long-Duration Personal Audio Recordings. In Proc. of the International Conference on Acoustics, Speech and Signal Processing, (Hawaii, USA: IEEE), pp. 233–236.

Oliveira, C., Paiva, S., Moutinho, L., and Teixeira, A. (2004). Um Novo Sistema de Conversão Grafema-Fone para PE Baseado em Transdutores. In Actas Do II Congresso Internacional de Fonética E Fonologia, (São Luís, Maranhão, Brasil),.

Oliveira, L.C., Viana, M.C., and Trancoso, I. (1992). A Rule-Based Text-to-Speech System for Portuguese. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, (California, USA), pp. 73–76.

Oliveira, L.C., Viana, M.C., Mata, A.I., and Trancoso, I. (2001). Progress Report of Project DIXI+: A Portuguese Text-to-Speech Synthesizer for Alternative and Augmentative Communication (Lisbon, Portugal: Fundação para a Ciência e a Tecnologia).

Papoulis, A. (1991). Probability, Random Variables and Stochastic Processes (McGraw-Hill Companies).

Park, A.S., and Glass, J.R. (2008). Unsupervised Pattern Discovery in Speech. IEEE Trans. Audio Speech Lang. Process. *16*, 186–197.

Pinquier, J., and André-Obrecht, R. (2004). Jingle Detection and Identification in Audio Documents. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, (Montreal, Canada), pp. 329–332.

Platt, J. (1998). Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines (Microsoft Research).

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The Kaldi Speech Recognition Toolkit. In

Proc. of the Workshop on Automatic Speech Recognition and Understanding, (Hawaii, US: IEEE Signal Processing Society),.

Proença, J., Celorico, D., Veiga, A., Candeias, S., and Perdigão, F. (2013). Acoustical Characterization of Vocalic Fillers in European Portuguese. In Proc. of the 6th Workshop on Disfluency in Spontaneous Speech, (Stockholm, Sweden: ISCA), pp. 63–66.

Público (2013). Jornal Público, http://www.publico.pt.

Rabiner, L.R. (1990). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. In Readings in Speech Recognition, A. Waibel, and K.-F. Lee, eds. (San Francisco, California, U. S. A.: Morgan Kaufmann), pp. 267–296.

Reynolds, D.A., and Rose, R.C. (1995). Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions on Audio, Speech, and Language Processing *3*, 72–83.

Riccardi, G., and Hakkani-Tur, D. (2003). Active and Unsupervised Learning for Automatic Speech Recognition. In Proceedings of the European Conference on Speech Communication and Technology, (Geneva, Switzerland), pp. 1825–1828.

Riedmiller, M., and Braun, H. (1993). A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In Proc. of the IEEE International Conference on Neural Networks, (California, USA), pp. 586–591.

Rohlicek, J.R., Russell, W., Roukos, S., and Gish, H. (1989). Continuous Hidden Markov Modeling for Speaker-Independent Word Spotting. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, pp. 627–630.

Rose, R.C., and Paul, D.B. (1990). A Hidden Markov Model Based Keyword Recognition System. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, pp. 129–132.

Rybach, D., Gollan, C., Heigold, G., Hoffmeister, B., Lööf, J., Schlüter, R., and Ney, H. (2009). The RWTH Aachen University Open Source Speech Recognition System. In Proc. of the 10th Annual Conference of the International Speech Communication Association, (Brighton, United Kingdom: ISCA), pp. 2111–2114.

Sanchez, M.H., Vergyri, D., Ferrer, L., Richey, C., Garcia, P., Knoth, B., and Jarrold, W. (2011). Using Prosodic and Spectral Features in Detecting Depression in Elderly Males. In Proc. of the 12th Annual Conference of the International Speech Communication Association, (Florence, Italy: ISCA), pp. 3001–3004.

Santos, D., and Rocha, P. (2001). Evaluating CETEMPúblico, a Free Resource for Portuguese. In Proc. of the 39th Annual Meeting on Association for Computational Linguistics, (Toulouse, France), pp. 450–457.

Shriberg, E. (1995). Acoustic Properties of Disfluent Repetitions. In Proc. of the XIIIth International Congress of Phonetic Sciences, (Stockholm, Sweden), pp. 384–387.

Silva, B., Mendes, H., Lopes, C., Veiga, A., and Perdigão, F. (2009). A Fast Discriminative Training Algorithm for Minimum Classification Error. In Proc. of the I Joint SIG-IL/Microsoft Workshop on Speech and Language Technologies for Iberian Languages, (Porto Salvo, Portugal), pp. 53–56.

Simon (2013). Simon listens, http://www.simon-listens.org.

SpeechDat (1998). Databases for the Creation of Voice Driven Teleservices, http://www.speechdat.org/SpeechDat.html.

Stolcke, A. (2002). SRILM – an Extensible Language Modeling Toolkit. In Proceedings of of the International Conference on Spoken Language Processing, (Denver, USA), pp. 901–904.

Stouten, V., Demuynck, K., and Van hamme Hugo (2008). Discovering Phone Patterns in Spoken Utterances by Non-Negative Matrix Factorization. IEEE Signal Process. Lett. *15*, 131–134.

Taylor, P. (2005). Hidden Markov Models for Grapheme to Phoneme Conversion. In Proc. of the 9th European Conference on Speech Communication and Technology, (Lisbon, Portugal), pp. 1973–1976.

Teixeira, J.P. (2004). A Prosody Model to TTS Systems. PhD. Universidade do Porto, Faculdade de Engenharia.

Teixeira, A., Oliveira, C., and Moutinho, L. (2006). On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme-Phone Conversion. In Computational Processing of the Portuguese Language, R. Vieira, P. Quaresma, M. das G.V. Nunes, N.J. Mamede, C. Oliveira, and M.C. Dias, eds. (Springer Berlin Heidelberg), pp. 212–215.

Teixeira, J.P., Freitas, D.R., Gouveia, P., Olaszy, G., and Németh, G. (1998). MULTIVOX - Conversor Texto Fala Para Português. In Anais Do III Encontro Para O Processamento Computacional de Português Escrito E Falado, (Porto Alegre, Brazil), pp. 88–98.

Thomas, S., Seltzer, M.L., Church, K., and Hermansky, H. (2013). Deep Neural Network Features and Semi-Supervised Training for Low Resource Speech Recognition. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, (Vancouver, Canada),.

Trancoso, I., Viana, M.C., Silva, F.M., Marques, G.C., and Oliveira, L.C. (1994). Rule-Based vs Neural Network-Based Approaches to Letter-To-Phone Conversion for Portuguese Common and Proper Names. In Proc. of the 3rd International Conference on Spoken Language Processing, (Yokohama, Japan), pp. 1767–1770.

Tree, J.E.F., and Clark, H.H. (1997). Pronouncing “the” as “thee” to Signal Problems in Speaking. Cognition *62*, 151–167.

Varadarajan, B., Khudanpur, S., and Dupoux, E. (2008). Unsupervised Learning of Acoustic Sub-word Units. In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies, (Ohio, USA), pp. 165–168.

Veiga, A., Lopes, C., and Perdigão, F. (2010a). Speaker Diarization Using Gaussian Mixture Turns and Segment Matching. In Proc. of the GTM, RTTH and SIG-IL VI Jornadas En Tecnología Del Habla and II Iberian SLTech Workshop, (Vigo, Spain),.

Veiga, A., Candeias, S., Sá, L., and Perdigão, F. (2010b). Using Coarticulation Rules in Automatic Phonetic Transcription. In Proc. of the International Conf. on Computational Processing of Portuguese, (Porto Alegre, Brazil),.

Veiga, A., Candeias, S., and Perdigão, F. (2011a). Conversão de Grafemas para Fonemas em Português Europeu - Abordagem Híbrida com Modelos Probabilísticos e Regras Fonológicas. Linguamática *3*, 39–51.

Veiga, A., Candeias, S., and Perdigão, F. (2011b). Generating a Pronunciation Dictionary for European Portuguese Using a Joint-Sequence Model with Embedded Stress Assignment. In Proc. of the 8th Brazilian Symposium in Information and Human Language Technology, (Cuiabá, Mato Grosso, Brazil), pp. 144–153.

Veiga, A., Candeias, S., Lopes, C., and Perdigão, F. (2011c). Characterization of Hesitations Using Acoustic Models. In Proc. of the 17th International Congress of Phonetic Sciences, (Hong Kong, China), pp. 2054–2057.

Veiga, A., Celorico, D., Proença, J., Candeias, S., and Perdigão, F. (2012a). Prosodic and Phonetic Features for Speaking Styles Classification and Detection. In Advances in Speech and Language Technologies for Iberian Languages, D. Torre Toledano, A. Ortega Giménez, A. Teixeira, J. González Rodríguez, L. Hernández Gómez, R. San Segundo Hernández, and D. Ramos Castro, eds. (Springer Berlin Heidelberg), pp. 89–98.

Veiga, A., Candeias, S., Celorico, D., Proença, J., and Perdigão, F. (2012b). Towards Automatic Classification of Speech Styles. In Computational Processing of the Portuguese Language, H. Caseli, A. Villavicencio, A. Teixeira, and F. Perdigão, eds. (Springer Berlin Heidelberg), pp. 421–426.

Veiga, A., Proença, J., Celorico, D., Candeias, S., and Perdigão, F. (2012c). Vocalic Filler Characterization using Acoustical Properties Derived from Cepstrum. In Proc. of the VII Jornadas En Tecnología Del Habla and III Iberian SLTech Workshop, (Madrid, Spain), pp. 260–268.

Veiga, A., Perdigão, F., and Sá, L. (2012d). Using Phoneme Acoustic Models for Fast Word Spotting. In Proc. of the Portuguese Conf. on Pattern Recognition, pp. 77–78.

Veiga, A., Candeias, S., and Perdigão, F. (2013a). Generating a Pronunciation Dictionary for European Portuguese Using a Joint-Sequence Model with Embedded Stress Assignment. Journal of the Brazilian Computer Society *19*, 127–134.

Veiga, A., Candeias, S., and Perdigão, F. (2013b). Developing a Hybrid Grapheme to Phoneme Converter for European Portuguese. In Proc. of the Conf. on Telecommunications, (Castelo Branco, Portugal), pp. 297–300.

Viana, M.C. (1989). Para a Síntese da Entoação em Português. PhD. Universidade de Lisboa.

Viterbi, A.J. (1967). Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. IEEE Transactions on Information Theory *13*, 260–269.

VOP (2010). Vocabulário Ortográfico do Português, http://www.portaldalinguaportuguesa.org.

Wang, A. (2006). The Shazam Music Recognition Service. CACM Communications of the ACM *49*, 44–48.

Wang, D., and King, S. (2011). Letter-to-Sound Pronunciation Prediction Using Conditional Random Fields. IEEE Signal Processing Letters *18*, 122–125.

Wang, L., Gales, M.J.F., and Woodland, P.C. (2007). Unsupervised Training for Mandarin Broadcast News and Conversation Transcription. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (Hawaii, USA), pp. 353–356.

Wang, Z., Schultz, T., and Waibel, A. (2003). Comparison of Acoustic Model Adaptation Techniques on Non-native Speech. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, (Hong Kong, China), pp. 540–543.

Weintraub, M. (1993). Keyword-Spotting Using SRI's DECIPHER Large-Vocabulary Speech-Recognition System. In Proc. of the International Conference on Acoustics, Speech, and Signal Processing, pp. 463–466.

Wells, J.C. (1997). SAMPA Computer Readable Phonetic Alphabet. In Handbook of Standards and Resources for Spoken Language Systems, D. Gibbon, R. Moore, and R. Winski, eds. (Berlin and New York: Mouton de Gruyter), pp. 60–107.

Wessel, F., and Ney, H. (2005). Unsupervised Training of Acoustic Models for Large Vocabulary Continuous Speech Recognition. IEEE Trans. Speech Audio Process. *13*, 23–31.

Witten, I.H., and Bell, T.C. (1991). The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. IEEE Transactions on Information Theory *37*, 1085–1094.

Woodland, P.C. (2002). The Development of the HTK Broadcast News Transcription System: An Overview. Speech Communication *37*, 47–67.

Young, S., Russell, N.H., and Thornton, M. (1989). Token Passing: a Simple Conceptual Model for Connected Speech Recognition Systems (Cambridge, United Kingdom: Cambridge University Engineering Department).

Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., et al. (2006). The HTK Book (for HTK Version 3.4) (Cambridge, United Kingdom: Cambridge University Engineering Department).

Yu, D., Varadarajan, B., Deng, L., and Acero, A. (2010). Active Learning and Semi-Supervised Learning for Speech Recognition: A Unified Framework Using the Global Entropy Reduction Maximization Criterion. Computer Speech & Language *24*, 433–444.

Zavaliagkos, G., Siu, M.-H., Colthurst, T., and Billa, J. (1998). Using Untranscribed Training Data to Improve Performance. In Proc. of the 5th International Conference on Spoken Language Processing, (Sydney, Australia: ISCA),.

Zdansky, J., and David, P. (2004). Automatic Audio Segmentation of Tv Broadcast News. In Proceedings of the International Conference Radioelektronika 2004, (Bratislava, Slovakia), pp. 358–361.

Zelenák, M., Schulz, H., and Hernando, J. (2010). Albayzin 2010 Evaluation Campaign: Speaker Diarization. In Proc. of the GTM, RTTH and SIG-IL VI Jornadas En Tecnología Del Habla and II Iberian SLTech Workshop, (Vigo, Spain),.

Zhang, R., and Rudnicky, A.I. (2006). A New Data Selection Approach for Semi-Supervised Acoustic Modeling. In Proc. of the IEEE International Conference on Acoustics, Speech and Signal Processing, (Toulouse, France), pp. 421–424.

Zweig, G., and Picheny, M. (2004). Advances in Large Vocabulary Continuous Speech Recognition. Advances in Computers *60*, 249–291.

Anexo I – Alfabeto SAMPA, extensões Unicaráter e IPA

nº	Símbolo SAMPA	Unicaráter (SAMPAuc)	Símbolo IPA	Exemplo
Vogais				
1	6	6 | ª	ɐ	d**a**, cr**â**nio, v**e**nho, **a**m**ê**ijo**a**
2	a		a	p**a**la, p**á**, **à**
3	@		ə	d**e**
4	e		e	d**e**do, v**ê**
5	E		ɛ	p**e**le, p**é**
6	i		i	v**i**, a**í**, r**e**al, henr**y**
7	o		o	**o**co, av**ô**, l**ou**co
8	O		ɔ	p**o**te, p**ó**
9	u		u	t**u**, ba**ú**, at**o,** ki**w**i
10	6~	ã	ɐ̃	v**ã**, **an**ca, **ân**sia, **am**pla, **âm**bar,
11	e~	ë	ẽ	ag**ên**cia, p**en**te, **êm**bolo, **em**pate
12	i~	ï	Ĩ	**in**ca, s**im**, **ím**pio, **ín**dio
13	o~	õ	õ	i**õe**s, c**ôn**sul, t**ôm**bola, p**on**te
14	u~	ü	ũ	n**ún**cio, **un**s, at**um,** c**úm**plice
Semivogais				
15	j		j	bo**i**
16	w		w	ca**u**le
17	j~	ì	j̃	feijõ**e**s
18	w~	ù	w̃	cã**o**
Plosivas/Oclusivas				
19	b		b	**b**e**b**er
20	d		d	**d**a**d**o
21	g		g	**g**ato
22	p		p	**p**ato
23	t		t	**t**oca
24	k		k	**q**uando, **c**asa, **k**iwi

Fricativas				
25	f		f	**f**é
26	s		s	**s**ol, ca**ç**a, au**x**ílio, **c**ima, a**ss**im
27	S		ʃ	**ch**ave, pá**s**, pa**z**, **x**á
28	v		v	**v**ida
29	z		z	ca**s**a, **z**ebra, e**x**emplo
30	Z		ʒ	**j**á, **g**ira, me**s**mo, e**x**-líder
Líquidas				
31	l		l	**l**ua
32	L		ʎ	ve**lh**o
33	r		ɾ	ca**r**o
34	R		ʀ	ca**rr**o, **r**ato
Consoantes nasais				
35	m		m	**m**ão
36	n		n	**n**ada
37	J		ɲ	se**nh**a