Conversão de Grafemas para Fonemas em Português Europeu – Abordagem Híbrida com Modelos Probabilísticos e Regras Fonológicas

Arlindo Veiga
Instituto de Telecomunicações,
Polo de Coimbra
DEEC, Universidade de Coimbra
aveiga@co.it.pt

Sara Candeias
Instituto de Telecomunicações,
Polo de Coimbra
saracandeias@co.it.pt

Fernando Perdigão
Instituto de Telecomunicações,
Polo de Coimbra
DEEC, Universidade de Coimbra
fp@co.it.pt

Resumo

A conversão de grafema para fonema diz respeito à tarefa de encontrar a pronúncia de um vocábulo dado na sua forma escrita, a qual tem uma forte componente de aplicação em sistemas de reconhecimento e de síntese de fala. Uma nova abordagem na conversão de grafema para fonema é proposta, aplicando um modelo híbrido para o qual concorrem regras fonológicas e decisões estatísticas. Os resultados mostram que a incorporação de regras fonológicas em algoritmos de informação estatística melhora acentuadamente o desempenho do conversor. Para este trabalho, foi construído um dicionário de pronunciação com mais de 40000 vocábulos derivados do corpus CETEMPúblico. Os dicionários fonológicos de pronunciação para o português europeu, bem como outros recursos produzidos durante este trabalho, estão disponibilizados publicamente. O sistema que aqui se descreve foi aplicado à língua portuguesa escrita, sem e com o Acordo Ortográfico de 1990, e, ainda que aplicado ao português na sua vertente europeia, observa características que permitem a sua aplicação a outras línguas românicas.

1. Introdução

Comummente designado por G2P[1], o mapeamento entre grafemas e fonemas tem por objetivo converter um texto escrito numa sequência de símbolos que representam os sons da fala de uma determinada língua, de uma forma inequívoca. Numa atualidade motivada para o uso de várias aplicações tecnológicas ativadas pela fala, a conversão de G2P tem sido alvo de estudos e de desenvolvimento. Apesar de ser já sólida e madura a investigação na área, em português europeu (PE), 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 sido já diagnosticada, e que envolve questões de natureza essencialmente morfológica e sintática (cf., a título de exemplo, (Braga *e* 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. O presente estudo levou à disponibilização, em (SPL, 2011), de *i*) dicionários de pronunciação; *ii*) modelos de sequências de conjuntos grafema-fonema e de *iii*) programas que permitem converter grafemas em fonemas.

No sentido de encontrar uma solução para os problemas da conversão G2P para o PE, são várias as abordagens que têm vido a ser propostas, de entre as quais destacamos as seguintes: *i*) por regras linguísticas, expostas em (Braga *et al.*, 2006), (Oliveira *et al*., 1992) e (Teixeira, 2004); *ii*) por regras inferidas a partir dos dados (Teixeira *et al*., 2006); *iii*) por máquinas de estados finitos (Caseiro *e* Trancoso, 2002), (Oliveira *et al.*, 2004); *iv*) por máxima entropia (Barros *e* Weiss, 2006); *v*) baseadas em redes neuronais (Trancoso *et al.*, 1994); e *vi*) por CARTs - Classification and Regression Trees (Oliveira *et al.*, 2001). Uma das técnicas a referenciar, e que tem sido aplicada essencialmente em línguas que não apresentam uma clara correspondência entre grafema e fone(ma), como é o caso do inglês, é a abordagem por modelos probabilísticos, apresentado por (Demberg *et al.*, 2007) e por (Bisani *e* Ney, 2008). Contrastando com as abordagens baseadas em regras, as quais são suportadas por um conhecimento linguístico da língua, que se pretende exaustivo, a abordagem estatística baseia-se no pressuposto de que a pronunciação de um vocábulo é possível de ser prevista, por analogia, a partir de exemplos de sequências suficientes de grafonemas – unidades identificativas da associação entre grafema e respetivo fonema (cf. (Bisani *e* Ney, 2008)). Uma das vantagens apontada pelos modelos probabilísticos é de não implicar uma verificação constante da interdependência das

[1] Do inglês *Grapheme to Phoneme*; por vezes também designado L2S: *Letter to Sound*.

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 G2P proveniente de modelos probabilísticos não capta um contexto suficientemente amplo de forma a impedir que a estrutura fonológica da língua seja violada. 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. Estas características explicam o sucesso da aplicação de regras linguísticas tais como a marcação de sílaba tónica (descrito em (Candeias *e* Perdigão, 2008), (Braga *et al.*, 2006), (Almeida *e* Simões, 2001) e (Teixeira *e* Freitas, 1998), como exemplos para o PE).

O estudo que aqui se apresenta propõe uma abordagem híbrida ao problema da conversão G2P, onde se utiliza um modelo probabilístico, no qual são incorporadas regras fonológicas.

Este artigo encontra-se estruturado da seguinte forma: a Secção 2 descreve o modelo probabilístico de sequências conjuntas e a Secção 3 descreve a criação do modelo híbrido usando regras fonológicas. Na Secção 4 são apresentados os resultados da contribuição de cada componente do módulo. A Secção 5 apresenta as conclusões. Estudos futuros são igualmente referidos. Parte deste trabalho foi apresentado parcialmente em (Veiga *et al.*, 2011).

2. Modelo probabilístico de sequências conjuntas

A tarefa de conversão de grafemas para fonemas pode ser formulada na 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 uma sequência de N grafemas e $F = F_1^M = \{f_1, f_2, \cdots, f_M\}$ como uma sequência de M fonemas, a determinação da sequência ótima de fonemas, F^*, é descrita da seguinte forma:

$$F^* = \arg\max_F P(F \mid G). \qquad (1)$$

Não sendo fácil determinar F^* calculando diretamente a probabilidade *a posteriori* $P(F \mid G)$ para todas as sequências F possíveis, podemos usar o teorema de Bayes e rescrever o problema como:

$$F^* = \arg\max_F \frac{P(G \mid F)}{P(G)} P(F). \qquad (2)$$

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:

$$F^* = \arg\max_F P(G \mid F) P(F). \qquad (3)$$

A estimação de $P(F)$ é feita usualmente recorrendo aos modelos "n-grama". Quanto à determinação de $P(G|F)$, 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 $P(G|F)$ pode ser decomposto da seguinte forma:

$$P(G \mid F) = \prod_{n=1}^{N} P(g_n \mid F). \qquad (4)$$

Esta simplificação parte de princípio que a dependência entre fonemas é suficiente para modelar o problema e que os contextos de fonemas replicam os contextos de grafemas (cf. (Taylor, 2005), (Demberg, 2006) e (Jiampojamarn *e* Kondrak, 2009)).

Existem outras abordagens que propõem a utilização de modelos de probabilidade conjunta, $P(F,G)$, para determinar a sequência ótima de fonemas usando diretamente a expressão da probabilidade conjunta em (1) no lugar da probabilidade condicional (em (Bisani *e* Ney, 2002) e (Galescu *e* Allen, 2001)). Estas abordagens possibilitam a modelação da dependência entre grafemas, a dependência entre fonemas e a dependência entre grafemas e fonemas.

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. Em comparação com a que usa modelos de Markov, a abordagem que reclama de um algoritmo de alinhamento apresenta um melhor desempenho, pelo que é a seguida neste trabalho.

2.1 Alinhamento entre grafemas e fonemas

Muitos grafemas do português têm uma correspondência unívoca com os fonemas, situação na qual a conversão G2P é direta. É o caso de muitas das consoantes, como o <p> e o <t>, em <**p**or**t**uguês> que são diretamente convertidos nos

fonemas /p/ e /t/, respetivamente. 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 <por**r**tug**u**ês>) e o estatuto morfológico[2], 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/[3] ou /E/ e /o/ ou /O/, respetivamente: <s**e**lo> (nome) →/s**e**lu/, <s**e**lo> (verbo) → /s**E**lu/; <**o**lho> (nome) → /**o**Lu/, <**o**lho> (verbo) → /**O**Lu/). Existem situações em que um único grafema pode originar vários fonemas, assim como existem situações em que vários grafemas podem originar um único fonema (como <g> e <j> → /Z/, conversão esta igualmente dependente de contexto). Todas as abordagens estatísticas deparam-se com este problema, sendo necessário, durante o processo de treino, segmentar e alinhar as duas sequências 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 em dois tipos:

1) **"um-para-um"**

Neste tipo de alinhador cada grafema é associado a apenas um fonema, originando segmentos com apenas um símbolo. Ainda assim, é necessário utilizar um símbolo nulo ('_') 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 PE, uma vez que ocorrem em pouquíssimos contextos, facilmente identificados, como é o caso do iode que ocorre em algumas das estruturas, tais como em <extra> /6iStr6/. 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. Na literatura da área, denomina-se "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".

2) **"muitos-para-muitos"**

Neste tipo 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 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 *e* Ney, 2008). Este tipo de associação é também conhecido como alinhamento "*m*–*n*".

2.2 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 *e* Ney, 2002). Mostra-se um exemplo com o vocábulo <compõem> no qual os grafonemas estão entre parênteses retos. Neste exemplo considerou-se, tanto quanto possível, um alinhamento de "um-para-um", mas onde se admitem casos de alinhamento de "2-para-1" e de "1-para-2".

$$\begin{matrix} \textit{Grafemas} \\ \textit{Fonemas} \end{matrix}\begin{bmatrix} \text{c} \\ \text{k} \end{bmatrix}\begin{bmatrix} \text{om} \\ \text{o\textasciitilde} \end{bmatrix}\begin{bmatrix} \text{p} \\ \text{p} \end{bmatrix}\begin{bmatrix} \tilde{\text{o}} \\ \text{o\textasciitilde i\textasciitilde} \end{bmatrix}\begin{bmatrix} \text{e} \\ \text{6\textasciitilde} \end{bmatrix}\begin{bmatrix} \text{m} \\ \text{i\textasciitilde} \end{bmatrix}.$$

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

$$F^* = \arg\max_F P\big(Q(F,G)\big). \qquad (5)$$

Dada uma sequência de *K* grafonemas, $Q(F,G)$, e não admitindo independência entre símbolos, a probabilidade da sequência, $P\big(Q(F,G)\big)$, pode ser calculada como:

$$P\big(Q(F,G)\big) = P(q_1)P(q_2 \mid q_1)P(q_3 \mid q_1 q_2)\cdots P(q_K \mid q_1 q_2 \cdots q_{K-1}) \qquad (6)$$

No modelo estatístico é frequente limitar-se o contexto (ou história) dos grafonemas utilizando os chamados modelos "*n*-grama", que correspondem a

[2] Em inglês, PoS – *part-of-speech*.
[3] Alfabeto SAMPA; cf. Tabela 1.

sequências limitadas a um comprimento até n símbolos. Deste modo, a equação (6) pode ser aproximada a:

$$P\big(Q(F,G)\big) \approx \prod_{i=1}^{K} P(q_i \mid q_{i-n+1} \cdots q_{i-1}) \,. \qquad (7)$$

2.3 Estimação do modelo

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

$$P(q_i \mid q_{i-n+1} \cdots q_{i-1}) = \frac{C(q_{i-n+1} \cdots q_i)}{C(q_{i-n+1} \cdots q_{i-1})} \,, \qquad (8)$$

onde

$$C(q_{i-n+1} \cdots q_{i-1}) = \sum_j C(q_{i-n+1} \cdots q_{i-1} q_j) \,. \qquad (9)$$

Ainda que esta probabilidade seja de cálculo simples, e acarreta o problema de atribuir probabilidade zero aos "n-grama" que não estão presentes no dicionário de treino. Além disso, podem existir "n-grama" que estão presentes no dicionário mas em número sem significado estatístico. Para evitar estes constrangimentos, é preciso precaver a existências de sequências que nunca foram encontradas no dicionário de treino (usando os chamados "descontos"), ou que são pouco frequentes (a "suavização")[4]. Assim, uma pequena massa de probabilidade é retirada dos "n-grama" mais frequentes e é reservada para os "n-grama" ausentes ou pouco frequentes no dicionário de treino.

Existem vários algoritmos propostos para resolver o problema da redistribuição da massa de probabilidade. São exemplos os algoritmos de desconto (de Good-Turing (Good, 1953), de Witten-Bell (Witten *e* Bell, 1991), de Kneser-Ney (Kneser *e* Ney, 1995)), de desconto absoluto de Ney (Ney *et al.*, 1994) e de suavização de Katz (Katz, 1987).

A estimação da probabilidade de "n-grama" com frequência inferior a um dado limiar é feita à custa de $(n-1)$-gramas (*backoff*).

O algoritmo implementado neste trabalho é igual ao utilizado em (Demberg *et al.*, 2007). Faz a "suavização" por interpolação e utiliza a versão modificada do algoritmo de Kneser-Ney (Chen *e* Goodman, 1998). O valor de n varia de 2 a 8 conforme será descrito na secção 4.

3. Criação do modelo híbrido

Nesta secção descreve-se um modelo híbrido em que se opera uma transformação da sequência de grafemas, introduzindo novos símbolos com significado fonológico preciso, proporcionando desta forma uma integração de regras fonológicas no modelo estatístico. O modelo estatístico não é alterado; apenas passam a existir mais símbolos em que a associação entre grafema e fonema é mais precisa.

3.1 Vocabulário

Originando o sistema de conversão um dicionário de pronunciação, foi necessário, numa primeira fase, ter disponível como base de trabalho uma listagem de vocábulos atuais e representativos do PE. O material utilizado para esse fim foi o corpus CETEMPúblico (Santos *e* Rocha, 2001), o qual contem 180 milhões de palavras[5], advindas de uma coleção de extratos do jornal Público de entre os anos 1991 e 1998.

O processo de criação dessa listagem consistiu em tomar todas as cadeias de carateres anotadas como palavras, obedecendo simultaneamente aos seguintes critérios: *i*) começar com um grafema do alfabeto português (a-z, A-Z, á-ú, Á-Ú); *ii*) não conter dígitos; *iii*) não apresentar todos os grafemas em maiúscula (caso de siglas); *iv*) não conter o caráter '.' (caso de URLs); *v*) terminar com um grafema do alfabeto português ou com '-'; *vi*) o lema correspondente não conter o caráter '=' (caso de nomes compostos).

A partir do resultado obtido, formou-se uma lista de cerca de 50k 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 cerca de 50k vocábulos. Por fim, foram retirados quer vocábulos estrangeiros quer estrangeirismos, usando, em primeiro lugar, critérios automáticos e, seguidamente, uma confirmação manual. A pesquisa automática excluiu todos os vocábulos que apresentavam grafemas ou sequências grafemáticas que não

[4] Em inglês, *discount* e *smoothing*, respectivamente.

[5] Por palavras entendem-se, aqui, todos os átomos do corpus que contêm, pelo menos, um grafema ou dígito.

fazem parte do sistema do PE, tais como <k>, <w> e <y>; <sh> e <pp>; e <b>, <d> ou <p> em posição final de vocábulo. Alguns destes dados serão depois base de constituição de um dicionário de pronunciação de estrangeirismos (cf. descrito em 3.2). Como resultado final deste processo, consistiu-se uma lista de cerca de 40k vocábulos, os quais correspondem ao vocabulário de referência tomado para este trabalho, referenciado infra 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 (AO), houve necessidade de se constituir uma listagem adicional com vocábulos grafados de acordo com o AO. Com esse fim, usou-se a ferramenta Lince (Lince) para converter os vocábulos na nova grafia. Dos 41586 vocábulos utilizados no vocabulário pré-AO, 915 sofreram alterações de grafia, nomeadamente a eliminação das consoantes mudas (<c> e <p>), a eliminação dos hífenes e alteração de acentuação gráfica. De acordo com a possibilidade de coexistirem duas grafias, este novo vocabulário apresenta pares de vocábulos ditos 'parónimos', tais como <conceptual> e <consetual> ou <desconectar> e <desconetar>. O vocabulário pós-AO é constituído por 41598 vocábulos, sendo referenciado como "voc_CETEMP_40k_ao".
Nas secções seguintes não é feita a distinção entre estes dois vocabulários, a não ser quando pertinente, como é o caso da secção dos resultados.

3.2 Transcrição fonológica

A transcrição fonológica do vocabulário de referência foi feita por um processo iterativo. Em primeiro lugar, foi feito um modelo estatístico, conforme descrito em 2.2, tendo por base o dicionário de pronunciações da base de dados SpeechDat (SpeechDat) com cerca de 15k vocábulos. Para a constituição do dicionário foram retirados os estrangeirismos e foram feitas algumas correções de pronunciação. Este dicionário foi também convertido para notação SAMPA (Wells, 1997), convencionando-se que os símbolos representativos das glides [j] e [w] fossem notados como as vogais correspondentes, por razões de uniformização que ultrapassam o âmbito deste artigo. Com o mesmo princípio de uniformização, 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]. Admitiu-se, igualmente, a necessidade de inclusão do iode, como foi já observado em 2.1, nomeadamente para nos aproximarmos, o mais possível, do PE padronizado. Neste ponto, requer-se esclarecer que os símbolos SAMPA adotados (cf. Tabela 1) resultaram de uma ponderação cuidada sobre representatividade do PE falado. A observação atenta dos alfabetos fonéticos SAMPA para o Português (SAMPA-PT) e X-SAMPA, dá 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]).
Relativamente à transcrição, uma dilucidação acerca da opção pela transcrição fonológica, e não fonética, é ainda devida. No que concerne o binómio fonética/fonologia, é comummente aceite pela comunidade linguística, que a fonética diz respeito às propriedades físicas e articulatórias de todos os sons que ocorrem na produção linguística, cabendo à fonologia o estudo da função de cada som pronunciado numa dada língua, a qual permite ao falante distinguir significados. É igualmente aceite que qualquer opção metodológica no que à análise da fala diz respeito, liga, inevitavelmente, as duas faces do binómio, uma vez que lida tanto com a relação que existe entre as unidades e a sua pertinência na língua falada (i.e., os fonemas) como com a realidade física que resulta na pronunciação dessas mesmas unidades (i.e., os fones e alofones) (cf. definições dos termos em (Crystal, 2001)). Tem sido frequente a alternância, muitas vezes não claramente justificada, entre os termos fone e fonema nos vários estudos efetuados no âmbito do G2P (cf., a título de exemplo, que a unidade fone é a adotada em (Caseiro *e* Trancoso, 2002) enquanto que (Barros *e* Weiss, 2006) apresenta o fonema como o resultado da conversão do grafema). Neste estudo, consideramos 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 (vulgo 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, aceitamos a unidade fonológica, ou fonema, como uma classe à qual pode corresponder um fone ou um feixe de realizações alofónicas disponíveis no PE (acolhendo-se, assim, a possível inserção de pronúncias alternativas). A transcrição fonológica resultante corresponde ao PE que admitimos como padronizado e não representa qualquer arquifonema ou neutralização de oposições. A transcrição é registada entre barras oblíquas, fazendo uso do alfabeto SAMPA, conforme descrito supra.
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.

Tabela 1 – Símbolos SAMPA e símbolos unicaráter (uc) associados a grafemas possíveis com vocábulos exemplificativos.

SAMPA	uc	Grafemas possíveis	Exemplos
6		a, e, â	c**a**ma, s**e**nha, c**â**mara
a		a, á, à	p**á**, p**a**la
@		e	d**e**
e		e, ê	v**ê**, d**e**do
E		e, é,	p**é**, p**e**le
i		i, í, e	v**i**, a**í**, r**e**al
o		o, ô, ou	**o**co, av**ô**, l**ou**co
O		o, ó	p**ó**, p**o**te
u		u, ú, o	t**u**, ti**o**, at**o**, ba**ú**
6~	ã	ã, an, ân, em, am, e, âm, é	br**an**co, **ân**cora, c**am**po, t**em**, l**âm**pada, al**ém**
e~	ë	ên, en, em	p**en**te, ag**ên**cia, **em**pate
i~	ï	i, in, im, ím, ín, m	mu**i**to, tr**in**car, s**im**, **ím**pio, **ín**timo, hom**em**
o~	õ	õ, ôn, ôm, on, om	p**õe**, c**ôn**sul, c**ôm**puto, p**on**te, p**om**bo
u~	ü	u, ún, un, um, úm	m**ui**to, an**ún**cio, at**um,** c**úm**plice
b		b	**b**e**b**er
d		d	**d**a**d**o
g		g, gu	**g**ato, **gu**elra
p		p	**p**ato
t		t	**t**oca
k		qu, c	a**qu**ela, **c**asa
f		f	**f**é
s		s, ç, x, c[eiéí], ss	**s**ol, ca**ç**a, trou**x**e, **c**éu, **c**ima, a**ss**im
S		ch, s, z, x	**ch**ave, pá**s**, pa**z**, **x**á
v		v	**v**ida
z		z, s, x	ca**s**a, **z**ebra, e**x**emplo
Z		j, g, s, z, x	**j**á, **g**ira, de**s**viar, e**x-**bar
l		l	**l**âmpada
L		lh	ve**lh**o
r		r	ca**r**o
R		r, rr	ca**rr**o, **r**ato
m		m	**m**ão
n		n	**n**ada
J		nh	se**nh**a

Seguiu-se então um processo moroso de confirmação 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 representatividade do PE. Deste processo resultou o dicionário de transcrição fonológica que referenciamos 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 subjacentes, 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> /@St"e~su/, não se transcrevendo a sequência <ex> como /6iS/ em certos contextos de atonicidade.

Tabela 2 – Símbolos unicaráter (uc) convencionados a partir de sequências SAMPA, exemplificados com vocábulos. Os primeiros 13 símbolos representam fonemas vocálicos em posição tónica; os restantes 7 representam sequências específicas de fonemas.

SAMPA	uc	Exemplos
"6	â	c**a**ma → /k**"6**m6/ → /k**â**m6/
"a	á	c**a**sa → /k**"a**z6/ → /k**á**z6/
"e	ê	t**e**ma → /t**"e**m6/ → /t**ê**m6/
"E	É	s**e**te → /s**"E**t@/ → /s**É**t@/
"i	í	t**i**ro → /t**"i**ru/ → /t**í**ru/
"o	ô	**o**vo → /**"o**vu/ → /**ô**vu/
"O	Ó	l**o**go → /l**"O**gu/ → /l**Ó**gu/
"u	ú	**u**va → /**"u**v6/ → /**ú**v6/
"6~	Ã	c**am**po → /k**"6~**pu/ → /k**Ã**pu/
"e~	Ë	c**en**tro → /s**"e~**tru/ →/s**Ë**tru/
"i~	Ï	c**in**co →/s**"i~**ku/ →/s**Ï**ku/
"o~	Õ	c**on**to → /k**"o~**tu/ → /k**Õ**tu/
"u~	Ü	ass**un**to → /6s**"u~**tu/ → /as**Ü**tu/
6i	æ	**e**xtrair → /**6i**Str6"ir/ → /**æ**Str6ír/
"6i	Æ	**e**xtra → /**"6i**Str6/ → /**Æ**Str6/
6~i~6~	Ê	t**êm** → /t"**6~i~6~**i~/ → /t**Êï**/
o~i~	¤	p**õe**m → /p"**o~i~**6~i~/→ /p**¤ã**ï/
ks	K	a**x**ila → /a**ks**"il6/ → /a**K**íl6/
ai	Å	c**ae**m → /k"**ai**6~i~/ → /k**Å**ãï/
Oi	®	constr**oe**m→/ko~Str**Oi**6~i~/→ /kÖStr**®**ãï/

Ao longo do desenvolvimento deste trabalho, o dicionário sofreu um processo constante de revisão e de correção. Apesar de admitirmos a presença de alguns erros de transcrição, estamos confiantes na sua precisão, pelo que acreditamos que o dicionário "dic_CETEMP_40k" constitui uma base de trabalho interessante para estudos na língua portuguesa, em especial na área da fonética e da fonologia. Acrescente-se que do processo que acompanhou este estudo, resultou igualmente um

Tabela 3.1 – Possíveis fonemas associados a cada grafema para alinhamento – caso de vogais. A coluna "G" indica os grafemas e "F" os fonemas possíveis. A coluna "Alt" indica os fonemas alternativos, segundo a Tabela 2. A coluna à direita mostra exemplos de transcrição sem e com marcação de acentuação.

G	F	Alt	Exemplos
a	6	â	c**a**ma → /k**6**m6/ → /k**â**m6/
	a	á	m**a**la → /m**a**l6/ → /m**á**l6/
	ã	Ã	c**a**nto → /k**ã**tu/ → /k**Ã**tu/
á,à	a	á	**à**s → /**a**S/ → /**á**S/
			p**á**s → /p**a**S/ → /p**á**S/
ã	ã	Ã	an**ã**o → /6n**ã**ü/ → /6n**Ã**ü/
â	ã	Ã	atl**â**ntico→/6tl**ã**tiku/→ /6tl**Ã**tiku/
	6	â	c**â**mara → /k**6**m6r6/ → /k**â**m6r6/
e	6	â	des**e**nho →/d@z**6**Ju/→ /d@z**â**Ju/
	@,e	ê	aqu**e**le → /6k**e**l@/ → /6k**ê**l@/
	E	É	p**e**le → /p**E**l@/ → /p**É**l@/
	i		ar**e**al → /6r**i**al/ → /6r**i**ál/
	ë	Ë	v**e**nto → /vëtu/ → /vËtu/
	ï		visõ**e**s → /vizõïS/ → /vizÕïS/
	6i	æ	**e**xtrair → /**6i**Str6ir/ → /**æ**Str6ir/
	6i	Æ	**e**xtra → /**6i**Str6/ → /**Æ**Str6/
	ã	Ã	t**e**ns → /tãïS/ → /tÃïS/
			vot**e**m → /vOtãï/ → /vÓtãï/
é	E	É	caf**é** → /k6f**E**/ → /k6f**É**/
	ã	Ã	cont**é**m → /kõnt**ã**ï/ → /kõt**Ã**ï/
ê	6	â	am**ê**ijoa→/6m**6**iZu6/→/6m**â**iZu6/
	e	ê	voc**ê** → /vOs**e**/ → /vOs**ê**/
	ë	Ë	p**ê**ndulo → /p**ë**dulo/ → /p**Ë**dulo/
	6i	Æ	**ê**xito → /**6i**zitu/ → /**Æ**zitu/
	ãïã	Ê	t**ê**m → /tãïã**ï**/ → /tÊï/
i,í	i	í	c**i**ma → /s**i**m6/ → /s**í**m6/
			sa**í** → /s6**i**/ → /s6**í**/
	ï	Ï	c**i**nco → /sïku/ → /s**Ï**ku/
			l**í**mpida → /l**ï**pid6/ → /l**Ï**pid6/
o	o	ô	c**o**rpo → /k**o**rpu/ → /k**ô**rpu/
	O	Ó	c**o**po → /k**O**pu/ → /k**Ó**pu/
	õ	Õ	c**o**nto → /k**õ**tu/ → /k**Õ**tu/
	ü		visã**o** → /vizã**ü**/ → /vizÃ**ü**/
	u	ú	p**o**rque → /p**u**rk@/ → /p**ú**rk@/
ó	O	Ó	c**ó**pia → /k**O**pi6/ → /k**Ó**pi6/
ô	o	ô	av**ô** → /6v**o**/ → /6v**ô**/
	õ	Õ	c**ô**nsul → /k**õ**sul/ → /k**Õ**sul/
õ	õ	Õ	p**õ**e → /p**õ**ï/ → /p**Õ**ï/
	õï	¤	p**õ**em → /p**õ**ïã**ï**/ → /p**¤**ãï/
u,ú	u	Ú	ap**u**ra → /6p**u**r6/ → /6p**ú**r6/
			t**ú**nel → /t**u**nEl/ → /t**ú**nEl/
	ü		**u**nto → /**ü**tu/ → /**Ü**tu/
		Ü	an**ú**ncio → /6n**ü**siu/ → /6n**Ü**siu/
		_	q**u**ente → /këthough/ → /k_Ët@/

dicionário de pronunciação de cerca de 1300 estrangeirismos. Este dicionário de estrangeirismos será incorporado no sistema final de conversão G2P, como tabela de exceções.

Tabela 3.2 – Continuação da Tabela 3.1 – caso de consoantes.

G	F	Alt	Exemplos
c	k		**c**apa → /**k**ap6/ → /**k**áp6/
	s		**c**edo → /**s**edu/ → /**s**êdu/
	ks	K	fi**cc**ional →/fi**ks**iunal/ → /fi**K**iunál/
		_	a**c**tuar → /6tuar/ → /6_tuár/ (pré-AO)
g	g		**g**atu → /**g**atu/ → /**g**átu/
	Z		**g**irafa → /**Z**iraf6/ → /**Z**iráf6/
l	l		**l**ado → /**l**adu/ → /**l**ádu/
	L		ma**l**ha → /ma**L**6/ → /má**L**_6/
m	m		**m**al → /**m**al/ → /**m**ál/
	ü		apela**m** → /6pElã**ü**/ → /6pÉlã**ü**/
	ï		vote**m** → /vOtã**ï**/ → /vÓtã**ï**/
		_	ca**m**po → /kãpu/ → /kÃ_pu/
n	n		ca**n**a → /k6**n**6/ → /k**â**n6/
	ï		te**n**s → /tã**ï**S/ → /tÃ**ï**S/
	J		ma**n**ha →/m6**J**6/→/mâ**J**6/→/mâ**J**_6/
		_	ca**n**to → /kãtu/ → /kÃ_tu/
p	p		**p**ar → /**p**ar/ → /**p**ár/
		_	ó**p**tica → /Otik6/ → /Ó_tik6/ (pré-AO)
r	r		ca**r**o → /ka**r**u/ → /ká**r**u/
	R	_	ca**rr**o → /ka**R**u/→ /ká**R**u/→ /ká**R**_u/
s	s	_	ma**s**sa → /ma**s**6/→ /má**s**6/→ /má**s**_6/
	S		à**s** → /a**S**/ → /á**S**/
	z		ca**s**a → /ka**z**6/ → /ká**z**6/
	Z		abi**s**mo → /6bi**Z**mu/ → /6bí**Z**mu/
x	s		má**x**imo → /ma**s**imu/ → /má**s**imu/
	S		**x**adrez → /**S**6dreS/ → /**S**6drêS/
	z		e**x**ame → /e**z**6m@/ → /e**z**âm@/
	Z		ex-diretor→/6i**Z**dirEtor/→/Æ**Z**dirEtôr/
	ks	K	fi**x**o → /fi**ks**u/ → /fí**K**u/
z	S		arro**z** → /6Ro**S**/ → /6R_ô**S**/
	z		a**z**ar → /6**z**ar/ → /6**z**ár/
	Z		feli**z**mente→/f@li**Z**mët@/→ /f@li**Z**mËt@/

3.3 Alinhador de grafemas com fonemas

Um passo importante na criação do modelo estatístico, no qual cada grafema dá origem a zero ou a um fonema (vertente "1-01"; cf. 2.1, 1)), consiste num passo inicial de alinhamento entre grafemas e fonemas. A opção pelo modelo de "1-01" foi desde logo tomada pela verificação de que em apenas 7 casos um grafema pode dar origem a mais do que um fonema. Esses casos, e respetivos contextos de ocorrência, estão indicados na Tabela 2, nas últimas 7 linhas.

O problema aportado pelo alinhamento segundo a vertente "1-01" foi resolvido definindo símbolos que correspondessem a mais do que um fonema

(por exemplo, definimos o símbolo /Ê/ para representar /6~i~6~/ em "têm";cf. Tabela 2). Com esta solução, cada grafema pode dar origem a um fonema ou a zero fonemas (sendo este último o caso do <n> em <canto> → /k"6~_tu/ ou do <h> inicial em <homem> → /"_Om6~i~/, onde '_' representa o símbolo nulo). A cada fonema foi, então, associado um único símbolo, do conjunto de carateres ISO Latino (ISO-8859-1), tal como se apresentam nas Tabela 1 e Tabela 2.

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). Para tal, foi necessário definir uma distância entre cada grafema e cada fonema. Esta distância, ou custo de associação, foi definida através da equação

$$d(g,p) = -\log_2\left(P(p \mid g)\right), \qquad (10)$$

onde a probabilidade condicional do fonema *p* dado o grafema *g*, $P(p \mid g)$, é estimada a partir de um dicionário de transcrições 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 alternativas de transcrição para cada grafema estão documentadas na Tabela 3.1 e Tabela 3.2 (focando casos de vogais e de consoantes, respetivamente), tendo em conta o alinhamento individual entre grafemas e fonemas. Exemplos para cada alternativa estão também nelas apresentados.

Observando as Tabelas 3.1 e 3.2, pode verificar-se que os grafemas suscetíveis de serem apagados são <u|c|m|n|p|r|s>. Os grafemas <r> e <s> apenas podem sofrer apagamento quando o alinhamento é feito sem o uso da convenção de transformação de dígrafos, explicitados de seguida, em 3.4.1, e indicados na Tabela 4. Os grafemas <c> e <p> são apagados apenas quando se converte o vocabulário grafado na forma prévia à aplicação do AO.

3.4 Regras fonológicas

Apresentando o PE uma certa regularidade fonética e fonológica e uma ortografia de base fonológica, adicionamos ao módulo de G2P constrições linguísticas do PE pertinentes à tarefa de transcrever o grafema em fonema. Assim, foram propostos algoritmos baseados em 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.

Tabela 4 – Grafemas especiais ("G") para dígrafos, associados a fonemas possíveis ("F") e a exemplos convencionados. Os primeiros 7 símbolos representam consoantes; os restantes 11 representam sequências específicas de vogal ditongada e de vogais nasais.

G	Dígrafos	F	Exemplo
C	cc	s, ks	fi**cc**ional → fi**C**ional
Ç	cç	s, ks	fi**cç**ão → fi**Ç**ão
R	rr	R	ca**rr**o → ca**R**o
§	ss	S	ma**ss**a →ma**§**a
L	lh	L	mo**lh**o → mo**L**o
J	nh	J	u**nh**a → u**J**a
S	ch	S	**ch**ave → **S**ave
º	ou	o	d**ou**rada → dºrada
Ä	an, am	ã, Ã	c**an**to →cÄto, c**am**po →cÄpo
Ë	en, em	ë, Ë	s**en**te →sËte, s**em**pre → sËpre
Ï	in, im	ï, Ï	l**im**bo →lÏbo
Ö	on, om	õ, Õ	c**on**to→cÖto,
Ü	un, um	ü, Ü	ass**un**to → a§Üto
Â	ân, âm	Ã	p**ân**tano →pÂtano,
Ê	ên, êm	Ë	**ên**fase → Êfase
Í	ín, ím	Ï	**ín**dio→Índio, l**ím**pido→lÍpido
Ô	ôn, ôm	Õ	c**ôn**sul→cÔsul
Ú	ún, úm	Ü	den**ún**cia→denÚcia, c**úm**plice→cÚplice

3.4.1 Dígrafos

Um dígrafo ocorre quando dois grafemas são pronunciados apenas por um único som. Na Tabela 4 apresentam-se símbolos para representar esses dígrafos.

A nossa proposta altera 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 consonânticas e sequências vocálicas. No âmbito das sequências vocálicas, consideramos a sequência oral <ou>, a qual, seguindo a pronúncia padronizada do PE, corresponde ao fonema singular /o/, e as sequências

nasais <a|e|i|o|u> +<m|n>, admitidas em contexto silábico de <*V*+C_{nasal}> (cf. Tabela 4). Na medida em que o modelo implementado recebe igualmente informação sobre a vogal tónica (cf. 3.4.2), às vogais que no âmbito dos dígrafos apresentam acento gráfico foi-lhes igualmente atribuído um único símbolo (uni caráter).

3.4.2 Marcação de tonicidade

Seguindo os pressupostos teóricos discutidos em (Mateus e d'Andrade, 2000), admitimos tratar-se 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 ($V_{tónica}$) tem sido reconhecida em trabalhos prévios de conversão de G2P, quer para a implementação de regras de transmutação do grafema em fone(ma), 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 $V_{tónica}$ traduziu-se num melhoramento ao modelo estatístico, uma vez que permitiu definir grafonemas de forma unívoca. Assim como em (Andrade *e* Viana, 1985), a nossa proposta considerou ser pertinente a marcação da $V_{tónica}$ (identificada com o símbolo SAMPA ' " ') e não da respetiva unidade silábica.

O processo de identificação de $V_{tónica}$ foi conseguido de uma forma que, tanto quanto nos é dado a perceber, não é usual noutros trabalhos. Atendendo ao contexto vocálico grafemático, de cada vocábulo, se alguma vogal (<*V*>) recebe um acento gráfico, essa <*V*> é identificada como $V_{tónica}$ (cf. Tabela 5, regra 1). Caso não apresente graficamente qualquer marca de tonicidade, analisamos a penúltima <*V*> nos vocábulos terminados em <a>, <o>, <e> ou <m> e nas correspondentes determinações de plural (cf. regra 2, Tabela 5). Excluindo os casos de presença de sequência ditongada, os quais são analisados à parte (regras 5 e 6, Tabela 5), essa <*V*> passa a ter a indicação de tonicidade (cf. regra 6, Tabela 5). Os restantes vocábulos (sem grafemas acentuados), recebem indicação de $V_{tónica}$ em posição oxítona (regras 3 e 4, Tabela 5). A aplicação das regras descritas são suficientes para não marcar tonicidade nos vocábulos com uma única <*V*> não acentuada graficamente, como é o caso: *i*) das preposições <com>, <de>, , <sem>, <sob> e das contrações <do(s)>, <no(s)>; *ii*) dos pronomes pessoais oblíquos <me>, <te>, <se>, <nos>, <vos>, <lhe(s)>, <o(s)> e <a(s)>, <lo(s)>, <no(s)>, <vo(s)> e das contrações <mo(s)>, <to(s)>, <lho(s)>; *iii*) do pronome relativo <que>; das conjunções <e>, <nem>, <que>, <se>; as quais se agregam frequentemente a um grupo de força acentual no âmbito do sintagma prosódico.

Tabela 5 - Regras para acentuação de vogais (<*V*>)

	Regra	**Exemplo**
1	**Se** vocábulo apresenta alguma <*V*> acentuada graficamente, e**ntão** <*V*> → <$V_{tónica}$> [6].	aux**í**lio, an**á**lise, avaliaç**ã**o, **à**s, t**ú**nel
2	**Se** vocábulo não apresenta acento gráfico e termina em <a>, <e> ou <o>, seguido ou não de <m|n|s>, e**ntão** <*V*> anterior a <a>, <e> ou <o> → <$V_{tónica}$>.	c**a**rta, d**a**nça, d**a**nçam, cont**e**nte(s), h**o**mem, h**o**mens, est**u**do(s)
3	**Se** vocábulo não apresenta acento gráfico e termina em <l>, <r>, <x> ou <z>, e**ntão** <*V*> anterior → <$V_{tónica}$>.	cant**a**r, emit**i**r, dev**e**r, can**a**l, pap**e**l, fun**i**l, cet**i**m, telef**a**x, dupl**e**x, cab**a**z
4	**Se** vocábulo não apresenta acento gráfico e termina em <i> ou <u>, seguidas ou não de <m|n|s>, e**ntão** <*V*> <i> ou <u> → <$V_{tónica}$>.	delf**i**m, bot**i**ns, par**i**s, alg**u**m, com**u**ns, jes**u**s
5	**Se** em 2, 3 e 4, a <*V*> <i> ou <u> é precedida de outra <*V*>, e**ntão** essa outra <*V*> → <$V_{tónica}$>.	p**a**i(s), r**e**i(s), l**e**u, m**a**u(s), decid**i**u, c**a**ixa(s),ad**e**us, p**e**ixe, p**a**uta(s)
6	**Se** em 5 a <*V*> <i> ou <u> é seguida de <ch>, <nh>, <m + *C*|#> ou <n + *C*>, e**ntão** <*V*> <i> ou <u> → <$V_{tónica}$>.	sandu**i**che, vento**i**nha, ra**i**nha, amendo**i**m, co**i**mbra

Um problema levanta-se quando nos confrontamos com vocábulos derivados morfologicamente, como é o caso dos advérbios de modo cuja terminação é <mente>, em especial quando a forma adjetival da qual derivam é marcada por um acento gráfico (exemplos: <rápido> → <rapidamente>, <dócil> → <docilmente>). O processo para a marcação da $V_{tónica}$ nos advérbios de modo terminados em <mente> passa pela seguinte solução: implementou-se um algoritmo que pesquisa os vocábulos com esse perfil e os divide em duas

[6] São exceções a esta regra, palavras como órfão(s), órfã(s), órgão(s), sótão(s), ímã(s), as quais, embora apresentem mais do que um acento gráfico, apenas têm uma sílaba tónica (em posição paroxítona).

partes (<*RAIZ*+mente>. A <*RAIZ*> passa por um módulo de tratamento específico, o qual apresenta uma lista de sequências grafemáticas em posição final, segundo padrões específicos, já com a determinação específica de $V_{tónica}$. Este método resolveu todos os casos presentes no vocabulário de referência, embora se admita que possam surgir casos remanescentes não resolvidos.

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/ (em <**a**lmoçar> → /**a**lmus"ar/); a definição de <*V*+s+*V*> resulta na notação de <s> em /z/ (em <ca**s**a> → /k"a**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 <V_{orais}> <u> e <i> passam a /V_{nasais}/.

4. Resultados

Todas as experiências foram baseadas no dicionário de pronunciação de 41586 vocábulos da língua portuguesa, descrito na subsecção 3.1. Aplicando diferentes formas de pré-processamento ao dicionário base, foram criados vários outros dicionários, nomeadamente:

1- Dicionário alinhado: nele apresenta-se a correspondência de "um-para-um" entre grafemas e fonemas. Todos os fonemas são representados com um único símbolo, incluindo as vogais com marcação de tónica (cf. Tabela 2 e Tabela 3). É introduzido um fonema especial (símbolo '_') para indicar o apagamento de um grafema, embora não existam inserções de fonemas (cf. subsecção 3.3).

2- Dicionário com símbolos para dígrafos: vocábulos em que os dígrafos são convertidos nos símbolos representados na Tabela 4. Tendo sido o objetivo da conversão de dígrafos num único símbolo facilitar a correspondência "um-para-um" entre grafemas e fonemas, este dicionário é alinhado.

3- Dicionário com acentuação: presença da marcação da vogal tónica em cada pronunciação.

4- Dicionário com acentuação e com símbolos para dígrafos: composição dos dois anteriores, usando alinhamento "um-para-um" entre grafemas e fonemas.

No total são tomados 5 dicionários, os 4 descritos mais o dicionário base. Estes dicionários estão disponibilizados com os seguintes nomes:
- dic_CETEMP_40k;
- dic_CETEMP_40k_alinhado;
- dic_CETEMP_40k_acentuado;
- dic_CETEMP_40k_alinhado_digrafos;
- dic_CETEMP_40k_acentuado_alinhado_digraf.

Para testar o modelo estatístico, cada um destes dicionários foi particionado em 5 dicionários de treino e 5 dicionários de teste, de forma rotativa. O dicionário inicial foi dividido em 5 partes, cada uma com 20% dos vocábulos (8317), escolhidos de forma aleatória. Os vocábulos foram mutuamente exclusivos em cada uma das 5 partes. Cada uma das partes deu origem a um dicionário de teste e os restantes 4 partes (33269 vocábulos) a um dicionário de treino. A rotação das partes deu origem a 5 ciclos de treino e teste dos modelos estatísticos para validação cruzada. Os resultados indicados correspondem à média dos 5 resultados parciais.

O desempenho do sistema de conversão de grafemas para fonemas é expresso em duas taxas médias de erros de conversão verificados nos dicionários de teste: taxa média de erro de fonemas (PER – "phoneme error rate") e taxa média de erro de vocábulos (WER – "word error rate"). A Tabela 6 sumariza os resultados obtidos usando "*n*-grama" entre 2 e 8 e utilizando o dicionário alinhado (dic_CETEMP_40k_alinhado), enquanto a Tabela 7 sumariza os resultados obtidos usando regras fonológicas.

Os gráficos da Figura 1 e da Figura 2 ilustram o contributo de cada etapa de pré-processamento fonológico no desempenho do sistema de conversão, apresentando as percentagens da taxa de erro de conversão de vocábulos. Como se pode observar, a marcação da vogal tónica é o processamento que mais contribui para o melhoramento do desempenho do sistema.

Tabela 6 – Resultados com modelo base (sem regras fonológicas)

n-grama	2	3	4	5	6	7	8
WER (%)	35.1	15.5	7.90	5.96	6.08	6.51	7.01
PER(%)	4.69	1.86	0.95	0.72	0.74	0.79	0.86

Tabela 7 – Resultados com modelo base (com todas as regras fonológicas)

n-grama	2	3	4	5	6	7	8
WER (%)	9.73	4.70	2.60	2.31	2.42	2.58	2.82
PER (%)	1.27	0.60	0.33	0.30	0.31	0.33	0.36

É de notar que, ao contrário do que se poderia supor, a utilização de "*n*-grama" com grandes contextos (*n* maior que 5) não aumenta o desempenho, verificando-se, ao invés, um ligeiro aumento das taxas de erros. Isto pode ser explicado pela falta de amostras suficientes para estimar convenientemente "*n*-grama" com grandes contextos. Verifica-se, pois, que o valor ideal de *n* fica dependente da dimensão da base de dados usada para treino.

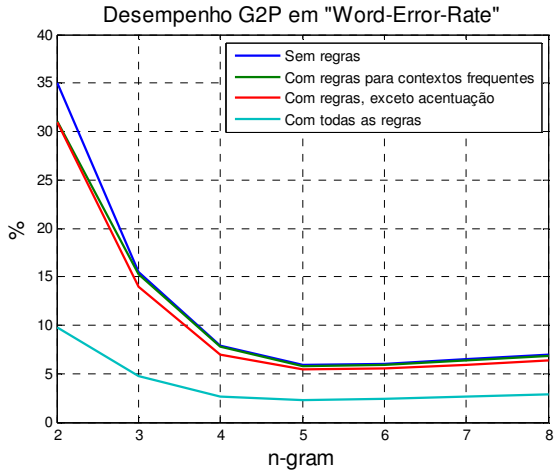

Figura 1 – Taxas de erro de palavras em função do comprimento do "*n*-grama" e da inclusão de regras fonológicas.

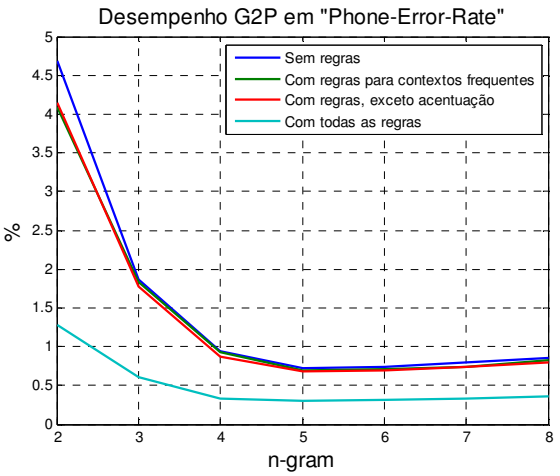

Figura 2 – Taxas de erro de fonemas em função do comprimento do "*n*-grama" e da inclusão de regras fonológicas.

O desempenho do sistema para o caso do dicionário pós-AO está representado na Figura 3 e na Figura 4. Observando os dados, é possível afirmar que o desempenho com este novo dicionário é ligeiramente inferior ao desempenho dos modelos anteriores ao AO (pré-AO), em todas as combinações de pré-processamento. Analisando os erros dos modelos com os melhores desempenhos (5-grama com marcação da vogal tónica e demais regras) observa-se que a soma dos erros dos 5 conjuntos de validação cruzada dos modelos pré-AO totaliza 958, enquanto nos modelos pós-AO totaliza 1022. Deve ainda ser referido que os dois modelos partilham 644 erros (resultantes de vocábulos que foram mal convertidas em ambos os modelos, especialmente no que diz respeito à conversão de <e> em /Ele/ e de <o> em /Olo/), sendo 313 e 378 os erros em vocábulos diferentes provocados pelos modelos pré- e pós-AO, respetivamente.

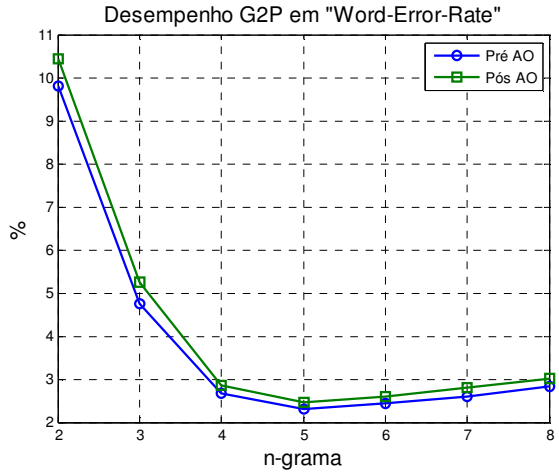

Figura 3 – Comparação do desempenho dos modelos com os dicionários pré- e pós-AO em termos de erros de vocábulos.

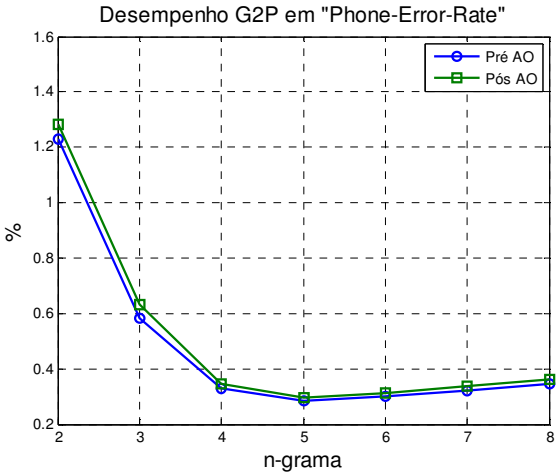

Figura 4 – Comparação do desempenho dos modelos com os dicionários pré- e pós-AO em termos de erros de fonemas.

Mais de 80% 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/, /6/ ou /a/. Nos modelos pós-AO, não existem erros provocados pelas consoantes mudas <c> e <p>, tendo-se verificado um acréscimo expressivo de erros provocados pela conversão do grafema <a>. 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 pronúncia aberta e 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 determinação da vogal tónica, sendo este um pré-processamento que, como vimos, contribui muito para o aumento do desempenho dos modelos.

Em termos de implementação, o conversor de grafema para fonema utiliza o dicionário como uma tabela de exceções que é carregado para uma tabela "*hash*" ("*hash table*"). A conversão com o modelo estatístico só é evocada para vocábulos que não constam no dicionário. Esta implementação requer mais recursos de memória mas, por outro lado, é muito mais célere e mais precisa. Uma tabela "*hash*" com 100k entradas para o dicionário base de 40k vocábulos necessita de cerca de 2MB de memória, apresentando cerca de 7500 colisões.

O ritmo de conversão é de cerca de 1M vocábulos por segundo, contra 20k vocábulos por segundo usando a conversão com um modelo estatístico de 2-grama (num PC "*quad core*" a 2,8GHz).

5. Conclusões e trabalho futuro

Neste trabalho pretendemos mostrar uma nova abordagem na tarefa de converter grafemas em fonemas em português europeu. Propomos um modelo de base estatística, imbuído de regras fonológicas. Sequências de grafonemas foram modeladas através de um algoritmo de alinhamento entre grafemas e fonemas, nas quais foram também consideradas informações advenientes do contexto fonológico da língua portuguesa, tais como a digrafia, 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*-grama" foram treinados e testados usando a grafia pré-AO e pós-AO, tendo-se verificado um ligeiro, mas consistente, decréscimo de desempenho dos modelos pós-AO.

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.

Os diferentes dicionários, bem como os modelos de "*n*-grama" estão livremente disponíveis em (SPL, 2011). O dicionário de estrangeirismos e o dicionário de múltipla pronunciação de homógrafos serão incluídos no sistema, a breve prazo. A pronunciação de adjetivos, de verbos e de nomes flexionados encontra-se em estudo, também com o objetivo de vir a integrar o sistema.

Agradecimentos

Os autores agradecem o contributo dos revisores deste artigo, pelas sugestões e comentários apresentados. Agradecem igualmente ao Instituto de Telecomunicações e à FCT as bolsas de doutoramento (Arlindo Veiga) e de pós-doutoramento (Sara Candeias, SFRH/BPD/ 36584/2007). Este trabalho recebeu ainda o apoio financeiro do projeto FCT - PTDC/CLE-LIN/11 2411/2009.

Referências

Almeida, J. J.; Simões, A. 2001. Text to Speech – "A Rewriting System Approach". *Procesamiento del Lenguaje Natural*, 27, pp. 247–255.

Andrade, E.; Viana, M. C. 1985. Corso I - Um Conversor de Texto Ortográfico em Código Fonético para o Português. *Technical report*, CLUL-INIC, Lisboa.

Barros, M. J.; Weiss, C. 2006. Maximum Entropy Motivated Grapheme-To-Phoneme, Stress and Syllable Boundary Prediction for Portuguese Text-to-Speech, *IV Jornadas en Tecnologías del Habla*, pp. 177-182. Zaragoza, España.

Bisani, M.; Ney, H. 2008. Joint-Sequence Models for Grapheme-To-Phoneme Conversion, *Speech Communication*, vol. 50 (5), pp. 434–451.

Bisani, M.; Ney, H. 2002. Investigations on Joint-Multigram Models for Grapheme-to-Phoneme Conversion, *Proc. 7th International Conference on Spoken Language Processing (ICSLP'02)*, pp. 105–108. Denver, USA.

Braga, D.; Coelho, L.; Resende Jr., F. 2006. A Rule-Based Grapheme-to-Phone Converter for TTS Systems in European Portuguese, *VI Int. Telecommunications Symposium*, pp. 328-333. Fortaleza-CE, Brazil.

Braga, D.; Marques, M. A. 2007. Desambiguação de Homógrafos para Sistemas de Conversão Texto-Fala em Português, *Diacrítica*, 21.1 Série Ciências da Linguagem, pp. 25-50. Braga, CEHUM/Universidade do Minho.

Candeias, S.; Perdigão, F. 2008. Conversor de Grafemas para Fones Baseado em Regras para Português, Costa, L.; Santos, D.; Cardoso, N. (Eds.). *Perspectivas sobre a Linguateca / Actas do encontro Linguateca: 10 anos*, cap. 14. Linguateca. Lisboa.

Caseiro, D. A.; Trancoso, I. 2002. Grapheme-to-Phone Using Finite-State Transducers, *Pro. 2002 IEEE Workshop on Speech Synthesis*, USA.

Chen, S.; Goodman, J. 1998. An Empirical Study of Smoothing Techniques for Language Modeling, *Tech. Report TR–10–98*. Center for Research in Comp.Tech., Harvard Univ.

Crystal, D. 2001. *A Dictionary of Linguistics and Phonetics*. Blackwell, Oxford.

Demberg, V.; Schmid, H.; Möhler, G. 2007. Phonological Constraints and Morphological Preprocessing for Grapheme-to-phoneme Conversion, *Proc. 45th Annual Meeting of the Association for Computational Linguistics* (ACL-07), pp. 96-103. Prague, Czech Republic.

Demberg, V. 2006. *Letter-to-Phoneme Conversion for a German Text-to-Speech System*. PhD Thesis. Stuttgart University, Germany,

Galescu, L.; Allen, J. 2001. Bi-directional Conversion Between Graphemes and Phonemes Using a Joint N-gram Model, *Proc. 4th ISCA Workshop on Speech Synthesis*, Scotland.

Good, I. 1953. The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika*, vol. 40 (3,4), 237-264.

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

Jiampojamarn, S.; Kondrak, G.; Sherif, T. 2007. Applying Many-to-Many Alignments and Hidden Markov Models to Letter-to-Phoneme Conversion, *HLT-NAACL*, pp. 372-379. Rochester, New York.

Jiampojamarn, S.; Kondrak, G. 2009. Online Discriminative Training for Grapheme-to-Phoneme Conversion, *Proc. INTERSPEECH*, pp. 1303-1306, Brighton, UK.

Katz, S. 1987. Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer, *IEEE Trans. Acoustics, Speech and Signal Processing*, 35(3), 400-401.

Kneser, R.; Ney, H. 1995. Improved Backing-Off for M-gram Language Modeling, *Proc. IEEE ICASSP*, vol. 1, pp. 181–184.

Lince. Lince - Conversor para a Nova Ortografia, http://www.portaldalinguaportuguesa.org/lince.php

Mateus, M. H.; d'Andrade, E. 2000. *The Phonology of Portuguese*. Oxford University Press.

Ney, Hermann; Essen, Ute; Kneser, Reinhard. 1994. On Structuring Probabilistic Dependences in Stochastic Language Modelling, *Computer Speech and Language*, vol. 8 (1), pp. 1-38.

Oliveira, C.; Moutinho, L.; Teixeira, A. 2004. Um Novo Sistema de Conversão Grafema-Fone para PE Baseado em Transdutores, *Actas II Congresso Int. Fonética e Fonologia*, Brasil.

Oliveira, L.; Viana, M. C.; Mata, A. I.; Trancoso, I. 2001. *Progress Report of Project Dixi+: A Portuguese Text-to-Speech Synthesizer for Alternative and Augmentative Communication*. Technical Report, FCT.

Oliveira, L.; Viana, M. C.; Trancoso; I. 1992. A Rule-Based Text-to-Speech System for Portuguese, *Proc. ICASSP'92*, San Francisco, USA.

Santos, D.; Rocha, P. 2001. Evaluating CETEMPúblico, a Free Resource for Portuguese, *Proc. 39th Annual Meeting of the Association for Computational Linguistics*, pp.442-449. Toulouse, France.

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

SPL, 2011. Material disponibilizado no âmbito deste artigo, http://lsi.co.it.pt/spl/resources.htm

Taylor, P. 2005. Hidden Markov Models for Grapheme to Phoneme Conversion, *Proc. INTERSPEECH*, pp. 1973-1976, Lisbon, Portugal.

Teixeira, A.; Oliveira, C., Moutinho, L., 2006. On the Use of Machine Learning and Syllable Information in European Portuguese Grapheme-Phone Conversion, *Proc. PROPOR'2006*, pp. 212-215.

Teixeira, J. P. 2004. *A Prosody Model to TTS Systems*. PhD Thesis, Faculdade de Engenharia da Universidade do Porto.

Teixeira, J. P.; Freitas, D. 1998. MULTIVOX-Conversor Texto-Fala para Português, *Proc. PROPOR'98*, Porto Alegre, Brasil

Trancoso, I.; Viana, M. C.; Silva, F.; Marques, G.; Oliveira, L. 1994. Rule-based vs. Neural Network Based Approaches to Letter-to-Phone Conversion for Portuguese Common and Proper Names", *Proc. ICSLP'94*, Yokohama, Japan. pp. 1767-1770.

Veiga, A.; Candeias, S.; Perdigão, F. 2011. Generating a Pronunciation Dictionary for European Portuguese Using a Joint-Sequence Model with Embedded Stress Assignment, *Proc. Brazilian Symposium in Information and Human Language Technology - STIL*, Cuiabá, Brazil, pp. 144 – 153.

Wells, J. C. 1997. SAMPA Computer Readable Phonetic Alphabet. Gibbon, D., Moore, R. and Winski, R. (Eds.), *Handbook of Standards and Resources for Spoken Language Systems*, part IV. Mouton de Gruyter, Berlin and New York.

Witten, I.; Bell, T. 1991. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression, *IEEE Trans. Information Theory*, vol. 37 (4), pp. 1085-1094.