Generating a Pronunciation Dictionary for European Portuguese Using a Joint-Sequence Model with Embedded Stress Assignment

Arlindo Veiga[1,2], **Sara Candeias**[1], **Fernando Perdigão**[1,2]

[1]Instituto de Telecomunicações, pole of Coimbra – Portugal

[2]Dept. of Electrical and Computer Engineering, FCTUC,
Universidade de Coimbra - Portugal

{aveiga,saracandeias,fp}@co.it.pt

Abstract. *This paper addresses the problem of grapheme to phoneme conversion in order to create a pronunciation dictionary from a vocabulary of the most frequent words in European Portuguese. A system based on a mixed approach funded on a stochastic model with embedded rules for stressed vowel assignment is described. The model can generate pronunciations from unrestricted words; however, a dictionary with the 40k most frequent words was constructed and corrected interactively. The vocabulary was defined using the CETEMPúblico corpus. The model and dictionary are publicly available.*

1. Introduction

The grapheme to phone(me) conversion (G2P), also called letter-to-sound conversion, maps a written text into a string of symbols which represent the speech sounds exactly and unequivocally. Several frameworks have been proposed to tackle the G2P conversion, among which linguistically ruled based modules [Kaplan and Kay 1994] and statistical approaches [Chotimongkol 2000] can be mentioned. Mainly in the languages in which orthography is roughly phonologically based, such as the Portuguese and other Romanic Languages, rule-based systems should provide a good coverage of the association between letters and sounds [Braga et al. 2006], [Oliveira et al. 1992], [Teixeira 2004]. However, probably no natural human-language satisfies this assumption exactly because exceptions from the G2P conversion can be found perhaps in every language. The most common irregularity covers situations when the association between grapheme and phoneme is not quite one-to-one but can be to some extent ambiguous and greatly dependent on the neighbor-contexts. To deal with this problem, rule based systems have been adopted to list all the exceptions. But this solution turns the development and the maintenance of the system very complex, hard and tiresome. Moreover, the rule based G2P is more likely to make mistakes for new words. In
contrast to the ruled based systems outlined above, a number of authors have addressed the G2P conversion from a stochastic perspective. This approach to G2P conversion is based on the idea that using pronunciation examples it could be possible to predict the pronunciation of unseen words by analogy. This method was already implemented by [Caseiro et al. 2002] and [Barros and Weiss 2006], among others, for Portuguese. In this paper we use a new statistical approach for which outstanding results have been

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 144–153,
Cuiabá, MT, Brazil, October 24–26, 2011. ©2011 Sociedade Brasileira de Computação

reported, named the joint-sequence model, [Bisani and Ney 2008]. In this model graphemes and phonemes are combined into a single state, giving rise to "graphonemes".

Although the joint-sequence model has shown to be a powerful tool, we also show in this paper that for the case of the Portuguese the determination of the stressed vowel leads to a substantial improvement in the system performance, as was also reported in [Caseiro et al. 2002]. Thus, we included a linguistically rule based pre-processing stage, for stress assignment, which marks and disambiguates most of the pronunciations.

The vocabulary used to generate the pronunciation dictionary is in its previous form of the current "Acordo Ortográfico" (AO). However, we think that this mixed-based G2P can also achieve good performance for European Portuguese (EP) with the AO. The inherent flexibility in dealing with the EP could be extended to other Romanic languages, which makes this an advantageous approach.

The remainder of the paper is organized as follows. In Section 2, the joint-sequence model is briefly discussed. Section 3 presents how the vocabulary and dictionary were generated while Section 4 describes the linguistic model. In Section 5 experimental results are presented, the main conclusions are summarized and future work directions are foreseen.

2. Joint-Sequence Model

Given a sequence of *N* graphemes defined by $G = G_1^N = \{g_1, g_2, \cdots, g_N\}$, the goal is to find a sequence of *M* phonemes, $F = F_1^M = \{f_1, f_2, \cdots, f_M\}$, that best describes the phonetic transcription of the original sentence. The statistical approach to this problem corresponds to the determination of the optimal sequence of phonemes, F^*, that maximizes the conditional probability of phonemes, *F*, given a sequence of graphemes, *G*:

$$F^* = \arg\max_F P\left(F \mid G\right). \tag{1}$$

It is difficult to determine F^* directly by calculating $P(F \mid G)$ for all possible sequences *F*. However, using the Bayes theorem, we can rewrite the problem as:

$$F^* = \arg\max_F P\left(F \mid G\right) = \arg\max_F \left\{P\left(G \mid F\right) \cdot P(F) / P(G)\right\}. \tag{2}$$

Since *P*(*G*) is common to all sequences *F*, the problem can be simplified in the following way:

$$F^* = \arg\max_F P\left(G \mid F\right) \cdot P(F). \tag{3}$$

Using a phonological dictionary, previously created, it is possible to estimate *P*(*G*|*F*) and the *a priori* probability, *P*(*F*), for all sequences *F* and *G* found in this dictionary.

The Markov based approaches estimate a model for each phoneme and use n-gram models to compute *P*(*F*). These approaches model the dependency between graphemes and phonemes and the dependency between phonemes, but do not model dependencies between graphemes [Taylor 2005], [Demberg 2006], [Jiampojamarn and

Kondrak 2009]. Due to these constraints, other statistical approaches emerged proposing joint probability models $P(F,G)$ to determine the optimal sequence of phonemes [Bisani and Ney 2002], [Galescu and Allen 2001], directly using the expression of the joint probability in (1) in place of the conditional probability. In this approach, all the dependencies present in the dictionary were modeled, resulting in improved performances than those obtained by the other models.

2.1 Alignment between graphemes and phonemes

Some graphemes have a univocal correspondence with the phonemes. However, for other graphemes the correspondence to phonemes depends on several factors, such as the grapheme context and the part-of-speech. There are also cases where several graphemes may lead to a single phoneme, and where a single grapheme can lead to several phonemes. All statistical approaches face this problem, being necessary, during the training process, segment and align the two sequences (a phoneme sequence and the corresponding grapheme sequence) with an equal number of segments. The solution is not always trivial or unique and depends on how the alignment algorithms associate graphemes to phonemes of a given word. Alignment can be classified as follows [Jiampojamarn et al. 2007]:

1) "**one-to-one**" - Each grapheme relates with only one phoneme (segments with one symbol only). A null symbol ('_') is used to deal with the cases in which a grapheme can originate more than one phoneme (the insertion of phonemes) or the cases where more than one grapheme originate only one phoneme (the deletion of phonemes). This alignment is easy to implement using the Levenshtein algorithm, [Navarro 2001]. In the literature these algorithms are called alignment "01-01" if insertions and deletions of phonemes are allowed, or "1-01" if only deletion of phonemes are allowed. This last case corresponds to the alignment used in this work.

2) "**many-to-many**" - The segments are composed of various symbols, which allow the association of several graphemes to several phonemes. This alignment is more generic and can be used without any prior knowledge of mapping between graphemes and phonemes. It handles insertions and deletions of phonemes without using any special symbol. On the other hand, the resulting model is more difficult to estimate and its performance is generally lower than the model with alignment "one-to-one". These alignments are also known as "n-to-m".

2.2 Statistical model

After the alignment, the sequences of graphemes and phonemes have the same number of segments. So, a new entity, born from the association of a segment of graphemes and phonemes can be defined, and is called "graphone(me)" [Bisani and Ney 2002]. A sequence of K graphonemes is annotated as $Q(F,G) = \{q_1, q_2, \cdots, q_K\}$. Given a sequence of K graphonemes, $Q(F,G)$, rather than assuming independence between symbols, the probability of the joint-sequence, $P\left(Q(F,G)\right)$, can be estimated using the so-called "n-grams" (sequences limited to n symbols).

2.3 Model estimation

The n-gram models are used to estimate the probability of symbols knowing the previous *n*–1 symbols (history). The estimation of the probability of an n-gram is based on the number of its occurrences. This probability is easy to compute, but there is a problem in assigning a zero probability to the n-grams not seen or with limited number of training examples. To overcome this limitation, it is necessary to model unseen examples (using a discount) or uncommon examples (using smoothing). Thus, a small probability mass must be reserved from the most frequent n-grams to the absent or uncommon n-grams. There are several proposed algorithms to solve this problem of probability mass redistribution, such as Good-Turing [Good 1953], Witten-Bell [Witten and Bell 1991], Kneser-Ney [Kneser and Ney 1995], Ney's absolute discount [Ney et al. 1994] and Katz's smoothing [Katz 1987]. In this work we adopted the algorithm implemented by [Demberg et al. 2007], which uses a modified version of Kneser-Ney algorithm [Chen and Goodman 1998].

3. Pronunciation Dictionary

In this work we intend to create a pronunciation dictionary from a given vocabulary. The vocabulary derives from the CETEMPúblico corpus [Santos and Rocha 2001], that corresponds to a collection of newspaper extracts published from 1991 to 1998, annotated in terms of sentences and containing 180 million words in European Portuguese. The process of generating the vocabulary starts by taking all the strings annotated as words, which obey simultaneously to the following criteria: i) start with a letter (a-z, A-Z, á-ú, Á-Ú); ii) do not contain digits; iii) are not all upper case (e.g. acronyms); iv) do not have the character '.' (e.g. URLs); v) end with a letter; vi) the corresponding lemmas do not contain '=' (e.g. compound nouns).

From the resulting list, we took the sub-list of words that occur more than 70 times in the corpus, totaling about 50k different words. Foreign words were then removed, using an automatic criteria followed by manual verification. This process results on a vocabulary of 41,586 words.

3.1 Transcription

The transcription of the vocabulary words is a result of an iterative procedure. First, a statistical model was estimated, as described in 2.2, using the SpeechDat pronunciation dictionary, [SpeechDAT 1998]. This dictionary contains about 15k entries, from which foreign words were deleted. Some SAMPA transcriptions [Wells 1997] were substituted according to the following directions: 1) we did not use the velar /l~/ and the semivowels /j/ and /w/; and 2) some standardization in the pronunciations was done.

The result of applying the statistical model to CETEMPúblico vocabulary was fairly accurate, although with some significant flaws. Then we followed a long procedure of manual verification and correction of the transcriptions. The next pass was to compare the transcriptions with other ones, generated by a commercial speech synthesizer. This comparison allowed us to rely on our results since the majority of the transcriptions agreed. All different transcriptions were analyzed one by one and we found that the transcriptions from our dictionary were the right ones most of the times.

This has led to the phonological transcription dictionary referred to as "dic_CETEMP_40k".

With the "dic_CETEMP_40k", a new statistical model was built. The test of this model on the training dictionary, allowed us to correct some remaining errors as well as to standardize and regulate some transcription procedures. Throughout the development of this work, the dictionary had been revised and corrected. Although it may still contain some errors, we are confident on its accuracy. We think that this dictionary could be an interesting resource for studies about phonetics and phonology of Portuguese.

3.3 Graphoneme alignment

An important step for establishing the statistical model is the alignment between graphemes and phonemes in the form "1-01" (one grapheme leads to zero or one phoneme; see § 2.1). The option "1-01" was chosen from the beginning, because we had identified only six cases where a grapheme could give rise to more than one phoneme. Some cases were the insertion of a yod in some words beginning with <ex->; others were the cases of non-common pronunciations such as <põem> → /po~i~6~i~/ and <têm> → /t 6~i~6~i~/. Defining symbols corresponding to more than one phoneme solved this problem of phoneme insertion. The problem of the phoneme deletions still remains, because there are always graphemes that do not give rise to any phoneme.

The alignment between graphemes and phonemes was, then, obtained using the known edit distance or Levenshtein algorithm [Navarro 2001]. This required defining a distance between each phoneme and grapheme. This distance or cost of association was defined using the log probability of this association, which was estimated from an aligned dictionary.

4. Phonetic-phonological restrictions

Since the EP is a language with much phonological regularity, we added to the G2P module some linguistic restrictions, which were pertinent to convert graphemes into phonemes. Before any regard on the linguistic rules, an aspect concerning the phonetic/phonological binomial must be clarified. While phonetics gives us the physical and articulatory properties of the sound pronounced (it means the surface structure) phonology studies the sound that has a given role in the pronunciation (the underlying structure). However, any methodological perspective concerning the speech transcription links these two linguistic fields since it deals with the inter-relationship between the units and its distinctive character (phonemes) and the physical reality of those units (phones and allophones) [Crystal 2002].

The studies on the G2P often alternate between the term phone: [Caseiro et al. 2002], [Oliveira and al. 2004] with the term phoneme; [Barros and Weis 2006], without any clarification on the perspective followed. We justify our option to adopt the term phoneme mainly because the procedure to convert the letter into the sound brings us information that derives from the structure of the language (such as, both left and right context which imply the choice of a single unit excluding all other units available in the language). The phoneme that corresponds to the grapheme is well accepted as a class to which may group all allophonic realizations able in EP (which could include all the multi pronunciations). We also considered that the phoneme conversion corresponds to the EP-standard. The phonological neutralization of oppositions is not described in this study and phonemes do not represent any archiphonemes.

Algorithms have been constructed based on practical linguistic rules, such as stress marking of the vowel (the syllable nuclei) of any single word and identifying short contexts in which the correspondence between grapheme and phoneme has a good stability.

4.1 Rules for stress assignment

Following the theoretical assumptions discussed in [Mateus and d'Andrade 2000], we adopted to mark all vowels, which are stressed (the syllable nuclei) within a word. The importance of the stressed vowel ($V_{stressed}$) has been recognized in previous G2P works, such as in [Caseiro et al. 2002]. Since the n-grams context is short and cannot, most of the times retain information about the syllable structure, marking the $V_{stressed}$ improves the statistical model by expressing graphoneme classes unequivocally. As in [Andrade and Viana 1985], our proposal considered to mark the $V_{stressed}$ (with the symbol ' " ') and did not require the identification of the syllabic unit. However, the process of identifying the $V_{stressed}$ that is described in this study was achieved in a very simple way. In the following Table 1, a set of rules for stressing vowels is presented with examples. All contexts were considered, including those without a stressed vowel, such as the prepositions <com>, <de>, <em>, <sem>, <sob>, <do(s)>, <no(s)>; the personal pronouns <me>, <te>, <se>, <nos>, <vos>, <lhe(s)>, <o(s)>, <a(s)>, <lo(s)>, <no(s)>, <vo(s)>, <mo(s)>, <to(s)>, <lho(s)>; the relative pronoun <que>; and the conjunctions <e>, <nem>, <que>, <se>; which are often added to a stressed nuclei within the prosodic unit.

Table 1: Rules for stress assignment of the vowels (V)

	Rules	**Example**
1	**If** the word has a *V* with a graphic stress mark, **Then** $V \rightarrow V_{stressed}$.	aux"ílio, an"álise, avaliaç"ão, "às, s"ót"ão
2	**If** the word has not a graphic stress mark and ends in <a>, <e> or <o> followed (or not) by <m|n|s>, **Then** prior *V* to <a>, <e> or <o> $\rightarrow V_{stressed}$.	c"arta, d"ançam, cont"ente(s), h"omem, h"omens, est"udo(s)
3	**If** the word has not a graphic stress mark and ends in *C* <l>, <r>, <x> or <z>, **Then** the last $V \rightarrow V_{stressed}$.	defens"or, cant"ar, emit"ir, dev"er, can"al, pap"el, fun"il, cet"im, telef"ax, dupl"ex, cab"az, fel"iz, arr"oz,
4	**If** the word has not a graphic stress mark and ends in *V* <i> or <u>, followed (or not) by <m|n|s>, **Then** <i> or <u> $\rightarrow V_{stressed}$.	delf"im, bot"ins, par"is, alg"um, com"uns, jes"us
5	**If** in 2, 3 and 4, the *V* <i> or <u> are preceded by other *V*, **Then** $V \rightarrow V_{stressed}$.	p"ai(s), r"ei(s), m"au(s), l"eu, decid"iu, c"aixa(s), ad"eus, p"eixe(s), p"auta(s), l"ouça(s), natur"ais
6	**If** in 5 *V* <i> or <u> are followed by <ch>, <nh>, <m + *C*|#> or <n + *C*>, **Then** <i> or <u> $\rightarrow V_{stressed}$.	sandu"iche, vento"inha, amendo"im, co"imbra

A problem arises with words, which are morphologically derived, such as the adverbs ending in <mente>, especially when the adjectival form, from which they derive, has a stress mark (e.g. <rápido> → <rapidamente>; <dócil> → <docilmente>). The solution adopted was the following: we implemented an algorithm that divides the word into two parts, <*ROOT*> and <mente>. The <*ROOT*> part undertakes a specific module, which compares it with a list of graphematic patterns which have the $V_{stressed}$ identified. This method solved all the cases present in the dictionary of 40k words.

This pre-processing module attributes a special symbol to all stressed vowels generating a univocal graphoneme.

5. Results and conclusions

All experiments were based on the pronunciation dictionary of 41,586 Portuguese words as described in Section 3.1. There are two cases, corresponding to the dictionary with and without stress marking.

To train and test the statistical model, each one of these two dictionaries was partitioned into five folds for a cross-validation procedure. The initial dictionary is divided into five folds, each one with 8317 (20%) randomly chosen words. The words are mutually exclusive in each of the five folds. Each fold gives rise to a training and testing run. Final results were obtained by evaluating the average of the five partial results.

The performance of the G2P conversion system was expressed in two average error rates: average error rate of phonemes (PER) and average error rate of words (WER). The following figures summarize the results obtained using n-grams with *n* between 2 and 8.

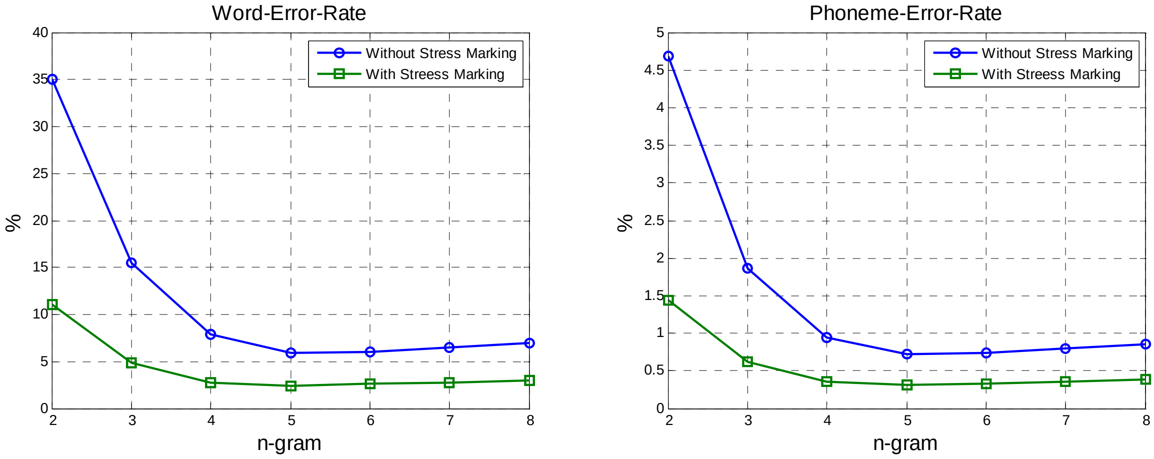

Figure 1. Word and Phoneme Error Rates for the two models.

As can be seen in Figure 1, the marking of the stressed vowel contributed to a significant improvement in the system performance. Note that, contrary to what we would expect, the use of n-grams with large contexts (*n* greater than 5) did not improve the system. In fact, there was a slight increase in the error rates. This can be explained by the lack of samples to estimate properly the n-grams with large contexts. The optimal length of n-grams was 5 in this case, but it depends on the size of the training dictionary. For example, the optimal context for the SpeechDat pronunciation vocabulary was *n*=4.

As a general conclusion, we can emphasize that the joint-sequence model achieved good results. In fact, inspecting the test errors, we observed that most of them resulted from uncommon grapheme patterns or compound words without graphic stress marks. However, the most frequent errors resulted from the pronunciation of the stressed <e> and <o> since they could be pronounced as /E/ vs /e/ and /O/ vs /o/ without any systematic rule.

It is our purpose to extend this work with the inclusion of other linguistic pre-processing stages for dealing with digraphs (both oral and consonantal) as well as with rules for regular contexts.

Our system is freely available through the site http://lsi.co.it.pt/spl/ and includes the models, dictionaries and the G2P module.

References

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. Zaragoza, Spain, pp. 177-182.

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

Bisani, M. and Ney, H. (2008) "Joint-Sequence Models for Grapheme-to-Phoneme Conversion", in Speech Communication, vol. 50(5), pp. 434–451.

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

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

Chen, S. and Goodman, J. (1998). An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report TR-10-98, Center for Research in Computing Technology (Harvard University).

Chotimongkol, A. and Black, A. (2000) "Statistically Trained Orthographic to Sound Models for Thai", Porc. of ICSLP, Beijing, China, vol. 2, pp. 551-554.

Crystal, D. (2002), A Dictionary of Linguistics and Phonetics, Oxford: Blackwell, 5th edition.

Demberg, V. (2006), Letter-to-Phoneme Conversion for a German Text-to-Speech System, Stuttgart University, published as book by Verlag Dr. Müller (VDM), ISBN: 978-3-8364-6428-4 (from Amazon.com)

Demberg, V., Schmid, H. and Möhler, G. (2007) "Phonological Constraints and Morphological Preprocessing for Grapheme-to-phoneme Conversion", Proc. of the

45th Annual Meeting of the Association for Computational Linguistics (ACL-07), Prague, Czech Republic, pp. 96-103.

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

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

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

Jiampojamarn, S. and Kondrak, G. (2009) “Online Discriminative Training for Grapheme-to-Phoneme Conversion”, Proc. of INTERSPEECH, Brighton, UK, pp. 1303-1306.

Kaplan, R. M. and Kay, M. (1994) “Regular Models of Phonological Rule Systems”, in Computational Linguistics, MIT Press, Cambridge, vol. 20(3), pp. 331-378.

Katz, S. (1987) “Estimation of Probabilities from Sparse Data for the Language Model Component of a Speech Recognizer”, IEEE Transactions on Acoustics, Speech and Signal Processing, vol. 35(3), pp. 400-401.

Kneser, R. and Ney H, (1995) “Improved Backing-Off for M-gram Language Modeling”, Proc. of ICASSP, vol. 1, pp. 181-184.

Mateus, M. H. and d’Andrade, E. (2000), The Phonology of Portuguese. Cambridge University Press, vol. 18(2), pp. 309-312.

Navarro, G. (2001) “A Guided Tour to Approximate String Matching”, ACM Computing Surveys, vol. 33(1), pp. 31-88.

Ney, H., Essen, U. and Kneser, R. (1994) “On Structuring Probabilistic Dependences in Stochastic Language Modelling”, Computer Speech and Language, vol. 8(1) pp. 1-38.

Oliveira, C., Moutinho, L. and Teixeira, A. (2004) “Um Novo Sistema de Conversão Grafema-Fone para PE Baseado em Transdutores”, Actas do II Congresso Internacional de Fonética e Fonologia, Maranhão, Brazil.

Oliveira, L. C., Viana, M. C., Trancoso, I. M. (1992) “A Rule-Based Text-to-Speech System for Portuguese”, Proc. of ICASSP. San Francisco, USA, vol. 2, pp. 73-76.

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

SpeechDAT (1998) Portuguese SpeechDat(II) FDB-4000, European Language Resources Association, http://www.elda.org/catalogue/en/speech/S0092.html

Taylor, Paul (2005) “Hidden Markov Models for Grapheme to Phoneme Conversion”, Proc. INTERSPEECH, Lisbon, Portugal, pp. 1973-1976.

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

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

Witten, I. and Bell, T. (1991) “The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression”, IEEE Transactions on Information Theory, 37(4), pp.1085-1094.