# Phonetically Transcribed Speech Corpus Designed for Context Based European Portuguese TTS

Maria Barros[1] and Bernd Möbius[2]

[1] Eurescom GmbH, Wieblinger Weg 19/4,
69123 Heidelberg, Germany
[2] University of Bonn, Am Hof 1,
53113 Bonn, Germany
barros@eurescom.eu, moebius@ifk.uni-bonn.de

**Abstract.** This paper presents a speech corpus for European Portuguese (EP), designed for context based text-to-speech (TTS) synthesis systems. The speech corpus is intended for small footprint engines and is composed by one sentence dedicated to each sequence of two phonemes of the language, incorporating as many language contexts as possible at diphone and word levels. The speech corpus is presented in three forms: its orthographic sentences, its phonetic transcriptions considering the words coarticulation, typical phenomenon in EP and its phonetic transcriptions considering the words coarticulation and the vocalic reduction effects, which is another typical phenomenon of EP language. The paper describes the corpus and presents the results in total number of phonemes and their distribution in the word and the sentence, and the total number of diphones, considering and not considering the vocalic reduction for the phonetic transcription.

**Keywords:** Speech corpus, European Portuguese corpus, text-to-speech corpus, context based corpus.

## 1 Introduction

In a TTS system, the speech synthesis quality is strongly affected by the quality of the speech corpus, the quality of the natural language processing module and the quality of the speech generation module. The speech corpus quality, on its turn, depends on the quality of the corpus design, the quality of the annotations and transcriptions and the quality of the recording system.

Nowadays TTS systems are usually corpus based statistical systems. This means that they rely on the quality of the corpus used to produce better or worse results. Besides that, they are often systems using the language contexts for choosing the best speech units for a particular target. This makes it extremely important that the speech corpus design includes as many contexts as possible. Usually, statistical methods are used to search in large amounts of texts, in the intended speech units in the considered language contexts. For this reason the speech corpus for generic vocabulary systems are very large, sometimes too large to be considered.

Z. Vetulani (Ed.): LTC 2009, LNAI 6562, pp. 12–23, 2011.

The contextual factors that can be used are many. For example, the contextual factors used for the EP HMM-based speech synthesis system (Barros et al, 2005), considering the hierarchical structure, i.e., phone, syllable, word, phrase and utterance, were the following:

At phone level: current, previous and next phones; phones before previous phone and after next phone; and positions (forward and backward) of current phone in current syllable;

At syllable level: stress condition of current, previous and next syllables; number of phones in previous, current and next syllable; positions (forward and backward) of current syllable in current word and in current phrase; number of stressed syllables before and after current syllable in the current phrase; syllable counts between the previous stressed syllable and the current syllable and between the current syllable and the next stressed syllable in the utterance; and vowel of the syllable;

At word level: number of syllables in the current, previous and next words; and positions (forward and backward) of current word in current phrase;

At phrase level: number of syllables and number of words in current, previous and next phrases; and positions (forward and backward) of current phrase in the utterance;

At utterance level: number of syllables, words and phrases in the utterance.

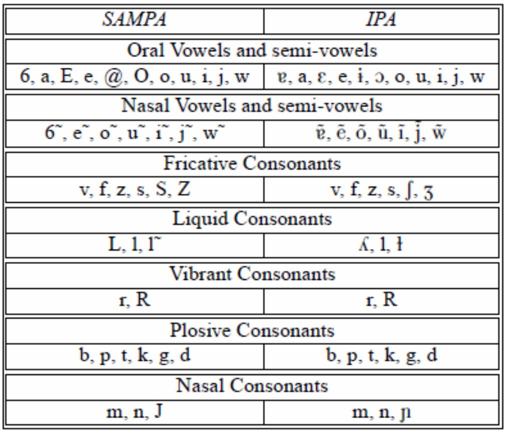
As a solution towards the best quality language context based systems with small footprint engines, efforts were concentrated in a new, manually designed, speech corpus. The speech corpus intends to achieve a large amount of important language contexts within a limited size. To do so, one sentence was constructed targeting each sequence of two phonemes of the language. In this sentence, it was tried to include as many contexts as possible. Through statistical searches it is possible to obtain as many features as it is needed, but it would need a very large corpus and that is difficult for languages without many resources.

This speech corpus is designed for European Portuguese, although the idea can be extrapolated to other languages. The same data should be used to train the natural language processing (NLP) tasks and the synthesis engine, as this way achieves a better synchronization between the units selected by the NLP and the units that can be found in the synthesizer database.

## 2 Methodology

The methodology used to design the corpus took into consideration the 38 phonemes from EP, presented in figure 1, and the silence unit, to construct one sentence dedicated to each of these 39 units combined with each other and themselves. Excluding the combination of silence with itself there were 1520 possible sentences to be constructed. The result is a speech corpus with 1436 sentences, once the rest of the combinations were not possible due to language rules that will be explained in the description section.

The sentences were all manually constructed, trying to take into account a number of language context factors and considering, but not limiting to, the most used words in the lists relating to European Portuguese usage of the words, collected from the first Morpholympics for Portuguese (Linguateca, 2003). These lists were extracted from a set of 613 different texts, from different fields (literature, politics, advertising, informal chats, general news, etc.), collected through different sources (newspapers, books, net, advertising, etc.), containing 80.903 text units correspondent to 17.128 different units. More about the description of the texts and the lists, with statistics, is available in the site (Linguateca, 2003).



| *SAMPA* | *IPA* |
|---|---|
| Oral Vowels and semi-vowels | |
| 6, a, E, e, @, O, o, u, i, j, w | ɐ, a, ɛ, e, ɨ, ɔ, o, u, i, j, w |
| Nasal Vowels and semi-vowels | |
| 6~, e~, o~, u~, i~, j~, w~ | ɐ̃, ẽ, õ, ũ, ĩ, j̃, w̃ |
| Fricative Consonants | |
| v, f, z, s, S, Z | v, f, z, s, ʃ, ʒ |
| Liquid Consonants | |
| L, l, l~ | ʎ, l, ɫ |
| Vibrant Consonants | |
| r, R | r, R |
| Plosive Consonants | |
| b, p, t, k, g, d | b, p, t, k, g, d |
| Nasal Consonants | |
| m, n, J | m, n, ɲ |

**Fig. 1.** EP Phoneme Inventory, in SAMPA and in IPA

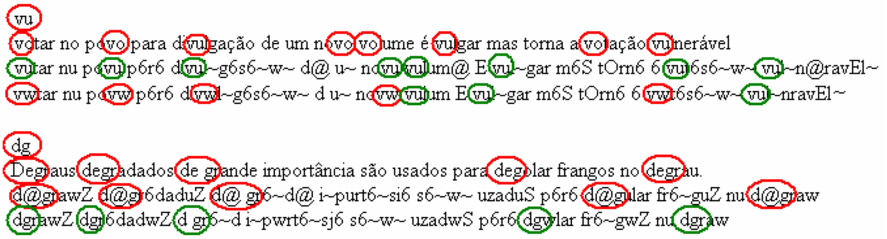
The contextual factors that were considered when building the sentences are:

At the diphone level, occurrences of the diphone: positioned at the beginning/end of the sentence; positioned at the beginning/middle/end of the word; and between two words.

At the word level, occurrences of the: words containing the target diphone at the beginning/middle/end of the sentence; and combinations with the target diphone between two words positioned at the beginning/middle/end of the sentence.

The speech corpus is presented in three forms:

- The orthographic sentences;
- A phonetic transcription considering the words coarticulation (that is a natural effect in EP continuous speech), following the grapheme-to-phoneme conversion (G2P) rules for the EP language;
- A phonetic transcription considering the words coarticulation and the vocalic reduction effects, common in EP language.



vu
votar no povo para divulgação de um novo volume é vulgar mas torna a violação vulnerável
vutar nu povu p6r6 dvul~g6s6~w~ d@ u~ novu vulum@ Evul~gar m6S tOrn6 6vul6s6~w~vul~n@ravEl~
vwtar nu povw p6r6 dvwl~g6s6~w~ d u~ novw vulum Evul~gar m6S tOrn6 6vwl6s6~w~vul~nravEl~

dg
Degraus degradados de grande importância são usados para degolar frangos no degrau.
d@grawZ d@gr6daduZ d@ gr6~d@ i~purt6~si6 s6~w~ uzaduS p6r6 d@gular fr6~guZ nu d@graw
dgrawZ dgr6dadwZ d gr6~d i~pwrt6~sj6 s6~w~ uzadwS p6r6 dgwlar fr6~gwZ nu dgraw

**Fig. 2.** Example of the sentences in the Speech Corpus

Figure 2 shows two examples of sentences, for the phonetic sequences /vu/ and /dg/, marking with circles all the possibilities of the sequence occurrence, independently of have been verified or not. In the example it is possible to see the difference in the number of the sequence occurrences for both types of phonetic transcription and verify that some diphones don't exist in the language but can appear when considering vocalic reduction.

## 3 Description

The speech corpus presented here has a total of 1436 sentences, comprising 5853 different words, with a total of 21500 words occurrences.

Two phonetic transcriptions for the orthographic sentences are provided, one following the G2P rules for EP language and the other considering the vocalic reduction effect that is common in the EP language. The number of words in the phonetic speech corpus transcribed following the G2P rules is 6264 and in the one taking the vocalic reduction into consideration, it is 6415. The difference in the number of words is due to the different transcriptions for some words when considering the vocalic reduction effect.

The effect of coarticulation between words is present in EP continuous speech. Regarding to this effect the same word can have different transcriptions, because there are graphemes that have different phonetic transcriptions according to the following one. For example, the grapheme <s> is transcribed as a /S/, if at the end of a sentence or followed by an unvoiced consonant (/p, t, k, s, S, f/ for EP); as a /Z/, if followed by a voiced consonant (/b, d, g, m, n, J, z, v, Z, l, l~, L, r, R/ for EP); or as a /z/, if followed by a vowel. Another example is the grapheme <l>, which is transcribed as a /l~/, if at the end of a sentence or followed by a consonant; or as a /l/, if followed by a vowel. One example of these situations in EP words is the word <vais>, which is transcribed as /vajS/, if it is followed by a word starting with an unvoiced consonant or if it is at the end of a sentence; as /vajZ/, if it is followed by a voiced consonant; and as /vajz/, if it is followed by a vowel. Another example is the word <mal>, which is transcribed as /mal~/, if it is followed by a word starting with a consonant or if it is at the end of a sentence; and as /mal/, if it is followed by a vowel.

The vocalic reduction is another effect present in EP continuous speech. It can be reflected by the suppression of /@/, by the phoneme /u/ reduction or suppression, or by the phonemes /u~/, /i/ and /i~/ reduction. Considering this effect, some language contextual factors are present that otherwise would not be found:

The phoneme /@/ suppression can happen in the middle of the words or at the end of the words or sentences. Some examples of this phenomenon are: <de> that would be transcribed as /d/ instead of /d@/; <dedica> that would be transcribed as /ddik6/ instead of /d@dik6/; <pote> that would be transcribed as /pOt/ instead of /pOt@/; and <apetece> that would be transcribed as /6ptEs/ instead of /6p@tEs@/.

The phoneme /u/ reduction or suppression happens in the middle of the words or at the end of the words or sentences. When the vowel /u/ suffers reduction it gives place to the semi-vowel /w/. An example of a word which usually presents /u/ suppression and reduction is the word <séculos>. Following the G2P rules this word is transcribed as /sEkuluS/, but is commonly found as /sEklwS/ due to vocalic reduction. The paper (Barros et al, 2001) covers this phenomenon.

The phonemes /u~/, /i/ and /i~/ can suffer reduction, giving place to the semi-vowels /w~/, /j/ and /j~/, respectively, but they cannot suffer suppression. Examples of these situations are: <conjuntura>, which would be transcribed as /ko~Zw~tur6/ instead of /ko~Zu~tur6/, <litoral>, which would be transcribed as /ljtwral~/ instead of /litural~/, and <inventa>, which would be transcribed as /j~ve~t6/ instead of /i~ve~t6/.

In EP many of the consonants can not be found next to each other in the same word. Considering the case of the phoneme /@/ vocalic reduction means to include almost all the combinations between consonants that in other ways would not exist.

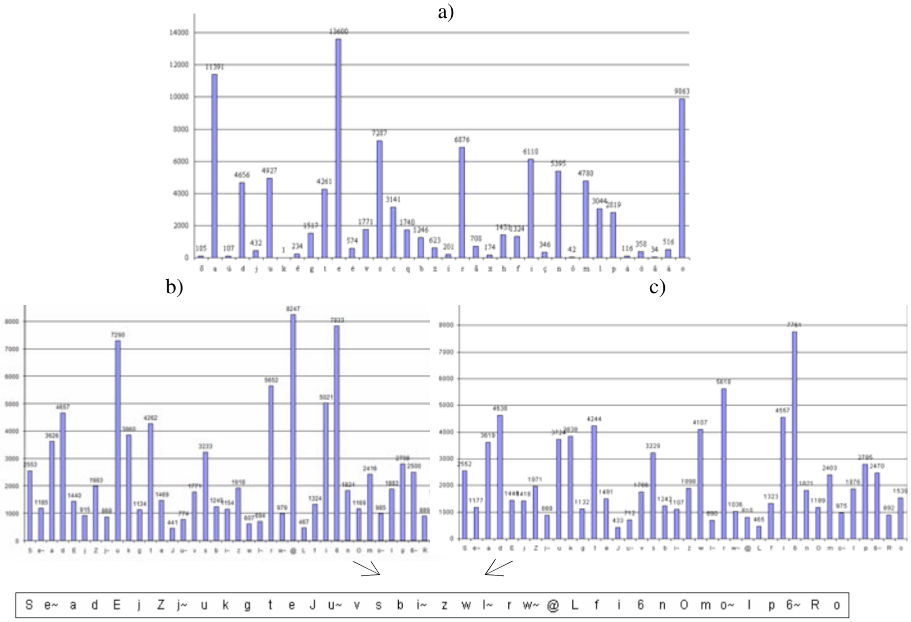


**Fig. 3.** Number of: a) Grapheme occurrences; b) Phoneme occurrences, by rules; c) Phoneme occurrences, with vocalic reduction

## 3.1 The Speech Corpus by Graphemes

The 1436 sentences speech corpus comprises of 36 graphemes (considering the graphemes with accents - à, á, â, ã, é, ê, í, ó, ô, õ, ú - as individual units) and a total of 101750 graphemes occurrences. The total number of occurrences of each grapheme is shown in figure 3 a).

Although <k> is not a Portuguese grapheme it can appear in some foreigner words adopted in Portuguese vocabulary. In this case it appears in the word <snack>.

The <à> is a particular case in Portuguese language because it only appears with the contraction between <a a> and it's not possible in the middle of a word. For example, the sentence <Dou à Joana.> (meaning <I give to Joana.>) has a grammar construction that comes from <Dou a a Joana.> (that would mean: <I give to the Joana.>). The examples present in the speech corpus are found in the words: <à>, <àquela>, <àquelas>, <àquele> and <àqueles>.

The <ç> is the only consonant that cannot appear in the beginning of a Portuguese word. This grapheme is always phonetically transcribed as /s/, which in the beginning of a word is produced by <s>, or <c> if followed by <(e, i)>.

The <ô>, <õ> or <ã> are the vowels that cannot appear in the beginning of a word. Relating to the last two, the nasal vowels, the <~> is used to nasalize the vowel and in cases with the vowel in the beginning of the word, the vowel is nasalized by using the consonants <(n,m)> in front of it.

The <o>, <a> and <s> are the most common endings for masculine words, feminine words and plurals, respectively. The infinitive form of Portuguese verbs always ends with <r>, what justifies the large amount of this grapheme's occurrences in the end of words and sentences.

From the Portuguese grammar rules, besides the consonants <s> and <r>, already mentioned above, only the consonants <m>, <z>, <l> and <b> can appear at the end of a word.

There are no cases of words ending with <â> or <õ>. In the last case, in EP the nasal sound in the end of the word is achieved by following the vowel with <m>.

The <h> is a special consonant in EP, once by itself does it not have a sonorous transcription, as it lost its aspiration with the evolution of the language. It is followed only by vowels, not by consonants, and it is used for the combinations <ch>, <nh> and <lh> that produce the phonemes /S/, /J/ and /L/, respectively.

### 3.2 Phonetic Transcription by Rules

The first phonetic transcription presented follows the G2P rules for the Portuguese language taking into consideration the words coarticulation natural in EP continuous speech, but not the vocalic reduction effect.

The 1436 sentences speech corpus comprises 38 phonemes and 92618 phonemes occurrences.

The total number of occurrences of each phoneme is shown in figure 3 b).

In European Portuguese language it is not possible to have words starting with the phonemes /l~/, /r/ or /J/, even considering phonetic effects inherent to continuous speech, unless it is some particular word like words imported from other Portuguese language dialects. This is the case of the word <nhangue>, imported from Angola Portuguese dialect to represent the name of an African bird and a place in Angola.

When considering the phonetic transcription following the G2P rules it is not possible to have words starting with a semivowel, although this can happen when considering the vocalic reduction effect.

All the vowels except /e~/ can be found at the end of a word or sentence. In the consonants case, the /d/, /k/, /g/, /t/, /J/, /v/, /s/, /L/, /f/, /n/, /m/, /l/, /p/ and /R/ are never found at the end of a word or sentence, unless due to particular cases of foreign words imported to European Portuguese vocabulary like, for instance, <internet>, <snack>, <camping>, <homeless> and <slogan>.

The consonants /Z/, /z/ and /l/ cannot be found at the end of a sentence or of an isolated word, but they can be found at the end of a word in continuous speech due to the words' coarticulation effect. The consonants /S/, /l~/ and /r/ can be found at the end of a sentence or isolated word.

Figure 5 presents a table with the total number of occurrences of each diphone when considering the phonetic transcription by rules, where the columns present the first phoneme of each diphone and the rows present the second one.

Any non-nasal vowel can only be next to a nasal vowel, or vice versa, if there is a small silence between words, in cases of slow rhythm of speech, because due to the words coarticulation effect a word that starts with a nasal vowel nasalizes the vowel from the precedent word and vice versa.

Most of the diphones composed by two consonants are not present in European Portuguese language, but there are some cases that are common: /pr/, /pl/, /br/, /bl/, /fr/ and /fl/; /Z/ followed by a voiced consonant; /S/ followed by an unvoiced consonant; /ks/ that transcribes the <x> for some grammatical cases (Barros and Weiss, 2006); /ps/, /pn/, /pt/, /kt/, /kn/, /gn/, /tn/, /gm/ and /tm/; /bt/, /bS/, /bZ/, /bz/, /bs/, /bv/ and /bm/; /dZ/, /dv/, /dm/ and /dr/; and /l~S/.

The <l> followed by a vowel is always transcribed as /l/, even if between words, unless there is a small silence in between the words, caused for instance by a comma, when then it is transcribed by a /l~/. It is also not possible to have /l/ or /l~/ followed by /r/, not because of the construction, that exists, but because in this case the <r> is transcribed by /R/ if in the same word, and there are no words starting with a /r/. Another case related to <l> not possible in EP is having /l/ or /l~/ followed by /J/, because <lnh> is not an allowed construction and there are no words starting with a /J/. Double "l", <ll>, doesn't exist in European Portuguese words, what makes the diphones among /l~/, /l/ and /L/ in the same word not possible. The /l~/ and the /L/ can never be preceded by consonant or a nasal vowel in the same word, as this would implicate a construction of three following consonants in a way that is not possible in EP language. Having a nasal vowel before a /l~/ implicates <(m,n)> after the vowel to nasalize it. So, with these cases, as well as with the consonants cases, it would be needed to have a consonant before the <l> and another after, to turn it in /l~/. It is also not possible to find these cases between words because there are no words starting with /l~/.

It's not possible to have /J/ preceded by a consonant, as it would lead to another type of construction of three consonants that is not allowed in European Portuguese language, the <nh> that gives the /J/ and any other consonant before. The same applies to nasal vowels preceding a /J/, because to have a nasal vowel before a <nh> it would have to be constructed by following the vowel with a <m> or <n>.

It's not possible to have/r/ in the same word after another /r/ or /R/, because the double <r> is always read as /R/. It is also not possible to have the /r/ after a nasal vowel in the same word, because it's the <n>, or the <m>, after a vowel that makes the nasal sound, but after a consonant the <r> is always read as /R/. There is no possibility of having these cases between words because there are no words starting with /r/.

It's not possible to have /n/ or /m/ preceded by a vowel and followed by a consonant, because in these cases the <m> and the <n> are used to nasalize the precedent vowel. Following this rule, it's not possible to have /n/ or /m/ followed by a consonant, because it would need a word construction of three consonants in a way that is not allowed in European Portuguese. There is one exception to these statements, which is the case of <m> followed by <n> and preceded by <a,o>, for example in the word <amnistia> or any word using the Portuguese prefix <omni>, which are transcribed as /6mniSti6/ and /Omni/, respectively.

The phoneme /e~/ is not found in the end of a word or after /e~/, /6~/ and /o~/, unless by influence of some regional accents/dialects, for instance in <lêem>, <têm> and <voem>, which are transcribed as /le~6~j~/, /t6~j~6~j~/ and /vo~6~j~/, but could be found in some regional accents as /le~e~/, /t6~e~/ and /vo~e~/, respectively.

### 3.3 Phonetic Transcription with Vocalic Reduction

The other phonetic transcription implemented to present the speech corpus considers the vocalic reduction effect that is common in the EP language, besides the words coarticulation. Due to the vocalic reduction there are less phonemes occurrences in this transcription than in the phonetic transcription by rules. The 1436 sentences speech corpus comprises of 38 phonemes and 84846 phonemes occurrences.

The total number of occurrences of each phoneme is shown in figure 3 c).

As it was explained before, in EP language it is not possible to have words starting with the phonemes /l~/, /r/ or /J/, even considering phonetic effects inherent to continuous speech, unless it is some particular word like words imported from other Portuguese language dialects.

Following the G2P rules it is not possible to have words starting with a semivowel, but this can happen when considering the vocalic reduction effect.

Due to the vocalic reduction effect, the vowel /@/ can be suppressed. This makes it possible to find almost all the consonants combinations. Also the consonant /S/ can be found before a voiced consonant from the transcription of <che> or <xe> that becomes /S/ instead of /S@/.

In European Portuguese language, the diphones /Ou/, /Ej/, /@j/, /aE/, /@E/, /aa/, /6a/, /ea/, /Ea/, /oa/ and /Oa/ aren't found in the same word, but can be found between words, both considering or not considering the vocalic reduction effect.

Figure 6 presents a table with the total number of occurrences of each diphone when considering in the phonetic transcription the vocalic reduction effect, where the columns present the first phoneme of each diphone and the rows present the second one.

The /@/ is suppressed in most of the cases of continuous speech, if considering the vocalic reduction effect, in any context of word or sentence.

## 4 The Vocalic Reduction Influence

To better understand the influence of the vocalic reduction in EP continuous speech, a comparison between both phonetic transcriptions, following the G2P rules and considering the vocalic reduction, in different contexts at word and sentence levels is presented.

Figure 4 a) presents a chart with the number of occurrences of each phoneme when in the middle of a word, for both phonetic transcriptions, following the G2P rules and considering the vocalic reduction. The same kind of chart is presented for the phonemes at the beginning of a word, in figure 4 b), at the end of a word, in figure 4 c), at the beginning of a sentence, in figure 4 d), and finally at the end of a sentence, in figure 4 e).

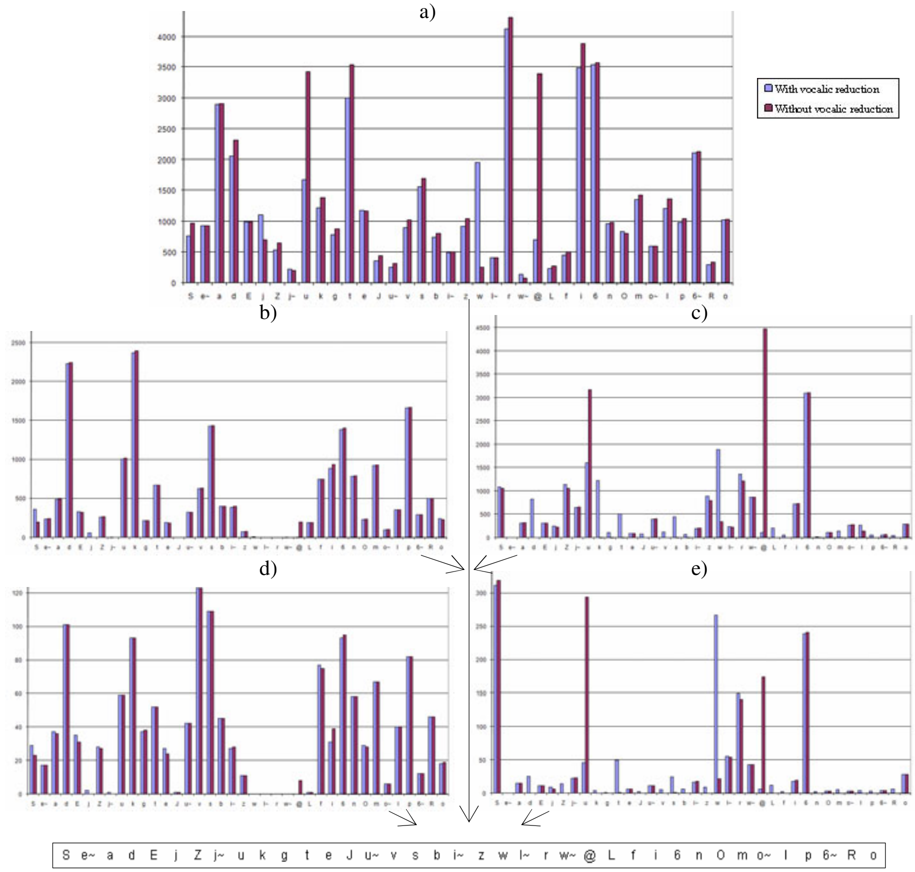


**Fig. 4.** Phonemes occurrences at the: a) Middle of word; b) Beginning of word; c) End of word; d) Beginning of sentence; e) End of sentence

From the figures it is possible to see that the biggest differences are in the number of occurrences of the phoneme /@/, the semi-vowels and the correspondent vowels. It is also possible to observe that the number of occurrences of some consonants in a particular context change if considering or not the vocalic reduction.

| | S | e~ | a | d | E | j | Z | j~ | u | k | g | t | e | J | u~ | v | s | b | i~ | z | w | l~ | r | w~ | @ | L | f | i | 6 | n | O | m | o~ | l | p | 6~ | R | o |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| S | 14 | 3 | 37 | 0 | 31 | 0 | 0 | 0 | 39 | 514 | 0 | 612 | 35 | 0 | 15 | 0 | 233 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 66 | 0 | 129 | 19 | 49 | 0 | 7 | 0 | 0 | 0 | 375 | 31 | 0 | 15 |
| e~ | 9 | 0 | 0 | 85 | 0 | 0 | 15 | 0 | 0 | 46 | 9 | 678 | 0 | 0 | 1 | 19 | 114 | 24 | 0 | 6 | 0 | 0 | 9 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 98 | 55 | 1 | 0 |
| a | 51 | 3 | 11 | 606 | 6 | 255 | 101 | 0 | 8 | 75 | 45 | 110 | 3 | 2 | 12 | 151 | 92 | 76 | 9 | 141 | 239 | 354 | 817 | 0 | 7 | 37 | 47 | 11 | 15 | 17 | 8 | 38 | 3 | 155 | 50 | 5 | 47 | 4 |
| d | 0 | 60 | 261 | 0 | 60 | 14 | 3 | 1 | 1323 | 0 | 0 | 0 | 115 | 0 | 11 | 7 | 0 | 0 | 35 | 0 | 0 | 0 | 25 | 0 | 1299 | 0 | 0 | 448 | 780 | 0 | 26 | 8 | 6 | 0 | 0 | 55 | 0 | 120 |
| E | 63 | 4 | 14 | 49 | 6 | 0 | 22 | 0 | 40 | 64 | 86 | 101 | 6 | 0 | 23 | 62 | 118 | 59 | 16 | 19 | 8 | 75 | 258 | 0 | 3 | 6 | 39 | 16 | 23 | 25 | 13 | 35 | 3 | 104 | 28 | 3 | 33 | 5 |
| j | 181 | 4 | 41 | 29 | 0 | 0 | 90 | 0 | 64 | 25 | 5 | 59 | 1 | 1 | 4 | 3 | 21 | 5 | 8 | 66 | 7 | 0 | 128 | 0 | 3 | 14 | 8 | 9 | 58 | 12 | 7 | 10 | 5 | 7 | 14 | 5 | 8 | 7 |
| Z | 0 | 31 | 87 | 536 | 27 | 4 | 33 | 0 | 110 | 0 | 23 | 0 | 10 | 0 | 29 | 93 | 0 | 52 | 19 | 7 | 1 | 0 | 2 | 0 | 185 | 6 | 0 | 129 | 92 | 123 | 13 | 183 | 0 | 47 | 0 | 57 | 63 | 21 |
| j~ | 64 | 11 | 16 | 52 | 4 | 0 | 54 | 0 | 26 | 131 | 8 | 73 | 5 | 0 | 12 | 23 | 52 | 15 | 8 | 30 | 0 | 0 | 0 | 0 | 5 | 8 | 23 | 24 | 38 | 28 | 6 | 23 | 7 | 9 | 44 | 28 | 11 | 7 |
| u | 565 | 27 | 122 | 534 | 54 | 1 | 407 | 0 | 126 | 615 | 126 | 247 | 21 | 42 | 23 | 154 | 275 | 140 | 40 | 382 | 0 | 96 | 615 | 0 | 23 | 58 | 135 | 255 | 240 | 302 | 40 | 428 | 7 | 223 | 420 | 47 | 158 | 48 |
| k | 0 | 4 | 200 | 0 | 62 | 3 | 0 | 0 | 493 | 0 | 0 | 16 | 51 | 0 | 34 | 0 | 17 | 1 | 26 | 0 | 45 | 0 | 114 | 62 | 1291 | 0 | 0 | 97 | 361 | 13 | 76 | 0 | 549 | 57 | 0 | 140 | 0 | 148 |
| g | 2 | 0 | 132 | 0 | 7 | 0 | 0 | 0 | 178 | 1 | 0 | 0 | 13 | 0 | 54 | 0 | 0 | 0 | 17 | 2 | 41 | 0 | 130 | 2 | 121 | 0 | 0 | 41 | 213 | 16 | 43 | 3 | 3 | 20 | 0 | 55 | 0 | 40 |
| t | 2 | 98 | 391 | 1 | 109 | 11 | 0 | 3 | 693 | 0 | 0 | 0 | 110 | 0 | 8 | 0 | 0 | 0 | 63 | 1 | 1 | 0 | 492 | 0 | 974 | 0 | 1 | 465 | 401 | 9 | 57 | 7 | 14 | 4 | 0 | 248 | 0 | 98 |
| e | 53 | 3 | 4 | 31 | 3 | 0 | 73 | 0 | 6 | 12 | 42 | 25 | 2 | 0 | 4 | 24 | 76 | 12 | 8 | 100 | 198 | 0 | 372 | 0 | 0 | 3 | 3 | 7 | 2 | 48 | 3 | 33 | 2 | 297 | 4 | 4 | 7 | 2 |
| J | 0 | 4 | 39 | 0 | 4 | 0 | 0 | 0 | 68 | 0 | 0 | 0 | 3 | 0 | 22 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 92 | 0 | 0 | 8 | 126 | 0 | 7 | 0 | 0 | 0 | 0 | 48 | 0 | 19 |
| u~ | 23 | 32 | 10 | 100 | 8 | 0 | 23 | 39 | 14 | 33 | 17 | 98 | 8 | 0 | 7 | 7 | 59 | 47 | 24 | 9 | 0 | 0 | 0 | 0 | 4 | 3 | 28 | 19 | 18 | 9 | 11 | 24 | 5 | 11 | 45 | 11 | 11 | 6 |
| v | 0 | 62 | 170 | 0 | 129 | 0 | 0 | 0 | 108 | 0 | 0 | 0 | 143 | 0 | 0 | 0 | 0 | 0 | 71 | 0 | 1 | 0 | 16 | 0 | 264 | 0 | 0 | 339 | 219 | 0 | 38 | 0 | 8 | 0 | 0 | 103 | 0 | 100 |
| s | 0 | 85 | 164 | 1 | 119 | 0 | 0 | 0 | 303 | 0 | 0 | 0 | 156 | 0 | 12 | 0 | 0 | 0 | 138 | 0 | 0 | 0 | 0 | 0 | 786 | 0 | 0 | 526 | 235 | 1 | 111 | 0 | 91 | 6 | 1 | 355 | 0 | 142 |
| b | 5 | 10 | 101 | 0 | 61 | 1 | 3 | 0 | 113 | 0 | 0 | 6 | 55 | 0 | 7 | 19 | 8 | 0 | 14 | 1 | 0 | 0 | 233 | 0 | 181 | 0 | 0 | 102 | 104 | 0 | 27 | 2 | 67 | 26 | 0 | 69 | 0 | 30 |
| i~ | 32 | 45 | 10 | 157 | 12 | 0 | 45 | 0 | 18 | 67 | 64 | 153 | 6 | 0 | 27 | 47 | 51 | 16 | 14 | 20 | 0 | 0 | 0 | 0 | 1 | 5 | 54 | 14 | 9 | 7 | 7 | 7 | 11 | 9 | 153 | 59 | 8 | 8 |
| z | 0 | 73 | 143 | 0 | 71 | 3 | 0 | 0 | 182 | 0 | 0 | 0 | 83 | 0 | 25 | 0 | 0 | 0 | 64 | 0 | 0 | 0 | 0 | 0 | 269 | 0 | 0 | 370 | 385 | 0 | 35 | 0 | 23 | 0 | 0 | 127 | 0 | 65 |
| w | 30 | 3 | 47 | 33 | 5 | 0 | 25 | 0 | 20 | 46 | 6 | 24 | 3 | 0 | 4 | 9 | 27 | 4 | 7 | 34 | 0 | 0 | 39 | 0 | 3 | 9 | 13 | 12 | 58 | 33 | 7 | 19 | 1 | 8 | 45 | 3 | 8 | 1 |
| l~ | 4 | 0 | 0 | 73 | 0 | 0 | 6 | 0 | 0 | 61 | 64 | 129 | 0 | 0 | 0 | 41 | 36 | 8 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 3 | 32 | 0 | 0 | 13 | 0 | 103 | 0 | 4 | 49 | 0 | 9 | 0 |
| r | 6 | 52 | 342 | 230 | 88 | 11 | 32 | 2 | 635 | 340 | 51 | 257 | 83 | 0 | 35 | 56 | 149 | 31 | 101 | 7 | 0 | 0 | 0 | 0 | 515 | 19 | 44 | 510 | 1015 | 115 | 74 | 171 | 21 | 14 | 125 | 316 | 32 | 32 |
| w~ | 9 | 11 | 26 | 128 | 17 | 0 | 3 | 0 | 52 | 90 | 3 | 32 | 5 | 0 | 11 | 30 | 61 | 13 | 11 | 5 | 0 | 0 | 0 | 0 | 15 | 6 | 28 | 36 | 72 | 40 | 3 | 35 | 1 | 8 | 73 | 76 | 21 | 15 |
| @ | 652 | 50 | 96 | 551 | 94 | 0 | 371 | 0 | 238 | 483 | 227 | 342 | 66 | 68 | 75 | 342 | 648 | 204 | 93 | 338 | 14 | 0 | 501 | 0 | 47 | 120 | 299 | 152 | 299 | 326 | 57 | 299 | 28 | 293 | 407 | 48 | 187 | 58 |
| L | 0 | 4 | 32 | 0 | 13 | 0 | 0 | 0 | 63 | 0 | 0 | 0 | 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 234 | 0 | 0 | 14 | 45 | 0 | 34 | 0 | 2 | 0 | 0 | 9 | 0 | 7 |
| f | 0 | 11 | 146 | 0 | 38 | 0 | 0 | 0 | 117 | 0 | 0 | 0 | 15 | 0 | 29 | 0 | 0 | 0 | 57 | 0 | 0 | 0 | 66 | 0 | 152 | 0 | 0 | 266 | 139 | 0 | 91 | 0 | 4 | 33 | 0 | 14 | 0 | 146 |
| i | 187 | 12 | 128 | 464 | 32 | 0 | 81 | 0 | 287 | 426 | 127 | 247 | 15 | 130 | 11 | 272 | 276 | 70 | 26 | 247 | 44 | 54 | 237 | 0 | 8 | 51 | 95 | 25 | 560 | 244 | 21 | 223 | 4 | 155 | 137 | 10 | 50 | 46 |
| 6 | 462 | 27 | 58 | 487 | 54 | 417 | 439 | 0 | 134 | 562 | 97 | 323 | 19 | 180 | 49 | 194 | 522 | 212 | 55 | 345 | 7 | 2 | 595 | 0 | 31 | 54 | 191 | 181 | 219 | 278 | 28 | 518 | 12 | 218 | 428 | 41 | 132 | 21 |
| n | 1 | 9 | 153 | 1 | 44 | 2 | 0 | 0 | 385 | 0 | 0 | 0 | 33 | 0 | 48 | 0 | 0 | 0 | 35 | 0 | 0 | 0 | 0 | 0 | 206 | 0 | 0 | 225 | 339 | 1 | 53 | 0 | 0 | 0 | 0 | 243 | 0 | 46 |
| O | 45 | 6 | 5 | 40 | 5 | 3 | 28 | 0 | 9 | 49 | 20 | 67 | 2 | 1 | 5 | 29 | 45 | 65 | 9 | 60 | 0 | 52 | 267 | 0 | 1 | 48 | 30 | 21 | 4 | 50 | 5 | 43 | 2 | 56 | 53 | 3 | 35 | 3 |
| m | 0 | 324 | 241 | 0 | 45 | 5 | 0 | 0 | 348 | 0 | 0 | 0 | 92 | 0 | 64 | 0 | 0 | 0 | 40 | 0 | 0 | 0 | 0 | 0 | 350 | 0 | 0 | 170 | 535 | 1 | 48 | 0 | 12 | 0 | 1 | 111 | 0 | 29 |
| o~ | 27 | 7 | 6 | 130 | 5 | 0 | 32 | 103 | 48 | 43 | 14 | 144 | 7 | 0 | 14 | 14 | 109 | 38 | 7 | 26 | 0 | 0 | 0 | 0 | 1 | 2 | 25 | 10 | 43 | 6 | 6 | 11 | 2 | 3 | 80 | 6 | 7 | 6 |
| l | 0 | 34 | 130 | 0 | 84 | 10 | 0 | 1 | 326 | 0 | 0 | 0 | 52 | 0 | 29 | 0 | 0 | 0 | 41 | 0 | 0 | 0 | 0 | 0 | 361 | 0 | 0 | 310 | 297 | 0 | 60 | 0 | 25 | 0 | 0 | 59 | 0 | 64 |
| p | 0 | 35 | 168 | 0 | 77 | 3 | 0 | 2 | 533 | 0 | 0 | 12 | 185 | 0 | 4 | 0 | 6 | 0 | 19 | 1 | 1 | 0 | 434 | 0 | 302 | 0 | 0 | 76 | 602 | 3 | 95 | 0 | 37 | 54 | 0 | 19 | 0 | 130 |
| 6~ | 14 | 2 | 3 | 168 | 3 | 0 | 24 | 717 | 8 | 27 | 39 | 343 | 3 | 0 | 3 | 3 | 80 | 30 | 22 | 15 | 0 | 0 | 0 | 915 | 0 | 1 | 4 | 4 | 2 | 1 | 3 | 1 | 2 | 2 | 49 | 4 | 2 | 2 |
| R | 0 | 14 | 45 | 0 | 26 | 18 | 0 | 0 | 82 | 0 | 0 | 1 | 26 | 0 | 9 | 0 | 0 | 0 | 9 | 0 | 0 | 0 | 0 | 0 | 436 | 0 | 0 | 61 | 84 | 0 | 15 | 0 | 20 | 0 | 0 | 12 | 0 | 31 |
| o | 29 | 8 | 11 | 70 | 6 | 154 | 46 | 0 | 34 | 57 | 23 | 111 | 3 | 16 | 12 | 49 | 49 | 78 | 9 | 40 | 0 | 61 | 302 | 0 | 3 | 13 | 13 | 10 | 57 | 45 | 6 | 125 | 2 | 12 | 37 | 9 | 13 | 6 |

**Fig. 5.** Total number of diphone occurrences, by rules

| | S | e~ | a | d | E | j | Z | j~ | u | k | g | t | e | J | u~ | v | s | b | i~ | z | w | l~ | r | w~ | @ | L | f | i | 6 | n | O | m | o~ | l | p | 6~ | R | o |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **S** | 43 | 4 | 34 | 2 | 32 | 0 | 2 | 0 | 11 | 504 | 31 | 612 | 34 | 0 | 10 | 2 | 233 | 0 | 11 | 0 | 28 | 0 | 0 | 6 | 11 | 0 | 132 | 19 | 52 | 1 | 5 | 2 | 0 | 0 | 370 | 32 | 1 | 16 |
| **e~** | 9 | 0 | 0 | 85 | 0 | 0 | 15 | 0 | 0 | 45 | 9 | 678 | 0 | 0 | 1 | 18 | 112 | 23 | 0 | 6 | 0 | 0 | 1 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 | 0 | 8 | 98 | 54 | 8 | 0 |
| **a** | 53 | 2 | 10 | 606 | 6 | 250 | 97 | 0 | 8 | 78 | 45 | 105 | 3 | 1 | 11 | 149 | 89 | 79 | 9 | 152 | 238 | 353 | 812 | 0 | 0 | 36 | 47 | 13 | 15 | 17 | 6 | 46 | 3 | 161 | 49 | 5 | 45 | 5 |
| **d** | 89 | 68 | 287 | 83 | 68 | 64 | 64 | 2 | 581 | 122 | 16 | 76 | 125 | 3 | 27 | 31 | 82 | 31 | 44 | 74 | 758 | 0 | 28 | 2 | 67 | 4 | 65 | 431 | 818 | 39 | 37 | 75 | 10 | 32 | 99 | 64 | 21 | 126 |
| **E** | 59 | 4 | 14 | 48 | 6 | 1 | 24 | 0 | 37 | 64 | 85 | 103 | 6 | 0 | 23 | 64 | 122 | 58 | 16 | 14 | 8 | 75 | 267 | 0 | 0 | 8 | 39 | 15 | 23 | 25 | 14 | 36 | 3 | 106 | 30 | 3 | 32 | 5 |
| **j** | 287 | 4 | 113 | 36 | 4 | 1 | 104 | 0 | 95 | 34 | 5 | 70 | 5 | 1 | 4 | 3 | 30 | 16 | 12 | 72 | 36 | 0 | 131 | 0 | 2 | 16 | 14 | 15 | 177 | 15 | 16 | 14 | 4 | 25 | 22 | 5 | 14 | 7 |
| **Z** | 12 | 32 | 88 | 541 | 30 | 16 | 39 | 0 | 45 | 4 | 24 | 12 | 10 | 11 | 29 | 96 | 12 | 52 | 28 | 13 | 68 | 0 | 21 | 2 | 9 | 7 | 3 | 117 | 97 | 130 | 15 | 183 | 2 | 56 | 5 | 57 | 69 | 22 |
| **j~** | 68 | 31 | 16 | 53 | 4 | 0 | 59 | 0 | 25 | 130 | 8 | 73 | 4 | 0 | 12 | 24 | 51 | 17 | 10 | 29 | 0 | 0 | 1 | 0 | 0 | 8 | 24 | 25 | 38 | 27 | 7 | 23 | 6 | 9 | 43 | 24 | 11 | 6 |
| **u** | 245 | 19 | 43 | 237 | 13 | 39 | 197 | 0 | 39 | 359 | 67 | 136 | 6 | 13 | 8 | 82 | 136 | 99 | 31 | 182 | 12 | 76 | 389 | 0 | 0 | 20 | 81 | 79 | 144 | 163 | 35 | 320 | 3 | 100 | 174 | 16 | 91 | 24 |
| **k** | 41 | 15 | 220 | 79 | 113 | 7 | 19 | 0 | 210 | 59 | 9 | 70 | 81 | 0 | 16 | 90 | 191 | 21 | 40 | 6 | 398 | 0 | 117 | 84 | 24 | 21 | 46 | 111 | 457 | 80 | 96 | 42 | 547 | 92 | 92 | 152 | 28 | 160 |
| **g** | 4 | 3 | 137 | 9 | 13 | 0 | 5 | 0 | 49 | 5 | 3 | 6 | 14 | 0 | 58 | 5 | 7 | 4 | 14 | 13 | 179 | 0 | 129 | 4 | 0 | 4 | 5 | 43 | 213 | 17 | 44 | 7 | 3 | 21 | 6 | 56 | 9 | 42 |
| **t** | 74 | 105 | 407 | 72 | 115 | 29 | 54 | 3 | 184 | 44 | 10 | 23 | 118 | 0 | 11 | 9 | 58 | 21 | 69 | 61 | 534 | 0 | 564 | 3 | 26 | 3 | 20 | 491 | 437 | 29 | 60 | 49 | 15 | 49 | 53 | 256 | 38 | 101 |
| **e** | 58 | 3 | 4 | 30 | 3 | 1 | 68 | 0 | 8 | 18 | 42 | 24 | 2 | 0 | 4 | 27 | 75 | 13 | 8 | 108 | 197 | 0 | 370 | 0 | 0 | 6 | 3 | 5 | 2 | 53 | 3 | 32 | 1 | 299 | 4 | 4 | 8 | 2 |
| **J** | 2 | 5 | 39 | 4 | 5 | 0 | 5 | 0 | 45 | 3 | 2 | 4 | 4 | 0 | 24 | 4 | 14 | 4 | 4 | 4 | 27 | 0 | 0 | 0 | 0 | 3 | 4 | 11 | 132 | 2 | 6 | 2 | 1 | 3 | 3 | 40 | 3 | 22 |
| **u~** | 27 | 10 | 11 | 98 | 8 | 1 | 22 | 40 | 13 | 32 | 17 | 93 | 12 | 0 | 7 | 7 | 55 | 39 | 17 | 9 | 0 | 0 | 0 | 0 | 0 | 3 | 28 | 14 | 17 | 9 | 10 | 23 | 5 | 10 | 37 | 8 | 11 | 7 |
| **v** | 28 | 63 | 170 | 13 | 133 | 59 | 14 | 1 | 36 | 6 | 3 | 6 | 142 | 0 | 5 | 6 | 17 | 5 | 52 | 4 | 77 | 0 | 31 | 0 | 55 | 5 | 5 | 305 | 219 | 13 | 48 | 7 | 8 | 9 | 8 | 106 | 8 | 94 |
| **s** | 19 | 90 | 182 | 57 | 126 | 178 | 18 | 10 | 129 | 75 | 63 | 22 | 164 | 12 | 16 | 24 | 53 | 15 | 139 | 12 | 184 | 0 | 12 | 4 | 64 | 12 | 12 | 361 | 262 | 46 | 114 | 35 | 92 | 35 | 44 | 353 | 15 | 156 |
| **b** | 9 | 8 | 100 | 10 | 61 | 8 | 9 | 0 | 21 | 12 | 2 | 13 | 56 | 0 | 4 | 22 | 16 | 24 | 17 | 5 | 93 | 0 | 233 | 3 | 43 | 3 | 4 | 91 | 104 | 11 | 25 | 3 | 68 | 46 | 3 | 71 | 5 | 34 |
| **i~** | 32 | 24 | 10 | 157 | 12 | 1 | 41 | 0 | 16 | 67 | 64 | 150 | 6 | 0 | 26 | 48 | 49 | 14 | 13 | 19 | 0 | 0 | 0 | 0 | 0 | 2 | 54 | 10 | 9 | 6 | 7 | 7 | 12 | 9 | 151 | 60 | 7 | 8 |
| **z** | 37 | 73 | 140 | 20 | 77 | 22 | 17 | 0 | 125 | 12 | 3 | 8 | 82 | 5 | 27 | 5 | 10 | 8 | 67 | 15 | 55 | 0 | 1 | 0 | 12 | 4 | 7 | 366 | 379 | 8 | 38 | 5 | 24 | 24 | 12 | 127 | 8 | 66 |
| **w** | 378 | 11 | 129 | 322 | 49 | 6 | 237 | 0 | 89 | 295 | 65 | 132 | 17 | 27 | 19 | 79 | 158 | 41 | 16 | 221 | 6 | 21 | 251 | 0 | 5 | 47 | 67 | 145 | 148 | 173 | 10 | 122 | 5 | 129 | 282 | 34 | 76 | 28 |
| **l~** | 8 | 0 | 0 | 71 | 0 | 0 | 3 | 0 | 0 | 62 | 66 | 129 | 0 | 0 | 0 | 38 | 36 | 8 | 0 | 5 | 0 | 0 | 0 | 0 | 0 | 3 | 31 | 0 | 0 | 13 | 0 | 103 | 0 | 4 | 46 | 0 | 9 | 0 |
| **r** | 77 | 53 | 341 | 242 | 89 | 46 | 80 | 9 | 417 | 352 | 58 | 281 | 83 | 0 | 35 | 83 | 202 | 33 | 97 | 59 | 248 | 0 | 0 | 2 | 39 | 21 | 62 | 479 | 1029 | 124 | 77 | 178 | 20 | 21 | 140 | 320 | 35 | 37 |
| **w~** | 23 | 34 | 27 | 130 | 20 | 4 | 4 | 1 | 51 | 87 | 3 | 36 | 3 | 0 | 11 | 30 | 65 | 21 | 20 | 5 | 0 | 0 | 0 | 0 | 0 | 6 | 28 | 34 | 68 | 40 | 4 | 35 | 1 | 8 | 78 | 80 | 21 | 15 |
| **@** | 16 | 2 | 2 | 60 | 3 | 0 | 42 | 0 | 3 | 34 | 39 | 19 | 4 | 29 | 2 | 61 | 36 | 24 | 2 | 23 | 0 | 0 | 290 | 0 | 0 | 9 | 31 | 3 | 2 | 12 | 2 | 8 | 3 | 11 | 17 | 1 | 13 | 1 |
| **L** | 5 | 6 | 32 | 25 | 14 | 1 | 11 | 0 | 17 | 11 | 3 | 12 | 11 | 0 | 7 | 4 | 8 | 8 | 8 | 5 | 64 | 0 | 1 | 0 | 2 | 2 | 16 | 21 | 66 | 6 | 37 | 7 | 5 | 5 | 8 | 7 | 9 | 9 |
| **f** | 16 | 12 | 146 | 11 | 32 | 10 | 4 | 0 | 38 | 5 | 4 | 8 | 23 | 0 | 28 | 6 | 13 | 4 | 61 | 2 | 78 | 0 | 85 | 1 | 10 | 1 | 5 | 264 | 138 | 7 | 91 | 6 | 5 | 38 | 2 | 15 | 4 | 148 |
| **i** | 106 | 12 | 52 | 453 | 29 | 0 | 76 | 0 | 118 | 411 | 128 | 235 | 10 | 130 | 12 | 273 | 273 | 60 | 25 | 237 | 165 | 53 | 225 | 0 | 4 | 52 | 92 | 26 | 437 | 250 | 19 | 222 | 5 | 128 | 129 | 11 | 43 | 38 |
| **6** | 498 | 27 | 60 | 488 | 51 | 407 | 440 | 0 | 132 | 558 | 97 | 321 | 20 | 176 | 51 | 195 | 521 | 208 | 51 | 332 | 5 | 2 | 593 | 0 | 0 | 49 | 189 | 186 | 219 | 276 | 30 | 500 | 14 | 208 | 427 | 42 | 129 | 20 |
| **n** | 8 | 9 | 153 | 2 | 41 | 3 | 7 | 3 | 308 | 4 | 5 | 1 | 38 | 0 | 47 | 1 | 16 | 3 | 33 | 8 | 77 | 0 | 15 | 1 | 112 | 0 | 10 | 221 | 337 | 3 | 55 | 4 | 0 | 4 | 1 | 243 | 2 | 44 |
| **O** | 45 | 6 | 5 | 43 | 6 | 22 | 26 | 0 | 9 | 37 | 20 | 66 | 2 | 1 | 5 | 29 | 47 | 65 | 9 | 60 | 0 | 61 | 272 | 0 | 0 | 46 | 33 | 2 | 4 | 57 | 3 | 53 | 2 | 57 | 52 | 3 | 33 | 5 |
| **m** | 10 | 325 | 247 | 30 | 48 | 9 | 10 | 0 | 140 | 8 | 1 | 28 | 95 | 0 | 66 | 4 | 15 | 5 | 44 | 7 | 206 | 0 | 20 | 0 | 27 | 37 | 8 | 176 | 538 | 48 | 55 | 15 | 13 | 10 | 12 | 104 | 5 | 32 |
| **o~** | 28 | 7 | 7 | 130 | 5 | 0 | 32 | 103 | 47 | 43 | 14 | 143 | 7 | 0 | 15 | 15 | 109 | 38 | 6 | 21 | 0 | 0 | 0 | 0 | 0 | 2 | 25 | 10 | 40 | 6 | 6 | 11 | 2 | 2 | 80 | 5 | 7 | 6 |
| **l** | 20 | 35 | 133 | 9 | 89 | 29 | 9 | 1 | 233 | 17 | 10 | 9 | 54 | 0 | 20 | 47 | 12 | 10 | 44 | 9 | 96 | 0 | 2 | 11 | 101 | 2 | 7 | 296 | 301 | 2 | 66 | 25 | 26 | 4 | 14 | 60 | 6 | 63 |
| **p** | 8 | 35 | 166 | 27 | 78 | 9 | 5 | 0 | 310 | 17 | 1 | 24 | 187 | 7 | 6 | 5 | 38 | 5 | 22 | 10 | 220 | 0 | 456 | 1 | 106 | 2 | 3 | 70 | 601 | 9 | 94 | 5 | 38 | 57 | 5 | 20 | 4 | 131 |
| **6~** | 13 | 0 | 2 | 167 | 0 | 0 | 23 | 712 | 8 | 27 | 38 | 341 | 3 | 0 | 2 | 3 | 80 | 30 | 20 | 15 | 0 | 0 | 0 | 910 | 0 | 1 | 4 | 3 | 2 | 1 | 3 | 2 | 0 | 2 | 49 | 1 | 2 | 2 |
| **R** | 39 | 15 | 45 | 17 | 25 | 44 | 15 | 2 | 34 | 34 | 12 | 12 | 30 | 0 | 9 | 6 | 28 | 10 | 12 | 28 | 50 | 0 | 0 | 2 | 89 | 3 | 25 | 57 | 88 | 6 | 16 | 11 | 21 | 39 | 18 | 10 | 1 | 33 |
| **o** | 29 | 8 | 10 | 70 | 5 | 149 | 46 | 0 | 34 | 70 | 23 | 111 | 3 | 15 | 12 | 48 | 51 | 82 | 9 | 42 | 0 | 49 | 301 | 0 | 2 | 16 | 10 | 11 | 55 | 39 | 6 | 118 | 2 | 15 | 40 | 9 | 15 | 6 |

**Fig. 6.** Total number of diphone occurrences, considering the vocalic reduction

## 5 Conclusions

The paper presents a speech corpus for EP, for small footprint context based text-to-speech TTS systems, with a reduced size, but composed by sentences incorporating each diphone of the language in as many language contexts as possible. Besides considering diphones, it also considers sequences of phonemes that would not constitute a diphone of EP, but appear in EP continuous speech due to the vocalic reduction effect.

The speech corpus is presented in its orthographic sentences and in its phonetic transcriptions considering the words coarticulation only and also the vocalic reduction effects. An analysis between both phonetic transcriptions is presented, in order to understand the importance of this effect in EP. The contextual factors considered for the corpus design were the occurrences of each diphone positioned at the beginning and end of the sentence, or at the beginning, middle and end of the word, and between two words, at the diphone level, and the occurrences of the words containing the target diphone at the beginning/middle/end of the sentence, or combinations with the target diphone between two words positioned at the beginning/middle/end of the sentence, at the word level.

## References

1. Barros, M.J., Maia, R., Tokuda, K., Freitas, D., Resende, F.G.: HMM-based European Portuguese Speech Synthesis. In: Interspeech 2005, Lisbon, Portugal (2005)
2. Linguateca: The first evaluation contest for morphological analysers of Portuguese (2003), http://www.linguateca.pt/Morfolimpiadas/ (last visited on 02-09-2009)
3. Barros, M.J., et al: Backclose Nonsyllabic Vowel [u] Behavior in European Portuguese: Reduction or Supression. In: ICSP 2001, Taejon, South Korea (2001)
4. Barros, M.J., Weiss, C.: Maximum Entropy Motivated Grapheme-to-Phoneme, Stress and Syllable Boundary Prediction for Portuguese Text-to-Speech. In: IV Biennial Workshop on Speech Technology, Zaragoza, Spain (2006)