Phonetic Events from the Labeling the European Portuguese Database for Speech Synthesis, FEUP/IPB-DB

João Paulo Teixeira[1], Diamantino Freitas[2], Daniela Braga[2], Maria João Barros[2], Vagner Latsch[2]

[1]ESTiG – Instituto Politécnico de Bragança
[2] Faculdade de Engenharia da Universidade do Porto
joaopt@ipb.pt, dfreitas@fe.up.pt, dbraga@fe.up.pt, mjbarros@fe.up.pt, vagner@fe.up.pt

Abstract

In this paper a labeled new speech signal database (FEUP/IPB-DB) in Standard European Portuguese (hereafter SEP) is presented. The objective of this work is, on one hand, to provide phonetic material for Text-to-Speech (TTS) systems construction, either from the start or to improve the quality of existing ones, and, on the other hand, to place at service of the SEP scientific community a phonetically and prosodically valuable speech corpus, essential for Speech Synthesis or Phonetics research. Our purpose is to make it available for the scientific community, since there isn't any other DB of its kind for EP. The main features of the database will be described as well as some basic statistical aspects. A discussion of some methodological problems and some observed phenomena in experimental phonetics deriving from the speech signal labeling is also done. The approach in our work is to produce a resource that can be further improved in subsequent steps with minimal re-work. The phonetic, linguistic and technical consistency are guaranteed through the involvement of a multidisciplinary team.

1. Introduction

Text-to-Speech systems play an increasing importance in man-machine communication devices. It becomes more and more usual to find internet sale systems using a speech interface, e-mail readers based on TTS, internet browsers with synthetic speech for users with special needs, speech interface for car drivers and many others.

For some languages there are already TTS systems with high enough quality to satisfy some of these applications, however there are other languages which still need a higher development level, this is the case of European Portuguese (EP).

In spite of some synthesis models having reached a technically stabilized state of art, that were and still are useful to develop engines, they all still need a source of language data to construct the TTS bases for new languages.

Different synthesis models, either parametric or non-parametric, time-domain [1], frequency domain [2] and [3], cepstral-domain, rule-based or diphone, triphone or syllable concatenation, require different structures of speech data. For all these structures, however, the minimum unit is the phoneme or even the phone segment, considered to be the largest homogeneous time signal frame.

For Standard European Portuguese in particular, we already know of two published TTS systems [2], [3], both based on formant synthesizer models and with a relatively low level of prosody. However, until now, there isn't any public phonetically labeled European Portuguese DB. The FEUP/IPB-DB is described below, in section 2. It is being used in our present project, that aims developing a new high quality SEP TTS, for two purposes. The first is to provide a phonetically rich and natural database of Portuguese phonemes and articulations specifically recorded from the high quality voice of a skilled professional speaker. This database is phonetically segmented, labeled and annotated in a way that allows it to be used for quasi-automatic construction of the segmental base of a TTS system, in respective of its structural organization. The second purpose is to supply word and phrase level annotations that are used to study and built prosody models for SEP read speech.

In section 3, we present some relevant phonetic aspects in our point of view, that resulted from the phonetic inspection, segmentation and labeling of the database. The motivation here is to present some observed phonetic phenomena that should be taken into consideration when producing TTS synthesis in order to increase the naturalness of the resulting speech. Some statistical results are presented in this section as well. In section 4 some conclusions are drawn.

2. Description of the Portuguese Speech DB

In this section we report all relevant technical data about the Database. The voice recordings were done in an acoustically treated professional studio of RDP, the public national radio broadcast company. The professional male speaker read some text materials and speech was digitally recorded using the regular studio equipment. A careful preparation of the session had been done with text preparations and trial readings. Different text materials serve different purposes of the database and the speaker was carefully instructed in accordance. After some edition treatment of the digital sound records, such as cutting main mistakes, material with a total duration of approximately 100 minutes was produced, organized in a set of sound tracks with a duration between 2 and 3 minutes each. An audio CD in cda format and a set of wave files in 44.1 KHz sampling rate, 16 bits, mono, were produced.

An example track is available in the web site http://www.portugues.mct.pt/Repositorio/EuroSpeechIPB/ where the wave, the original text and three other files with the phonetic, word and phrase labelling, can be found.

2.1. Text Corpus

The text corpus of the speech database consists of 9 text excerpts from different articles published in biggest nationwide newspaper in November 1999, 2 additional texts from another article and one interview published in the weekly biggest newspaper also in the same month, 2 sets of specially prepared interrogative sentences, with and without interrogative pronouns (who – "quem", which – "qual", how many – "quantos", how – "como", where – "onde", etc.), and 1 set of phonetically engineered log-atoms carrying all standard Portuguese diphones and several triphones in a

congruent context. Some text readings, due to their extensions are divided into two or more sound tracks.

The set of log-atoms consists of vowels and nasal vowels and diphthongs, read in a continuous way in concatenative alternation between vocalic sounds or between vocalic and consonantal sounds. This is divided into 3 tracks. The main purpose of this set is to guarantee that some specimens of each diphone are present in the data base, for speech synthesis, spoken in an as monotonous as possible way.

2.2. Sound data segmentation and labeling

Every track has been carefully examined and segmentation marks placed using the Speech Filing System (SFS) software tool from UCL [4]. Cool Edit and PRAAT [5] have been incidentally used as well. A log-book of events was maintained. Phrase, word and phoneme labels were then attached. The tonic syllable was also identified and labeled just before the first phoneme of the syllable. All annotations reside in a text file together with the time label of the instant of beginning of the element. The phonetic level labels are based in the SAMPA code [6] extended with some other necessary codes presented in table 1. The segmentation labels used at the word and phrase levels are presented in the final rows.

Table 1: Labels used at the phonetic, word and phrase levels.

Code	Meaning
p, b, t, d, k, g	Burst segment of plosive consonants in SAMPA code
!	Stop segment of plosive consonants
f, v, s, z, S, Z	Fricatives in SAMPA code
m, n, J	Nasals in SAMPA code
L, l, R, r	Liquids in SAMPA code
l*	l at the end of syllable (velar l)
i, e, E, a, 6, O, o, u, @	Vowels in SAMPA code
i~,e~,6~,o~,u~, w~,j~	Nasal vowels in SAMPA code
w, j	Semi-vowels in SAMPA code
X	Silence
XX	Aspiration
“	Beginning of tonic syllable
Word level	
i	Beginning of word
f	End of word
Phrase level	
i	Beginning of phrase
.	End of phrase
, ! () - ; : ... “	All punctuation marks that occur in the text

When one word starts right after the previous symbol without a break, the code of start of word was used to simultaneously label the end of one word and the beginning of the next. The same procedure is used for phrases boundaries.

All work of word and phrase labeling and about half of the phonetic labeling were manually done. This task was accomplished by a professional phonetician and production rate is about 1 day for 1 minute of sound material. The other half phonetic labeling was done using an automatic alignment tool from University of Gent [7] and the result was subsequently manually reviewed and corrected. This automatic alignment tool starts from the wave file and the phonetic transcription of the text, as well as the word and phrase labels in the phonetic transcription, to finally produce the phonetic labeling, inserting or removing some phonemes due to the reduction phenomenon. This process is strongly encouraged because there are benefits in time consumption.

3. EP Phonetic changing phenomena

The distance between a phonetic and a phonological transcription is rather close to the practice/theory binomial. Phonetics, as it is defined by Fromkin and Rodman, “gives us the means that lead to the spoken sound description”, while Phonology “studies how the sounds of language form systems and models” [8]. Therefore, any methodological options concerning speech labeling lays between these two classic linguistic fields: a phonological transcription, involving phonemes’ interactions, their distinctive and semantic importances; and a closer report of the corporeal reality of the language, which is, its phonetics.

Modern Phonology and Phonetics are still ruled by Martinet’s structuralist achievements [9] and Chomsky’s generativist theories [10]. Our methodology is necessarily based on this inheritance, supplemented with the theoretical Portuguese Phonology. Nevertheless, language changing issues were anyhow taken into consideration in the construction of this DB, in particular those related to dialectical or geographic varieties, as well as those concerning individual tendencies, style or habits. These aspects will be described below.

It is also important to stress that this DB allows us to extract segmental and supra-segmental features for European Portuguese (EP), what means that it represents the basis for a broader knowledge on EP Phonetics and Prosody.

Before any regard on phonetic transcription, two main aspects must be considered: in one hand, the inherent subjectivity of the transcriptor her(him)self when making her report of the speech signal, and, on the other, the linguistic changing factors. Therefore, being aware of these conditions, a trial to carry out an accurate and close phonetic transcription of the DB, following coherent criteria was done. Some of the questions that have to taken into consideration when labeling the speech signal are now going to be described. These are of great importance to the quality of the synthetic speech subsequently produced, because of their strong impact in phonetic co-articulation events and specially in prosodic aspects.

3.1. Dialectal Changing

In spite of Linguistics’ open and tolerant attitude towards any dialectal variety or accent, the fact is that when one is learning a language she(he) always studies its official and prestigious angle. Social-linguistics explains that each language has a range of regional varieties, that may differ in phonetic, morphological, syntactic or even lexical aspects, though they still belong to the same language. Political, sociological and historical reasons decide which variety is elected to be the standard and prestigious one. J. Andrade Peres [11] emphasizes that the Standard Portuguese is in fact a dictatorship imposed by cultivated and powerful classes. Hence, regional varieties are understood by these classes as deviations, outsiders or outcasts. Considering language as a social phenomena, it was decided to choose the standard Portuguese, for its official, institutional and academic

importance and extension. Nevertheless, some of the "*dialectal slips*" that are legitimate and interesting in a certain way are going to be described, but that must be taken into consideration when recording, using or studying any speech signal DB.

3.1.1. "Dialectal slips"

These "dialectal slips" originate in relaxed articulation habits that sometimes happen even in a professional speaker. In table 2 some of these habits that can be identified in Oporto region are presented.

Table 2: Examples of dialectal slips in EP.

Example	**Standard EP**	**Dialectal change**
d**ou**tores	/o/	/ow/ diphtongization
h**o**je **e**le	/o/ /e/	/oj/, /je/ diphthongization before palatal consonant
regressou	/R/	/r/ multiple alveolar trill
embora	/e~/	/6~j/

3.2. **Contextual Changing**

Linguistic changing is also related to phonetic context and inter-segmental co-articulation phenomena. Despite the classic well-known EP distribution features of the phonemes /l/ or /s/, this DB allows an experimental and faithful report of Portuguese phonetic reality, specially concerning suppressions, additions and allophones.

3.2.1. Suppressions or reductions

From the labeling of this DB, it can be observed that the vowels [@] and [u] are often practically omitted, at every possible position in the word (beginning, middle, or end), except in a tonic syllable position. Anyhow, these phenomena occur in non stressed syllables, thus producing unexpected consonant meetings (table 3).

Table 3: Examples of non-tonic vowels suppressions in EP.

Suppressions	**[@]**	**[u]**
In the beginning	<explorado> - [Splu"radu]	Not/Available
In the middle	<decisão> - [dsi"z6~w]	<português> - [prt"geS]
In the end	<deve> - ["dEv]	<Porto> - ["port]

3.2.2. Vowel quality transformations

These phenomena occur when two vowels of different qualities get together in an utterance. Two events are expected:

- the two vowels melt and experience a quality change; this occurs between non-closed vowels (e.g. <fica admirado> [fikadmiradu]; <contra o> [kõtrO]).
- one of the vowels, the closed one, [@] or [i], is reduced and becomes a semivowel; the result is a diphthong [e.g. <se aprende> [sj6pre~d]>; <na idade> [n6jdad])

The above-described events are ancient and have always existed in a conscious domain since Latin literature, which always used this knowledge with metrical and rhythmic purposes.

3.2.3. Additions

It is also common to produce reduced vocalic sounds so-called "schwas" between relaxed consonantal groups such as the pair plosive/lateral (pl, tl, kl, bl, dl, gl) or plosive/trill (pr, tr, kr, br, dr, gr): e.g. <branco> [b@rãku].

3.2.4. Allophones

Using the common definition of an *allophone*, in Phonology, as a variant of a phone, when analysing the speech signal's physical and acoustical characteristics, it can be observed that two equal phones cannot be found; they all have a certain degree of dispersion which allows them to vary according to the speaker's mood, age, health, condition or other factors. Anyway, there are some essential features that remain intact and that carry out the information conveyed. Additionally, there are some contextual interferences from the neighbour phones that change phones so much that they can only be recognized by the phonological structure of the word and its connections to the psycho-cognitive meaning. Some of those changes motivated by the articulatory context are listed and explained below:

- <-te> syllable in a word final position followed by a pause: the closed reduced vowel [@] is acoustically weak and its presence is not absolutely necessary for the communication success, which causes its reduction; the plosive "fricatizes" with the voiceless fricative consonant that is closer to its articulatory point – [s]; we can observe this phenomenon in the DB.
- <-r> in word final position followed by a pause: it's a different [r], longer in duration and usually voiceless.
- <l> in closed syllable: as this phoneme's contextual variant is already assumed by Portuguese phonetics, we labelled it with a stipulated code [l*], because of its big distinctive acoustical importance.
- The "fricatization" of voiced plosives in an intervocalic context (< -b- >→ **β**; < -d- >→ **ð** ;< -g- >→ **γ)** can also be observed.

3.2.5. Phonetic Changes

Co-articulation phenomena and compensatory mechanisms sometimes commit mistakes in the physical plan, though assuring communication success in a perceptive way. That's what happens when a voiced sound, like a vowel, transmits its voiced characteristics to the neighbour voiceless consonant. This phenomenon, called sonorization, is one of a wide range of phonetic assimilations, which are responsible for diachronic linguistic change when it becomes a habit. In this DB some of these occurrences (e.g. <ao contrário> [aw**g**o~"trariu]; <quarenta e cinco> [kware~**d**6jsi~ku]) can be found.

3.3. Phonetic statistics

In this DB it was extracted and studied some statistical information, using some dedicated programs such as the phonetic syllable division [12]. This study was done in a total of 21 minutes of speech, which consists of 18.647 phones segments and 15.633 phonemes. The DB has a speech rate of 12,2 phonemes/s. For each considered phone segment or phoneme, was determined his relative frequency (in %), average duration and standard deviation when they are in a general position and in the tonic syllable position. The table 4 reports the result.

Table 4: percentage of occurrences and average duration of all phones, considering all positions and tonic syllable position.

Phone	General Position			Tonic Syllable Position		
	%	Av. (ms)	std	%	Av. (ms)	std
a	4.0	110	34	2.2	121	32
6	10.0	68	28	0.8	75	33
E	1.7	97	29	1.0	102	27
e	1.8	95	40	1.0	102	38
@	1.7	53	38	0.03	33	15
i	5.2	69	28	1.5	85	28
O	1.4	106	33	0.8	116	29
o	1.6	97	34	0.9	103	34
u	5.1	57	29	0.7	65	33
j	2.8	49	26	0.8	53	24
w	2.5	44	27	0.7	47	31
j~	0.1	64	20	0.03	61	21
w~	0.04	53	31	0.02	55	34
6~	2.9	75	35	0.9	97	38
e~	1.2	107	31	0.6	117	33
i~	0.7	109	42	0.2	132	49
o~	0.9	98	36	0.3	119	41
u~	0.6	86	45	0.2	77	43
p	3.3	20	9	1.0	18	6
!	3.3	64	19	1.0	70	19
t	5.3	29	19	1.3	23	10
!	5.3	48	20	1.3	49	20
k	4.1	37	16	1.0	36	11
!	4.1	59	17	1.0	61	16
b	1.3	17	18	0.5	15	7
!	1.3	43	16	0.5	44	15
d	4.7	20	17	0.8	15	5
!	4.7	41	17	0.8	39	15
g	1.3	20	13	0.6	19	7
!	1.3	44	13	0.6	43	12
m	2.8	62	19	0.7	63	19
n	2.0	54	19	0.4	51	15
J	0.4	68	18	0.1	67	19
l	1.8	52	20	0.4	53	20
l*	0.9	68	30	0.4	78	32
L	0.4	56	21	0.1	43	15
r	6.5	32	16	2.1	34	17
R	0.7	73	21	0.1	78	20
v	1.4	65	22	0.3	69	20
f	1.2	93	27	0.4	99	25
z	1.6	70	18	0.4	74	19
s	4.2	103	31	1.1	100	28
S	4.1	89	33	0.6	83	26
Z	1.9	78	25	0.4	79	25
XX	2.4	320	173			
X	3.6	165	219			

Comparing the phone segment duration's in general position with the phone segment duration's in tonic syllable position, we can conclude that all vowels and the phoneme [l*] are longer in tonic position, the phoneme [L] are shorter, and all other consonants including the stops of plosives [!] are not affected by the tonic syllable.

4. Conclusions

In the present paper the main features and preparatory work of a first high-quality SEP speech database for Speech Synthesis, FEUP/IPB-DB, were described, together with a justification of the methodology used for manual segmentation and labeling at the word and phrase level and partially manual at the phonetic level. The resulting corpus is useful for basic as well as prosodic TTS development. A set of example observations and results are presented and explained, at the statistical level, for frequencies of occurrence and basic segmental durations, and at the level of phonetic changing phenomena, namely "dialectal slips", suppressions or reductions, vowel quality transformations, additions, allophones and phonetic changes. All these aspects have a partial contribution to speech naturalness, therefore needing to be considered in the operation of a TTS system. The work described takes a step in this direction.

Development of the present database is defined in the directions of full technical analysis classification and indexing of the records and labels and prosodic modeling.

5. Acknowledgements

We would like to thank to Jean Pierre Martens, Gent, for his work in the automatic segmentation, to Mark Huckvale, from UCL, for the usability of SFS and to Diamantino Guedes, from RDP, for his voice. This work was possible thank to the collaboration of IPB, FEUP, Antígona Project and COST 258.

6. References

[1] Moulines, E., Charpentier, F. "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", 1990, Speech Communication, 453-467.

[2] Oliveira, L., Viana, M., Trancoso, I., "DIXI: Sistema de Síntese de Fala a Partir de Texto para o Português", 1993, EPLP, 153-158.

[3] Teixeira, J. P., Freitas, et all, "MULTIVOX - Conversor Texto Fala para Português", 1998, III PROPOR, Porto Alegre.

[4] Mark Huckvale, Speech Filing System Tools for Speech Research http://www.phon.ucl.ac.uk/resource/sfs/

[5] Paul Boersman and D. Weenink "Praat doing Phonetics by Computer", http://www.fon.hum.uva.nl/praat/

[6] Wells J. SAMPA computer readable phonetic alphabet, 2000. http://www.phon.ucl.ac.uk/home/sampa/home.htm

[7] Vorstermans, A., Martens, J.P. and Bert Van Coile, "Automatic segmentation and labeling of multi-lingual speech data",1996, Speech Communication, 271-293.

[8]] Fromkin, V; Rodman, R.:1983, Introdução à Linguagem, Coimbra, Almedina, 1993.

[9] Martinet, André: 1960, *Elementos de Linguística Geral,* portuguese translation, J. Morais Barbosa, Lisboa, Livraria Sá da Costa, 1984.

[10] Chomsky, N., "Syntactic Strutures", 1957, The Hague: Mouton and Co.

[11] Peres, J. Andrade; Móia, Telmo: *Áreas Críticas da Língua Portuguesa*, Lisboa, Caminho1995

[12] Gouveia, P., Teixeira, J.P., Freitas, D. "Divisão Silábica Automática do Texto Escrito e Falado", 2000, V PROPOR, S. Paulo.