Universidade de São Paulo

Biblioteca Digital da Produção Intelectual - BDPI

Departamento de Ciências de Computação - ICMC/SCC

Comunicações em Eventos - ICMC/SCC

2014-08

A method for the extraction of phonetically-rich triphone sentences

International Telecommunications Symposium, 2014, São Paulo.
http://www.producao.usp.br/handle/BDPI/48603

Downloaded from: Biblioteca Digital da Produção Intelectual - BDPI, Universidade de São Paulo

A Method for the Extraction of Phonetically-Rich Triphone Sentences

Gustavo Mendonça*, Sara Candeias†‡, Fernando Perdigão†, Christopher Shulby*,
Rean Toniazzo§, Aldebaro Klautau¶ and Sandra Aluísio*
*Instituto de Ciências Matemáticas e de Computação,
Universidade de São Paulo – São Carlos, Brazil.
†Instituto de Telecomunicações,
Universidade de Coimbra – Coimbra, Portugal.
‡Microsoft Language Development Center – Lisbon, Portugal.
§Departamento de Engenharia de Materiais,
Universidade Federal de São Carlos – São Carlos, Brazil.
¶Laboratório de Processamento de Sinais,
Universidade Federal do Pará – Belém, Brazil.
Email: gustavom@icmc.usp.br, saracandeias@co.it.pt, fp@co.it.pt, chrisshulby@gmail.com,
reantoniazzo@gmail.com, a.klautau@ieee.org, sandra@icmc.usp.br

***Abstract*—A method is proposed for compiling a corpus of phonetically-rich triphone sentences; i.e., sentences with a high variety of triphones, distributed in a uniform fashion. Such a corpus is of interest for a wide range of contexts, from automatic speech recognition to speech therapy. We evaluated this method by building phonetically-rich corpora for Brazilian Portuguese. The data employed comes from Wikipedia's dumps, which were converted into plain text, segmented and phonetically transcribed. The method consists of comparing the distance between the triphone distribution of the available sentences to an ideal uniform distribution, with equiprobable triphones. A greedy algorithm was implemented to recognize and evaluate the distance among sentences. A heuristic metric is proposed for pre-selecting sentences for the algorithm, in order to quicken its execution. The results show that, by applying the proposed metric, one can build corpora with more uniform triphone distributions.**

I. Introduction

In what regards to speech technology, although there are some studies which employ words [1], syllables [2] and monophones [3] to develop Automatic Speech Recognition (ASR) and Text to Speech (TTS) systems, most of the current research widely makes use of contextual phone units, such as triphones and diphones.

The issue of developing a phonetically-rich triphone sentences corpus is of great significance for many areas of knowledge. In many applications of ASR and speech synthesis, for instance, rich speech databases are important for properly estimating the acoustic models [4]. In speech therapy, phonetically-rich sentences are often employed in reading aloud tasks so as to assess the speech production of patients in various phonetic/phonological contexts [5]. Laboratory phonologists are also interested in such corpora in order to develop prompts for analyzing speech production and variability [6].

Formally, the task discussed in this work can be described as follows: given a corpus K with s sentences, find a subset P containing s_p sentences, such that the triphones that compose s_p holds a uniform distribution as much as possible. Despite its apparent simplicity, in what concerns to computational complexity, the task cannot be considered a simple one. Since it has a combinatorial nature, it lacks a polynomial-time solution and should be regarded as an intractable problem [7].

We evaluate the proposed method in building a phonetically-rich triphone sentences corpus for Brazilian Portuguese. The sentences come from the Portuguese Wikipedia dump [8], which was converted into plain text, segmented and phonetically transcribed. The algorithm employs a greedy approach to select sentences, in a way such that the triphone distribution in the selected sentences is as uniform as possible. In order to expedite its execution, a heuristic metric is proposed to pre-select sentences for the algorithm, favoring the least frequent triphones over the most frequent ones.

The remainder of this paper is organized as follows. In Section II, we briefly describe the related work available in the literature. In Section III, we describe the method proposed. In Section IV, we evaluate it by building a phonetically-rich corpus for Brazilian Portuguese. The final remarks are outlined in Section V.

II. Related Work

Speech can be analyzed in a myriad of forms. The phonetic or phonological structure of a language can be described through phones, phonemes, syllables, diphones, triphones, feet, etc. For languages such as Mandarim, in which tones have a phonological value, one must even posit units such as tonemes in order to properly describe speech phenomena [9].

Many methods have been proposed for extracting phonetically-balanced corpora, that is to say corpora made of sentences which reproduce the triphone distribution of a given language [10][11][12][13].

It is known that many linguistic phenomena, including triphone sets, show a Zipfian distribution [14]. A phonetically-balanced corpus, for this reason, is a corpus which follows

978-1-4799-3743-1/14/$31.00 ©2014 IEEE

Zipf's law in representing each triphone inversely proportional to its rank in the frequency table. These kinds of corpora are important specially for Large Vocabulary Continuous Speech Recognition (LVCSR), where unbalanced triphone representations can achieve better Word Error Rates (WER). However, phonetically-balanced corpora are not adequate for many other tasks, even regarding speech recognition. When building a system to assess one's pronunciation quality or to synthesize speech, for instance, more accurate results can be attained by using uniform triphone representations, i.e. phonetically-rich corpora.

Phonetically-rich corpora in our work are those which show sentences with a high variety of triphones, distributed in a uniform fashion regardless their representation in the language. In other words, in order to build such corpora, Zipf's law must be nullified, by favoring less frequent triphones and disfavoring more frequent ones. However, there are studies that consider other definitions and even other basic units to build phonetically-rich corpora.

In Abushariah et al. [10], the concept of "rich" is used in the sense that the set must contain all the phonemes of Arabic language (the chosen language for their study) but without a need for a uniform distribution. The set of sentences was handmade developed by linguists/experts. They used a set of 663 words, also defined by hand, and then Arabic independent sentences have been written using the 663 phonetically-rich words. The final database consists of 367 sentences with 2 to 9 words per sentence.

Arora et al. [15] considered syllables as the basic unit to extract, in an automatic way, phonetically-rich sentences from a large text corpus from Indian languages, justifying their choice because a syllable is the smallest segment of the uterance. In their process to extract the sentences for a given corpus, the chosen set should have the same distribution of syllabic words and also the same distribution of consonant, vowel and other symbols.

Nicodem et al. [16] deals specifically with Brazilian Portuguese and proposed a method based on genetic algorithms to select a set of sentences for a speech syntesis system. Their goal was to select a recording corpus that would improve the phonetic and prosodic variability of the system. They tried to fulfill the gap of phonetically-balanced corpora available for Brazilian Portuguese, since the available corpora disregards prosodic features. They evaluated it through the CETENFolha corpus (www.linguateca.pt/cetenfolha/) which has circa 1,5 million sentences in order to gather 4,000 sentences phonetically- and prosodically- rich. Their approach is composed of 4 stages, including grapheme-to-phoneme conversion, prosodic annotation, feature vector representation, and selection. The authors obtained prosodic features based on the pitch, therefore identifying tone events for each syllable (N, H+, H-, H, L, and L-, where H and L stands for high and low, respectively, and N for neutral). Using these features to represent each sentence, they developed a genetic algorithm (GA) to select a subset. Their paper, however, does not discuss how the GA fitness function meets both constraints (phonetic and prosodic).

III. Method

A. Unit of analysis

Contextual phone units are extensively applied to speech technology systems given their ability to encompass allophonic variation and coarticulation effects, specially triphones. A triphone is represented as a sequence $(p_{left}$ - p - $p_{right})$, where p_{left} is the phone which precedes p and p_{right} is the one which follows it. Table I presents a comparison of the word *speech* transcribed using monophones and triphones.

Word	Monophone Form	Triphone Form
speech	[s p i tʃ]	[#-s-p s-p-i p-i-tʃ i-tʃ-#]

TABLE I. A COMPARISON BETWEEN MONOPHONE AND TRIPHONE TRANSCRIPTION.

As one might observe, triphones are capable of describing the surrounding environment of a given phone and this has a huge impact in the performance of acoustic models for speech recognition or speech synthesis. Given the above reasons, we chose triphones as the unit of analysis for our algorithm.

B. Heuristic Metric

For the expedition of the sentence extraction through the greedy algorithm, due to its high time complexity order, we set a heuristic metric to pre-select sentences and rank them according to the triphones they contained. The metric uses the probability of the triphones in the corpus in order to favor the least frequent triphones over the most frequent ones. It consists of a summation of the reciprocal probability for each triphone in the sentence.

Formally, this can be defined in the following way. Consider a corpus K consisting of a set of sentences $S = \{s_1, s_2, s_3, ..., s_n\}$. Each sentence s is formed by m triphones, represented as $T = \{t_1, t_2, t_3, ..., t_m\}$. The *a priori* probability of the triphones can be calculated straightforwadly: let $P_K(t_i)$ be the probability of the triphone t_i in the corpus K, then $P_K(t_i)$ is the number of times t_i occur divided by the total number of triphones in K. For that matter, a sentence s can be considered phonetically-rich if it possess many triphones with low probability of occurence. Therefore, we define the phonetic richness of a sentence s as the summation of its triphones' reciprocal probabilities:

$$\varrho(s) = \sum_{i=1}^{m} \frac{1}{P_K(t_i)} \tag{1}$$

C. Algorithm

Our algorithm for extracting rich sentences was implemented in Python and follows a greedy strategy. The distance metric is calculated through the SciPy library [17].

Greedy algorithms have been widely used in Computer Science, when the optimum solution of the problem can not be guaranteed [18]. Greedy strategies make locally optimal choices hoping to find the global optimum. Notwithstanding, in many cases, greedy algorithms have been notorious for jams at local maxima, since the best solution for a given problem may not concur with the sum of each partial best choice.

978-1-4799-3743-1/14/$31.00 ©2014 IEEE

However, for the extraction of phonetically rich sentences, this approach is suitable, owing to the fact that it is computationally intractable to analyze all possible sets of sentences.

We initialize the algorithm by applying the heuristic metric described in Section III-B to all sentences in the corpus. After this, all sentences are ranked in descending order and the first 50,000 sentences with the best values are selected. This metric was proposed because the algorithm has an order of $O(mn^2)$ time complexity, where n is the number of sentences and m the number of selected triphones, and its execution was slow considering all the sentences available in the corpus. Afterwards, the algorithm loops through 50,000 sentences and calculates the euclidean distance between the triphone distribution of the set made up with the selected sentences and the current sentence to an ideal corpus, containing equiprobable triphones. The sentence with the minimum value is appended to a list of selected sentences and removed from the corpus. Then the loop starts over, considering for the calculation of the distance not just each sentence in isolation, but a set comprising each remaining sentence in the corpus together with the sentences already selected in the last step. When the list reaches n selected sentences, the execution is suspended. The pseudocode for the algorithm is described below.

```
Corpus <- List of available sentences
Selected <- [] // List of selected sentences
Metrics <- [] //List made of tuples with sentences
              and euclidean distance values
Ideal <- Ideal corpus, with all equiprobable triphones

while length(Selected) < n do:
  for Sentence in Corpus:
    calculate distance between Sentence+Selected and Ideal
    append Sentence and its metric in the list Metrics
  BestSentence <- select the sentence in the loop with the
                  minimum distance
  append BestSentece to Selected
  clear the Metrics list
end.
```

IV. Example Evaluation

A. Corpus

As a proof-of-concept we evaluated our method by building a phoneticaly-rich corpus for Brazilian Portuguese. The original database of sentences consisted of the Wikipedia dump produced on 23rd January 2014. Table II summarizes the data.

Articles	Word Tokens	Word Types
∼820,000	168,823,100	9,688,039

TABLE II. Portuguese Wikipedia Summary – Dumped on 23rd January 2014.

In order to obtain only plain text from Wikipedia articles, we used the software WikiExtractor [19], to strip all of the MediaWiki markups and other metadata. Then, we segmented the output into sentences, by applying the Punkt sentence tokenizer [20]. Punkt is a language-independent tool, which can be trained to tokenize sentences. It is distributed together with NLTK [21], where it already comes with a model for Portuguese, trained on the Floresta Sintá(c)tica Treebank [22].

Following, each sentence was transcribed phonetically by using a pronunciation dictionary for each language variety. We employed the UFPAdic 3.0 [23], developed for Brazilian Portuguese, which contains 38 phones and 64,847 entries. Triphones were generated dynamically, based on the transcription registered in the dictionary. Cross-word triphones were considered in the analysis along with cross-word short pause models. Given its encyclopedic nature, many sentences in Wikipedia present dates, periods, percentages and other numerical information. For this reason, we decided to supplement the dictionary, by introducing the pronunciation of numbers from 0 to 2014. The pronunciations were defined manually and embedded into the dictionary. The transcription task was carried out in the following way: a Python script was developed to loop over each sentence and check if all its belonging words were listed on the dictionary. If all the words were listed, the sentence was accepted, otherwise rejected. Due to the fact that many words which occur in Wikipedia were not registered in the pronunciation dictionary, a large number of sentences had to be discarded. Details are described in Table III.

Total Sentences	Used	Used/total
7,809,647	1,229,422	15.7%

TABLE III. Sentences' summary after WikiExtractor and Punkt.

Some pilot experiments showed that the metric benefited sentences which were too long, as they had more triphones; or too short, as some of them had very rare triphones. The problem with long sentences is that they can be too complex for a recording prompt, inducing speech disfluencies such as pauses, false starts, lenghtenings, repetitions and self-correction [24]. In addition, the short sentences selected by the algorithm were usually only nominal, containing titles, topics or proper names; therefore, they would not be adequate for sentence prompts. For this reason, we filtered the sentences, selecting only those which had an average size (i.e. between 20 and 60 triphones, and more than four words). Further information is given in Table IV. After that, we applied the heuristic metric described in Section IV-A, and the top 50,000 sentences were selected (= 2,340,237 triphone tokens and 10,237 triphone types).

Total Sentences	Short	Average	Long
1,229,422	15,581	873,546	340,295

TABLE IV. Sentences' summary after the length filter.

B. Discussion

For this example evaluation, we discuss the extraction of 250 phonetically-rich sentences. Table V describes some triphone statistics for different sets of sentences extracted with the method proposed. The first column presents the number of extracted sentences; the second number of different triphones or triphone types; the third the number of triphone tokens; and the last the triphone type/token ratio which can be used to measure the method's performance. Owing to the fact that no other methods for the extraction of phonetically-rich triphone sentences were found in the literature, we established a list of random sentences as the baseline for comparison. Table VI contains the data regarding sentences selected randomly. The list of random sentences derives from the pool of 50,000 sentences described in Section IV-A. Ten different seed states were used in order to ensure randomness, the average of these results are presented.

Sentences	Triphone Types	Triphone Tokens	Type/Token
25	923	928	0.99
50	1485	1541	0.96
75	1965	2151	0.91
100	2389	2774	0.86
125	2736	3384	0.81
150	3091	4075	0.76
175	3390	4736	0.72
200	3715	5477	0.68
225	3991	6200	0.64
250	4189	6908	0.61

TABLE V. TRIPHONE RESULTS FROM THE EXTRACTION OF SENTENCES THROUGH OUR METHOD.

Sentences	Triphone Types	Triphone Tokens	Type/Token Ratio
25	774	1121	0.69
50	1318	2037	0.65
75	1713	3093	0.55
100	1917	3968	0.48
125	2352	5166	0.46
150	2564	6110	0.42
175	2820	7375	0.38
200	2961	8000	0.37
225	3211	9578	0.34
250	3335	10482	0.32

TABLE VI. TRIPHONE RESULTS FROM THE SENTENCES TAKEN RANDOMLY.

As it can be seen through the type/token triphone ratio, the method is capable of extracting sentences in a much more uniform way. For 250 sentences, our method was capable of extracting 4189 distinct triphones (40,9% of all types in the corpus), as opposed to 3335 (32,5%) in the random set; a difference of 854 novel distinct triphones. Furthermore, this higher number of distinct triphones was achieved with less triphone tokens (6908 *vs.* 10482), in a way that the type/token ratio for the method we propose was almost double the baseline: 0.61 in contrast to 0.32. Considering sets with different numbers of sentences, the method outperformed the random selection in all experiments. A Kolmogorov-Smirnov Test (K–S Test) confirms that the sentences selected through our method are closer to a uniform distribution than the ones extracted randomly.

One can observe that, as the number of selected sentences increases, the type/token ratio decreases. It may be the case that, after a huge number of sentences, the method's output converges to a limit such that no statistical significance can be noticed while comparing to a random selection. However, given time limitations, it was not feasible to analyze such a situation. As the number of selected sentences increases so does the number of triphones for comparison. After a while, the number of triphones for comparison becomes so large that the algorithm's execution time might not be proper for practical applications.

Additionally, the algorithm's output needs to be revised. Despite all our caution in the data preparation process, we noticed that some of the sentences selected by the algorithm were, in fact, caused by mistakes from the pronunciation dictionary. Foreign and loan words are known to be a problem for grapheme to phoneme conversion because they do not follow the orthographic patterns of the target language [25]. Several sentences selected by our algorithm contained foreign words which were registered in the dictionary with abnormal pronunciations, such as *Springsteen* [sprĩgsteẽ], *hill* [iww], *world* [wohwdʒ]. Since no other words are registered with the triphones [e-e+ẽ] or [e-ẽ+#] except for *Springsteen*, the algorithm ends up by selecting the sentence in which it occurs. Seeing that our method of comparing triphone distributions is greedy, our algorithm is fooled into believing that these are rare jewels. While this may be the case either way, the algorithm cannot function properly with incorrect transcriptions. A corpus with 100 revised sentences extracted by this method can be found in the Appendix.

V. FINAL REMARKS

We proposed a method for compiling a corpus of phonetically-rich triphone sentences. It was evaluated for Brazilian Portuguese. All sentences considered come from the Portuguese Wikipedia dumps, which were converted into plain text, segmented and transcribed. Our method consisted of comparing the distance between the triphone distribution of the sentences to a uniform distribution, with equiprobable triphones. The algorithm followed a greedy strategy in evaluating the distance metric. The results showed that our method is capable of extracting sentences in a much more uniform way, while comparing to a random selection. For 250 sentences, we were able to extract 854 new distinct triphones, in a set of sentences with a much higher type/token ratio. However, the method has its limitations. As discussed, it depends entirely on the quality of the pronunciation dictionary. If the pronunciation dictionary has some incorrect words, it might be the case that the algorithm favors them, if they possess triphone types not registered in other words. As a future work, we intend to define a method that recognizes foreign words and excludes them from the selected sentences. We also plan in applying the method to others corpora, e.g. CETENFolha, in order to make the results comparable with other studies for Brazilian Portuguese, such as Nicodem et al. [16]. All resources developed in this paper are freely available on the web[1].

REFERENCES

[1] R. Thangarajan, A. M. Natarajan, and M. Selvam, "Word and triphone based approaches in continuous speech recognition for Tamil language," *WSEAS Trans. Sig. Proc.*, vol. 4, no. 3, pp. 76–85, 2008.

[2] A. Ganapathiraju, J. Hamaker, J. Picone, M. Ordowski, and G. Doddington, "Syllable-based large vocabulary continuous speech recognition," *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, vol. 9, pp. 358–366, 2001.

[3] A. Kumar, M. Dua, and T. Choudhary, "Article: Continuous Hindi speech recognition using monophone based acoustic modeling," *IJCA Proceedings on International Conference on Advances in Computer Engineering and Applications*, vol. ICACEA, no. 1, pp. 15–19, March 2014.

[4] L. Rabiner and R. Schafer, "Introduction to digital speech processing," *Foundations and trends in signal processing*, vol. 1, pp. 1–194, 2007.

[5] A. P. Mendes, A. N. d. Costa, A. D. Martins, A. F. O. Fernandes, S. M. D. d. R. Vicente, and T. C. S. Freitas, "Contributos para a construção de um texto foneticamente equilibrado para o Português-Europeu," *Revista CEFAC*, vol. 14, pp. 910–917, 10 2012.

[6] J. B. P. Pierrehumbert, M. E. Beckman, and D. R. Ladd, "Conceptual foundations of phonology as a laboratory science," in *Phonological knowledge: Conceptual and empirical issues*. Oxford University Press., 2000, pp. 273–304.

[7] R. Sedgewick and P. Flajolet, *An introduction to the analysis of algorithms*. Addison-Wesley-Longman, 2013.

[8] Wikimedia, "Portuguese Wikipedia database dump backup," http://dumps.wikimedia.org/ptwiki/20140123/, 2014.

[1] http://nilc.icmc.usp.br/listener

[9] X. Lei, M. yuh Hwang, and M. Ostendorf, “Incorporating tone-related MLP posteriors in the feature representation for Mandarin ASR,” in *In Proc. Eur. Conf. Speech Communication Technology*, 2005, pp. 2981–2984.

[10] M. A. M. Abushariah, R. N. Ainon, R. Zainuddin, M. Elshafei, and O. O. Khalifa, “Phonetically rich and balanced text and speech corpora for Arabic language,” *Language Resources and Evaluation*, vol. 46, no. 4, pp. 601–634, 2012.

[11] J.-L. Shen, H.-M. Wang, R.-Y. Lyu, and L.-S. Lee, “Incremental speaker adaptation using phonetically balanced training sentences for Mandarin syllable recognition based on segmental probability models.” in *ICSLP*. ISCA, 1994. [Online]. Available: http://dblp.uni-trier.de/db/conf/interspeech/icslp1994.html#ShenWLL94

[12] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren, “DARPA TIMIT acoustic phonetic continuous speech corpus cdrom,” 1993.

[13] E. Uraga and C. Gamboa, “VOXMEX speech database: Design of a phonetically balanced corpus,” in *LREC*. European Language Resources Association, 2004.

[14] C. D. Manning and H. Schütze, *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[15] K. Arora, S. Arora, K. Verma, and S. S. Agrawal, “Automatic extraction of phonetically rich sentences from large text corpus of Indian languages.” *INTERSPEECH*, 2004.

[16] M. Nicodem, I. Seara, R. Seara, D. Anjos, and R. Seara-Jr, “Seleção automática de corpus de texto para sistemas de síntese de fala,” *XXV Simpósio Brasileiro de Telecomunicações - SBrT 2007*, 2007.

[17] E. Jones, T. Oliphant, and P. Peterson, “SciPy: Open source scientific tools for Python,” http://www.scipy.org/, 2014.

[18] B. Coppin, “Inteligência artificial,” *Rio de Janeiro: LTC*, 2010.

[19] Medialab, “Wikipedia extractor,” http://medialab.di.unipi.it/wiki, 2013.

[20] T. Kiss and J. Strunk, “Unsupervised multilingual sentence boundary detection.” *Computational Linguistics*, vol. 32, no. 4, pp. 485–525, 2006.

[21] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python*. O’Reilly Media, Inc., 2009.

[22] C. Freitas, P. Rocha, and E. Bick, “Floresta sintá(c)tica: Bigger, thicker and easier,” in *Computational Processing of the Portuguese Language*. Springer, 2008, pp. 216–219.

[23] N. Neto, C. Patrick, A. Klautau, and I. Trancoso, “Free tools and resources for Brazilian Portuguese speech recognition,” *Journal of the Brazilian Computer Society*, vol. 17, no. 1, pp. 53–68, 2011.

[24] M. Watanabe and R. Rose, “Pausology and hesitation phenomena in second language acquisition,” *The Routledge Encyclopedia of Second Language Acquisition*, pp. 480–483, 2012.

[25] J. Steigner and M. Schrder, “Cross-language phonemisation in German text-to-speech synthesis.” in *INTERSPEECH 2007*. ISCA, 2007, pp. 1913–1916.

Appendix: Examples of the extracted sentences

Number of Sentences: 100; **Number of Triphone Types:** 2307; **Number of Triphone Tokens:** 2959; **Type/Token Ratio:** 0.78.

1) A ilha fica tão próxima da praia que, quando a maré baixa, pode ser atingida a pé.
2) Diadorim é Reinaldo, filho do grande chefe Joca Ramiro, traído por Hermógenes.
3) A Sicília tem alguns moinhos ainda em bom estado de conservação que lhe dão beleza e encanto.
4) Em geral, chegaram ao Brasil como escravos vindos de Angola, Congo, e Moçambique.
5) A sardinha é um peixe comum nas águas do mar Mediterrâneo.
6) Possuem esse nome pois costumam viver na plumagem dos pombos urbanos.
7) É brilhante, doce e muito harmônico, sem presença de metal na voz.
8) Para fechar Alessandro Del Piero fez outro aos 121’.
9) Roman Polanski dirige Chinatown com Jack Nicholson.
10) A atriz sabe falar fluentemente espanhol.
11) Eles achavam Getúlio Vargas um problema.
12) Oppenheimer captura cavalo com peão.
13) Um bago tem tamanho médio não uniforme.
14) Segundo relatório da força aérea belga há confrontos com a União Soviética.
15) É irmão do também antropólogo Gilberto Velho.
16) Ganhou sete Oscar e oito Emmy.
17) Qual é minha perspectiva agora?
18) Ela é um fantasma verde, feminino!
19) Justin em seguida volta no tempo.
20) Nós fizemos um álbum do Korn.
21) Desde então Edílson é fã dessas bandas.
22) Há um só senhor uma só fé um só batismo.
23) Ivan Lins faria um show em Mossoró à noite.
24) Cresceram maior que um gato.
25) Há locações disponíveis em Tóquio no Japão.
26) Preso a um tronco nenhum lugar é seguro!
27) Hoje é professor emérito da UFBA.
28) Veio até aqui e não vai mergulhar?
29) Luís Jerônimo é um jovem rico.
30) Na hora pensei: ”tenho que fazer isso?”
31) A campanha teve coordenação de Sanches.
32) A mulher que você me deu, fugiu.
33) Eu nunca tive um encontro com Bianca.
34) Homer jura vingança a Burns.
35) Beijo, me liga e amanhã sei lá!
36) Um colégio é como um ser vivo.
37) Sophie é filha de um amigo gay de Alan Greg.
38) Xuxa guarda rancor e é ambiciosa.
39) No mesmo ano conhece Aldir Blanc em Viena.
40) É um imenso painel reunindo um elenco famoso.
41) A Sé integra três belos órgãos.
42) Em ambos, Shannon conquistou medalha.
43) A terra é abundante em recursos como vinagre e óleo vegetal.
44) Faça sua escolha e bom jogo!
45) Quem é que poderia sonhar com algo assim?
46) Ela é ruiva com olhos azuis.
47) Deu a louca na Chapeuzinho!
48) De onde venho e para onde vou?
49) Eu choro e sofro tormentas!
50) Um falcão pousa em um pedregulho.
51) Ninguém tenha medo, nem fraqueza!
52) É membro do grupo Monty Python.
53) A sondagem de Senna pela Benetton e a chegada à kart.
54) Isto é um negócio e a única coisa que importa é ganhar.
55) Robert é um forte glutão da equipe.
56) Um bárbaro no exército romano?
57) Infância e juventude em Linz.
58) Já ir à Argentina era muito bom!
59) Fiquei com inveja dele.
60) Há dragões ao redor do mundo!
61) Edmond é pai do biólogo Jean.
62) A mãe lhe telefonava às vezes.
63) Tonho é tímido, humilde e sincero.
64) André Jung ocupa um lugar central no fórum.
65) Lois pergunta: ”você é um homem ou um alienígena?”
66) Sua voz é um assobio fino e longo.
67) Por isso é sempre bom conferir!
68) Celso Lafer recuperou a jóia e devolveu-lhe.
69) É próxima ao Rio Parnaíba.
70) Lendo aquilo fica bem difícil.
71) A faculdade de John Oxford até hoje possui fãs fiéis.
72) Existe uma crença moderna no dragão chinês.
73) Sean Connery já sugeriu que Gibson fosse James Bond.
74) A raiz dos dentes é longa.
75) Essa noite produziu um feito singular.
76) Fim da Segunda Guerra Mundial.
77) –No Zorra, eu fazia humor rasgado.
78) Charles vê um homem ser morto em um tiroteio.
79) Tinham um novo senhor agora.
80) É comum ocorrerem fenômenos ópticos com estas nuvens.
81) Era um cão de pelo escuro e olhos negros.
82) Há títulos na região tcheca da Tchecoslováquia.
83) Raquel Torres vai investigar a área.
84) Clay foge e leva a jovem Jane como refém.
85) Djavan jogou futebol e hóquei no gelo na infância.
86) A origem do fagote é bastante remota.
87) Um jedi nunca usa a força para lucro ou ganho pessoal.
88) Chamavam José Alencar de Zezé.
89) Um código fonte é um sistema complexo.
90) A igreja tem um altar barroco.
91) Luís Eduardo pronunciou a senha: ”esgoto”.
92) Quanto ao sexo: macho ou fêmea?
93) A rádio Caxias cumpriu esse papel.
94) Roger Lion é um campeão orgulhoso que ama boxe.
95) Um outeiro é menor que um morro.
96) Hitoshi Sakimoto nasceu em Yokohama.
97) Nenhum isótopo do urânio é estável.
98) Chicago é um bairro tranquilo e festivo.
99) Hong Kong continua a utilizar a lei comum inglesa.
100) Só cinco funcionam como museus.

978-1-4799-3743-1/14/$31.00 ©2014 IEEE