HMM-based Brazilian Portuguese TTS

Daniela Braga, Pedro Silva, Manuel Ribeiro, Mário Henriques, Miguel Sales Dia

MLDC – Microsoft Language Development Center, Porto Salvo, Portugal
i-dbraga; i-pedros; i-manrib; i-marioh; midias}@microsoft.com

Abstract

In this paper, a Hidden Markov Models-based Text-to-Speech system (hereafter HTS) for Brazilian Portuguese (BP) was presented. The Brazilian Portuguese Text-to-Speech system (hereafter BP TTS), within other languages, will be enabled in several Microsoft products whose release date is still being planning. The architecture of the system was described as well as the main issues arising from this specific language development. The test plan and procedures were presented and its results were discussed. The comprehension tests carried out by 7 listeners gave rise to 98.915% of overall intelligibility rate distributed by the following domains: e-mail/sentences, addresses, person's names, date and time and news/paragraphs. Another comprehension test was carried out with a competitor commercial system and the obtained overall intelligibility rate was 91.95%.

Introduction

Speech Technology development companies are more and more interested in providing speech technology in Portuguese to a increasing market of around 235 millions of Portuguese native speakers spread all over the world. Brazilian Portuguese, due to its obvious market dimensions[2], is more required than any other Portuguese varieties. Amongst the academic Brazilian Portuguese TTS systems, we can find the *Aiuruetê*, a concatenative-based TTS system produced by LPDF-DECOM (Laboratório de Processamento Digital da Fala Departamento de Comunicações) of the UNICAMP (Campinas University) [1], [2], [3], [4], [5]; the concatenative-based TTS system developed by LINSE (Laboratório de Circuitos e Processamento de Sinais da Universidade Federal de Santa Catarina) [6], [7] and the most recent HMM-based TTS system developed by LPS (Laboratório de Processamento de Sinais) in collaboration with the Nagoya Institute of Technology [8], [9].
Companies have also started driving their attention to the development of Brazilian Portuguese synthetic voices. Among them we can find *Raquel* of Nuance [10], *Fernanda, Gabriela* and *Felipe* of Loquendo [11], *Paola, Joana, Pedro and Carlos* of Acapela [12] and three Brazilian dialects of Aculab [13]. Microsoft recently decided to develop a Brazilian Portuguese synthetic voice, together with other languages for mobile and desktop interfaces. This paper accounts for this development focusing the test process and the obtained results of the internal beta version release.

System workflow

The front-end of the system is dictionary-based, being composed by a lexicon with around 140 thousand words, phonetically annotated with phonetic transcriptions, stress marks and syllable boundaries, and with Part-of-Speech (POS) information. The stress and syllable marking was automatically assigned using linguistic rule-based algorithms specially developed for Brazilian Portuguese language [14]. The front-end is also composed by the text analysis, which involves the sentence separator and word breaker modules and includes a couple of other files, such as phone set and features and the POS tags set. It also includes the TN (Text Normalization) rules, the homograph ambiguity (also polyphony) resolution algorithm [14], [15] a stochastic-based LTS (Letter-to-Sound) converter to predict phonetic transcriptions for out-of-vocabulary words and the prosody models, which are data-driven using a prosody tagged corpus of 2000 sentences.
The front-end outputs phonetic transcriptions that are subsequently input of the TTS runtime engine or back-end, which then outputs synthetic voice. Figure 1 illustrates the system workflow.

[2] Brazilian population is already 191 908 598, according to CIA world factbook, available at: https://www.cia.gov/library/publications/the-world-factbook/geos/br.html#People, last updated in July 2008)

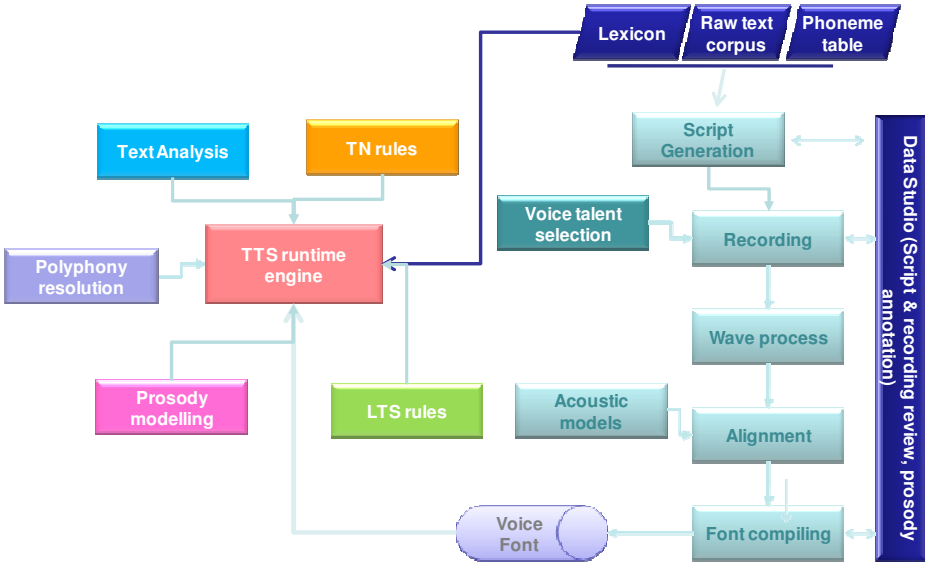

Figure 1: Brazilian Portuguese TTS pipeline.

The voice font building is also a very complex and demanding process that requires the following steps: script selection (using different text genders, phonetically balanced, with a total of 11 500 prompts and nearly 13 hours of speech), recording process according to TC-Star's specifications (2 channels, audio and EGG (Electroglottograph) signal, at 96 kHz, 24 bits of sampling rate) [16], edition of the prompts, recording quality control, re-recording and edition of the prompts which failed in the quality control, wave process, automatic alignment and quality validation, font compiling and conversion of the original recorded waves to 8khz, 8 bits sample rate. Figure 2 depicts the voice font building process.

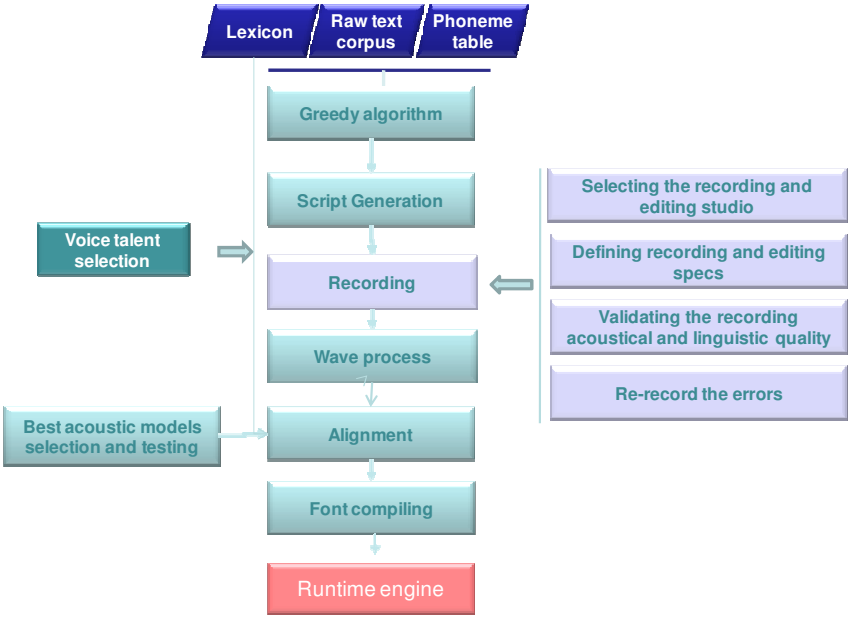

Figure 2: Voice font building.

Test description and results

The BP TTS test plan design addressed the following components:

1. Front-end modules: sentence separator, word breaker, text normalization module, pronunciation (homograph ambiguity resolution, LTS conversion);
2. Back-end and overall system functionality: compliance tests, performance tests and BVT (Build Verification Tests);
3. Overall system voice quality: comprehension tests and MOS (Mean Opinion Score) tests.

In a first stage, the assessment affects module by module. In a second stage, the goal is the whole system in terms of its intelligibility. In a third stage, the system voice quality is assessed by an external group of 40 native people, allowing us to obtain a more realistic feed-back of the quality of the product. In the final stage and before it is released to market, the product is distributed internally and within Microsoft partners, who provide their feedback and report remaining bugs.

For the front-end modules, a significant number of test cases were created with a given input and the expected result, which was compared with the TTS system's output in real time. The tests should pass 100% for cases defined in the specifications as being P0 or P1 for a certain language (P means priority). The definition of the priority level is related with the impact that a certain rule or failure of a rule might have in terms of the final intelligibility of the synthetic voice. The *liaison* module, which is basically a rule-based module that predicts *sandhi* phenomena in Portuguese, was considered producing less impact in the system's intelligibility, and therefore classified with a P2 tag.

For the back-end functionality, a similar procedure was carried out. Several tests must be run in order to assure the compliance, stability and overall quality of the build and the system functionality. In Table 1, the list and results of these tests can be seen.

Comprehension tests are conducted after all the front-end and back-end tests for P0 and P1 tests cases pass 100%. The purpose of these tests is to make a first complete assessment of the system's intelligibility and identify remaining bugs that might have not been found in the first set of tests done to each module independently. This is basically a comprehension test covering domains which reflect the product end-users daily usage. As Lemmety [17] points out, "In comprehension tests a subject hears a few sentences or paragraphs and answers to the questions about the content of the text, so some of the items may be missed [18]. It is not important to recognize one single phoneme, if the meaning of the sentence is understood, so the 100% segmental intelligibility is not crucial for text comprehension and sometimes even long sections may be missed [19]."

The tests were conducted by 7 native listeners (3 female and 4 male) who were exposed to a set of 410 prompts divided by six domains: addresses, date and time, phone numbers (where TN module could be checked), single e-mail/news sentences, e-mail/news paragraphs and proper nouns (where different aspects of grapheme-to-phone conversion could be checked). The test scripts were carefully selected from real texts. The listeners were asked to give a rate to their level of understanding. The classification of the intelligibility was a given in a five-scale rate measured in percentages. The overall intelligibility rate was 98.915%.

Test item	Priority level	current results
Sentence Separator tests	P0	100%
	P1	100%
	P2	78,38%
Word Breaker tests	P0	100%
	P1	100%
	P2	89%
TN tests	P0	100%
Pronunciation test	P0	100%
Liaison test	P2	100%
SpeechFX BVT test	P0	100%
Speech FXFunctional	P0	100%
SSML compliance test	P0	100%
SAPI compliance test	P0	100%
SAPI TTS BVT	P0	100%
SAPI TTS Functional	P0	100%
SAPI security test	P0	100%
MS Intelligibility tests (comprehension)	P0	98,92%
MS Competitor tests (comprehension)	P0	91,90%

Table 1: BP TTS tests results.

A similar test was conducted for a commercial available TTS system in BP, using the same test set and listeners, but in different days. The overall intelligibility rate was 91.9%.

Conclusions and future work

The success rates of our system's comprehension results, when compared with other available commercial systems, may be explained by mainly two reasons: the HTS technology enabled in the back-end, which largely increases intelligibility by making the segmental phone transitions smoother, and the application of several rule-based modules in the front-end, which allows a better accuracy rate in the grapheme-phoneme conversion (namely the homograph disambiguation and the *liaison* modules [14] and a more efficient way of lexicon expansion using an automatic syllable and stress marker [14]). However, naturalness in HTS technology is still

not as good as when compared with unit selection-based systems. More research on naturalness improvement using HTS is ongoing.

MOS tests [17] on our BP TTS are currently being conducted by an external vendor. The listeners are 40 Brazilian native speakers (20 male and 20 female). Forty different tests packages were randomly selected from a set of new prompts and include natural voice wave samples, Microsoft BP TTS synthetic wave samples and other commercial BP TTS synthetic wave samples. This allows each listener to have a different test and it covers more aspects of the language. The assessed domains are basically the same conducted in the comprehension tests. Results will be obtained very soon.

References

1. Boeffard, O.; Violaro, F. 1994. "Using a Hybrid Model in a Text-To-Sppech System to Enlarge Prosodic Modifications", *Proceedings of ICSLP 94*. Yokohama, Japan. pp. 727-730.
2. Violaro, F. & Böeffard, O. 1998. "A Hybrid Model for Text-to-Speech Synthesis", *IEEE Transactions on Speech and Audio Processing*, Vol. 6, no 5, pp. 426-434.
3. Gomes, L.C.T. 1998. *Sistema de conversão texto-fala para a língua portuguesa utilizando a abordagem de síntese por regras*. Dissertação de Mestrado. Campinas: Unicamp.
4. Barbosa, P.; Violaro, F.; Albano, E.; Simões, F.; Aquino, P.; Madureira, S.; Françozo, E. 1999. "Aiuruetê: A High-Quality Concatenative Text-to-Speech System for Brazilian Portuguese with Demisyllabic Analysis-Based Units and a Hierarchical Model of Rhythim Production", *Eurospeech'99 - 6th European Conference on Speech Communication and Technology*. Budapest, Hungria. Volume 5, pp. 2059-2062.
5. Simões, F. O., Violaro, F., Barbosa, P. A. e Albano, E. C. 2000. "Um Sistema de Conversão Texto-Fala para o Português Falado no Brasil", *Revista da Sociedade Brasileira de Telecomunicações*. Vol. 15, no 2, pp. 70-77, dezembro/2000.
6. Nicodem , M. V.; Seara, R.; Pacheco, F. S. 2005. "Reducing the Natural Click Effect within Database for High Quality Corpus-Based Speech Synthesis", *8th IEEE International Symposium on Signal Processing and its Applications*. Sydney, Austrália. pp. 607-610.
7. Nicodem; M. V. , Kafka , S. G.; Seara Junior , R.; Seara, R. 2007. "Refinamento da Segmentação Fonética em Aplicações de Síntese de Fala", *XXV Simpósio Brasileiro de Telecomunicações (SBrT 2007)*. pp.1-6.
8. Maia, R.; Zen, H.; Tokuda, K.; Kitamura T. and Resende, Jr., F. G. V. 2006. "An HMM-based Brazilian Portuguese speech synthesizer and its characteristics", *IEEE Journal of Communication and Information Systems*. No. 2, vol. 21, pp. 58-71.
9. Maia. R. 2006. Speech Synthesis and Phonetic Vocoding for Brazilian Portuguese Based on Parameter Generation form Hidden Markov Models. PhD thesis. Department of Computer Science and Engineering, Nagoya Institute of Technology, Nagoya, Japan.
10. Available at: http://www.nuance.com/realspeak/languages/ (07/07/2008)
11. Available at: http://www.loquendo.it/en/demos/demo_tts.htm (07/07/2008)
12. Available at: http://www.acapela-group.com/portuguese-brazil-46-text-to-voice.html (07/07/2008)
13. Available at: http://www.aculab.com/Support/ttsasr/VoiceStyles.html (07/07/2008)
14. Braga, D. 2008. *Algoritmos de Processamento da Linguagem Natural para Sistemas de Conversão Texto-Fala em Português*. PhD Thesis. A Coruña University, Spain.
15. Braga, D.; Coelho, L.; Resende Jr., F. G. V. 2007. "Homograph Ambiguity Resolution in Front-End Design for Portuguese TTS Systems", Proceedings of Interspeech 2007. Antwerpen, Belgium. pp. 1761-1764.
16. Bonafonte, A.; Höge, H., Tropf, H., Moreno, A.; Heuvel, H.; Sündermann, D.; Ziegenhain, U.; Pérez, J.; Kiss, I. 2004. *Deliverable no.: D8. TTS Baselines and specifications.* TC-Star (Technology and Corpora for Speech to Speech Translation) Report. FP6-506738.
17. Lemmety, S. 1999. *Review of Speech Synthesis Technology*. Ms Thesis. Helsinki University of Technology.
18. Allen J.; Hunnicutt S.; Klatt D. 1987. *From Text to Speech: The MITalk System*. Cambridge University Press, Inc.
19. Bernstein J.; Pisoni D. 1980. "Unlimited Text-to-Speech System: Description and Evaluation of a Microprocessor Based Device" *Proceedings of ICASSP 80* (3): 574-579.