In: *Proceedings of the Eurospeech'99*, Budapest, Hungary, 5-9 september. v. 5, 2059-2062.

AIURUETÊ: A HIGH-QUALITY CONCATENATIVE TEXT-TO-SPEECH SYSTEM FOR BRAZILIAN PORTUGUESE WITH DEMISYLLABIC ANALYSIS-BASED UNITS AND A HIERARCHICAL MODEL OF RHYTHM PRODUCTION

Plínio A. Barbosa[*], *Fábio Violaro*[*], *Eleonora C. Albano*[*], *Flávio Simões*[*], *Patrícia Aquino*[*], *Sandra Madureira*[†] *and Edson Françozo*[*]
*Universidade Estadual de Campinas and †Pontifícia Universidade Católica de São Paulo, Brazil

ABSTRACT

Aiuruetê is a high-quality concatenative TTS system for Brazilian Portuguese. Its name (pronounced [ajuˌɾueˈte]) illustrates the challenges we have fixed as a research paradigm: to feed the system with the specificities of our language, highlighted by an up-to-date discussion of the Phonology/Phonetics and prosody/segments interfaces, without a huge computational cost. The choice for the concatenative method of synthesis was determined by a trade-off between scientific (the desired human-like naturalness of the acoustic output) and practical (mainly reduced staff and tight schedule) constraints. Procedural and declarative modules are described here: the ortofon, the unit inventory, the rhythm model and the synthesis techniques. Aiuruetê is still being evaluated, but when compared to the previous system, adopted by the national telephony company, its superior quality is apparent.

1. INTRODUCTION

This work summarizes the results of five years of joint efforts from linguists and engineers in order to build a high-quality concatenative TTS system for Brazilian Portuguese (henceforth BP) which we have coined after the Tupi-Guarani name of the most talkative of the Psittacidae: aiuruetê (true parrot, species Amazona æstiva).

Since 1994, these two groups of speech scientists, linguists from the Laboratório de Fonética Acústica e Psicolingüística Experimental (LAFAPE) and engineers from the Laboratório de Processamento Digital da Fala (LPDF), both institutions of a public university, have faced and overcome different kinds of academic and financial obstacles to pursue the goal of building a high-quality concatenative TTS system that aims at preserving the main specificities of Brazilian-spoken Portuguese. From the very beginning, technology was never an end in itself, but a by-product of decisions taken from our linguistic and engineering insights.

The choice for the synthesis method was guided by an engineering goal: to build a natural speech synthesis system without a high computational cost. The condition of great naturalness and our tight schedule have excluded a synthesis-by-rule method from the very beginning. To invest in articulatory synthesis would be a very costly choice considering our limited resources. The concatenative method has naturally emerged from these considerations.

The grapheme-to-phone converter (henceforth Ortofon), the unit inventory and the rhythm model have inherited most of their high-quality features from discussions concerning the Phonetics-Phonology interface and the prosody-segments interaction, both in the light of a dynamical-system perspective (cf. Browman and Goldstein's *Articulatory Phonology* [7][8] and Port's *Temporal Phonology* [19]). Before setting the system into operation, it is possible to select one from two pitch-synchronous synthesis techniques: TD-PSOLA or the hybrid model.

All modules and interfaces, mostly implemented in C++, form a classical concatenative TTS system layout which was installed in a PC with a friendly DELPHI interface for the user [21]. All modules were independently developed by different research teams.

2. THE ORTOFON

Due to research convenience, the Ortofon was separated into two parts: the preprocessor and the grapheme-to-phone converter itself. The preprocessor transduces grapheme to grapheme by treating classical problems associated with the pronunciation of abbreviations, numbers, special symbols and acronyms. The second part reads the previous output and transduces the grapheme set to the phoneme-like notation of BP sounds [3].

Our notation explores the ASCII table of characters (for ease of implementation among the different modules) and differentiates reduced (and reduceable) from plain phones. It is important to stress that this notation does not represent allophonic variation that is intrinsic to the concatenative units (for instance, the affrication of [t] before [i] and [ɪ] is present within the "ti" unit and does not need to be redundantly indicated in the notation). Reduced and reduceable forms were represented in capital letters, and plain forms in minuscules: in the word "casa" (house), for instance, our phonic notation transcribes it [kazA], which is equivalent to the IPA transcription [ˈkazɐ]. Sequences of up to two characters may stand for one sound. In this regard, we can highlight the use of the character "h" for marking greater vowel aperture (e.g. [eh] corresponds to [ɛ]) and for representing some IPA consonantal sounds as [ʃ],[ʒ],[ʎ] and [ɲ] (respectively equivalents to [sh], [zh], [lh] and [nh]). In BP, there are only four consonants that can occupy the coda position: /s/, /r/, /l/ and the archiphoneme /N/. They are all realized as reduced forms and then signalled in our notation by a capital letter: [S], [R], [L] and [N]. The liquids are also reduced after plosives and [f]/[v] within onset clusters.

BP vowel nasalization deserves some clarification. It is phonologically characterized by a vowel followed by the nasal archiphoneme /N/ (whose point of articulation is traditionally assimilated to the next consonant, if present). We decided to maintain this convention even if the phonetic reality is more complex (cf. next section). In fact, this decision has allowed phonetically transparent concatenation rules as far as nasalized vowels are concerned (cf. section 5).

This module works with less than 4% of errors, tested on databases of Brazilian newspapers, even without a complete parsing operation. The parser is being studied and implemented in Prolog by our team of psycholinguists, headed by Françozo [20]. This team needs to solve some tricky problems as the vowel aperture distinction in the pair "sede" (/sede/, *thirst*) vs "sede" (/sɛde/, *establishment*). In this regard, the ortofon is already able to treat the most frequent cases of verb vs adjective pair (e.g. "seca" /sɛka/ - *(he/she) dries* - vs "seca" /seka/ - *dried*, feminine) distinctions by including simple statistics-based syntactic analyses. In the next lines, some samples of input/output examples of this module are shown.

input: "Eu seco o carro que não está seco." (*I have wiped the car which had got wet.*).
output: [eU sehkO O karO ke naNU eSta sekO]

input: "O canto é um belo exercício." (*Singing is a wonderful exercise*.)
output: [O kaNtO eh uN behlO ezeRsisIO]

input: "O enxame de abelhas feriu o rapaz." (*The swarm of bees stung the young man*.)
output: [O eNshamE dE abelhAS feRiU O rapaS]

3. THE UNIT INVENTORY

The polyphone database has circa 2,500 units, from demisyllables to five-segment sequences. The phonotactly most common within-word phone sequences were embedded within non-sense words and the sequences naturally pronounced across word boundaries were embedded in two-word sequences. In both cases, non-sense and real words were inserted in carrier sentences. Our speaker was selected from a group of possible candidates according to some criteria as good distinction between vowels pairs, capacity of perfectly executing the control instructions, a modal voice quality, among others. At the time of recording, he was about 35 years old. He is natural of Recife, capital of the State of Pernambuco, in Northeast Brazil.

It is important to note that before retrieving the to-be-concatenated units in the inventory, an algorithm, developed by Simões [21], reads the ortofon output and determines the correct concatenative polyphone structure. This is necessary because we have units of different sizes: from two to five segment sequences. For instance, to synthesize the sentence "O canto da arara" (The macaw's singing), the ordered polyphones [_O, Ok, kA, aNt, tO, Od, dA, Aa, aRa, aRA, A_] are the correct choices for retrieval at the inventory. Note particularly the sequence [kA, aNt]. In the case of [aN], the reduced oral /a/, [A], is used to concatenate with the VNC next sequence [aNt] as a consequence of the demisyllabic analysis and the nasality specificities of BP.

In BP, tautosyllabic VC segment sequences are strongly coarticulated with fast formant transitions, frequently without a clear steady state. Nasalized vowels and rhymes ending with /s/ illustrate these facts [2].

The VN sequence is phonetically triphasic [22]. It begins with an oral, reduced version of the phonological vowel, turns into a fully nasalized vowel and ends with a nasal murmur of reduced amplitude.

Coda /s/ are weak and tightly coproduced with the previous vowel. Sometimes this vowel has no steady state or can exhibit the phenomenon of iotization [1] ([j] epenthesis between the vowel and the alveolar fricative).

Diphthongs, which are usually offgliding in BP, were preserved by setting the left boundary at the transition between the previous segment and the next vowel, and the right boundary at the transition between the semivowel and the next segment. Sequences of two or three vowels also constitute a large set of units.

The complex nature of these examples indicates that diphones are not the best choice for BP. Instead, a demisyllabic analysis [12] was adopted to split typical CVC sequences into a very short CV unit (preserving the sole transition) and a larger rhyme (VC) unit. In this kind of analysis, the cutting points which define the concatenative unit boundaries are set at the vicinity of the so-called P-center [17].

4. THE PROSODY MODULE

Research on intonation is in progress [14][16] and has pointed out some specificities of BP f_0 patterns. It is unlikely that f_0 rising can be triggered by post-stressed syllables. Instead, they function as a medium where f_0 rising/levelling-off and rising/falling movements, triggered by lexically stressed syllables, can unfold [15]. For oxytons, this movement extends to the first pre-stressed syllable after the phrase boundary. But for the moment, as a default, f_0 contours are considered to be simple declarative declination lines.

Another work, this one on rhythm, has in fact confirmed the limits of BP phrase stress domain by showing that stress degree extends to post-stressed syllables and includes at least one onset consonant after the phrase boundary [6]. This work is based on the assumption that the speech chain is twofold: a continuous vowel flow (probably represented by the alternate movement of jaw openings and closings) on which consonant gestures are superimposed [11][18]. This means that the minimal unit of rhythmic organization (the so-called rhythmic programming unit, henceforth RPU) has the syllable size. Elsewhere we have shown that BP needs two RPUs to charaterize its duration patterns: the syllable and the inter-p-center group (IPCG) [5].

Marcus and colleagues [17] have coined the term P-center for designating acoustic anchor points in the speech signal which are used by listeners to perceive sequences of syllables as occurring isochronously. More ecological data [13] have confirmed Marcus's intuitions about the P-center location: the vowel onset. It seems to us that P-centers would delimitate the minimal RPU for the phonetic implementation of rhythm. Port and colleagues' experiments confirm this assumption by choosing vowel onsets as rhythmic beats in their speech cycling task [19].

These assumptions form the basis of a two-stage model of segmental duration assignment. In the first stage, a neural network automatically generates z-score values characterizing syllable and IPCG lengthening (or shortening) from a rich prosodic phonetic description at the input. This input is constituted by the ortofon output enriched by additional information as the number of vowels in the sentence, the clitic words and the phrasal stresses, whose placement follows eurhythmic constraints. In the next stage, a statistical procedure shares the syllable-size duration amount among the segments, based on their respective statistical mean and standard-deviation durations [4]. These segmental durations are used by both synthesis techniques to generate sound.

5. THE SYNTHESIS TECHNIQUES

Two pitch-synchronous techniques can be used to recode pre-stored units: TD-PSOLA [9] or the hybrid model [23]. The TD-PSOLA technique is well known in the literature, but the situation is not the same for the hybrid model.

The hybrid model codes all pitch-marked concatenative units with a set of parameters which shows two advantages over the TD-PSOLA technique: (1) the synthesizing of the unit with any value of duration, and not with only the pitch-quantized characteristic of the TD-PSOLA technique, and (2) modifications in f_0 can be done without the need of readusting durations. These alterations are done by previously separating the noisy component from the harmonic one (that is why the name *hybrid*) in the voiced chunks of the signal and only a noisy component, in the unvoiced chunks. The harmonic component is analysed by using a procedure based on DFT that computes a maximum frequency for each corresponding speech signal chunk (normally two pitch periods). This frequency can take a value from 2 to 5 kHz in steps of 1 kHz. This off-line phase is made once. Several evaluations have shown that this technique is very sensitive to the voiced/unvoiced distinction.

6. CONCLUSION

Comparing the TTS synthesis system adopted by Telebrás (National Telephony Company) [10], which has a 1200-polyphone database, with Aiuruetê, the preference for our system is apparent (both systems using the TD-PSOLA technique).

Some samples of synthesized sentences will be presented during the Conference, including the presentation sentence, "Meu nome é Aiuruetê" (My name is Aiuruetê), which has a very complex word to pronounce due to its diphthong/hiatus/tap/hiatus sequence: [ajuˌɾueˈte].

We have shown that with reduced financial and human resources it is possible to build a high-quality TTS synthesis system. This quality would be impossible without basic research on BP linguistic specificities carried out under a dynamical segmental and prosodic phonological/phonetic perspective.

ACKNOWLEDGEMENTS

This work is financed in part by the project *Jovem Pesquisador em Centro Emergente*, for the first author, (proc. n° 95/09708-6) from Fundação de Amparo à Pesquisa do Estado de São Paulo-FAPESP and by grants (procs. n° 524110/96-4) from the Conselho Nacional de Desenvolvimento Científico e Tecnológico-CNPq. During the five years it was also financed by the *Projeto Temático* (proc. n° 93/00565-2) from FAPESP. We also especially thank our speaker for his voice.

REFERENCES

[1] Albano, E.C. (1999), Gestural solutions for some glide epenthesis problems. To be published in *Proceedings of the XIVth International Congress of Phonetic Sciences*, August, San Francisco, Berkeley: University of California.

[2] Albano, E.C. and Aquino, P. A. (1997), Linguistic criteria for building and recording units for concatenative speech synthesis in Brazilian Portuguese. *Proceedings of Eurospeech'97*, v.2, 725-728. Rhodes, Greece.

[3] Albano, E.C. and Moreira, A.A. (1996), Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese. *Proceedings of the ICSLP'96*, October 3-6, v.3, 1708-1711.

[4] Barbosa, P.A. (1997), A model of segment (and pause) duration generation for Brazilian Portuguese text-to-speech synthesis. *Proceedings of Eurospeech'97*, v.2, 2655-2658. Rhodes, Greece.

[5] Barbosa, P.A. (1996), At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration. *Proceedings of the first ESCA Tutorial and Research Workshop on speech production modeling*, 85-88.

[6] Barbosa, P.A. and Madureira, S. (1999), Toward a hierarchical model of rhythm production: evidence from phrase stress domains in Brazilian Portuguese. *Proceedings of the XIVth International Congress of Phonetic Sciences*, August, San Francisco, Berkeley: University of California.

[7] Browman, C. and Goldstein, L. (1992), Articulatory Phonology: an overview. *Phonetica*, 49. 155-180.

[8] Browman, C. and Goldstein, L. 1990. Tiers in Articulatory Phonology with some implications for casual speech. In: Kingston, J. and Beckman, M.E. (Eds.) *Papers in Laboratory Phonology I.* Cambridge: Cambridge University Press, 341-376.

[9] Charpentier, F. and Moulines, E. (1990), Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9 (5/6), 453-467.

[10] Egashira, F. and Violaro, F. (1995), Conversor Texto-Fala para a Língua Portuguesa. *13^{o} Simpósio Brasileiro de Telecomunicações*, September 3-6, Brazil, pp. 71-76.

[11] Fujimura, O. (1995). Prosodic organization of speech based on syllables: the C/D model. *Proceedings of the XIIIth International Congress of Phonetic Sciences,* 3, 10-17.

[12] Fujimura, O. (1979), An analysis of English syllables as cores and affixes. *Zs. für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, 32, 471-476.

[13] Janker, P. (1995). On the influence of the internal structure of a syllable on the P-center perception. *Proceedings of the XIIIth International Congress of Phonetic Sciences*,2, 510-513.

[14] Madureira, S. (1994), Pitch patterns in Brazilian Portuguese: an acoustic-phonetic analysis. *Proceedings of the fifth Australian International Conference on Speech Science and Technology*, Perth, Australia, v.1, 156-161.

[15] Madureira, S., Barbosa, P.A., Fontes, M., Spina, D. and Crispim, K. (1999), Post-stressed syllables in Brazilian Portuguese as markers. To be published in *Proceedings of the XIVth International Congress of Phonetic Sciences*, August, San Francisco, Berkeley: University of California.

[16] Madureira, S. and Fontes, M. (1997), Fundamental contours in Brazilian Portuguese words. In: Botinis, A. Kouroupetroglou, G. and Carayiannis, G. (Eds.) *Proceedings of the ESCA workshop Intonation: Theory, Models and Applications*. September 18-20, Athens, Greece. Univ. of Athens. pp. 211-214.

[17] Morton, J., Marcus, S. and Frankish, C. (1976), Perceptual centers (p-centers*). Psychological revue*. 83 (5), 405-408.

[18] Öhman, S. (1966), Coarticulation in VCV utterances: spectrographic measurements. *J. Acoustic. Soc. Am*., 39, 151-168.

[19] Port, R., Cummins, F. and Gasser, M. (1995), A Dynamic approach to rhythm in language: toward a Temporal Phonology. *Proceedings of the Chicago Linguistics Society*. Luka, B. and need, B. (Eds.), 375-397.

[20] Rosa, J.L.G and Françozo, E. (1999), Hybrid Thematic Role Processor: Symbolic Linguistic Relations Revised by Connectionist Learnings, To be published in *Proceedings of the IJCAI'99 - Sixteenth International Joint Conference on Artificial Intelligence*, Stockholm, Sweden.

[21] Simões, F. O. (1999), Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil. *Unpublished master's thesis*. Campinas: DECOM/FEEC/UNICAMP, Brazil.

[22] Sousa, E.G. (1994), Para a caracterização fonético-acústica da nasalidade no português do Brasil. *Unpublished master's thesis.* Campinas: LAFAPE/IEL/UNICAMP, Brazil.

[23] Violaro, F. and Böeffard, O. (1998), A hybrid model for text-to-speech synthesis. *IEEE Transactions on Speech and Audio Processing*, September, 6 (5), 426-434.