Uma Nova Abordagem sobre Modelagem Prosódica para um Sistema Texto-Fala Concatenativo

Manuel Leonel da Costa Neto, Benedito Guimarães Aguiar Neto e Maria Auxiliadora Bezerra

***Resumo*— Este artigo apresenta uma nova abordagem sobre modelagem prosódica para um sistema texto-fala concatenativo para o Português Brasileiro. O modelo é baseado em regras e na tonicidade de palavras para determinar os contornos de entonação. São utilizadas sílabas e demissílabas como unidades acústicas do dicionário do sistema texto-fala. A base de dados de unidades acústicas, relativamente elevada, incorpora informações prosódicas das mais relevantes, bem como das características articulatórias correspondentes aos fenômenos de coarticulação, para a obtenção de uma fala sintetizada de qualidade.**

***Palavras-Chave*— Prosódia, Sistema Texto-Fala, Síntese da Fala.**

***Abstract*— This article presents a new approach on prosody modelling in a concatenative text-speech system for Brazilian Portuguese. The model is based on rules and on stressed syllables in the words, to determine the intonation contours. Syllables and demissylables are used as acoustic units of the dictionary. The database of units acoustic, relatively high, incorporates the most important prosodic information, as well as the articulatory characteristics corresponding to the coarticulation phenomena, in order to obtain a good synthesized speech.**

***Keywords*— Prosody, Text-Speech System, Speech Synthesis.**

I. INTRODUÇÃO

Na comunicação vocal homem-máquina, vários sistemas de produção da fala têm sido desenvolvidos, porém os sistemas de síntese da fala apresentam várias vantagens devido a maior flexibilidade e capacidade de compressão da informação. De forma geral, esses sistemas podem ser divididos em dois grupos [1], [2]: sistemas de reprodução vocal e sistemas de conversão texto-fala. Os sistemas de reprodução vocal se baseiam no processo de concatenação de palavras isoladas ou partes de sentenças. Nesse caso, tem-se um melhor controle da naturalidade e inteligibilidade da fala, mas em contrapartida só podem ser usados em aplicações que envolvam um vocabulário restrito. Por outro lado, os sistemas de conversão texto-fala produzem a fala automaticamente a partir de um texto com base em algoritmos de processamento lingüístico e de sinais. Nesse caso, tem-se maior flexibilidade, pois o vocabulário pode ser irrestrito, de modo que não há necessidade de armazenar todas as possíveis palavras de uma língua. Na realidade são usadas unidades acústicas, através das quais as palavras são construídas. Dessa forma, tal sistema se torna mais complexo que o anterior, pois inclui um tratamento lingüístico e prosódico para a obtenção de uma fala inteligível e natural.

Manuel Leonel da Costa Neto¸ Departamento de Engenharia de Eletricidade, Universidade Federal do Maranhão, São Luís, Maranhão, Brasil; Benedito Guimarães Aguiar Neto, Departamento de Engenharia Elétrica, Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brasil; Maria Auxiliadora Bezerra, Departamento de Letras, Universidade Federal de Campina Grande, Campina Grande, Paraíba, Brasil, e-mails: leonel@dee.ufma.br, bganeto@dee.ufcg.edu.br, cidabezerra@uol.com.br.

Dentre os aspectos envolvidos na síntese da fala tem-se dado grande destaque atualmente à prosódia, que limita-se ao estudo de três elementos: o acento de energia (ligado à maior ou menor força com a qual o ar é expelido dos pulmões), o acento de entonação (ligado à maior ou menor freqüência do som fundamental), e a duração (ligada à sustentação maior ou menor do fonema) [3]. Portanto, a sua principal atribuição é destacar as sílabas predominantes (sílabas tônicas) na produção das palavras, o ritmo e a entonação na produção de frases.

Para aplicar a prosódia em um sistema de síntese texto-fala é necessário definir um modelo prosódico, que determina a evolução temporal dos parâmetros prosódicos, de forma que seja possível identificar na fala a acentuação, o ritmo e a entonação. Normalmente é utilizado um modelo de duração e um modelo de entonação, ou apenas este último. No modelo de duração é realizado um tratamento automático no qual as durações dos fones de um enunciado possam ser determinadas [4]. No modelo de entonação é aplicado um contorno padrão em todas as frases geradas no estágio de síntese, a partir de um conjunto de regras baseadas em informações sintáticas e do conhecimento de fronteiras entre palavras e sílabas [5], [6], ou extraído de elocuções naturais, que são ajustadas ao texto que se deseja sintetizar [7], [8].

Vários métodos tem sido utilizados para a geração da prosódia em sistemas de síntese texto-fala, principalmente os baseados em regras [9], [10], em técnicas de aprendizagem com redes neurais [11], [12], em árvores de classificação e regressão [13], [14], ou em modelos de Markov escondidos (HMMs) [15], [16]. Cada método apresenta vantagens e desvantagens de modo que se torna difícil um estudo comparativo entre eles, principalmente em termos práticos. O método baseado em regras, por exemplo, depende da língua com a qual se está trabalhando e as regras podem atuar em nível de fonema, sílaba, palavra ou frase. Quanto mais completo for o conjunto de regras melhor serão os resultados [17]. Por outro lado, o método baseado em HMMs produz bons resultados segundo os autores, mas tem a desvantagem de ser mais complexo que o de regras e de eventualmente, produzir erros bastante grosseiros principalmente quando forem encontrados contextos fonético-prosódicos mais raros, não contemplados no *corpus* da base de dados [16], [17].

Este artigo apresenta uma nova abordagem sobre modelagem prosódica para um sistema texto-fala concatenativo para o Português Brasileiro. O modelo é baseado em regras

e na tonicidade de palavras para determinar os contornos de entonação. A base de dados de unidades acústicas, relativamente elevada, incorpora informações prosódicas das mais relevantes, bem como das características articulatórias correspondentes ao fenômeno de coarticulação. O modelo aqui proposto de tonicidade é levado a efeito por palavras, considerando-se a complexidade que seria em uma abordagem por frases, permitindo assim que aspectos fonético-fonológicos sejam melhor explorados no âmbito de palavras, e posteriormente possam servir de suporte a futuros estudos de modelagens mais abrangentes. O objetivo é obter uma simplificação nos estágio de processamento do sinal de um sistema texto-fala e a produção de uma fala sintetizada inteligível e natural.

Na Seção 2 são apresentados os estágios do sistema de síntese concatenativo, adapatado ao modelo prosódico. Na Seção 3 é apresentado o modelo. Na Seção 4 são apresentados os resultados obtidos com o modelo aplicado ao sistema. Na Seção 5 são apresentadas as conclusões e sugestões sobre o desenvolvimento de trabalhos futuros.

II. Sistema de Síntese

O sistema de síntese texto-fala concatenativo para o Português Brasileiro, no qual o modelo prosódico é aplicado, é apresentado na Figura 1.

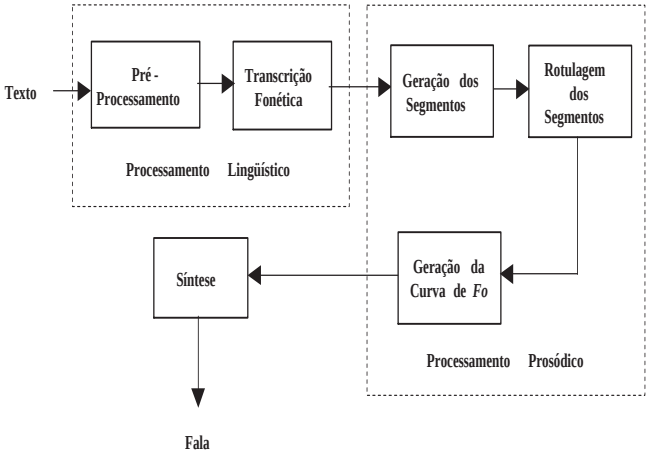

Fig. 1. - Sistema de síntese texto-fala concatenativo proposto para o Português Brasileiro.

No estágio de Pré-Processamento siglas, números e abreviaturas são escritos por extenso, como, por exemplo, Sr. → Senhor. Para tal, são utilizadas regras e um banco de dados contendo siglas, números e abreviaturas e os respectivos nomes escritos por extenso. No estágio de Transcrição Fonética, é realizada a troca de letras pelos respectivos fonemas, através de regras estabelecidas com base em conhecimentos fonéticos para o Português Brasileiro [18], [19], como, por exemplo, casa → kaza. No estágio de Geração dos Segmentos, são selecionados inicialmente os grupos CV (consoante-vogal). Posteriormente as vogais ou consoantes restantes são agrupadas ou não aos grupos CV, dependendo da estrutura das sílabas, como, por exemplo na palavra *substantivo* tem-se: su/b/s/ta/m/ti/vu → subs/tam/ti/vu. No estágio Rotulagem dos Segmentos, os segmentos fonéticos resultantes da etapa anterior são rotulados com valores conforme a tonicidade e a posição que ocupam dentro da palavra. Para isso, é usada uma estrutura de pesos para palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, apresentada na Tabela I.

TABELA I
VALORES ARBITRADOS AOS PESOS DAS SÍLABAS NO MODELO PROSÓDICO

Sílaba	Pre4	Pre3	Pre2	Pre1	Tônica	Pos1	Pos2
Peso	1	2	3	4	5	6	7

Onde: Pre1, Pre2 Pre3 e Pre4 correspondem às sílabas pretônica 1, pretônica 2, pretônica 3 e pretônica 4, respectivamente e Pos1 e Pos2 correspondem às sílabas postônica 1 e postônica 2, respectivamente.

No estágio Geração da Curva de F_0 é feito um mapeamento dos pesos sobre os padrões de F_0 de cada segmento fonético. Um exemplo de classificação para as sílabas da palavra *paroxítona* com o valor de F_0 (Hz) correspondente é apresentado na Tabela II. Após a classificação e rotulamento tem-se a seguinte transcrição fonético-prosódica: [pa 117 / ro 121 / xí 127 / tom 132 / na 135].

TABELA II
CLASSIFICAÇÃO DAS SÍLABAS DA PALAVRA *paroxítona* COM O VALOR DE F_0 (HZ) CORRESPONDENTE

Sílaba	pa	ro	xí	to	na
Peso	3	4	5	6	7
F_0 (Hz)	117	121	127	132	135

No estágio Síntese é realizada a concatenação de unidades acústicas (junção dos segmentos dos sinais de fala de forma seqüenciada), conforme os comandos de saída do processamento prosódico. As unidades estão contidas em um dicionário (base de dados) e rotuladas conforme o rotulamento resultante nas unidades fonético-prosódicas na saída do estágio do processamento prosódico. Por exemplo, se na saída do processamento for gerado o segmento fonético 'pa 117', onde 117 é a frequência em Hz da unidade 'pa', rotulada conforme o modelo prosódico definido para a sílaba em função da sua posição na palavra e da tonicidade, deve existir no dicionário a unidade acústica correspondente (arquivo .WAV) rotulado como 'pa 117', para que seja selecionado e seja gerada a fala.

III. Modelo Proposto

O modelo de prosódia proposto é baseado em regras e na tonicidade de palavras para determinar os contornos de entonação. Assim, são estabelecidas regras para a determinação da freqüência fundamental das unidades acústicas correspondentes às sílabas, em função do contexto em que se inserem. São consideradas as seguintes características da palavra:

1) Número de sílabas (monossílabos, dissílabos, trissílabos e polissílabos com até cinco sílabas);

2) Tonicidade: sílaba tônica e sílaba átona (postônica e pretônica);
3) Posição da sílaba tônica (oxítonas, paroxítonas, proparoxítonas);
4) Número de caracteres do segmento fonético correspondente a cada sílaba; e
5) Estrutura do segmento fonético correspondente a cada sílaba no que se refere à combinação das vogais com as consoantes (CV, CVC, CVV,...).

A. Padrões de Freqüência Fundamental das Unidades

Para a determinação dos padrões de freqüência fundamental das unidades acústicas a serem usadas no modelo, foi inicialmente elaborado e gravado um *corpus* constituído por palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, contendo as mais diversas combinações de fonemas para as sílabas. Para evitar o efeito da leitura em forma de lista de palavras e para dar mais naturalidade às palavras em análise, foram usadas, na etapa de gravação, frases veículo do tipo '*digo* (**palavra**) *baixinho*' [20]. Também foi gravado um *corpus* de 200 frases foneticamente balanceadas [1] elaboradas por Alcaim et al. [21]. As unidades acústicas correspondentes às sílabas foram segmentadas utilizando-se o editor de áudio *Sound Forge* e foram determinados os valores de freqüência fundamental (F_0), usando-se o detetor de *pitch* desenvolvido por Fechine em [22]. Analisando-se os valores de F_0 das unidades acústicas, correspondentes às sílabas do *corpus* estabelecido, são encontrados alguns aspectos importantes, os quais foram incorporados no modelo, tais como:

- em palavras oxítonas, as sílabas pretônicas apresentam valores de F_0 inferiores aos das sílabas tônicas (em torno de 10%). Um exemplo é observado na palavra *abafar*, onde os valores de F_0 medidos para as sílabas pretônicas **a** e **ba** são 135 e 140 Hz, respectivamente, enquanto que a sílaba **far** tem um valor de 144 Hz. Essa tendência ocorre em mais de 80% das palavras oxítonas no *corpus* considerado.
- em palavras paroxítonas, as sílabas postônicas têm valor de F_0 superior a F_0 das tônicas (em torno de 10%), e as sílabas pretônicas apresentam valores de F_0 inferiores aos das sílabas tônicas (em torno de 10%). Um exemplo é observado na palavra *discurso*, onde os valores de F_0 medidos para as sílabas: pretônica **dis**, tônica **cur** e postônica **so**, são 110, 122 e 130 Hz, respectivamente. Essa tendência ocorre para em mais de 85% das palavras paroxítonas no *corpus* considerado.
- em palavras proparoxítonas polissílabos, tem-se um crescimento de F_0 das pretônicas para as tônicas (em torno de 10%) e das tônicas para as postônicas (em torno de 5%). Um exemplo é observado na palavra *matemática*, onde os valores de F_0 medidos para as sílabas: pretônicas **ma** e **te**, para a tônica **má** e para as postônicas **ti** e **ka**, são 114, 128, 131, 137 e 138 Hz, respectivamente. Essa tendência ocorre em mais de 80% das palavras proparoxítonas no *corpus* considerado.

Estes aspectos também foram observados em estudos realizados por Madureira em [23], para trissílabos.

Portanto, é estabelecida uma relação de freqüência fundamental entre as sílabas tônicas, pretônicas e postônicas de palavras com até cinco sílabas, conforme mostrado na Tabela III, resultando em um modelo padrão para a freqüência fundamental das unidades acústicas usadas no modelo proposto.

TABELA III
MODELO GERAL DE FREQÜÊNCIA FUNDAMENTAL DAS UNIDADES CORRESPONDENTES ÀS SÍLABAS

Relação de Freqüência Fundamental entre Sílabas
(F_0 da sílaba tônica) $<$ (F_0 da sílaba Pos_1) $<$ ($1,10.F_0$ da sílaba tônica)
(F_0 da sílaba tônica) $<$ (F_0 da sílaba Pos_2) $<$ ($1,15.F_0$ da sílaba tônica)
($0,96.F_0$ da sílaba tônica) $<$ (F_0 da sílaba Pre_1) $<$ (F_0 da sílaba tônica)
($0,92.F_0$ da sílaba tônica) $<$ (F_0 da sílaba Pre_2) $<$ (F_0 da sílaba tônica)
($0,88.F_0$ da sílaba tônica) $<$ (F_0 da sílaba Pre_3) $<$ (F_0 da sílaba tônica)
($0,84.F_0$ da sílaba tônica) $<$ (F_0 da sílaba Pre_4) $<$ (F_0 da sílaba tônica)

A incorporação dos valores de F_0 para as unidades acústicas usadas no modelo é realizada através de ajustes na freqüência fundamental dessas unidades, usando-se o editor de áudio *Sound Forge*.

B. Duração das Unidades

Apesar de o modelo prosódico proposto ser baseado na entonação das palavras, de modo que a seleção de unidades acústicas no dicionário seja realizada com base em valores de F_0, determinadas características de duração relativas à tonicidade também são incorporadas nessas unidades. Assim, a partir de uma análise nos valores de duração das unidades acústicas, correspondentes às sílabas do *corpus* de palavras e frases estabelecido anteriormente, são encontrados alguns aspectos importantes, tais como:

- sílabas tônicas têm uma tendência de possuir duração maior do que sílabas pretônicas e sílabas postônicas. Isso pode ser observado no exemplo da forma de onda da palavra *benéfico*, apresentada na Figura 2. Neste caso a sílaba tônica *né* tem duração de 254 milisegundos, enquanto que as postônicas *fi* e *co* tem durações de 148 e 198 milisegundos, e a pretônica *be* tem duração de 138 milisegundos. Esse fato é corroborado em estudos realizados por Massini-Cagliari em [24], que conclui que a maioria das sílabas tônicas tem duração maior que as sílabas átonas, para um determinado conjunto de palavras. Em um total de 626 palavras analisadas, observou-se que mais de 80% mantêm-se nessa regra.
- sílabas tônicas com vogais abertas (tipo /a/) têm uma tendência de possuir uma duração maior do que tônicas com vogais fechadas (tipo /ê/). Isso pode ser observado, por exemplo, na forma de onda das palavras *página* e *pêsames*, apresentadas na Figura 3. A sílaba *pá* da palavra *página* tem uma duração de 204 milisegundos e a sílaba *pê* da palavra *pêsames* tem uma duração de 182 milisegundos. Esse fenômeno também ocorre com

[1] Uma lista de frases é foneticamente balanceada quando a freqüência de ocorrência dos fones se aproxima de modo significativo daquela com que ocorrem na língua falada [21].

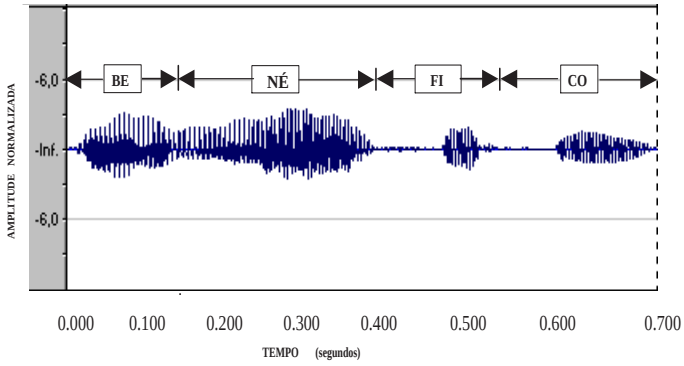

Fig. 2. - Forma de onda no tempo da palavra *benéfico*.

sílabas postônicas e pretônicas em posições equivalentes nas palavras, e é confirmado em estudos realizados por Massini-Cagliari em [24], a qual conclui que vogais abertas têm uma duração maior que as fechadas.

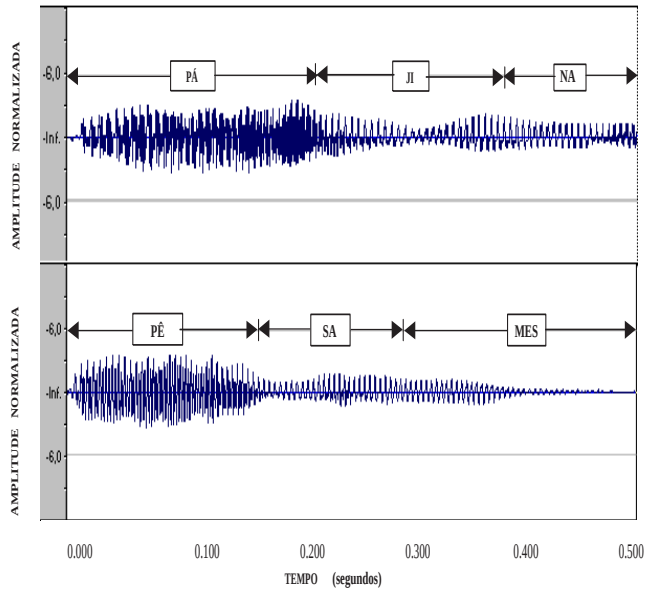

Fig. 3. - Formas de onda no tempo das palavras *página* e *pêsames*.

- sílabas com vogais nasais ou nasalisadas (tipo /am/) têm uma tendência de possuir duração maior do que sílabas com vogais orais (tipo /a/). Isso pode ser observado, por exemplo, nas formas de onda das palavras *abater* e *abanar*, apresentadas na Figura 4. A sílaba *ba* da palavra *abater* tem uma duração de 130 milisegundos e a sílaba *ba* da palavra *abanar* é nasalisada pela sílaba *nar*, e tem uma duração de 175 milisegundos. Esse fato também é corroborado em estudos realizados por Massini-Cagliari em [24] e por Moraes em [25].

Portanto, além de padrões de freqüência fundamental também são criados padrões de duração, para as unidades acústicas do dicionário, para atender a tonicidade de sílabas em palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas.

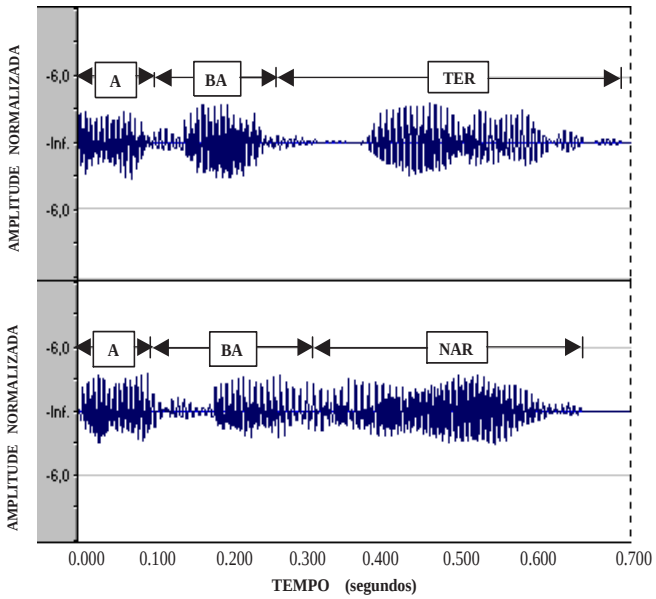

Fig. 4. - Formas de onda no tempo das palavras *abater* e *abanar*.

IV. Resultados Experimentais

Para verificar a eficiência do modelo prosódico proposto, foram realizados testes com o objetivo de analisar a tonicidade das palavras, em termos de freqüência fundamental e duração. Assim, o modelo foi avaliado através de testes informais de escuta em palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, obtendo-se bons resultados. Também foram realizados testes formais em um *corpus* de 20 frases foneticamente balanceadas, conforme relação apresentada na Tabela IV do Apêndice I, que constituem uma amostra representativa do universo de todos os fonemas e unidades do dicionário. Os testes com 20 frases foi realizado com 40 ouvintes e usando-se a escala MOS (*Mean Opinion Score*), obtendo-se um escore de 4,25, superior ao escore bom (4,0). Cada ouvinte atribuiu uma nota a cada frase escutada, conforme a escala e os critérios de classificação definidos a seguir:

1) Péssimo - **alta degradação da fala**. (frase onde todas as palavras não foram entendidas);
2) Pobre - **com muitas falhas** - (frase com a maioria das palavras sem inteligibilidade e naturalidade);
3) Regular - **com determinada quantidade de falhas, porém aceitável** - (frases com metade de palavras com pouca inteligibilidade e sem naturalidade);
4) Bom - **com poucas falhas** - (frase com no máximo duas palavras sem inteligibilidade e sem naturalidade);
5) - Excelente - **sem falhas notáveis** - (frase com palavras inteligíveis e com naturalidade).

A inteligibilidade está relacionada ao estado de um enunciado que pode ser ouvido distintamente e facilmente compreendido [3], e a naturalidade está relacionada a prosódia, ou seja, ao ritmo e a entonação, de maneira que a frase sintetizada se aproxime ao máximo da frase natural.

A partir dos resultados obtidos foi traçado um gráfico conforme mostrado na Figura 5. Neste gráfico observa-se que

os percentuais de excelente (45%) e bom (37%) prevalecem sobre regular (15%) e pobre (3%). Os percentuais dos escores regular e pobre decorrem principalmente da escuta das frases 1 e 19 da Tabela IV, onde determinadas unidades acústicas nasalizadas, como, por exemplo, [tim] na palavra *atingiremos* e [rem] na palavra *conferência* não apresentaram a qualidade desejada.

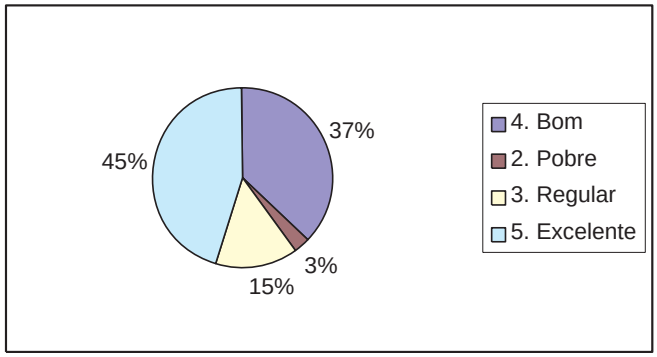

Fig. 5. - Avaliação MOS das 20 Frases utilizando o modelo prosódico proposto.

V. Conclusões

Este trabalho apresentou uma nova abordagem sobre modelagem prosódica para um sistema texto-fala concatenativo para o Português Brasileiro. O modelo é baseado em regras e na tonicidade das palavras e em um dicionário contendo unidades acústicas contemplando regras fonéticas e fonológicas, como também transições entre os fonemas, correspondentes ao fenômeno de coarticulação. Foram escolhidas sílabas e demissílabas como unidades acústicas do dicionário, considerando-se a redução do número de unidades em relação ao uso exclusivo de sílabas e a inclusão de um número maior de coarticulações comparando-se aos difones, que têm sido bastante utilizados na síntese da fala. Assim, foram determinadas 1994 unidades, com base na identificação dos fonemas e nas possíveis combinações desses, dentro da Língua Portuguesa falada no Brasil. Foi necessário fazer ajustes de pitch e duração em algumas unidades do dicionário, dependendo da posição e da tonicidade das sílabas dentro de cada palavra. O modelo foi avaliado através de testes informais de escuta de palavras oxítonas, paroxítonas e proparoxítonas com até cinco sílabas, obtendo-se bons resultados. Também foram realizados testes formais em um *corpus* de 20 frases declarativas, foneticamente balanceadas, com um escore de 4,25 na escala MOS, superior ao escore bom (4,0). Assim, os resultados obtidos demonstram o bom desempenho do modelo de prosódia proposto, considerando que as 20 frases utilizadas constituem uma amostra representativa do universo de todos os fonemas e unidades do dicionário.

Acrescenta-se que a partir dos dados obtidos, é possível expandir o sistema para que seja considerada a tonicidade das sílabas em nível de frases. Nesse caso, devem ser incluídos novos parâmetros, tais como, classificação do tipo de frase pela pontuação (frases declarativas, frases interrogativas, etc.) e a divisão de cada frase em grupos prosódicos, ou seja, regiões nas quais os parâmetros prosódicos apresentam certa medida independente do verificado nas demais regiões. Para tal, deve ser inserido no sistema um analisador gramatical na etapa do Processamento Lingüístico, para auxiliar a identificação das fronteiras prosódicas. A estrutura de pesos deve ser ampliada para atender aos diversos casos, e deve ser acrescentado um algoritmo de busca para que as melhores unidades acústicas sejam selecionadas na síntese.

VI. Agradecimentos

Agradecemos à CAPES (Coordenação de Aperfeiçoamento de Nível Superior), pela bolsa de estudos concedida durante o curso de doutorado na UFCG (Universidade Federal de Campina Grande). Agradecemos também a UFMA (Universidade Federal do Maranhão) pela liberação e apoio para realizarmos este trabalho.

Referências

[1] R. W. Sproat and J. P Olive. *Text-to-Speech Synthesis*. AT&T Technical Journal, pp. 35–44, March/April, 1995.

[2] C. H. da Silva. *Modelamento Prosódico para Conversão Texto-Fala do Português Falado no Brasil*. Dissertação de Mestrado, Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Dezembro, 1995.

[3] J. Dubois; M. Giacomo; L. Guespin; C. Marcellesi; J. B. Marcellesi; J. P. Mevel. *Dicionário de Lingüística*. Editora Cultrix, São Paulo, 1998.

[4] M. Eichner, M. Wolff and R. Hoffmann. *Improved duration control for speech synthesis using a multigram language model*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 417–420, Orlando, USA, May, 2002.

[5] Q. Yan and S. Vaseghi. *Analysis, modelling and synthesis of formants of British, American and Australian accents*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 712–715, USA, April, 2003.

[6] P. Seresangtakul and T. Takara. *A generative model of fundamental frequency contours for polysyllabic words of Thai tones*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 452–455, USA, April, 2003.

[7] Z. Sheng, T. Jianhua and D. Ling. *Chinese prosodic phrasing with extended features*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 492–495, USA, April, 2003.

[8] J. L. Rouas, J. Farinas, F. Pellegrino and R. Andre-Obrecht. *Modeling prosody for language identification on read and spontaneous speech*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 40–43, USA, April, 2003.

[9] W. Chen, F. Lin, J. Li and B. Zhang. *Generation of chinese prosodic phrasing rules by a extension matrix algorithm*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 489–492, Orlando, USA, May, 2002.

[10] R. Prudon, C. d'Alessandro and P. B. de Mareüil. *Prosody Synthesis by Unit Selection and Transplantation on Diphones*. Workshop on Speech Synthesis, pp. 119–122, California, USA, September, 2002.

[11] C. Erdem and H. G. Zimmermann. *A data-driven method for input feature selection within neural prosody generation*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 477–480, Orlando, USA, May, 2002.

[12] C. Erdem, F. Beck, D. Hirschfeld, H. Hoege and R. Hoffmann. *Robust Unit Selection Based on Syllable Prosody Parameters*. Workshop on Speech Synthesis, pp. 159–162, California, USA, September, 2002.

[13] C. Blouin, P. C. Bagshaw and O. Rosec. *A method of unit preselection for speech synthesis based on acoustic clustering and decision trees*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 692–695, USA, April, 2003.

[14] Y. Qian and F. Chen. *Assigning phrase accent to Chinese text-to-speech system*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 485–488, Orlando, USA, May, 2002.

[15] A. Venkataraman, L. Ferrer, A. Stolcke and E. Shriberg. *Training a prosody-based dialog act tagger from unlabeled data*. In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 272–275, USA, April, 2003.

[16] D. M. Milone and A. J. Rubio. *Prosodic and accentual information for automatic speech recognition*. IEEE Transactions on Speech and Audio Processing, 4(11):321–333, USA, July, 2003.

[17] F. O. Simões. *Implementação de um Sistema de Conversão Texto-Fala para o Português do Brasil*. Dissertação de Mestrado, Faculdade de Engenharia Elétrica, Universidade Estadual de Campinas, Maio, 1999.

[18] M. L. da Costa Neto. *Conversor Texto-Fala de Alta Qualidade para a Língua Portuguesa*. Exame de Qualificação, Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, Abril, 2000.

[19] T. C. Silva. *Fonética e Fonologia do Português*. Editora Contexto, São Paulo - SP, 2002.

[20] P. A. de Aquino. *O papel das vogais reduzidas pós-tônicas na construção de um sistema de síntese concatenativa para o português do Brasil*. Dissertação de Mestrado, Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, 1997.

[21] A. Alcaim, J. A. Solewicz and J. A. de Moraes. *Freqüência de Ocorrência dos Fones e Listas de Frases Foneticamente Balanceadas no Português Falado no Rio de Janeiro*. Revista da Sociedade Brasileira de Telecomunicações, 7(1):23-41, Dezembro, 1992.

[22] J. M. Fechine. *Reconhecimento Automático de Identidade Vocal Utilizando Modelagem Híbrida: Paramétrica e Estatística*. Tese de Doutorado, Departamento de Engenharia Elétrica, Universidade Federal da Paraíba, Dezembro, 2000.

[23] S. Madureira and M. A. S. Fontes. *Fundamental Contours in Brazilian Portuguese Words*. Proceedings of an ESCA Workshop on Intonation: Teory, Models and Applications, 211–214, Athens, Greece, September 1997.

[24] G. Massini-Cagliari. *Acento e Ritmo*. Coleção Repensando a Língua Portuguesa, Editora Contexto, São Paulo - SP, 1992.

[25] J. A. de. Moraes. *Um Algoritmo para a Correção-Simulação da Duração dos Segmentos Vocálicos em Português*. Estudos de Prosódia, pp. 69-84, Editora da UNICAMP, Campinas - SP, 1999.

APÊNDICE I
FRASES USADAS NOS TESTES MOS

Neste apêndice são relacionadas as 20 frases usadas nos testes MOS, para a avaliação qualitativa do modelo prosódico descrito na Seção 3, com base nas considerações feitas na Seção 4. As 20 frases foram selecionadas dentre as 200 frases foneticamente balanceadas determinadas em um trabalho realizado por A. Alcaim et al. em [21].

TABELA IV

RELAÇÃO DAS 20 FRASES USADAS NOS TESTES MOS

1	sei que atingiremos o nosso objetivo
2	o analfabetismo é a vergonha do país
3	a casa foi vendida sem pressa
4	isso se resolverá de forma tranqüila
5	as crianças conheceram o filhote de ema
6	a bolsa de valores ficou em baixa
7	uma garota foi presa ontem à noite
8	essa medida foi devidamente alterada
9	a mudança é lenta porém duradoura
10	muito prazer em conhecê-lo
11	parece que nascemos ontem
12	hoje eu acordei muito calmo
13	nosso telefone quebrou
14	queremos discutir o orçamento
15	ela tem muita fome
16	hoje dormirei bem
17	o termômetro marcava um grau
18	o discurso de abertura é bem longo
19	eu precisei de microfone na conferência
20	nossa filha é a primeira aluna da classe