Parâmetros Lingüísticos Utilizados para a Geração Automática de Prosódia em Sistemas de Síntese de Fala

Rui Seara Jr., Izabel C. Seara, Sandra G. Kafka, Fernando S. Pacheco, Rui Seara e Simone Klein

***Resumo*—Este artigo propõe um classificador automático de sílabas para aplicação em sistemas de síntese de fala. Através desse classificador silábico, obtemos automaticamente a classificação e divisão das sílabas de qualquer palavra do léxico do português brasileiro. Apresentamos também as regras para a identificação de grupos clíticos necessária ao estabelecimento da prosódia da fala sintetizada, pois esse tipo de constituinte prosódico força a ressilabação das palavras internas a esses grupos. O uso de contextos silábicos e grupos clíticos na escolha da melhor unidade-alvo mostra, em testes informais de escuta, uma sensível melhora na naturalidade da fala sintetizada.**

***Palavras-Chave*—Classificação silábica automática, grupos clíticos, naturalidade da fala sintetizada, padrão silábico.**

***Abstract*—This paper proposes an automatic syllabic classifier for use in speech synthesis systems. Through this syllabic classifier, we automatically obtain the classification and division of the syllables for any word of the Brazilian Portuguese lexicon. In addition, rules to identify clitic groups needed to generate the prosodic structure for high quality speech synthesis are established, since its identification requires the resyllabification of the inner words of those groups. The use of syllabic contexts and clitic groups for choosing the better target-unit improves the naturalness of the synthesized speech.**

***Keywords*—Automatic syllabic classification, clitic groups, speech synthesis naturalness, syllabic pattern.**

I. INTRODUÇÃO

Apesar dos avanços significativos na área de síntese de fala, os sistemas atuais que usam essa tecnologia ainda apresentam problemas de falta de naturalidade, quando comparados a um falante humano. Tais problemas estão associados sobretudo à entonação que, em geral, nos modelos prosódicos atualmente considerados, estão ligados a uma pobre representação lingüística [1]. Os estudiosos interessados em sistemas de síntese de fala com abordagem concatenativa têm empregado métodos de seleção de unidades de tamanho variável, o que tem contribuído significativamente para a qualidade da fala sintética produzida [2]. Tem-se observado também que a utilização de um modelo lingüístico mais eficiente, que leve em conta detalhes fonéticos sutis, é importante para a melhoria da qualidade de tais sistemas [2].

A busca por naturalidade na fala sintetizada tem direcionado os pesquisadores a procurarem entender detalhadamente os mecanismos que levam os ouvintes a diferenciarem fala sintética daquela produzida por humanos. Estudos perceptuais ([3], [4]) têm mostrado que, quando se manipulam parâmetros para obtenção de prosódia, essa ação conduz a uma perda significativa de qualidade. Assim, são favorecidas as técnicas que fazem uso mínimo de modificações no sinal, descartando também segmentos com parâmetros prosódicos extremos. Tais estudos apontam como parâmetros prosódicos mais relevantes aqueles relacionados à estruturação e tonicidade silábicas [5], pois a sílaba é a menor categoria prosódica. As palavras são então analisadas em função dos fonemas que as compõem, do contexto fonético anterior e posterior desses fonemas, da posição do fonema em relação à sílaba em que está inserido, da posição da sílaba em relação à palavra em que está incluída, da posição da palavra em relação ao constituinte e desse em relação à frase. Esses contextos são então utilizados em árvores de decisão responsáveis pela pré-seleção dos grupos de unidades candidatas à síntese.

As teorias acerca de regras de acento do português brasileiro (PB) mostram estreita relação do acento com o peso silábico, isto é, com as sílabas pesadas ou travadas (aquelas finalizadas por consoantes). Outra consideração dessas teorias é que o cabeça dos constituintes (elemento mais marcado prosodicamente, ou seja, com maior proeminência rítmica) localiza-se geralmente em palavras de conteúdo como nomes e verbos. É apontado em [1] a importância de considerar detalhes fonéticos para a síntese de fala, quando destaca, por exemplo, para o inglês, os efeitos de ressonância associados ao /r/[1] cujo espraiamento vai depender da tonicidade da sílaba, do número de consoantes no *onset* silábico, da qualidade vocálica e do número de sílabas de cada

Rui Seara Jr., Izabel C. Seara, Sandra G. Kafka, Fernando S. Pacheco, Rui Seara e Simone Klein, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mails: {ruijr, izabels, kafka, fernando, seara, klein}@linse.ufsc.br.

Este trabalho foi parcialmente financiado pela empresa Dígitro Tecnologia e pelo CNPq.

[1] Os símbolos fonéticos empregados neste artigo seguem a notação do Alfabeto Fonético Internacional (IPA).

constituinte. Para o PB, outros fatores interferentes na qualidade prosódica são o número de sílabas da palavra, posição da sílaba tônica na palavra e a sua classificação morfossintática.

Desta forma, a divisão e classificação silábicas e a identificação de grupos clíticos (unidades prosódicas que seguem a palavra fonológica) são importantes componentes que devem ser integrados a sistemas de síntese de fala.

Este artigo propõe assim um classificador automático de sílabas e as regras necessárias para a identificação de grupos clíticos as quais conduzem a uma ressilabação das palavras internas a esses grupos que serão usados na geração automática de prosódia de sistemas de síntese de fala. O uso de contextos silábicos e grupos clíticos na escolha da melhor unidade-alvo auxilia na melhora da naturalidade da fala sintetizada.

II. Padrões Silábicos do PB e os Sistemas de Síntese de Fala

A procura por uma melhor naturalidade, a partir de modelos lingüísticos, prescinde da análise do *corpus* nos diversos níveis gramaticais (fonêmico, fonético, morfológico, sintático, etc.). A obtenção desse *corpus* é feita através da leitura e gravação de um texto, no qual a voz do locutor selecionado será aquela a ser sintetizada. Essas gravações serão então etiquetadas e indexadas de forma a possibilitar a busca das melhores unidades no processo de síntese.

O processo de etiquetagem inicia-se pela etapa de transcrição grafema-fonema, em que é estabelecida a unidade sonora correspondente ao contexto em que essa unidade está inserida. Tal transcrição é denominada "canônica". Em uma segunda etapa, realiza-se a transformação dessa transcrição em uma transcrição restrita. Nessa fase, são explicitados os detalhes do ponto de vista acústico, considerando-se os aspectos condicionados por contexto ou características específicas do locutor [6] ou do PB. Por exemplo, a transcrição canônica da frase: "Tenho esperança de que a corda da ginástica não arrebente" é: [ˈtẽɲu espeˈɾẽsɐ dɪ kɪ a ˈkɔRdɐ dɐ ʒiˈnastikɐ nãw̃ aRɐˈbẽtɪ]. Já, sua transcrição restrita é: [ˈtẽɲu ispeˈɾẽsɐ dikjɐ ˈkɔRdɐ dɐ ʒiˈnaʃikɐ nãw̃ aRɐˈbẽtɪ].

Se compararmos as duas transcrições, observamos que muitos segmentos presentes na transcrição canônica desaparecem ou se modificam na restrita. Assim, com a etiquetagem do *corpus* baseada na transcrição restrita, teremos de considerar, durante o processo de síntese, por exemplo, que os vocábulos monossilábicos "que a" ([kɪ] [a]), em seqüência, serão etiquetados no banco por uma única sílaba [kjɐ] (CXV.)[2]. Para fazermos essa consideração, devemos observar algumas regras que delimitam grupos clíticos. Tais regras definiriam a seqüência "de que a corda" como um grupo clítico no qual se teria uma ressilabação das palavras "que a", cujas sílabas foram classificadas como CV.V pela transcrição canônica, e que passariam a uma única sílaba CXV. Com essa estratégia, obtemos uma aproximação da forma silábica do sintetizador com aquela efetivamente apresentada no *corpus* e que é característica não só do locutor, mas da língua a ser sintetizada, o PB. Nas Seções III e IV, são apresentadas tais regras.

O novo processo de transcrição canônica (denominado "transcrição canônica melhorada" – TCM) substitui o processo original usado na fase da geração da primeira transcrição grafema-fonema de etiquetagem do *corpus* como também no processo de síntese. A TCM tem como primeira fase a transcrição canônica original, com a qual se obtém uma transcrição grafema-fonema inicial. Essa transcrição é utilizada como base para as etapas seguintes. A partir daí, é feita a divisão silábica das palavras, sendo as sílabas resultantes classificadas conforme seus tipos. Essa classificação se faz necessária para identificar as sílabas leves e pesadas, a rima e o *onset* silábicos. Em uma próxima etapa, a classificação morfossintática das palavras que compõem as frases a serem sintetizadas [7] é estabelecida. Finalmente, em uma última etapa, identificam-se os grupos clíticos e reclassificam-se as sílabas internas, ajustando-se suas transcrições fonéticas.

Como os constituintes prosódicos contam com informações de diferentes tipos (fonológicas, morfológicas e sintáticas), a identificação de constituintes prosódicos, como sílabas, palavra fonológica e grupos clíticos, envolve essas várias etapas de processamento anteriormente mencionadas. As etapas de transcrição canônica e divisão silábica servem para determinar a palavra fonológica. A etapa de classificação morfológica serve para o agrupamento de clíticos. Essas etapas auxiliam na melhora da prosódia da fala sintetizada, pois são usadas como parâmetros de decisão na busca do segmento-alvo.

Para a construção do classificador silábico, consideramos o sistema fonotático do PB e elaboramos as regras necessárias para a classificação e divisão automática das sílabas.

De forma geral, conceitua-se sílaba como uma cadeia sonora composta de fonema(s) consonântico(s) e vocálico(s), sendo sua enunciação completa formada pelo *onset* (ou aclive), núcleo (ou ápice) e coda (ou declive). O núcleo e a coda silábicos constituem a rima.

Considerando-se, por exemplo, a palavra "casca", tem-se a transcrição /ˈkaska/, que é constituída de duas sílabas (CVC.CV): uma pesada (CVC) e uma leve (CV).

Pode-se dizer então que a sílaba é uma unidade fonológica composta de vogal (núcleo silábico) e consoantes ou semivogais (margens silábicas), unidas por um único acento culminativo, no caso do PB [8]-[13].

Para determinarmos que tipos de segmentos podem ocupar as margens silábicas (*onset* e coda), devemos fazer algumas observações a respeito da interpretação das vogais e semivogais. Consideramos como segmentos que podem

[2] C corresponde à consoante; X à semivogal; V à vogal. O ponto corresponde à divisão silábica.

ocupar o núcleo silábico: sete vogais orais [a e ɛ i o ɔ u] e cinco nasais [ẽ ẽ ĩ õ ũ]. Aqui evidencia-se a tendência de se considerar vogais nasais como fonemas do PB, no nível fonêmico. Dessa forma, em "canto", tem-se apenas 4 fonemas e duas sílabas CV: /'kãto/ (['kãtu]) e não 5 fonemas e duas sílabas (uma CVC e outra CV): /'kʌnto/ (['kaⁿto]).

Já, no que concerne às semivogais, elas podem ser vistas tanto como vogais assilábicas, quanto como segmentos consonantais. Se as considerarmos como semiconsoantes, elas devem ser incorporadas ao grupo de consoantes que travam sílabas como [s] ou [ʀ]. Caso contrário, teremos de considerá-las vogais assilábicas e assim aumentar nosso número de tipos silábicos, visto que essas vogais não se enquadrariam no tipo CVC, CCVC, etc., mas sim no tipo CVX, CCVX, etc.

Na tentativa de incorporá-las como consoantes, nossa metodologia de classificação silábica produzia tipos silábicos inexistentes. Dessa forma, tivemos de optar pela observação desses segmentos como semivogais, apesar de, com isso, tornar mais complexo o sistema fonotático a ser classificado. A partir dessas considerações, podemos dizer que, nas margens silábicas, encontraremos tanto consoantes quanto semivogais.

Quanto aos ditongos, tivemos de considerar tanto ditongos crescentes (semivogal + vogal), quanto decrescentes (vogal + semivogal), já que, na etapa de transcrição fonética, considerava-se a existência desses dois tipos de ditongos, apesar da colocação de [9] de que os verdadeiros ditongos seriam os decrescentes. Segundo esse autor, os ditongos crescentes seriam falsos ditongos, pois há uma variação livre entre ditongos, podendo também serem vistos como duas vogais em sílabas separadas. No entanto, nossa transcrição de palavras como "ilusório" e "causa" era realizada como [ilu'zɔɾyu] e ['kawzɐ], respectivamente. Assim, estabelecemos como seus tipos silábicos: {V**.**CV**.**CV**.**CXV} e {CVX**.**CV}, respectivamente. As estruturas silábicas que serão apresentadas por nosso Classificador são mostradas na Tabela 1.

TABELA 1
ESTRUTURAS SILÁBICAS DO PB E EXEMPLOS

V ([o] em *ovo*)	VX ([oj] em *oitavo*)
VC ([as] em *asma*)	VXC ([ejs] em *eis*)
VCC ([ads] em *adstringente*)	CXV ([ɾju] em *armário*)
CV ([ka] em *casas*)	CVX ([taw] em *cristal*)
CCV ([pɾɛ] em *prece*)	CXVC ([ɾjas] em *várias*)
CVC ([kɔʀ] em *corda*)	CXVX ([kwaw] em *qualquer*)
CCVC ([tɾas] em *traste*)	CCVX ([gɾaw] em *grau*)
CVCC ([pɛʀs] em *perspectiva*)	CCXV ([kɾja] em *criado*)
	CVXC ([tajs] em *metais*)
	CCVXC ([gɾaws] em *graus*)
	CXVXC ([kwajs] em *quaisquer*)

A partir das colocações acerca dos segmentos vocálicos e consonantais e das estruturas silábicas criadas a partir deles, podemos considerar o diagrama em árvore mostrado na Fig. 1 como o mais adequado para modelar o sistema fonotático aqui discutido.

No diagrama da Fig. 1, para cada sílaba σ, C_1 corresponde a qualquer uma das consoantes do PB; C_2 corresponde às consoantes [ɾ l n s]; X, às semivogais [j w j̃ w̃]; V, a qualquer vogal do PB; C_3, às consoantes do grupo [p b t d k g f z ʀ]; C_4, à consoante [s] em final de sílaba.

A análise aqui elaborada sobre as estruturas silábicas do PB difere daquelas apresentadas por análises tradicionais como de [9], [8] (*apud* [12]), dentre outras. No entanto, tivemos de basear nossas considerações em condições que permitissem a aplicabilidade desse sistema fonotático de forma a tornar viável a classificação silábica de qualquer palavra do léxico do PB, inserida no sistema de síntese de fala. Incorreções na transcrição fonética e conseqüentemente uma inadequada classificação silábica traria prejuízos à prosódia de nosso sistema de síntese de fala.

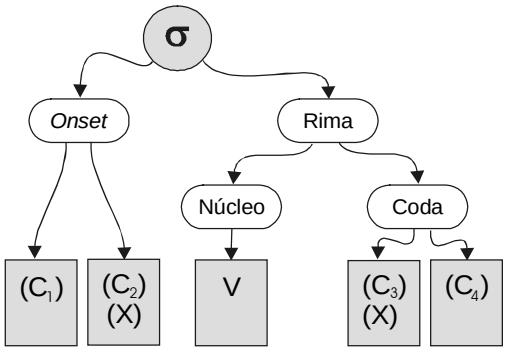

Fig. 1. Diagrama em árvore para a classificação das sílabas do PB (os símbolos entre parênteses são opcionais).

Um exemplo desta incompatibilidade entre os moldes tradicionais e o aqui aplicado pode ser visto na palavra *adstringente* [adsˑtɾĩ.'ʒẽˑtɪ] que, em nosso sistema de síntese, apresenta como sílabas VCC**.**CCV**.**CV**.**CV. A primeira sílaba VCC não apareceria nos moldes silábicos apresentados por [8][3] (*apud* [12]), já que esses autores consideram nesse caso a inserção de uma vogal epentética após as consoantes plosivas [a.disˑtɾĩ.'ʒẽˑtɪ], estabelecendo, para a sílaba VCC ([adsˑ]), as sílabas V**.**CVC ([a.dis]). A consideração dessa vogal epentética leva a uma seleção de unidades inadequadas para um sistema de fala sintética, pois, como o *corpus* é montado a partir da leitura cuidadosa de textos, as unidades geralmente têm duração maior se comparada à fala corrente. Dessa forma, se a vogal epentética fosse sintetizada a partir da vogal de uma sílaba CV, ela não seria apropriadamente curta para dar naturalidade à síntese. Também, para que nosso diagrama em árvore não estabelecesse tipos silábicos inexistentes, criamos algumas

[3] Cujo molde tem a forma: $\left(\left\{\begin{matrix}P & (L)\\ Z & \end{matrix}\right\}\right)(G)V\left(\left\{\begin{matrix}Z\\ \wedge V\end{matrix}\right\}\right)$
onde P representa as consoantes plosivas e fricativas bilabiais; L, as líquidas; Z, as soantes e sibilantes. G representa os *glides*; V, as vogais; ^V, as semivogais.

restrições como a de que, nessas sílabas do tipo VCC, só teríamos como última consoante de coda o fonema [s]. Foram essas restrições que nos levaram à criação de 14 diferentes conjuntos de segmentos que serão especificados na seção seguinte.

III. Regras para Classificação Automática das Sílabas do Português Brasileiro

O classificador silábico foi modelado por um autômato finito definido por suas transições. Essas transições (apresentadas na Tabela 2) são representadas por quádruplas (S_I; Σ_T; ξ_T; S_F), nas quais S_I é o estado inicial; Σ_T é o conjunto de símbolos (correspondente a um dos 14 grupos de segmentos, apresentados na Tabela 3) que disparam a transição; ξ_T é o símbolo de saída ([(C|V|X|.|∅)])[4] produzido pela passagem nesta transição e S_F será o novo estado corrente do autômato após a transição. Nosso classificador silábico é modelado por um autômato finito representado por 42 transições.

Para que pudéssemos iniciar a formalização das regras necessárias a cada estado do autômato de classificação das sílabas do PB, primeiramente estabelecemos grupos de segmentos que refletem os diferentes contextos restritivos.

A primeira restrição se deve à questão das rimas ramificadas, pois a coda silábica[5], que é dominada pela rima (Fig. 1), não é obrigatória no português. Dessa forma, a ramificação da rima em núcleo e coda impõe fortes restrições aos segmentos que, nesse caso, podem estar associados à coda silábica. Pode-se observar também que, em nosso sistema, a líquida lateral [l] é transcrita como a semivogal [w] (*sal* ['saw]) e a não lateral, como [ʁ] (*corda* ['kɔʁdɐ]). Esta última ([ʁ]) neutralizará na transcrição canônica as outras variantes desse fonema na posição de coda ([rɤɣhɦ]). As fricativas de coda [sz ʃ ʒ] também serão neutralizadas na transcrição canônica através da forma [s] (*casta* ['kastɐ]). Em relação às consoantes [t p d k b g f], possíveis em coda silábica, não consideramos, para a transcrição canônica, a vogal epentética [i] que normalmente é empregada em tais contextos. Para a coda complexa, formada pela seqüência de fonemas [ʁs], como em *perspectiva*, somente a consoante [s] pode ocupar a última posição dessa coda. Assim, essas restrições nos fizeram gerar os grupos CONS2, TRAVADOR_S e TRAVADOR_R, apresentados na Tabela 3.

O *onset* também não é um constituinte obrigatório na sílaba (Fig. 1) e, quando existente, pode ser preenchido por uma ou duas consoantes, sendo, neste último caso, chamado de complexo ou ramificado. Para dar conta do *onset* ramificado que tem como segundas consoantes [ɾ l s n], criamos os grupos CONS1 e CONS4. Com esses grupos, obtêm-se as sílabas de *onset* ramificado tradicionalmente constituído por uma consoante obstruinte [p b t d k g f v] e uma líquida [ɾ l], como em *prova* ['pɾɔvɐ] e *placa* ['plakɐ]. No entanto, da mesma maneira que não consideramos a epêntese da vogal [i] para a coda composta das consoantes [t p d k b g f], também não a consideramos no *onset* complexo em que as segundas consoantes sejam [n] ou [s] como em *pneu* e *psicologia*. Em nosso *corpus*, tais palavras são transcritas da seguinte forma: ['pnew] e [psikolo'ʒiɐ][6], respectivamente. Os grupos CONS_T1 e CONS3 constituem as consoantes de *onset* medial (internos às palavras) ou absoluto (inicial de palavras) em *onset* não ramificado. O grupo CONS_T2 compõe o grupo de consoantes de *onset* medial e os grupos VOGAL e CONS0, os segmentos que dão início ao processo de classificação no estado inicial do autômato. O grupo SEMIVOGAL é composto das semivogais [j w j̃ w̃] que tanto podem ser elementos de *onset* ou de coda. O núcleo silábico não é visto como ramificado, já que posicionamos a semivogal na coda silábica de acordo com [12]. Dessa maneira, têm-se, como núcleo silábico, somente os fonemas que compõem o grupo VOGAL. Os demais grupos (PONTO, ESPAÇO, OUTROS_SONS) referem-se aos símbolos que representam final de palavra e que levam o autômato a reiniciar o processo de classificação, a partir do estado inicial (S_0).

Assim, a partir das restrições anteriormente citadas, criamos 14 grupos de segmentos necessários para o estabelecimento dos estados que farão parte do autômato do classificador silábico, conforme Tabela 3.

A classificação silábica do texto de entrada é gerada então pelo histórico de símbolos de saída produzidos pelas transições ativadas até o último símbolo de entrada (Grupos). Na Tabela 2, apresentamos as transições do Autômato Finito e, na Tabela 4, pode-se ver um exemplo de seu emprego.

IV. Identificação de Grupos Clíticos

O grupo clítico é a unidade prosódica que segue a palavra fonológica. Algumas teorias não consideram esse nível por já incluírem o clítico como elemento da palavra fonológica. Em nosso caso, levamos em consideração esses dois níveis (palavra fonológica e clítico), pois estabelecemos regras para que se possa realizar uma ressilabação dentro dos grupos clíticos [14]. Os clíticos do português mostram propriedades de dependência e, ao mesmo tempo, de independência em relação à palavra seguinte [15], variando conforme o "dialeto".

[4] O símbolo ∅ corresponde a uma transição nula, isto é, não é produzido nenhum símbolo na saída deste estado.

[5] A coda nasal medial não ocorrerá em nossos dados, visto que será realizada através da nasalização da vogal. Assim, em *canto*, teremos ['kãtu] e não ['kaⁿtu]. Já a coda nasal final será produzida como uma semivogal, tornando-se um dos elementos do ditongo e estendendo a nasalidade para o núcleo silábico. Dessa maneira, em *catam* e *comem*, tem-se ['katãw̃] e ['komẽj̃], respectivamente.

[6] Se fosse considerada a vogal epentética, teríamos como transcrição [pi'new] e [pisikolo'ʒiɐ] e como tipos silábicos CV.CVX e CV.CV.CV.CV.CV.V, respectivamente.

Em nossos dados, grupos clíticos são formas dependentes e, quando identificados, transformam, por exemplo, vogais átonas finais de palavras seguidas de outra palavra pertencente ao mesmo grupo clítico em vogais átonas não finais. Dessa maneira, em "quero que me leve", observam-se dois grupos clíticos: *quero* e *que me leve*. Esse segundo grupo (*que me leve* [kɪ mɪ 'lɛvɪ]), constituído de três vocábulos: uma conjunção, um pronome e um verbo, vai se transformar em uma única palavra fonológica que terá como transcrição ([kimi'lɛvɪ]) e, nesse caso, as vogais átonas finais [ɪ] de *que* e *me* se transformarão em vogais átonas não finais [i].

TABELA 2
ESTADOS DO AUTÔMATO FINITO

(S_0; "ESPAÇO";"."; S_0)	(S_5; "TRAVADOR_S";"C."; S_0)
(S_0; "CONS0";".C"; S_2)	(S_5; "TRAVADOR_R";"∅"; S_{10})
(S_0; "VOGAL";"V"; S_1)	(S_5; "SEMIVOGAL";"X"; S_1)
(S_1; "SEMIVOGAL";"X"; S_4)	(S_5; "CONS_T2";".C"; S_2)
(S_1; "TRAVADOR_S";"C."; S_0)	(S_5; "ESPAÇO";"."; S_0)
(S_1; "TRAVADOR_R";"∅"; S_{10})	(S_5; "CONS2";"∅"; S_3)
(S_1; "CONS_T2";".C"; S_2)	(S_5; "VOGAL";".V"; S_1)
(S_1; "ESPAÇO";"."; S_0)	(S_6; "VOGAL";"∅"; S_7)
(S_1; "SEMIVOGAL";"X"; S_5)	(S_7; "TRAVADOR_S";"XVC.";S_0)
(S_1; "CONS2";"∅"; S_3)	(S_7; "SEMIVOGAL";"XVX"; S_1)
(S_1; "VOGAL";".V"; S_1)	S_7; "CONS2";"∅"; S_9) (
(S_2; "CONS1";"C"; S_4)	(S_7; "CONS_T2";"XV.C"; S_2)
(S_2; "SEMIVOGAL";"∅"; S_6)	(S_7; "VOGAL";"XV.V"; S_1)
(S_2; "ESPAÇO";"."; S_0)	(S_7; "ESPAÇO";"XV."; S_0)
(S_2; "CONS4";"C"; S_9)	(S_8; "VOGAL";"V"; S_5)
(S_2; "VOGAL";"V"; S_1)	(S_9; "ESPAÇO";"XVC."; S_0)
(S_3; "ESPAÇO";"C."; S_0)	(S_9; "SEMIVOGAL";"XV.CV"; S_4)
(S_3; "CONS1";".CC"; S_4)	(S_9; "VOGAL";"XV.CV"; S_1)
(S_3; "CONS_T1";"C.C"; S_2)	(S_9; "CONS_T1";"VC.C"; S_2)
(S_3; "SEMIVOGAL";".CX"; S_4)	(S_9; "CONS1";"XV.CC"; S_0)
(S_3; "VOGAL";".CV"; S_1)	(S_{10}; "ESPAÇO";"C."; S_0)
(S_3; "TRAVADOR_S";"CC."; S_0)	(S_{10}; "CONS0";"C.C"; S_2)
(S_4; "ESPAÇO";"."; S_0)	(S_{10}; "TRAVADOR_S";".CC"; S_0)
(S_4; "SEMIVOGAL";"X"; S_1)	
(S_4; "VOGAL";"V"; S_1)	

Determina-se o grupo clítico a partir da classificação morfológica que já está implementada [16]. Será no nível dos agrupamentos clíticos que se observará o sândi. Esse fenômeno serve justamente para atestar a proposta da existência de grupos clíticos e conceitua-se como a modificação que afeta foneticamente o início ou o fim de uma palavra ou morfema quando combinado com outro elemento na cadeia da fala. Por exemplo, em "cadernos especiais", tem-se esse fenômeno que leva a transcrevermos a consoante de coda da sílaba final *nos* da palavra *cadernos* como sendo também o *onset* da primeira sílaba *es* da palavra *especiais*, já que, na fala contínua, tem-se [ka'dɛʀnuzispesi'ajs].

Não consideramos todas as situações de sândi, pois os casos em que há apagamento de fonemas não serão levados em conta (casos denominados, na teoria fonológica, de degeminação). As regras para a criação dos grupos clíticos são baseadas em [15] que trabalha com o português brasileiro e [17] com o italiano.

Nosso classificador morfossintático automático é composto das seguintes classes: artigo (ART); preposição (PREP); locução prepositiva (LPREP); conjunção (CONJ); locução conjuntiva (LCONJ); substantivo e adjetivo (NOME); verbo na forma finita (VER); verbo na forma nominal (VERN); pronome adjetivo demonstrativo (DEM); pronome substantivo demonstrativo (DEM1); pronome pessoal do caso reto (RET); pronome pessoal do caso oblíquo átono (OBA); pronome pessoal do caso oblíquo tônico (OBT); pronome interrogativo (PER); pronome indefinido (IND); interjeição (INT); advérbio (ADV); locução adverbial (LADV); palavra denotativa (PDEN).

TABELA 3
GRUPOS GERADOS PARA O DIVISOR SILÁBICO

Grupo	Fonemas Associados
VOGAL	[a ɐ ɛ e i o ɔ u ɐ̃ ẽ ĩ õ ũ ɪ ʊ]
CONS0	[ɾ l ʀ p b t d k g s z ʃ ʒ f v ʎ m n ɲ]
CONS1	[ɾ l n s]
CONS2	[f p t d k g]
CONS3	[ʃ f v ʎ m ɲ s z]
CONS4	[p ʒ b t d k g f v]
CONS_T1	[ʀ p b t d k g z ʃ ʒ f v ʎ m ɲ]
CONS_T2	[ʀ ɾ l s z ʃ ʒ v m n ɲ]
SEMIVOGAL	[y w ỹ w̃]
PONTO	{ponto; vírgula; ponto e vírgula; dois pontos; exclamação; interrogação}
ESPAÇO (entre palavras)	{#}
TRAVADOR_S	[s]
TRAVADOR_R	[ʀ]
OUTROS_SONS	{silêncio; respiração; cliques; bucal; sopro}

TABELA 4
EXEMPLO DE USO DO AUTÔMATO: "TRASTE" [#'tɾasti#]

# (ESPAÇO)	$S_0 \rightarrow S_0$	.
t (CONS0)	$S_0 \rightarrow S_2$	C
ɾ (CONS1)	$S_2 \rightarrow S_4$	C
a (VOGAL)	$S_4 \rightarrow S_1$	V
s (TRAVADOR_S)	$S_1 \rightarrow S_{10}$	∅
t (CONS0)	$S_{10} \rightarrow S_2$	C.C
i (VOGAL)	$S_2 \rightarrow S_1$	V
# (ESPAÇO)	$S_1 \rightarrow S_0$	.
Resultado final	**.CCVC.CV.**	

Para a observação de grupos clíticos, primeiramente separamos as classes em palavras gramaticais (palavras destituídas de acento próprio) e de conteúdo (aquelas que possuem acento próprio). Às gramaticais correspondem as preposições (PREP, LPREP), os artigos (ART), as conjunções (CONJ, LCONJ), os pronomes oblíquos (OBA), os pronomes demonstrativos (DEM1 e DEM), os indefinidos (IND) e os do caso reto (RET). As demais classes correspondem às palavras de conteúdo.

Assim, caracteriza-se um grupo clítico como:

i) uma seqüência de palavras gramaticais e uma ou mais palavras de conteúdo precedendo uma palavra gramatical;

ii) uma seqüência de palavras gramaticais e uma ou mais palavras de conteúdo precedendo uma marca de pontuação do tipo vírgula, respiração, cliques, ruídos bucais ou sopro;

iii) uma seqüência de palavras gramaticais e uma palavra de conteúdo precedendo um ponto final, ponto e vírgula, uma exclamação ou uma interrogação.

A partir das regras acima, na frase: Ele escreveu uma crônica para os cadernos especiais do Jornal do Brasil, tem-se 5 grupos clíticos (conforme sublinhado) e observam-se mudanças em:

a) *Ele escreveu* que transforma a átona final [ɪ] em "ele" ([elɪ]) na átona não final [i] como pode ser visto na transcrição [elieskɾe'vew]. Essa vogal seguida de outra vogal transforma-se em uma semivogal, passando sua transcrição para [elyiskɾe'vew], com sílabas do tipo V.CXVC.CCV.CVX;

b) para os cadernos especiais que transforma as átonas finais de palavra [ʊ] em "os" ([ʊs]) e "cadernos" ([ka'dɛʀnʊs]) na vogal átona não final [u], como visto pela transcrição ['paɾauska'dɛʀnuzispesi'ajs];

c) do Jornal que transforma a átona final de palavra [ʊ] em "do" ([du]) na vogal átona não final [u], como pode ser visto pela transcrição [duʒoʀ'naw];

d) do Brasil que transforma a átona final [ʊ] em "do" ([du]) na vogal átona não final [u], como também pode ser visto pela transcrição [dubɾa'ziw].

Outras mudanças que refletem na ressilabação são relacionadas às palavras terminadas por sílabas travadas [s] ou [ʀ] como "cadernos" e que tem em seqüência dentro do mesmo grupo clítico uma palavra que inicia com uma vogal, como "especiais", conforme pode se observar no exemplo (b) anterior, cujo fonema final [s] da palavra "cadernos" passa para [z].

V. Conclusões e Futuras Implementações

Como vimos, muitas das teorias que explicam os fenômenos lingüísticos devem ser adaptadas para aplicação em sistemas de síntese de fala para que se tenha um resultado mais natural. A observação de discrepâncias entre a fala e a representação lingüística aqui apresentada, como por exemplo o caso da vogal epentética que, por sua curta duração, não deve ser levada em conta para a tipologia silábica, conduz a uma fala sintética mais natural – objetivo deste estudo.

O classificador silábico aqui especificado classifica e divide adequadamente as sílabas das palavras do léxico do PB. Foram testadas 5.000 palavras, coletadas aleatoriamente de um conjunto de notícias da Agência Brasil, que continha um total de 148.000 palavras. Todas as palavras que apresentaram uma correta transcrição fonética foram adequadamente divididas em sílabas. A implementação dos grupos clíticos levou a ressilabação das palavras que compunham cada grupo. Os grupos clíticos e a classificação silábica são utilizados para a busca da melhor unidade para a concatenação e, em testes informais de escuta, a inclusão desses parâmetros lingüísticos têm mostrado uma sensível melhora na fala sintetizada.

Referências

[1] R. Ogden, S. Hawkins, J. House *et al*., "Prosynth: An Integrated Prosodic Approach to Device-independent, Natural-sounding Speech Synthesis," *Computer Speech and Language*, vol. 14, pp. 177-210, 2000.

[2] A. P. Breen and P. Jackson, "Non-uniform Unit Selection and the Similarity Metric within BT's Laureat TTS System," *III ESCA Workshop on Speech Synthesis*, Jenolan Caves, Austrália, pp. 201-206, Nov. 1998.

[3] B. Bozkurt, T. Dutoit et V. Pagel, "Synthèse Vocale par Sélection D'unité: Une Méthode pour la Redéfinition de la Courbe Intonative," *XXIVèmes Journées d'Étude sur la Parole*, Nancy, France, pp. 121-124, Juin 2002.

[4] F. Courtois, P. Di Cristo, B. Lagrue et J. Véronis, "Un Modèle Stochastique des Contours Intonatifs en Français pour la Synthèse à partir de Textes," *4ème Congrès Français d'Acoustique*, Marseille, France, pp. 373-376, Avril 1997.

[5] M. Adda-Decker, P. Boula de Mareüil, G. Adda, and L. Lamel, "Investigating Syllabic Structure and its Variation in Speech from French Radio Interviews," *ISCA ITRW on Pronunciation Modeling and Lexicon Adaptation for Spoken Language Technology*, Aspen Lodge, Estes Park, Colorado, USA, pp. 89-94, Sept. 2002.

[6] P. Ladefoged, *A Course in Phonetics*, New York: Harcourt Brace Jovanovich, 1975.

[7] R. Ribeiro, L. Oliveira, and I. Trancoso, "Using Morphossyntactic Information in TTS Systems: Comparing Strategies for European Portuguese," *Proc. PROPOR'2003, VI Encontro para o Processamento Computacional do Português Escrito e Falado*, Faro, Portugal, pp. 143-150, Jun. 2003.

[8] E Lopes, *Fundamentos da Lingüística Contemporânea*, São Paulo: Cultrix, 1975.

[9] J. M. Câmara Jr., *Problemas de Lingüística Descritiva*, 12. ed., Petrópolis: Vozes, 1986.

[10] L. S. Cabral, *Introdução à Lingüística*, 7. ed., Rio de Janeiro: Globo, 1988.

[11] D. Callou e Y. Leite, *Iniciação à Fonética e Fonologia*, Rio de Janeiro: Zahar, 1990.

[12] G. Collischonn, "A Sílaba em Português," In: L. Bisol (Org.) *Introdução a Estudos de Fonologia do Português Brasileiro*, Porto Alegre: Edipucrs, pp. 95-130, 1996.

[13] T. C. Silva, *Fonética e Fonologia do Português: Roteiro de Estudos e Guia de Exercícios*, 6. ed., São Paulo: Contexto, 2002.

[14] P. Boula de Mareüil, M. Adda-Decker, and V. Gendner, "Liaisons in French: A Corpus-based Study Morpho-syntactic Information," *International Congress on Phonetic Science*, Barcelona, pp. 1329-1332, 2003.

[15] L. Bisol, "Constituintes Prosódicos," In.: L. Bisol. (Org.) *Introdução a Estudos de Fonologia do Português Brasileiro*, Porto Alegre: Edipucrs, pp. 247-261, 1996.

[16] I. C. Seara, S. G. Kafka, S. Klein e R. Seara, "Alternância Vocálica das Formas Verbais e Nominais do Português Brasileiro para Aplicação em Conversão Texto-fala," *Revista da Sociedade Brasileira de Telecomunicações*, vol. 17, no. 1, Junho 2002, pp. 79-85.

[17] P. Boula de Mareüil, "Linguistic-prosodic Processing for Text-to-speech Synthesis in Italian," *International Congress on Spoken Language Processing*, Pequim, China, pp. 697-700, 2000.