Seleção Automática de *Corpus* de Texto para Sistemas de Síntese de Fala

Monique V. Nicodem, Izabel C. Seara, Rui Seara, Daiana dos Anjos e Rui Seara Jr.

***Resumo*— Visando melhorar a naturalidade da fala sintética, este trabalho propõe um procedimento para selecionar o *corpus* de gravação de um sistema de síntese de fala concatenativa desenvolvido para o português brasileiro. O objetivo de tal seleção é atribuir uma maior variabilidade fonética e prosódica à fala sintética. Nesse procedimento, quatro etapas são consideradas: conversão grafema-fonema, anotação prosódica, representação em vetores de características e seleção propriamente dita. O procedimento de anotação prosódica de sentenças exclamativas e questões alternativas é uma contribuição original deste trabalho.**

***Palavras-chave*— Seleção de um *script* de gravação, Português brasileiro, Variabilidade fonética e prosódica, Algoritmos genéticos.**

***Abstract*— In order to improve the naturalness of synthetic speech, this work proposes a procedure to select the recording script for a speech synthesis system developed for the Brazilian Portuguese language. The objective of such a selection is to improve both phonetic and prosodic variability in synthetic speech. In this selection approach, four stages are included: grapheme-to-phoneme conversion, prosodic annotation, feature vector representation, and selection itself. The procedure of prosodic annotation of exclamative sentences and alternative questions is an original contribution of the current research work.**

***Keywords*— Recording script design, Brazilian Portuguese, Phonetic and prosodic variability, Genetic algorithms.**

I. Introdução e Formulação do Problema

Tecnologias de síntese de fala permitem converter um texto escrito em fala sintética. Essa conversão é usualmente realizada considerando duas etapas principais: análise lingüística e processamento de sinais. Na etapa de processamento de sinais, a maioria dos atuais sistemas comerciais utiliza uma arquitetura baseada em síntese concatenativa. Nesses sistemas, inicialmente realiza-se a gravação de um banco de fala, em uma sala acusticamente isolada, utilizando a voz de um locutor profissional [1], [2]. Durante a síntese propriamente dita, um procedimento automático é adotado para selecionar segmentos de tamanho não-uniforme contidos nesse banco. Embora a qualidade da fala advinda desses sistemas já se encontre muito próxima da fala humana, a sua naturalidade ainda está limitada por fatores tais como a qualidade das gravações, características da voz do locutor, o texto escolhido para gravação e algoritmos de modelagem entoacional [3]–[5].

Monique V. Nicodem, Izabel C. Seara, Rui Seara, Daiana dos Anjos e Rui Seara Jr., LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mails: {monique, izabels, seara, daiana, ruijr}@linse.ufsc.br.

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Financiadora de Estudos e Projetos (FINEP) e Dígitro Tecnologia Ltda.

Na última década, um grande esforço de pesquisa vem sendo realizado para a obtenção de modelos entoacionais robustos. Nesse caso, um modelo entoacional é previamente treinado para cada estilo expressivo (interrogativo, ênfase contrastiva, emocional com alegria ou tristeza) [6] e um estágio *online* de predição de parâmetros prosódicos (duração, *pitch* e energia) é efetuado [7]. Em seguida, ocorre um procedimento de busca dos segmentos do *corpus* de fala que satisfazem os requisitos desejados de prosódia. Por fim, uma etapa de alteração dos parâmetros prosódicos dos segmentos selecionados pode ser realizada com a finalidade de atingir a prosódia predita. Entretanto, essa etapa pode introduzir algumas degradações na fala sintética, especialmente quando importantes alterações são efetuadas [7]. Tais degradações seriam pouco significativas se apenas pequenas mudanças de *pitch* fossem realizadas ou se tais mudanças fossem evitadas, o que ocorreria se o *corpus* de fala acessado durante a síntese tivesse segmentos com características fonéticas e prosódicas muito similares às preditas [8].

Assim, um banco de fala deve idealmente ser otimizado para maximizar as variantes de um dado fone em distintos contextos prosódicos [2]. Uma solução para obter tal maximização consiste em adotar durante a síntese um grande *corpus* de fala. Entretanto, em algumas aplicações, o tamanho do *corpus* é restringido por requisitos de memória e complexidade de busca bem como pelo tempo demandado para gravação de grandes *corpora* [5]. Além do mais, em sistemas de síntese de fala expressiva, a criação de um grande *corpus* para cada estilo expressivo não seria viável. Nesses casos, o *script* de gravação deve ser cuidadosamente escolhido.

Em nosso conhecimento, bancos de fala foneticamente ricos (concebidos para o português falado no Brasil – PB) não são disponibilizados como uma ferramenta de pesquisa para aplicações em processamento de fala. Bancos de fala foneticamente balanceados são obtidos em [9] e [10]. Em [9], são apresentados 20 conjuntos de 10 sentenças foneticamente balanceadas obtidas por seleção manual. Em [10], um procedimento automático de busca baseado em algoritmos genéticos é adotado, fornecendo 1.000 sentenças declarativas foneticamente balanceadas. É importante mencionar que [9] e [10] lidam apenas com a representatividade fonética das sentenças escolhidas para gravação, sem considerar a representatividade prosódica.

Com o objetivo de melhorar a variabilidade fonética e prosódica de tais *corpora*, este trabalho propõe um procedimento para selecionar as sentenças a serem previamente gravadas e em seguida usadas durante a síntese. Nesse procedimento, quatro etapas principais são adotadas: conversão grafema-

fonema, anotação prosódica, representação na forma de vetores de características e seleção automática propriamente dita.

A primeira etapa (conversão grafema-fonema) é realizada para um grande *corpus* de texto sob análise usando regras de transcrição e um dicionário de exceções previamente definido, sendo algumas dessas regras descritas em [11]. A etapa de anotação prosódica é baseada em regras concebidas para descrever um padrão entoacional simplificado, sendo capaz de caracterizar as diferenças entoacionais entre sentenças declarativas, interrogativas e exclamativas do PB. Essas regras são especialmente desenvolvidas visando atribuir eventos tonais *low* e *high* (em níveis distintos de altura de *pitch*) a algumas sílabas-alvo das sentenças sob análise. Em uma próxima etapa, as informações fonética e prosódica geradas nos estágios de conversão e anotação prosódica são representadas na forma vetorial. Por fim, a seleção das sentenças que maximizam a variabilidade fonética e prosódica é realizada automaticamente por intermédio de algoritmos genéticos. Nesses algoritmos, a função de aptidão a ser maximizada consiste no número de vetores de características distintas. Como resultado de tal etapa de seleção, um conjunto de 4.000 sentenças do PB – incluindo sentenças declarativas, interrogativas parciais (wh), interrogativas totais (sim/não), interrogativas alternativas e exclamativas – é selecionado para compor o *corpus* de texto de nosso sistema de síntese de fala.

O trabalho aqui apresentado consiste em uma extensão do artigo apresentado em [12] com a inclusão das regras de anotação prosódica para sentenças exclamativas e questões alternativas como também do desempenho de nossos classificadores lexical e sintagmático.

Este trabalho está organizado como segue. A Seção II apresenta uma descrição da etapa de conversão grafema-fonema. O estágio de anotação prosódica é apresentado na Seção III. A representação em vetores de características é descrita na Seção IV. A Seção V apresenta o procedimento de seleção automática de sentenças. A Seção VI destaca os resultados experimentais. Por fim, as conclusões são dadas na Seção VII.

II. Conversão Grafema-Fonema

A conversão grafema-fonema é responsável por determinar a transcrição fonética de cada sentença contida no *corpus* sob análise. Tal transcrição é obtida considerando um léxico contendo pronúncias canônicas e um conjunto de regras de transcrição especialmente desenvolvido para o PB, sendo algumas dessas regras descritas em [11]. Nosso módulo de conversão grafema-fonema tem um desempenho bastante satisfatório, uma vez que apresenta problemas somente em algumas conversões de nomes próprios e palavras estrangeiras.

III. Anotação Prosódica

O procedimento proposto para anotação prosódica considera as seguintes etapas: classificação lexical de cada palavra existente em uma dada sentença, classificação sintagmática, classificação da tipologia da sentença e anotação prosódica propriamente dita.

A. Classificação Lexical

O primeiro estágio requerido para obtenção da anotação prosódica consiste na classificação lexical das palavras constituintes de uma dada sentença. Para tal classificação, consideramos as seguintes categorias: adjetivo (ADJ), advérbio (ADV), artigo (ART), conjunção (CONJ), locução adverbial (LADV), locução conjuntiva (LCONJ), locução prepositiva (LPREP) (nessa classe estão incluídas as contrações entre preposições e pronomes), locução pronominal relativa (LREL), número (NUM), pronome oblíquo átono (OBA), pronome oblíquo tônico (OBT), pronome demonstrativo (DEM), pronome indefinido (IND), pronome possessivo (POS), pronome interrogativo (PER), verbos em suas formas finitas (VER) e verbos em suas formas nominais (VERN).

É importante mencionar que as regras de classificação lexical não são aqui apresentadas por estarem fora do escopo deste trabalho. Nosso objetivo consiste em aumentar a variabilidade prosódica do *corpus* de texto. Por isso, apresentaremos apenas as regras de anotação prosódica de tal *corpus*.

Para mostrar o desempenho do classificador lexical aqui empregado, fizemos a etiquetagem automática das 200 primeiras frases de [10]. O desempenho obtido foi de 96,04%.

B. Classificação Sintagmática

Em lingüística, um sintagma consiste em um grupo de palavras que funciona como uma unidade, satisfazendo uma hierarquia de constituintes gramaticais. Cada sintagma apresenta uma palavra (chamada núcleo) cuja classificação lexical determina a categoria do sintagma no qual essa palavra está contida. Por exemplo, as categorias de sintagmas nominal, verbal, adverbial e preposicional (sendo aqui representadas por SN, SV, SADV e SPREP, respectivamente) têm como palavra núcleo um nome, um verbo, um advérbio e uma preposição, respectivamente [5], [13], [14].

Um exemplo de regra de classificação sintagmática é observado na frase “O garoto mora naquela casa”. As palavras que formam essa frase apresentam, repectivamente, as seguintes classificações lexicais: artigo, nome, verbo, locução prepositiva e nome. Dentre as regras existentes de classificação sintagmática, citamos aqui aquelas adotadas para classificar esse exemplo. Uma das regras estabelece que um artigo seguido por um nome constitui um SN. Essa regra permite classificar como SN o sintagma “o garoto”. Uma outra regra estabelece que um verbo constitui por si só um SV. Tal regra determina que a palavra “mora” é um SV. Outra regra estabelece que uma locução prepositiva seguida de um nome constitui um SPREP. Tal regra denota a classe sintagmática de “naquela casa” como SPREP.

Levando em consideração as categorias lexicais definidas, desenvolvemos um classificador sintagmático baseado em regras capazes de determinar as classes sintagmáticas constituintes de uma dada sentença.

Assim como a classificação lexical, a determinação da classe sintagmática não é foco desse trabalho. Por esse motivo, as regras não são aqui apresentadas. O desempenho do classificador foi também avaliado usando as 200 primeiras frases de [10]. Cada palavra constituinte dessas frases foi classificada e o desempenho obtido na classificação foi de 92,86%.

C. Classificação da Tipologia das Sentenças

Verificando que um distinto padrão é observado para cada classe de sentença, regras distintas de anotação prosódica são consideradas para cada classe [15]. Dessa forma, após a etapa de segmentação, o texto de entrada é processado visando determinar se uma dada sentença é declarativa, exclamativa ou interrogativa. Também foram consideradas as seguintes categorias para as sentenças interrogativas: parciais, totais e alternativas.

Interrogativas parciais (wh) são sentenças nas quais a resposta esperada é determinada por constituintes interrogativos (ou locuções wh) tais como: “quais”, “onde”, “quem”, “quando”, “por que”, dentre outros. Tais pronomes interrogativos podem estar localizados no início, meio ou fim de uma sentença. Em função de tal posicionamento, tais sentenças podem ser classificadas como interrogativas parciais iniciais, mediais e finais.

Em uma interrogativa total (sim/não), a resposta esperada é sim ou não. Por outro lado, em uma alternativa, tal resposta consiste em duas proposições reciprocamente excludentes.

Exemplos de interrogativas parciais (wh), totais (sim/não) e alternativas são, respectivamente, as sentenças “Qual é a sua idade?”,“Você gostaria de viajar comigo?” e “Ele gostaria de comer chocolate ou pudim?”.

D. Anotação Prosódica

Na etapa de anotação prosódica, as sentenças são anotadas de tal maneira que os padrões entoacionais de cada classe de sentença sejam representados de maneira simplificada. Nesse caso, adotamos aqui um conjunto de regras especialmente desenvolvido tendo como base a entoação do locutor de um sistema de síntese de fala do PB. Essas regras atribuem eventos prosódicos aos fonemas resultantes da conversão grafema-fonema e tomam como referência a análise do padrão entoacional das sentenças contidas em um banco de fala com 25 horas de duração. No banco de análise, os pontos correspondentes a picos ou vales no contorno de *pitch* de uma sentença podem ser também associados a características sintáticas comuns entre as sentenças analisadas, tais como classificação sintagmática, posição do sintagma na senteça, tonicidade, dentre outras.

Os símbolos adotados na etapa de anotação prosódica se baseiam nos símbolos da fonologia entoacional. Nesse caso, os eventos tonais *high* (H) e *low* (L) em diferentes níveis de altura de *pitch* (H+, H−, H, L, and L−) são atribuídos a algumas sílabas de uma dada sentença, tomando como referência os sintagmas aos quais pertencem. Os eventos mencionados são apresentados na Tabela I em ordem decrescente de freqüência fundamental ($H+ > H > H- > L > L-$) para as diferentes classes de sentenças. As sílabas que não se encaixam em qualquer das regras apresentadas nessa tabela são marcadas com o símbolo N (de neutro), indicando que os fonemas correspondentes podem apresentar qualquer contorno de *pitch* (ascendente, descendente ou neutro).

Para exemplificar as regras de anotação prosódica mostradas na Tabela I, apresentamos, na Fig. 1, o espectrograma e o contorno de *pitch* da frase “Você é do sexo masculino?” pronunciada pelo locutor de nosso sistema de síntese de fala. Essa frase constitui uma interrogativa total, por essa razão, verificamos que, assim como as regras evidenciam, há um aumento de *pitch* na sílaba tônica da palavra “você” que constitui o sintagma inicial da sentença. Verificamos também um aumento de *pitch* na sílaba tônica da última palavra da sentença (“masculino”). Além disso, as sílabas que antecedem os picos de *pitch* têm realmente valores baixos de *pitch*. Assim, tais sílabas são representadas através de um símbolo tonal baixo (L). Observamos, de forma geral, que esse padrão entoacional da sentença reafirma as regras estabelecidas para anotação prosódica de uma interrogativa total. Verificamos também que esse padrão é válido para a grande maioria de outras sentenças interrogativas totais. Assim, podemos estabelecer uma relação entre análises morfossintática e sintagmática e o padrão entoacional de sentenças.

O procedimento completo de anotação prosódica (classificação lexical, classificação sintagmática, classificação em sentenças e anotação prosódica propriamente dita) é aqui exemplificado para uma sentença declarativa, apresentada na Tabela II. Nesse caso, a sentença é dividida em sintagmas tomando como referência as categorias lexicais de suas palavras. As palavras são separadas em sílabas e, posteriormente, eventos prosódicos são associados a cada sílaba.

IV. Representação Vetorial

Após a etapa de anotação prosódica, as informações fonética e prosódica obtidas nos estágios anteriores são adotadas para representar cada fonema por um vetor de características.

Cada vetor possui os seguintes elementos: fonema anterior, fonema atual, fonema posterior e anotação prosódica do fonema atual. Para a sentença considerada na Tabela II, verificamos o seguinte conjunto de vetores: [silêncio a k N], [a k ˈe N], [k ˈe l N], [ˈe l i N], [l i b N], [i b ɐ̃ H+] e assim por diante.

V. Seleção Automática de Sentenças

Após a representação dos fonemas e de suas anotações prosódicas por um conjunto de vetores de características, é realizado o procedimento automático de seleção de sentenças. Tal seleção realiza uma busca, baseada em algoritmos genéticos, do conjunto de sentenças (população) que possui a maior quantidade de vetores de características, excluindo aqueles vetores contendo o símbolo prosódico “N”. Esse algoritmo só considera os vetores de características contendo os eventos tonais H+, H, H−, L e L−, uma vez que tais tons estão relacionados àquelas sílabas que constituem pontos característicos necessários (pontos em que contornos de *pitch* ascendentes ou descendentes ocorrem) à determinação do padrão entoacional simplificado (declarativo neutro, exclamativo, questão alternativa, parcial ou total) da sentença. Nesse caso, um conjunto maior de movimentos de *pitch* seria coberto pelo banco de dados de fala e uma maior variabilidade prosódica e fonética seria alcançada.

Este trabalho propõe uma possível solução para o problema de encontrar dentro de um grande *corpus* de texto (com aproximadamente 1.500.000 sentenças extraídas do banco de dados

TABELA I

REGRAS DE ANOTAÇÃO PROSÓDICA DAS SENTENÇAS

Sentenças declarativas	
H+	Primeira sílaba da última palavra do sintagma inicial da sentença
H	Primeira sílaba da última palavra de um SADV ou penúltima sílaba da sentença se não houver um SADV na sentença
H−	Penúltima sílaba do último sintgma da sentença se o sintagma antecessor for SADV
L	Sílaba que antecede o H ou H− do sintagma final de uma sentença
L−	Última sílaba do sintagma final da sentença
Sentenças exclamativas	
H	Primeira sílaba da sentença
L	Última sílaba do sintagma final da sentença
Questões parciais (wh) iniciais e mediais	
H+	Sílaba tônica da última palavra de uma locução wh (pronome interrogativo)
H	Sílaba tônica da última palavra da sentença
H−	Sílaba tônica da última palavra de um SADV e da última palavra de um sintagma intermediário[1]
L	Sílaba que antecede o H final e o H+ da locução wh da sentença
L−	Sílaba que sucede o H final
Questões parciais (wh) finais	
H+	Primeira sílaba de uma locução wh (pronome interrogativo)
H	Sílaba tônica da última palavra do sintagma inicial da sentença
H−	Sílaba tônica da última palavra de um SADV e da última palavra de um sintagma intermediário[1]
L	Sílaba que antecede o H+ da locução e o H do sintagma inicial
L−	Sílaba sucessora do H+ final da locução wh final
Questões totais (sim/não)	
H+	Sílaba tônica da última palavra da sentença
H	Sílaba tônica da última palavra do sintagma inicial da sentença
H−	Sílaba tônica da última palavra de um SADV e da última palavra de um sintagma intermediário[1]
L	Sílaba que antecede o H+ final ou sílaba que antecede o H do sintagma inicial
L−	Sílaba sucessora do H+ final
Questões alternativas	
H	Sílaba tônica que precede a conjunção “ou”
H−	Conjunção “ou”
L	Sílaba antecessora do H e sílaba sucessora do H se uma pausa ocorrer antes de tal conjunção
L−	Última sílaba tônica da sentença

[1] Sintagmas intermediários são anotados quando a sentença possuir mais de cinco sintagmas.

TABELA II

SENTENÇA QUE EXEMPLIFICA A ETAPA DE ANOTAÇÃO PROSÓDICA

Sentença	Aquele			bandido			foi	preso		ontem		à noite		
Classe lexical	DEM			NOME			VER	VERN		ADV		LADV		
Classe sintagmática	SN						SV			SADV		SADV		
Transcrição	[a	ˈke	li	b ɐ̃	ˈd i	d ʊ	ˈfoj	ˈpre	zu	ˈõ	tejˀ	a	ˈnoj	ʧi]
Símbolo prosódico	N	N	N	H+	N	N	N	N	N	H	N	L	H−	L−

CETENFolha[2]) 4.000 sentenças fonética e prosodicamente ricas, sendo 1.000 declarativas, 1.000 interrogativas parciais, 1.000 totais, 500 alternativas e 500 sentenças exclamativas. Considerando que sentenças exclamativas e interrogativas alternativas ocorrem a uma taxa muito menor tanto no banco de dados CETENFolha quanto no PB, 500 sentenças foram selecionadas para cada uma dessas classes.

[2]O banco CETENFolha é formado por uma compilação dos textos do jornal brasileiro “Folha de São Paulo”. Tal compilação foi feita pelo “Núcleo Interinstitucional de Lingüística Computacional (NILC)” localizado em São Carlos, Brasil.

A. Algoritmos Genéticos

Algoritmos genéticos são ferramentas de busca e otimização baseadas em seleção natural e herança genética. Tais algoritmos são recomendados para aplicações nas quais o espaço de busca da solução ótima é considerado grande o suficiente para tornar proibitivo um procedimento de busca exaustiva [10].

Em um algoritmo genético, primeiramente a função de aptidão é calculada para cada cromossomo da população inicial. Dois cromossomos (pais) são selecionados dentre os cromossomos existentes nessa população. Tais cromossomos se combinam através de uma operação genética de cruza-

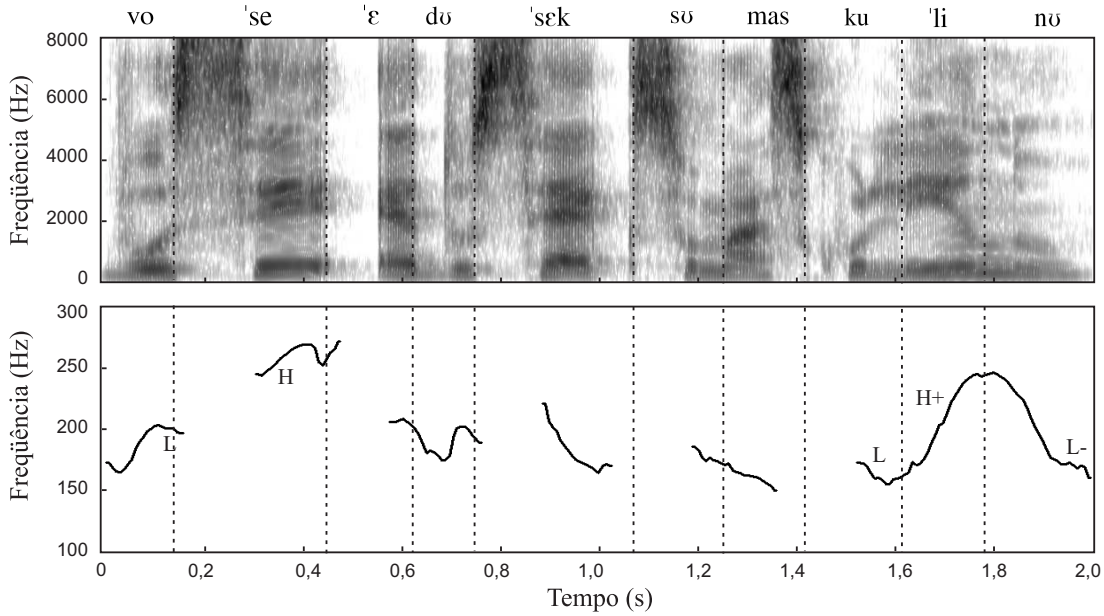

Fig. 1. Espectrograma e contorno de *pitch* para a sentença "Você é do sexo masculino?".

mento (*crossover*) com a finalidade de gerar cromossomos filhos. Uma operação de mutação pode ocorrer ao invés do cruzamento. Após a operação (de cruzamento ou mutação), a função de aptidão é calculada para os cromossomos resultantes da operação genética. Se os cromossomos resultantes (filhos) apresentarem uma função de aptidão de maior valor do que a de seus pais, os cromossomos filhos (ou cromossomo mutante) substituem seus pais. Caso contrário, os filhos são descartados e seus pais sobrevivem para a próxima geração. Esse ciclo se repete até que algum critério de parada seja atingido (por exemplo, um número pré-definido de gerações) [16].

VI. Resultados Experimentais

Em nossos experimentos, o *corpus* CETENFolha foi inicialmente dividido em sentenças declarativas, exclamativas e interrogativas. Tal *corpus* apresenta 1.390.000 sentenças declarativas, 909 exclamativas e 36.166 interrogativas (sendo 23.208 questões parciais iniciais/mediais, 392 questões parciais finais, 11.151 questões totais e 1.415 questões alternativas). Tal algoritmo é aqui executado separadamente para cada uma das seguintes classes: declarativa, exclamativa, interrogativa parcial, total e alternativa.

O conjunto de sentenças declarativas é inicialmente dividido em 40 grupos de 35.000 sentenças. Nesse caso, um algoritmo genético que seleciona as 1.000 sentenças de maior variabilidade fonética e prosódica é executado para cada grupo. Assim, 40.000 sentenças são obtidas. Essas 40.000 (40 grupos de 1.000) sentenças são tomadas como referência para outro algoritmo genético responsável por obter as 1.000 sentenças declarativas com o maior número de vetores de características.

Para o caso de sentenças exclamativas, as 500 sentenças mais aptas são selecionadas realizando apenas mutações (sem a ocorrência de qualquer *crossover*) de um único cromossomo.

No caso das sentenças interrogativas, 1.000 sentenças são selecionadas dentre as 23.600 interrogativas parciais (23 grupos de 1.000 sentenças mais 600 adotadas para mutação). Outras 1.000 sentenças são obtidas levando em consideração 11.151 interrogativas totais (11 grupos de 1.000 sentenças mais 151 para mutação). Por fim, 500 interrogativas alternativas são selecionadas dentre um total de 1.415 (sendo dois grupos de 500 considerados no cruzamento mais 415 para mutação).

Nos algoritmos considerados, cada grupo de sentenças corresponde a um único cromossomo. O número de vetores de características distintas (excluindo aqueles com o símbolo "N") é determinado para cada cromossomo. Em uma dada geração, uma operação genética (de mutação ou cruzamento) é realizada. Esse processo continua de forma iterativa até que o cromossomo mais apto da população permaneça inalterado por, no mínimo, 1.000 gerações.

É necessário mencionar que uma taxa de mutação igual a 10% é considerada para sentenças declarativas, interrogativas parciais e totais. Para as interrogativas alternativas e sentenças exclamativas, taxas de mutação de 50% e 100% (sem cruzamento) são adotadas, respectivamente. É importante notar que a taxa de mutação de interrogativas alternativas e sentenças exclamativas é maior do que aquela de outras classes de sentenças, visto que suas ocorrências no banco de dados CETENFolha conduzem a um reduzido conjunto de cromossomos. Além do mais, é possível observar que a operação de mutação apresenta melhor desempenho do que o cruzamento quando um conjunto reduzido de cromossomos é disponibilizado. Em um caso extremo, temos as sentenças exclamativas nas quais apenas um cromossomo é disponibilizado, o que impossibilita a operação de cruzamento sem repetição de sentenças.

A seleção de pais é baseada no método da roleta [16]. Nesse método, os cromossomos mais adaptados (com um maior número de vetores de características distintas) possuem uma maior probabilidade de serem selecionados como pais. O ponto de cruzamento é obtido de forma aleatória. Após a operação

de cruzamento, dois filhos são obtidos. Se o filho mais apto possuir um número maior de vetores de características distintas do que o pai mais apto, então os filhos substituem seus pais. Quando uma mutação ocorrer, o ponto de mutação e a quantidade de genes mutantes são obtidos de forma aleatória.

Os algoritmos aqui adotados são implementados na linguagem de programação Python. O resultado obtido utilizando a abordagem proposta é resumido na Tabela III. Por exemplo, o grupo inicial de 1.000 interrogativas totais com o maior número de vetores de características apresenta 5.720 vetores. No final do procedimento, o melhor conjunto de sentenças (cromossomo com 1.000 interrogativas totais) possui 6.191 vetores, indicando uma melhoria de 8,2%.

TABELA III
AUMENTO PERCENTUAL NO NÚMERO DE VETORES DE CARACTERÍSTICAS (EXCLUINDO AQUELES CONTENDO O SÍMBOLO PROSÓDICO N)

	Declarativas	**Parciais**	**Totais**	**Alternativas**	**Exclamativas**
Antes	7.268	6.124	5.720	3.963	1.069
Depois	8.413	6.469	6.191	4.145	1.150
Aumento	15,7%	5,6%	8,2%	4,6%	7,0%

Outra vantagem interessante do procedimento de seleção proposto consiste na possibilidade de reduzir o tamanho do banco de dados mantendo o mesmo número de vetores de características. Com a finalidade de determinar a quantidade de *pruning* do banco que poderia ser obtida, realizamos um segundo experimento. Sentenças foram selecionadas em ordem aleatória até que o número de vetores de características obtido após o procedimento de seleção (8.413, 6.469, 6.191, 4.145 e 1.150 vetores de características para declarativas, interrogativas parciais, interrogativas totais, interrogativas alternativas e sentenças exclamativas, respectivamente) fosse atingido para cada classe de sentença. Nesse caso, um número de 1.249 sentenças declarativas, 1.168 interrogativas parciais, 1.299 totais, 605 alternativas e 560 sentenças exclamativas são selecionadas, indicando que um *pruning* de, respectivamente, 19,9%, 14,4%, 23,0%, 17,4% e 10,7%, capaz de manter o mesmo número de vetores de características para cada classe de sentença, poderia ser atingido pela adoção do procedimento proposto.

VII. CONCLUSÕES

Neste trabalho, um conjunto de sentenças declarativas, exclamativas e interrogativas foi selecionado a partir de um grande banco de dados através da adoção de uma ferramenta de seleção baseada em algoritmos genéticos. Essa abordagem se mostrou útil para aumentar a variabilidade fonética e prosódica de um *corpus* de texto adotado em sistemas de síntese de fala. Para trabalhos futuros, pretendemos realizar tal seleção para obtenção de *corpora* com emoção visando integrá-lo ao atual sistema de síntese. Um teste perceptual de escuta também será realizado para comparar a fala sintética obtida a partir de dois bancos distintos de mesmo tamanho, sendo um fonética e prosodicamente rico e outro não.

REFERÊNCIAS

[1] A. J. Hunt and A. W. Black, "Unit selection in a concatenative speech synthesis system using a large speech database," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP'96)*, Atlanta, USA, May 1996, pp. 373–376.

[2] W. Zhu, W. Zhang, Q. Shi, F. Chen, H. Li, X. Ma, and L. Shen, "Corpus building for data-driven TTS systems," in *Proc. IEEE Workshop on Speech Synthesis (TTS'02)*, Santa Monica, USA, Sept. 2002, pp. 199–202.

[3] M. V. Nicodem, R. Seara, and F. S. Pacheco, "Reducing the natural click effect within database for high quality corpus-based speech synthesis," in *Proc. IEEE Int. Symp. Signal Processing and Its Applications (ISSPA'05)*, Sydney, Australia, Aug. 2005, pp. 607–610.

[4] M. V. Nicodem and R. Seara, "Natural click processing through wavelet analysis and extrapolation for speech enhancement," in *Proc. IEEE Int. Telecommunications Symposium (ITS'06)*, Fortaleza, Brazil, Sept. 2006, pp. 600–605.

[5] C.-H. Wu, C.-C. Hsia, T.-H. Liu, and J.-F. Wang, "Voice conversion using duration-embedded bi-HMMs for expressive speech synthesis," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 4, pp. 1109–1116, July 2006.

[6] J. F. Pitrelli, R. Bakis, E. M. Eide, R. Fernandez, W. Hamza, and M. A. Picheny, "The IBM expressive text-to-speech synthesis system for American English," *IEEE Trans. Speech Audio Processing*, vol. 14, no. 4, pp. 1099–1108, July 2006.

[7] H. Kawai, T. Toda, J. Ni, *et al*., "Ximera: A new TTS from ATR based on corpus-based technologies," in *Proc. ISCA Tutorial and Research Workshop on Speech Synthesis (SSW'04)*, Pittsburg, USA, June 2004, pp. 179–184.

[8] E. Klabbers and J. van Santen, "Control and prediction of the impact of pitch modification on synthetic speech quality," in *Proc. Europ. Conf. Speech Commun. Technol. (EUROSPEECH'03)*, Geneva, Switzerland, Sept. 2003, pp. 317–320.

[9] I. C. Seara, "Estudo Estatístico dos Fonemas do Português Brasileiro Falado na Capital de Santa Catarina para Elaboração de Frases Foneticamente Balanceadas," Dissertação de Mestrado, Universidade Federal de Santa Catarina, Florianópolis, Brasil, 1994.

[10] R. Cirigliano, C. Monteiro, F. Barbosa, *et al*., "Um conjunto de 1000 frases foneticamente balanceadas para o português brasileiro obtido utilizando a abordagem de algoritmos genéticos," *Anais do Simpósio Brasileiro de Telecomunicações (SBrT'05)*, Campinas, Brasil, Set. 2005, pp. 544–549.

[11] I. C. Seara, F. S. Pacheco, R. Seara Jr., S. G. Kafka, S. Klein e R. Seara, "Geração automática de variantes de léxicos do português brasileiro para sistemas de reconhecimento de fala," *Anais do Simpósio Brasileiro de Telecomunicações (SBrT'03)*, Rio de Janeiro, Brasil, Out. 2003, pp. 1–6.

[12] M. V. Nicodem, I. C. Seara, R. Seara, and D. dos Anjos, "Recording script design for a Brazilian Portuguese TTS system aiming at a higher phonetic and prosodic variability," in *Proc. IEEE Int. Symp. Signal Processing and Its Applications (ISSPA'07)*, Sharjah, United Arab Emirates, Feb. 2007, pp. 1–4.

[13] K. Yoon, "A prosodic phrasing model for a Korean text-to-speech synthesis system," *Computer, Speech, and Language*, vol. 20, no. 1, pp. 69–79, Jan. 2006.

[14] C. Mioto, M. C. Figueiredo Silva e R. E. Vasconcellos, *Manual de Sintaxe*. Florianópolis: Insular, 1999.

[15] J. A. Moraes, "Intonation in Brazilian Portuguese," in *Intonation Systems: A Survey of Twenty Languages*, D. Hirst and A. Di Cristo, Eds. Cambridge University Press, 1998, ch. 10, pp. 179–193.

[16] J. M. Johnson and V. Rahmat-Samii, "Genetic algorithms in engineering electromagnetics," *IEEE Antennas and Propagation Mag.*, vol. 39, no. 4, pp. 7–21, Aug. 1997.