Um Conjunto de 1000 Frases Foneticamente Balanceadas para o Português Brasileiro Obtido Utilizando a Abordagem de Algoritmos Genéticos

Ricardo J.R. Cirigliano, Clarisse Monteiro, Filipe Leandro de F. Barbosa, Fernando Gil Vianna Resende Jr., Letícia Rebollo Couto, João A. de Moraes

Resumo— **Este trabalho apresenta um algoritmo que é utilizado na obtenção de um conjunto de 1000 frases foneticamente balanceadas, com diversas aplicações em processamento de voz. Os conjuntos utilizados até então, além de serem consideravelmente menores, foram projetados para uma variedade específica da língua portuguesa. Neste trabalho, o conjunto de 1000 frases é obtido utilizando-se técnicas de algoritmo genético, visando a minimizar o número de unidades de síntese de voz não vistas na base. O espaço de busca de tal conjunto é um corpus eletrônico de um jornal brasileiro. Resultados mostram que o conjunto obtido é foneticamente balanceado para os fones representativos da variedade do português falado nos telejornais de abrangência nacional.**

Palavras-Chave— **Base de dados, foneticamente balanceado, síntese de voz, português brasileiro.**

Abstract— **This work presents an algorithm used to obtain a collection of 1000 sentences phonetically balanced, with several applications in speech processing. Until now, all sentences colletions were considerably smaller and designed for some variety of Portuguese. In this work, the 1000 sentences set is obtained through genetic algorithms techniques, looking for the minimization of the number of speech synthesis units not present in the collection. The search space is an electronic corpus from a Brazilian newspaper. The results show that the obtained collection is phonetically balanced to the phones of Brazilian Portuguese used in newspapers.**

Keywords— **Data base, phonetically balanced, speech synthesis, Brazilian Portuguese.**

I. Introdução

Bases de voz são de suma importância em sistemas de síntese de fala (TTS - *text-to-speech*). Nos sistemas TTS baseados em HMMs (*hidden Markov models*), por exemplo, quanto maior o número de ocorrências de uma dada unidade em uma base de dados de voz, mais acurado pode ser o modelo relacionado a tal unidade. Em [1] foi proposto um conjunto de 200 frases foneticamente balanceadas para o português falado no Rio de Janeiro. Tal conjunto vem sendo amplamente utilizado como base de voz para sistemas TTS para o português [3][4].

Ricardo J.R. Cirigliano, Fernando Gil Vianna Resende Jr. Programa de Engenharia Elétrica, COPPE, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, E-mails: rjcirig@lps.ufrj.br e gil@lps.ufrj.br. Filipe Leandro de F. Barbosa Departamento de Engenharia Eletrônica e de Computação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, E-mail: filipe@lps.ufrj.br. Clarisse Monteiro, Letícia Rebollo Couto, João A. de Moraes, Faculdade de Letras, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil, E-mails: clarissemontteiro@ig.com.br, leticiarcouto@yahoo.fr, jamoraes@gbl.com.br.

Neste trabalho é apresentado um algoritmo que auxilia na obtenção de um conjunto de 1000 frases foneticamente balanceadas, com diversas aplicações em processamento de voz, a partir de uma base de texto pré-selecionada de um extrato de jornal brasileiro. Tal conjunto foi projetado visando o balanceamento fonético, tomando como base o português falado em telejornais de abrangência nacional, caracterizado por uma pronúncia pouco marcada regionalmente e bem aceita em diversas regiões do país.

O presente trabalho está organizado em seções como segue. Na Seção II é apresentada uma breve descrição de sistemas de síntese de voz. Na Seção III são descritas as principais aplicações de algoritmos genéticos e é apresentado o algoritmo genético implementado para solucionar o problema da obtenção das 1000 frases. A Seção IV contém os resultados obtidos. Na Seção V são apresentados os descartes e substituições de frases realizados após a aplicação do algoritmo genético. A Seção VI apresenta os testes para validação do balanceamento fonético das 1000 frases. A Seção VII apresenta as conclusões a que chegamos.

II. Sintetizadores de Voz

Um sistema de síntese de voz pode ser entendido como um conversor texto-fala, ou seja, um sistema que recebe como entrada um texto e dá como saída a voz sintetizada. Existem vários métodos de síntese de voz, dentre os quais os mais comuns são os métodos de concatenação de formas de onda, como o PSOLA (*Pitch Synchronous Overlap and Add*). Outra técnica que vem se tornando popular é a baseada em HMMs [2].

Ao sintetizar voz, devemos escolher as unidades mínimas de síntese a serem utilizadas. Em alguns sintetizadores de voz, essas unidades são as sílabas. Outros conversores texto-fala utilizam os fones para realizar a síntese. A pesquisa em que este trabalho está inserido utiliza os trifones como unidades principais de síntese. Os trifones são representados por um fone central, com as transições do fone anterior e do posterior. A escolha de trifones como unidade de síntese representa um bom equilíbrio entre a co-articulação, refletida na transição entre fones, e o número de unidades existentes.

III. Algoritmos Genéticos

A. Descrição Geral

Algoritmos genéticos são algoritmos desenvolvidos com base na teoria da evolução [5]-[8]. Tais algoritmos são uti-

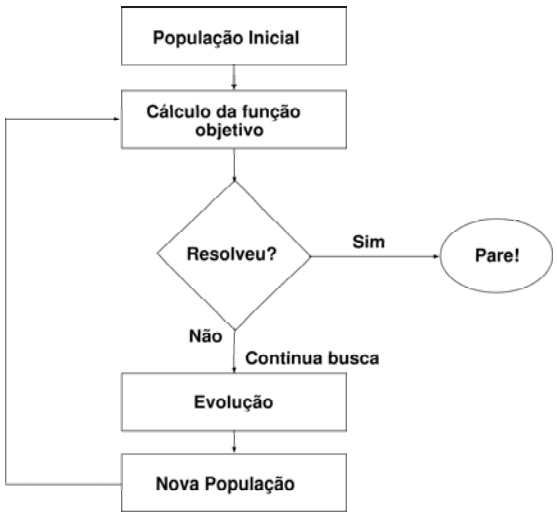

Fig. 1. Diagrama esquemático de um algoritmo genético.

lizados quando o espaço de busca da solução ótima para um dado problema é considerado grande o suficiente, para que torne proibitiva a procura com os recursos computacionais de hoje. Dependendo do tempo disponível para a obtenção da resposta para um dado problema, os algoritmos genéticos podem ser aplicáveis. Nesses algoritmos são utilizados conceitos de população, reprodução, seleção natural, mutações genéticas e *cross-over* [5].

A Figura 1 mostra um diagrama esquemático de um algoritmo genético simples. Podemos ver que a primeira fase do algoritmo é a geração da população inicial, feita aleatoriamente. Em seguida, calculando-se a função-objetivo do problema em questão, o membro mais apto da população é apontado. Se o objetivo de minimização ou maximização não foi alcançado, os membros mais aptos se reproduzem, e também sofrem mutações genéticas e *cross-over*, fatos resumidos, conforme mostrado na Figura 1, pelo termo Evolução. Os membros menos adaptados, ou seja, aqueles cuja função-objetivo não atingiu valores acima de limiares desejados, têm uma maior probabilidade de sofrerem alterações em seus genes. Após evoluir, a população passa para a próxima geração, na qual a função-objetivo será novamente calculada e o ciclo recomeçará, dando, por sua vez, origem a novas gerações.

B. Buscando a solução por algoritmos genéticos

O presente trabalho busca na teoria de algoritmos genéticos a solução para o problema de encontrar, em uma base de texto eletrônico de um extrato de jornal brasileiro, 1000 frases foneticamente balanceadas, minimizando o número de trifones não vistos na mesma, ou seja, fazendo com que, dentre todos os trifones existentes na língua portuguesa, essas 1000 frases contivessem o maior número possível. A base de dados de texto utilizada é a base do CETEN-Folha [9]. Essa base de dados foi obtida a partir das 365 edições do jornal brasileiro Folha de São Paulo no ano de 1994 e possui, aproximadamente, 24 milhões de palavras. Um conjunto menor de frases foi obtido, inicialmente, a partir da base do CETEN-Folha, selecionando apenas as frases que continham entre 9 e 12 palavras. Tais frases foram selecionadas por considerar-se este tamanho como representativo da média de grupos rítmicos e entonacionais do português. Além disso, frases curtas possuem poucos trifones e frases muito longas podem tornar o treinamento dos HMMs computacionalmente inviável. Cabe aqui ressaltar que a unidade palavra foi definida em função do espaço em branco. Segundo este critério "Rio de Janeiro" corresponde a três palavras e "lua-de-mel" a uma.

Aparentemente, o problema em questão parece ser de minimização, visto que queremos minimizar o número de trifones não vistos na base. Contudo, tal problema pode ser reescrito como um problema de maximização, em que o objetivo seria o de maximizar o número de trifones diferentes em um conjunto de 1000 frases. Assim, estaríamos automaticamente minimizando o número de unidades não vistas.

A função-objetivo do algoritmo genético desenvolvido é o número de trifones diferentes encontrados em cada conjunto de 1000 frases selecionado. Para quantificarmos a impossibilidade de uma computação ótima para resolver o problema em questão, vamos verificar o número total de combinações de 1000 frases que teríamos que obter para chegarmos ao conjunto ótimo, ou seja, ao conjunto de frases que tivesse o maior número de trifones diferentes.

Considerando que o arquivo de frases selecionado do CETEN-Folha possui 213000 frases, aproximadamente, teríamos que computar a combinação de 213000 frases, tomadas 1000 a 1000. Tal número saturaria uma variável de 64 bits.

Para que tal combinação fosse calculada, utilizou-se uma biblioteca implementada em C++, que realiza operações aritméticas com resolução arbitrária, a GMP (*Gnu Multiple Precision*) [10].

O número total de combinações encontrado, ou seja, de conjuntos de 1000 frases possíveis, foi $6,22.10^{2764}$. Considerando o fato de que o cálculo da função-objetivo, para cada conjunto de frases, demora aproximadamente 5 segundos em um computador com processador ATHLON 700 MHz, o número total de segundos para finalizar tal computação seria $3,1.10^{2765}$, que corresponde a 10^{2758} anos, aproximadamente.

Para efeito de comparação, desde o surgimento do Universo até hoje, transcorreram cerca de 10^{10} anos, o que deixa claro o fato do cálculo do conjunto ótimo ser inviável, considerando o poder computacional existente atualmente.

Para que tal solução ótima fosse possível precisaríamos de um computador capaz de computar, em um segundo, 3.10^{2757} funções objetivo para os conjuntos de 1000 frases. Ainda assim, o resultado levaria um ano para ser obtido. O poder de processamento deste computador precisaria ser 3.10^{2757} maior do que um computador 5 vezes mais rápido que o computador utilizado para computar a função-objetivo.

C. Descrição do algoritmo desenvolvido

Para poder solucionar o problema das 1000 frases em tempo hábil, foi desenvolvido um algoritmo baseado na teoria de algoritmo genético. Tal conjunto de procedimentos é descrito em passos, a seguir:

TABELA I
RESULTADO PARA DIVERSAS CONFIGURAÇÕES DE N E G.

N	G	Regiões	Total Trifones
2	5	152952	3208
2	10	21375	3206
10	30	174198	3231
30	50	27015	3205
40	50	120582	3217
50	60	174201	3229

1) Escolher N regiões aleatórias nesse espaço: essas regiões aleatórias serão representadas por números naturais que se referem às linhas do arquivo das 213000 frases, a partir das quais as 1000 frases serão selecionadas.
2) Obter as 1000 frases a cada região escolhida: a obtenção dessas frases é feita selecionando as 1000 frases subseqüentes à cada região selecionada no passo anterior.
3) Calcular o número total de trifones para cada grupo e verificar se existe algum trifone com ocorrência menor que 3. Se existir, esse pai gera um filho novamente e esse passo é repetido.
4) Dividir, para as próximas gerações, a população em duas, uniformemente: população adaptada (N/2) e população não adaptada (N/2). A primeira foi a população que deu origem aos melhores conjuntos de 1000 frases.
   a) Obter a próxima geração a partir da população adaptada. Para a próxima geração, é gerado 1 filho por pai com mutações proporcionais à colocação dos pais no que diz respeito ao número de trifones diferentes encontrados:
      - A constante de proporcionalidade será um valor inteiro K, que satura em 1000. O sinal de K é obtido aleatoriamente.
      - Se K é positivo: a nova região gerada será R + K*(colocação do pai), onde R é a região obtida na geração atual.
      - Se K é negativo: a nova região gerada será R - K*(colocação do pai), onde R é a região obtida na geração atual.

      Após a obtenção dos filhos, a função-objetivo é calculada, verificando se eles têm um maior número de trifones do que os pais. No caso positivo, os filhos são mantidos. Do contrário, os filhos são anulados e os pais sobrevivem para a próxima geração.
      O objetivo de tal procedimento é tentar encontrar uma nova região, melhor do que a região pai, na sua proximidade.
   b) Obter a próxima geração a partir da população não adaptada. A partir dessa população, com o objetivo de gerar populações mais adaptadas para as próximas gerações, são feitas mutações genéticas ou crossover, com taxas aleatórias [7]. Isto significa dizer que cada região R da população não adaptada irá gerar uma outra região R', por mutação genética (troca aleatória dos bits de R) ou *cross-over* (comutação de parte dos bits de R com parte dos bits de outra região não adaptada). As taxas aleatórias são justamente a quantidade de bits de R considerados na mutação genética ou no crossover. A decisão se o membro da população sofrerá mutação genética ou *cross-over* também é obtida aleatoriamente.
      Vale lembrar que os passos 4.a e 4.b serão ambos calculados a cada geração.
5) Retornar ao passo 1, percorrendo esse ciclo até que a região escolhida como a mais apta, isto é, com um maior número de trifones, não se altere por mais de G gerações. É importante notar que cada vez que a melhor região é atualizada por uma nova, o contador de gerações é zerado. A cada geração sem atualização da melhor região, o contador é incrementado. Quando o contador atinge G, o algoritmo se encerra.

IV. RESULTADOS

Na implementação do algoritmo genético desenvolvido, foram testadas várias configurações para os valores de N e G. Tais configurações estão listadas na Tabela I, mostrando as melhores regiões obtidas para cada configuração. Foram testados números de população de 2 até 50. A Tabela I sumariza tais resultados. Sendo executado na primeira configuração, o algoritmo levou poucos minutos para convergir, obtendo um total de 3208 trifones diferentes e apontando a região 152952 como sendo a melhor. Na segunda configuração, quando aumentamos o número G, não houve um bom resultado, já que o programa apontou uma região com ocorrência de 3206 trifones diferentes. Ao aumentar o número N, obtivemos um melhor resultado, 3231 trifones, com o algoritmo rodando por 3 horas, aproximadamente. Cabe lembrar que para essa configuração o número G também aumentou, provocando um maior número de buscas antes de convergir. Na quarta configuração do sistema, observamos uma piora no resultado, com a convergência do algoritmo após 4 horas de execução.

A quinta configuração obteve uma melhora em relação a quarta, com o algoritmo terminando após 5 horas de execução. Na última configuração, após 6 horas de execução do algoritmo, obtivemos a mesma região, somente havendo a troca de duas frases, obtida na configuração 3. A região 174198 foi a escolhida, e as 1000 frases foram obtidas a partir delas, levando a 3231 trifones diferentes.

O número de trifones diferentes obtido para as 1000 frases mostrou-se satisfatório, visto que foi testada a função-objetivo no conjunto de 213000 frases e foram obtidos cerca de 4200 trifones diferentes com ocorrência percentual maior ou igual à ocorrência do trifone que apareceu menos vezes nas 1000 frases.

V. DESCARTES, SUBSTITUIÇÕES E CORREÇÕES DE FRASES

Uma vez estabelecido o primeiro conjunto de 1000 frases, foram identificadas algumas frases que precisariam ser descartadas. Afim de obtermos um conjunto final de 1000, após todos os descartes, o segundo melhor conjunto de 1000 frases

foi unido ao primeiro, resultando em um total de 2000 frases. Esse segundo conjunto corresponde à região 120582.

Foram descartadas frases com as seguintes características:

- Frases com palavras, siglas ou abreviações que apresentassem, em sua grafia, seqüências grafemáticas não previstas nas convenções ortográficas do português (como Senna, shoppings), ou palavras, geralmente nomes próprios estrangeiros, cuja pronúncia corrente não obedece às regras, estabelecidas para o português, de conversão texto-fone (como Freud, Einstein) ou às regras prosódicas de atribuição do acento tônico com base na grafia (como em Roger);
- Frases com um número de palavras acima do estipulado, surgidas após a transcrição de numerais e datas por extenso ou, ao contrário, abaixo do previsto, após a correção de erros de transcrição na atribuição de espaço em branco no que diz respeito tanto a caracteres de pontuação, quanto a palavras compostas, pela omissão do hífen, como em "lua de mel";
- Frases incompletas ou que perderam o sentido fora do contexto;
- Frases que estivessem fora do padrão de estilo lido, assertivo ou continuativo. Foram eliminadas portanto frases com trechos de poemas, frases interrogativas e frases exclamativas. Igualmente foram descartadas frases introduzidas por marcadores conversacionais do tipo "mas" ou "e" (tais como em "Mas vamos fechar os olhos e pensar numa outra coisa."). Após alguns testes de leitura em voz alta, percebemos a importância de manter um estilo de leitura homogêneo, pois a alternância constante de registro e padrões prosódicos (leitura, fala espontânea, interrogações e exclamações), somada ao cansaço inerente à extensão da base de dados, causa hesitações e titubeios durante a oralização das frases.

Ao final o algoritmo genético foi novamente rodado para selecionar apenas 1000 frases. A região 2 foi a selecionada, apresentando 3223 trifones.

Na Tabela II são apresentadas as 30 primeiras frases, do total das 1000 frases obtidas, disponíveis na página virtual www.lps.ufrj.br / voz/1000frases.

VI. Balanceamento Fonético

A Tabela III mostra uma lista da freqüência relativa dos fones nas 1000 frases e no CETEN-Folha. As freqüências fonéticas do CETEN-Folha e das 1000 frases foram levantadas após utilizarmos o algoritmo de transcrição grafema-fone descrito em [11], que faz o mapeamento das letras do texto eletrônico, os grafemas, para suas respectivas leituras, os fones.

Com o objetivo de verificar se as 1000 frases são foneticamente balanceadas em relação à base CETEN-Folha, foi realizado um teste qui-quadrado como descrito em [1]. O teste considerou que, uma vez que a base de dados possui 38 fones, existem 37 graus de liberdade. Segundo a tabela apresentada em [1], para 35 graus de liberdade o qui-quadrado deve estar abaixo de 18,51 para considerarmos o conjunto de 1000 frases foneticamente balanceado em relação à base completa do CETEN-Folha. O teste realizado apresentou um qui-quadrado igual a 0,53, significativamente menor que 18,51, mostrando que as 1000 frases estão foneticamente balanceadas.

No entanto, um conjunto de 1000 frases é demasiadamente extenso para ser utilizado em testes subjetivos. Para tais testes, é interessante que as 1000 frases sejam utilizadas apenas na fase de treinamento do sistema, e conjuntos menores, de 10 ou 20 frases, para os testes. A base de 1000 frases foi então dividida em conjuntos de 10 e 20 frases, sem que fosse feita uma busca otimizada para o agrupamento. Os conjuntos foram retirados de forma seqüêncial das 1000 frases. Em cada conjunto foi realizado um teste qui-quadrado para verificar se ele era foneticamente balanceado em relação à base completa do CETEN-Folha. Todos os conjuntos de 10 e 20 frases tiveram os seus qui-quadrado abaixo de 18,51. Contudo, um grande número de conjuntos, apesar de foneticamente balanceados, não apresentavam alguns fones que possuem baixa probabilidade de ocorrência. Foi então definida a restrição que só seriam utilizados conjuntos com qui-quadrado abaixo de 18,51 e que apresentassem no mínimo duas ocorrências de cada fone. Sob esta restrição, nenhum conjunto de 10 frases se mostrou válido como conjunto de testes. Contudo, 17 conjuntos de 20 frases, apresentados na Tabela IV, foram identificados como válidos para realização de testes. Esse conjuntos foram movidos para o início do conjunto de 1000 frases e ordenados de forma crescente segundo seus qui-quadrados.

VII. Conclusões

Neste trabalho é obtido um conjunto de 1000 frases através de um algoritmo baseado na teoria de algoritmo genético. A abordagem baseada em algoritmo genético possibilitou a resolução de um problema cuja solução ótima seria proibitiva nos dias de hoje. O espaço de busca utilizado é o corpus eletrônico de um jornal brasileiro. Os testes realizados mostram que o conjunto de 1000 frases é foneticamente balanceado para português falado em telejornais, permitindo que estas frases sejam utilizadas em diversas aplicações de processamento de voz. De forma a facilitar a utilização do conjunto de 1000 frases para testes subjetivos, foram identificados 17 conjuntos de 20 frases que são foneticamente balanceados em relação ao conjunto total de 1000 frases, e que apresentam no mínimo duas ocorrências de cada um dos fones.

Referências

[1] A. Alcaim, J. A. Solewicz e J. A. 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, Rio de Janeiro. v. 7, n. 1, p. 23-41, 1992.

[2] K. Tokuda, "Text-to-Speech Synthesis: New Paradigms and Advances, Shrikanth Narayanan, Abeer Alwan (Eds.), An HMM-Based Approach to Multilingual Speech Synthesis", Prentice Hall, 2004.

[3] G. Pinto, F. L. F. Barbosa, and F. G. V. Resende Jr., "A Brazilian Portuguese TTS based on HMMs", Proc. of International Telecommunication Symposium, 2002. p. 868-872.

[4] R. S. Maia, R. Zen, K. Tokuda, T. Kitamura and F. G. V. Resende Jr., "Towards the development of a Brazilian Portuguese Text-to-Speech System Based on HMMs", In EUROSPEECH, p.2465-2468, 2003.

[5] "GENETIC Algorithms - An Intuitive Introduction", Disponível em: http://homepage .sunrise.ch/homepage/pglaus/gentore.htm. Acesso em: 03/06/2004.

[6] "GENETIC Algorithms Research And Application Group (The Garage)", Disponível em: http://garage.cps.msu.edu/demos/ flywheel /index.html. Acesso em: 31/05/2004.

[7] "A DEMONSTRATION of The Genetic Algorithm", Disponível em: http://math .hws .edu/xJava/GA/. Acesso em: 05/06/2004.

[8] "INTRODUCTION to Genetic Algorithms", Disponível em: http://cs.felk .cvut .cz/ xobitko /ga/. Acesso em: 10/06/2004.

[9] "CORPUS de Extractus de Textos Electrónicos Nilc/ Folha de São Paulo (Ceten-Folha)", Disponível em: http:// acdc.linguateca .pt /cetenfolha/. Acesso em: 06/12/2002.

[10] "BIBLIOTECA GMP (GNU MULTIPLE PRECISION)", Disponível em: http://www. swox.com/gmp/. Acesso em: 14/06/2004.

[11] F. L. F. Barbosa, G. O. Pinto, F. G. V. Resende Jr., C. A. Gonçalves, R. Monserrat and M. C. Rosa, "Grapheme-phone transcription algorithm for a Brazilian Portuguese TTS", In 6th. Workshop on Computational Processing of the Portuguese Language-Written and Spoken, 2003, Faro, Portugal.

TABELA II

AS 30 PRIMEIRAS FRASES DO CONJUNTO DAS 1000 FRASES.

1	Pesquisa é uma coisa que muda a toda hora.
2	No total, serão chamados vinte e seis mil candidatos.
3	O número de convocados por vaga é de doze candidatos.
4	Atualmente, esse abatimento é limitado a setenta por cento dos gastos .
5	Sandra Regina Machado: acho que ela enfim criou juízo.
6	Eles estão colocando armadilhas nas fazendas onde já ocorreram os ataques.
7	Dessas, somente umas trezentas e vinte foram inauguradas em território americano.
8	No total, sete mísseis foram disparados contra o encrave.
9	Em Florianópolis, foi registrado dois graus Celsius na manhã de domingo.
10	As situações ditas embaraçosas são resolvidas com os dados.
11	Itamar tem razão de estar exultante como nunca desde que virou presidente.
12	A mãe de todas as reformas é a reforma política.
13	Conseguiram eliminar áreas supérfluas ou que antes eram desperdiçadas.
14	Uma lata de leite em pó integral vale um ingresso.
15	A maioria dos passageiros do barco naufragado era de crianças.
16	A provável causa do acidente foi excesso de lotação a bordo.
17	Não prometo nada, porque não adianta eu prometer e não cumprir.
18	Se for eleito, vocês vão ver o meu trabalho.
19	Ele era um dos poucos atores negros que tinham espaço.
20	A secretaria estadual de saúde distribuirá cem mil preservativos no carnaval.
21	São essas qualidades que inspiraram o plano real desde a sua criação.
22	Todos os batizados são consagrados a Deus e devem tender à santidade.
23	"O homem Sem Qualidades" não é um livro comum.
24	Os problemas surgem nas importações diretas via catálogos, por exemplo.
25	Lula chegou ao Rio com duas horas de atraso.
26	No ano, a taxa é de três vírgula sete por cento.
27	Batizado de Heitor, o trabalho traz treze faixas compostas pelo guitarrista.
28	Já Viola levou o terceiro cartão amarelo domingo passado.
29	Inflação volta a subir no Rio e em São Paulo.
30	Quero estar no grupo e ajudar sempre que necessário.

TABELA III

LISTA DE FREQÜÊNCIAS RELATIVAS DE FONES NO CETEN-FOLHA E NAS 1000 FRASES.

CETEN-Folha			
a	12,17	dZ	1,81
i	8,76	R	1,61
u	6,70	v	1,46
s	6,54	o~	1,41
e	6,03	f	1,31
r	4,50	w~	1,22
o	3,68	i~	1,20
t	3,45	b	1,09
k	3,39	X	1,07
d	3,31	g	0,99
z	2,94	tS	0,95
p	2,91	j	0,78
m	2,87	u~	0,74
w	2,83	Z	0,71
n	2,77	E	0,66
a~	2,65	S	0,44
e~	2,13	J	0,40
j~	2,11	L	0,25
l	1,99	O	0,22
1000 Frases			
a	12,21	dZ	1,63
i	8,91	v	1,57
s	6,77	R	1,45
u	6,38	o~	1,33
e	6,09	j	1,23
r	4,67	i~	1,23
o	3,68	w~	1,21
d	3,48	f	1,20
t	3,39	x	1,04
z	3,29	tS	1,01
k	3,25	g	0,98
p	2,92	b	0,98
n	2,81	u~	0,75
m	2,71	Z	0,69
a~	2,65	E	0,67
w	2,54	S	0,39
e~	2,21	J	0,34
l	1,97	O	0,29
j~	1,87	L	0,22

TABELA IV

SUB-CONJUNTOS DE 20 FRASES RETIRADOS DO CONJUNTO DE 1000 FRASES VÁLIDOS PARA TESTES.

Núm. da Frase Inicial	Núm. da Frase Final	Qui quadrado
1	20	2,23
21	40	2,96
41	60	3,13
61	80	3,40
81	100	3,41
101	120	3,53
121	140	3,59
141	160	3,64
161	180	3,88
181	200	4,10
201	220	4,37
221	240	4,77
241	260	4,90
261	280	5,04
281	300	5,60
301	320	6,06
321	340	8,73