 • ISSN Print: 2447-0481 • ISSN Online: 2447-049X

Avaliação do Desempenho de um Sistema de Reconhecimento Automático de Voz em Português do Brasil Baseado em *Software* Livre para Geração de *Closed Caption*

Luiz Fausto de Souza Brito, Flávio Luis de Mello

***Abstract*— Automatic Speech Recognition (ASR) is a technique that may be applied to the generation of closed caption for live television programs with lower operating costs than stenotyping. However, there is a lack of satisfactory commercial systems that implement this technology for Brazilian Portuguese language. This paper presents the training and the comparative performance assessment of an ASR system for Brazilian Portuguese, using free software and public databases. The results analysis suggests that it is feasible to develop a system with superior performance compared to both stenotyping and currently available commercial Brazilian Portuguese ASR systems.**

***Index Terms*—Automatic Speech Recognition, Brazilian Portuguese language, Closed Caption.**

I. Introdução

A necessidade de utilização de legenda oculta (*closed caption*) na programação da televisão, tanto como recurso de acessibilidade, quanto para cumprimento de legislação específica, já foi discutida em um artigo anterior [1]. Naquele mesmo trabalho, foi apresentada a tecnologia de Reconhecimento Automático de Voz (processo de conversão do sinal de voz em uma transcrição textual correspondente), como uma alternativa à Estenotipia (digitação em tempo real utilizando símbolos fonéticos em teclado especial, convertidos em palavras de acordo com um dicionário) com menor custo operacional para viabilizar o uso de *closed caption* durante a ocorrência de fala espontânea ao vivo. Foram também apresentados alguns exemplos de utilização dessa abordagem relatados por diversos autores ao redor do mundo. Por fim, foram apresentadas as dificuldades específicas de implementação desse tipo de sistema utilizando o idioma Português do Brasil e alguns caminhos possíveis para a superação de tais dificuldades.

Entre as dificuldades relatadas está a falta de opções software comerciais de RAV em Português do Brasil otimizados para essa aplicação, citando em particular a obsolescência do IBM ViaVoice [2]. Desde então, surgiram algumas poucas opções de software comerciais de RAV utilizados para geração de legenda oculta no Brasil [3] [4], mas a acurácia relatada desses sistemas ainda está distante da meta de 98% de acerto para *closed caption* ao vivo estipulada pela norma ABNT NBR 15290 [5].

Como caminho possível para a superação dessas dificuldades, foi indicada a disponibilidade de diversas ferramentas de *software* livre que implementam os algoritmos de processamento de sinais e modelos estatísticos utilizados no Reconhecimento Automático de Voz, mas que para serem aplicadas ao Português do Brasil necessitam de treinamento com bases de dados brasileiras. Foram também indicadas algumas iniciativas que buscam disponibilizar publicamente as bases de dados necessárias para o Português do Brasil.

Este artigo apresenta os resultados de testes comparativos de desempenho de um sistema de RAV em Português do Brasil, desenvolvido a partir de ferramentas e recursos disponíveis publicamente, com o desempenho do IBM ViaVoice e da Estenotipia, uma síntese do trabalho desenvolvido como requisito parcial para obtenção do grau de Mestre em Computação Aplicada [6].

II. Proposta de Avaliação

Neste trabalho foi realizada uma avaliação de desempenho da transcrição textual realizada por Estenotipia e por Reconhecimento Automático de Voz tomando como objetos de estudo um telejornal, um programa jornalístico sobre saúde e um programa de auditório. Estes três tipos de programas apresentam níveis distintos de dificuldade no que tange a transcrição textual, sendo o telejornal um caso menos complicado, o programa jornalístico sobre saúde um representante de um caso intermediário e o programa de auditório o caso mais complicado. O Reconhecimento Automático de Voz empregou a relocução (repetição das falas por um locutor específico em um ambiente acusticamente controlado) e utilizou o aplicativo IBM ViaVoice [2] e um

Luiz Fausto de Souza Brito é Mestre em Computação Aplicada pela Universidade Estadual do Ceará e Engenheiro de Projetos de Telecomunicações da Rede Globo, Rio de Janeiro, RJ, Brasil (e-mail: luiz.fausto@tvglobo.com.br).

Flávio Luis de Mello, DSc, é Professor Adjunto da Universidade Federal do Rio de Janeiro da Escola Politécnica, do Departamento de Eletrônica e de Computação, Rio de Janeiro, RJ, Brasil (e-mail: flavio.mello@del.ufrj.br).

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

sistema baseado em um *software* livre (CMU Sphinx) [7] treinado para o Português do Brasil utilizando bases de dados disponíveis publicamente.

O uso da relocução é conveniente porque o sistema pode ser adaptado à voz do relocutor, reduzindo a complexidade do reconhecimento de voz, por não precisar considerar a variabilidade fonética entre indivíduos. Além disso, como a relocução se dá em um ambiente acusticamente controlado, proporciona uma razão sinal/ruído melhor para o sistema de reconhecimento. Outra vantagem de utilizar a relocução, é que o relocutor pode reformular a fala, corrigindo as disfluências comuns na linguagem oral, tornando-a mais adequada a uma transcrição para a linguagem escrita. Em caso de fala muito rápida (e.g. locução esportiva), o relocutor pode resumir a fala, para que a velocidade de exibição do texto seja suficientemente lenta para permitir a leitura [8] [9]. Desta forma, o uso da relocução é uma abordagem que mitiga uma série de problemas difíceis de serem tratados computacionalmente.

O IBM ViaVoice foi utilizado como referência por ter sido o primeiro *software* de Reconhecimento Automático de Voz em Português do Brasil utilizado na geração de Legenda Oculta, sendo ainda hoje muito utilizado nesta aplicação. Chama-se a atenção para o fato de que este software foi descontinuado pelo seu fabricante e que atualmente não há atualização ou suporte técnico para o mesmo, criando uma vulnerabilidade às operações que dependem dele.

O CMU Sphinx foi escolhido para esse teste por ser um *software* de RAV livre bem conceituado atualmente, relativamente maduro (desenvolvido desde 1986) e que conta com uma comunidade de desenvolvimento bastante ativa. Além disto, contém todas as ferramentas necessárias ao treinamento e teste de um sistema de RAV, está bem documentado para permitir a realização de todos os testes pretendidos e adota uma licença de uso que não restringe sua aplicação comercial.

As bases de dados necessárias para o treinamento do CMU Sphinx são de dois tipos: um conjunto de textos (*corpus* de texto) e gravações de voz com transcrição textual (*corpus* de voz). O *corpus* de texto é utilizado para gerar o dicionário, que define o vocabulário do sistema (utilizando as palavras mais frequentes do texto) e a sequência de fonemas correspondente a cada palavra (através de conversão grafema-fonema). O *corpus* de texto também é utilizado para gerar o modelo de linguagem, que define a probabilidade de ocorrência de sequências de palavras. O *corpus* de voz é utilizado para gerar o modelo acústico, que define a correspondência entre o sinal de voz e os fonemas.

III. Material de Teste

Uma representação esquemática da preparação do material de teste é apresentada na Fig. 1.

Fig. 1. Representação esquemática da preparação do material de teste

Para os testes foram escolhidos o telejornal "Bom Dia Brasil", o programa jornalístico sobre saúde "Bem Estar" e o programa de auditório "Domingão do Faustão", todos da Rede Globo, exibidos entre os dias 18 e 19 de novembro de 2012. O primeiro contém cinco blocos com duração total de 52 minutos e 50 segundos e 7903 palavras. O segundo contém dois blocos com duração total de 38 minutos e 55 segundos e 6333 palavras. O terceiro contém quatro blocos com duração total de 1 hora, 45 minutos e 2 segundos e 14686 palavras. Os referidos programas possuíam *closed caption* produzido por Estenotipia.

Neste sentido, foi extraído o texto do *closed caption* produzido por Estenotipia (utilizando o *software* livre CCExtractor [10]). Paralelamente, foi gravada a relocução e foram feitas transcrições manuais do áudio original e relocutado. O áudio da relocução foi utilizado para reconhecimento de voz pelo IBM ViaVoice e pelo CMU Sphinx. As transcrições manuais do áudio original e relocutado, os textos produzidos por Estenotipia e as transcrições realizadas pelo IBM ViaVoice e pelo CMU Sphinx para o áudio de relocução foram comparados de modo a avaliar acurácia da Estenotipia, da relocução e de cada um dos sistemas de RAV empregados. Além disso, foi verificada a latência da Estenotipia, da relocução e dos sistemas de RAV, além do consumo de recursos computacionais dos sistemas de RAV.

A acurácia é avaliada através da taxa de erro de palavras (*WER – Word Error Rate*) [11], definida como:

$$WER = (S + O + I)/N \quad (1)$$

onde S é o número de substituições, O é o número de omissões, I é o número de inserções e N é o número total de palavras na transcrição correta. Dessa forma, a acurácia é definida como (1 - *WER*). Note-se que a acurácia é distinta da taxa de acerto, que é definida como (C / N), onde C é o número de palavras corretas. Em geral, a acurácia apresenta um valor menor que a taxa de acerto. A taxa de acerto é a métrica utilizada na ABNT NBR 15290 [5], enquanto a acurácia é a métrica preferencialmente utilizada em todo o restante da literatura sobre o assunto, inclusive neste trabalho.

Para verificação dos erros de uma transcrição, o texto deve ser alinhado com o texto de referência de forma que possibilite uma avaliação mais acertada entre as transcrições. A necessidade de alinhamento está ilustrada na comparação dos

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

textos *A* (referência) e *B* (transcrição sendo avaliada) com e sem alinhamento na Tabela 1.

TABELA 1
COMPARAÇÃO DE TEXTOS COM E SEM ALINHAMENTO

Comparação de Textos sem Alinhamento									
A:	começa	a	funcionar	o	reforço	no	policiamento	das	divisas
B:	começa	a	funcionar	reforço	no	policiamento	da	rede	visa
	C	***C***	***C***	***S***	***S***	***S***	***S***	***S***	***S***

Comparação de Textos com Alinhamento											
A:	começa	a	funcionar	o	reforço	no	policiamento	das	divisas		
B:	começa	a	funcionar		reforço	no	policiamento	da	rede	visa	
	C	***C***	***C***	***O***	***C***	***C***	***C***	***S***	***S***	***I***	

Sem alinhamento dos textos, a omissão ou inserção de uma palavra pode fazer com que várias palavras subsequentes que foram transcritas corretamente sejam comparadas com palavras distintas, resultando na contagem indevida de diversos erros de substituição. No exemplo apresentado, sem alinhamento, considera-se a ocorrência de 6 erros por substituições de palavras, enquanto com alinhamento, considera-se a ocorrência de apenas 4 erros (1 omissão, 2 substituições e 1 inserção).

IV. AVALIAÇÃO DA ESTENOTIPIA

O *closed caption* gerado por Estenotipia foi extraído da programação gravada em dois formatos: um deles apenas com o texto, para avaliação da acurácia da transcrição, e outro contendo a marcação de tempo de cada caractere, para avaliação da latência. Um exemplo de um trecho de *closed caption* extraído nos dois formatos está ilustrado na Tabela 2.

TABELA 2
EXEMPLO DE TRECHO DE CLOSED CAPTION EXTRAÍDO EM DOIS FORMATOS: APENAS TEXTO E TEXTO COM MARCAÇÃO DE TEMPO DE CADA CARACTERE

Closed Caption **(apenas texto)**	***Closed Caption*** **(texto com marcação de tempo de cada caractere)**
>> CHICO PINHEIRO: BOM DIA,	00:00:14,179 | >> C 00:00:14,212 | >> CHI 00:00:14,246 | >> CHICO 00:00:14,279 | >> CHICO P 00:00:14,313 | >> CHICO PIN 00:00:14,346 | >> CHICO PINHE 00:00:14,379 | >> CHICO PINHEIR 00:00:14,413 | >> CHICO PINHEIRO: 00:00:14,447 | >> CHICO PINHEIRO: B 00:00:14,480 | >> CHICO PINHEIRO: BOM 00:00:14,513 | >> CHICO PINHEIRO: BOM 00:00:14,547 | >> CHICO PINHEIRO: BOM DI 00:00:14,613 | >> CHICO PINHEIRO: BOM DIA,

A. Acurácia

Nesta etapa de teste, foi realizada a transcrição manual do áudio dos programas, para ser comparada com o texto da Estenotipia. Em ambos os textos, todos os caracteres foram substituídos por letras minúsculas, todas as abreviações e números foram reescritos por extenso e foram removidos todos os sinais de pontuação e caracteres especiais. O texto da Estenotipia continha, em alguns casos, a identificação da pessoa cuja fala estava sendo transcrita (como no exemplo da Tabela 2, com o nome do apresentador do telejornal). Para possibilitar a comparação dos textos da Estenotipia com a transcrição manual das falas, tais identificações foram manualmente removidas. Embora a Estenotipia permita correção do *closed caption* em tempo real (apagando caracteres), a inserção e apagamento de caracteres podem ser observados apenas no texto com marcação de tempo. No texto sem marcação de tempo, que foi utilizado na avaliação da acurácia, aparecem apenas as linhas finalizadas após eventuais correções.

Os resultados de acurácia obtidos são apresentados na Tabela 3. Observe que as acurácias obtidas para o programa jornalístico sobre saúde e para o programa de auditório são significativamente inferiores àquela obtida no telejornal, um fenômeno que retrata a dificuldade da transcrição em situações onde os oradores possuem falas mais espontâneas.

TABELA 3
ACURÁCIA DA ESTENOTIPIA

Telejornal (Bom Dia Brasil)			
Duração:	52m50s		
Número de Palavras:	7903		
Acertos:	6710 (84,90%)		
Erros:	1420 (17,97%)	*Substituições:*	550 (6,96%)
		Omissões:	643 (8,14%)
		Inserções:	227 (2,87%)
Acurácia:	82,03%		
Programa Jornalístico sobre Saúde (Bem Estar)			
Duração:	38m55s		
Número de Palavras:	6333		
Acertos:	4240 (66,95%)		
Erros:	2313 (36,52%)	*Substituições:*	844 (13,33%)
		Omissões:	1249 (19,72%)
		Inserções:	220 (3,47%)
Acurácia:	63,48%		
Programa de Auditório (Domingão do Faustão)			
Duração:	01h45m02s		
Número de Palavras:	14686		
Acertos:	8934 (60,83%)		
Erros:	5917 (40,29%)	*Substituições:*	1324 (9,02%)
		Omissões:	4428 (30,15%)
		Inserções:	165 (1,12%)
Acurácia:	59,71%		
TOTAL			
Duração:	03h16m47s		
Número de Palavras:	28922		
Acertos:	19884 (68,75%)		
Erros:	9650 (33,37%)	*Substituições:*	2718 (9,40%)
		Omissões:	6320 (21,85%)
		Inserções:	612 (2,12%)
Acurácia:	66,63%		

B. Latência

A latência da Estenotipia foi avaliada comparando-se os tempos de fim de três frases no áudio do programa com os tempos de exibição do último caractere de cada uma dessas frases no *closed caption*. Desta forma, verificou-se a latência média de 4,034 segundos.

V. AVALIAÇÃO DA RELOCUÇÃO

A relocução do material de teste foi gravada em ambiente residencial, sem controle acústico e sem um sistema de áudio profissional, pelo próprio autor deste trabalho.

A. Acurácia

Neste teste foi realizada uma transcrição manual do áudio da relocução, para ser comparada com a transcrição manual do áudio original. Todos os caracteres foram substituídos por letras minúsculas, todas as abreviações e números foram

Avaliação do Desempenho de um Sistema de Reconhecimento Automático de Voz em Português do Brasil Baseado em Software Livre para Geração de Closed Caption
Luiz Fausto de Souza Brito, Flávio Luis de Mello

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

reescritos por extenso e foram removidos todos os sinais de pontuação e caracteres especiais. Os resultados de acurácia obtidos podem ser observados na Tabela 4.

TABELA 4
ACURÁCIA DA RELOCUÇÃO

Telejornal (Bom Dia Brasil)			
Duração:	52m50s		
Número de Palavras:	7903		
Acertos:	7833 (99,11%)		
Erros:	87 (1,10%)	*Substituições:*	38 (0,48%)
		Omissões:	32 (0,40%)
		Inserções:	17 (0,22%)
Acurácia:	98,90%		
Programa Jornalístico sobre Saúde (Bem Estar)			
Duração:	38m55s		
Número de Palavras:	6333		
Acertos:	6043 (95,42%)		
Erros:	317 (5,01%)	*Substituições:*	210 (3,32%)
		Omissões:	80 (1,26%)
		Inserções:	27 (0,43%)
Acurácia:	94,99%		
Programa de Auditório (Domingão do Faustão)			
Duração:	01h45m02s		
Número de Palavras:	14686		
Acertos:	13906 (94,69%)		
Erros:	1058 (7,20%)	*Substituições:*	461 (3,14%)
		Omissões:	319 (2,17%)
		Inserções:	278 (1,89%)
Acurácia:	92,80%		
TOTAL			
Duração:	03h16m47s		
Número de Palavras:	28922		
Acertos:	27782 (96,06%)		
Erros:	1462 (5,05%)	*Substituições:*	709 (2,45%)
		Omissões:	431 (1,49%)
		Inserções:	322 (1,11%)
Acurácia:	94,95%		

B. *Latência*

A latência da relocução foi avaliada comparando-se os tempos de fim de três frases no áudio do programa e no áudio da relocução. Desta forma, verificou-se a latência média de 1,059 segundo.

VI. AVALIAÇÃO DO IBM VIAVOICE

O IBM ViaVoice já vem treinado para o Português do Brasil, com um dicionário, um modelo de linguagem e um modelo acústico independente de locutor previamente definidos. Apenas os seguintes ajustes são possíveis no IBM ViaVoice: (1) adaptação do dicionário e do modelo de linguagem; (2) adaptação do modelo acústico (adaptação de locutor); (3) ajuste do desempenho de reconhecimento (rápido, equilibrado, preciso ou automático). Nestes testes foi utilizado o ajuste de desempenho de reconhecimento preciso.

A. *Adaptação do Dicionário e do Modelo de Linguagem*

A adaptação do dicionário e do modelo de linguagem do IBM ViaVoice requer um ou mais textos, dos quais são extraídos o vocabulário e a probabilidade de sequências de palavras.

O texto utilizado nessa adaptação foi extraído das páginas de Internet hospedadas no domínio globo.com em 17 de novembro de 2012. Em seguida, o texto passou por um processo de revisão manual, para remoção dos erros eventualmente presentes (sobretudo provenientes de comentários publicados por usuários nos *sites*). O texto resultante possuía 100.954 sentenças e 1.707.869 palavras, com vocabulário de 53.633 palavras distintas.

B. *Adaptação do Modelo Acústico*

Para a adaptação do modelo acústico (adaptação de locutor), o IBM ViaVoice utiliza um conjunto de textos padrão que vêm com o aplicativo, composto por 1.027 sentenças e 6.622 palavras, com vocabulário de 1.953 palavras distintas.

C. *Acurácia*

A transcrição da relocução realizada pelo IBM ViaVoice foi comparada com a transcrição manual da relocução. Os resultados de acurácia obtidos são apresentados na Tabela 5.

TABELA 5
ACURÁCIA DO IBM VIAVOICE

Telejornal (Bom Dia Brasil)			
Duração:	52m50s		
Número de Palavras:	7883		
Acertos:	6615 (83,91%)		
Erros:	1650 (20,93%)	*Substituições:*	1011 (12,82%)
		Omissões:	357 (4,53%)
		Inserções:	282 (3,58%)
Acurácia:	79,07%		
Programa Jornalístico sobre Saúde (Bem Estar)			
Duração:	38m55s		
Número de Palavras:	6278		
Acertos:	4488 (71,49%)		
Erros:	2020 (32,18%)	*Substituições:*	1419 (22,60%)
		Omissões:	371 (5,91%)
		Inserções:	230 (3,67%)
Acurácia:	67,82%		
Programa de Auditório (Domingão do Faustão)			
Duração:	01h45m02s		
Número de Palavras:	14641		
Acertos:	9497 (64,87%)		
Erros:	5617 (38,36%)	*Substituições:*	4006 (27,36%)
		Omissões:	1138 (7,77%)
		Inserções:	473 (3,23%)
Acurácia:	61,64%		
TOTAL			
Duração:	03h16m47s		
Número de Palavras:	28802		
Acertos:	20600 (71,52%)		
Erros:	9287 (32,24%)	*Substituições:*	6436 (22,34%)
		Omissões:	1866 (6,48%)
		Inserções:	985 (3,42%)
Acurácia:	67,76%		

D. *Latência*

A latência foi avaliada comparando os tempos de fim de três frases no áudio da relocução e na transcrição do IBM ViaVoice. Verificou-se a latência média de 2,098 segundos.

E. *Consumo de recursos computacionais*

O IBM ViaVoice, durante a operação de reconhecimento, ocupou menos de 20% de um único núcleo da CPU do computador (Intel Core i5 de 2,4 GHz) e menos de 70 MB de RAM.

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

VII. AVALIAÇÃO DE UM SISTEMA DE RAV BASEADO EM *SOFTWARE* LIVRE

Originalmente, o CMU Sphinx não possui nenhum dicionário, modelo de linguagem ou modelo acústico para o Português do Brasil. No presente teste, esses recursos foram desenvolvidos utilizando apenas bases de dados disponíveis publicamente.

A. Criação do Dicionário e do Modelo de Linguagem

A criação do dicionário e do modelo de linguagem é feita a partir de um *corpus* de texto. Foi gerado um *corpus* de texto de uso geral e um *corpus* de texto específico para cada programa utilizado no teste (Bom Dia Brasil, Bem Estar e Domingão do Faustão).

O *corpus* de texto de uso geral foi gerado pela concatenação do TextCorpora1.5 desenvolvido pelo grupo FalaBrasil da UFPA [12], com os textos extraídos do domínio globo.com para adaptação do dicionário e do modelo de linguagem do IBM ViaVoice. Todos os caracteres foram substituídos por letras minúsculas, todas as abreviações e números foram reescritos por extenso e foram removidos todos os sinais de pontuação e caracteres especiais. O *corpus* de texto foi organizado com uma sentença por linha e com marcações especiais de início (<s>) e fim (</s>) em cada sentença. Por fim, o *corpus* de texto resultante possuía 1.593.389 sentenças e 24.746.658 palavras, com vocabulário de 210.446 palavras distintas. Um trecho do *corpus* de texto pode ser visto na Fig. 2.

<s> o fim do mundo em dois mil e doze é um dos temas de maior sucesso na internet </s>
<s> uma infinidade de vídeos fotos textos e teorias defende que nós não vamos passar dessa data </s>
<s> o dia teria sido previsto pelos maias antigo povo do méxico e américa central que tinha grande conhecimento astronômico </s>
<s> vários fatores se apresentam como candidatos a acabar com nosso planeta o inimigo mais assustador e mais comentado pelos teóricos é chamado de nibiru </s>
<s> um planeta desconhecido do sistema solar que estaria vindo em direção à terra </s>
<s> e ao passar perto ou se chocar com a gente ocorreria uma explosão </s>

Fig. 2. Trecho do *corpus* de texto

O *corpus* de texto específico para cada programa utilizou, além dos textos utilizados no *corpus* de texto de uso geral, 50% das sentenças da transcrição manual da relocução do programa, selecionadas ao acaso. A transcrição manual foi segmentada também manualmente em sentenças correspondentes à segmentação do áudio da relocução de forma automática pela detecção de silêncios.

A Perplexidade é um conceito da Teoria da Informação relacionado à entropia, podendo ser definido por: $P = 2^E$, onde P é a perplexidade e E é a entropia. A perplexidade de um modelo de linguagem pode ser interpretada como o grau de dificuldade de prever a próxima palavra em um texto utilizando o referido modelo. Por exemplo, um valor de perplexidade de 200 corresponde à mesma dificuldade de prever uma palavra dentre uma lista de 200 palavras equiprováveis. Portanto, quanto menor o valor da perplexidade do modelo de linguagem mais ele contribui para a acurácia do sistema de Reconhecimento Automático de Voz [11].

O texto selecionado do telejornal possuía 82 sentenças e 4.011 palavras, com vocabulário de 1.333 palavras distintas. O texto selecionado do programa jornalístico sobre saúde possuía 76 sentenças e 3.056 palavras, com vocabulário de 939 palavras distintas. O texto selecionado do programa de auditório possuía 341 sentenças e 7.690 palavras, com vocabulário de 1.569 palavras distintas. Como o texto da transcrição manual da relocução dos programas era muito menor que o texto utilizado no *corpus* de texto de uso geral, ele poderia não modificar de forma muito significativa a probabilidade de sequências de palavras. Por isso, essas sentenças foram acrescentadas repetidamente ao *corpus* de texto de uso geral, até que não houvesse redução na perplexidade do modelo de linguagem resultante.

A variação observada da perplexidade do modelo de linguagem do programa de auditório com o número de repetições da transcrição de parte do programa no *corpus* de texto está ilustrado na Fig. 3.

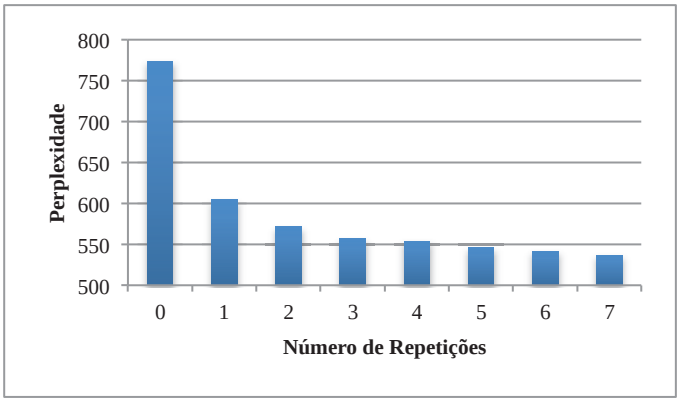

Fig. 3. Variação da perplexidade do modelo de linguagem do programa Domingão do Faustão com o número de repetições da transcrição de parte do programa no corpus de texto

De cada *corpus* de texto produzido, foram extraídas as 65535 palavras mais frequentes, para que o vocabulário fosse compatível com a utilização de índices de 16 bits do CMU Sphinx. Os modelos de linguagem foram gerados a partir de cada vocabulário, de cada *corpus* de texto e da lista de símbolos de contexto (<s> e </s>). Foram gerados modelos de linguagem trigrama (em que a probabilidade de uma palavra é condicionada às duas palavras anteriores), com vocabulário aberto (em que é atribuída uma probabilidade à ocorrência de palavras fora do dicionário).

Cada vocabulário gerado nas etapas anteriores (excluindo o cabeçalho e as entradas <s> e </s>) foi transformado em dicionário com transcrição fonética utilizando um *software* livre de conversão grafema-fonema do grupo FalaBrasil da UFPA [12] que utiliza um conjunto de 38 fonemas. Os fonemas utilizados por esse *software* são representados por combinações de letras maiúsculas, minúsculas e símbolos.

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

Essas representações precisaram ser alteradas para garantir a compatibilidade com o CMU Sphinx, que não diferencia letras maiúsculas e minúsculas e não aceita símbolos. Um dicionário separado foi construído com as entradas <s>, </s> e <sil>, mapeando todas para o mesmo símbolo fonético (SIL), correspondente ao silêncio. Foi gerada ainda uma lista dos símbolos fonéticos usados nos dicionários (nesse caso, 38 fonemas e um silêncio). A lista dos símbolos fonéticos utilizados pode ser vista na Fig. 4.

SIL	v	ww	ii	oo	sm	u	k
om	lm	z	e	aa	n	d	b
em	g	dz	t	uu	j	f	s
jm	jj	ee	rm	r	xm	i	a
p	zm	m	w	l	o	ts	

Fig. 4. Lista dos símbolos fonéticos utilizados

Um trecho ilustrativo do dicionário pode ser visto na Fig. 5.

diferencia	dz i f e r ee s i a
diferenciada	dz i f e r ee s i a d a
diferenciadas	dz i f e r ee s i a d a s
diferenciado	dz i f e r ee s i a d u
diferenciados	dz i f e r ee s i a d u s
diferenciais	dz i f e r ee s i a j s
diferencial	dz i f e r ee s i a w
diferenciam	dz i f e r ee s i aa ww
diferenciar	dz i f e r ee s i a xm
diferenciação	dz i f e r ee s i a s aa ww
diferencie	dz i f e r ee s i
diferenciou	dz i f e r ee s i o w

Fig. 5. Trecho do dicionário

B. Taxa de palavras fora do dicionário e perplexidade do modelo de linguagem

Os dicionários e os modelos de linguagem gerados foram avaliados, respectivamente quanto à taxa de palavras fora do dicionário e quanto à perplexidade, em relação à transcrição manual da relocução dos programas utilizados no teste. É importante ressaltar que foram consideradas apenas as sentenças não utilizadas nos *corpora* de texto específicos de cada programa.

Os resultados obtidos podem ser observados na Tabela 6. Para cada programa obteve-se a taxa de palavras fora do dicionário, considerando um dicionário de uso de geral e um dicionário específico de cada programa, bem como a perplexidade do modelo de linguagem, considerando um modelo de linguagem de uso geral e um modelo de linguagem específico de cada programa.

TABELA 6
TAXA DE PALAVRAS FORA DO DICIONÁRIO E PERPLEXIDADE DO MODELO DE LINGUAGEM

	Taxa de palavras fora do dicionário		**Perplexidade**	
	Dicionário de uso geral	**Dicionário específico do programa**	**Modelo de linguagem de uso geral**	**Modelo de linguagem específico do programa**
Bom Dia Brasil	0,57%	0,47%	266,22	246,74
Bem Estar	1,34%	1,27%	666,42	589,79
Domingão do Faustão	2,30%	1,37%	773,50	536,94

C. Treinamento do Modelo Acústico

Para treinamento do modelo acústico é necessário possuir um *corpus* de voz, que deve consistir em um conjunto de arquivos de áudio contendo a gravação da fala segmentada em trechos de curta duração (idealmente de 5 a 30 segundos) e uma transcrição textual dessas gravações.

Foram treinados três modelos acústicos: um utilizando apenas uma voz masculina, um utilizando diversas vozes masculinas e um utilizando vozes masculinas e femininas.

Para o primeiro modelo (treinado com *corpus* mais reduzido, apenas com uma voz masculina), foi utilizado apenas o *corpus* Constituição1.0 do grupo FalaBrasil da UFPA [12], com 1.238 sentenças, 68.575 palavras, vocabulário de 5.305 palavras distintas, 8 horas, 50 minutos e 12 segundos de gravação de um único locutor do sexo masculino, lendo textos da Constituição Federal em um ambiente controlado (estúdio).

Para o segundo modelo (treinado com *corpus* intermediário, com vozes masculinas apenas), foram incluídas mais gravações de vozes masculinas, provenientes do *corpus* LapsBenchMark1.4 do grupo FalaBrasil da UFPA com 500 sentenças, 5.166 palavras, vocabulário de 2.102 palavras distintas, 38 minutos e 10 segundos de gravação em ambiente não controlado, da voz de 25 homens com aproximadamente a mesma duração. Além disto, foi acrescido o material do *site* VoxForge, com 1.828 sentenças, 9.173 palavras, vocabulário de 584 palavras distintas, 1 hora, 51 minutos e 24 segundos de gravação em ambiente não controlado, da voz de 78 homens com duração variável [13].

Para o terceiro modelo (treinado com *corpus* maior, com vozes masculinas e femininas), foram incluídas também as gravações de vozes femininas provenientes do *corpus* LapsBenchMark1.4 do grupo FalaBrasil da UFPA, com 200 sentenças, 2.062 palavras, vocabulário de 1.064 palavras distintas, 15 minutos e 51 segundos de gravação em ambiente não controlado, da voz de 10 mulheres com aproximadamente a mesma duração. Neste caso também foi acrescido o material do *site* VoxForge, com 180 sentenças, 855 palavras, vocabulário de 351 palavras distintas, 9 minutos e 30 segundos de gravação em ambiente não controlado, da voz de 6 mulheres com duração variável, removendo as gravações em Português de Portugal e as gravações que estavam ininteligíveis (nível de áudio excessivamente baixo, ruído ou distorção excessivamente altos).

A maioria das configurações do treinamento do modelo acústico foram mantidas conforme o padrão do CMU Sphinx, exceto os seguintes itens:

- **LDA/MLLT**

Por padrão, o processamento digital do sinal de voz do CMU Sphinx utiliza um vetor de parâmetros MFCC (*Mel-Frequency Cepstral Coefficients*), que utiliza os 12 primeiros coeficientes da DCT do logaritmo do espectro de potência na escala de frequência Mel (escala de frequência subjetivamente linear), mais um coeficiente que representa a energia média do sinal, além da primeira e da segunda derivada desses 13 coeficientes, denominados de coeficientes dinâmicos, “delta” ou vetores de velocidade e aceleração, que ajudam a caracterizar os efeitos

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

coarticulatórios, formando um vetor de parâmetros ou de características com 39 coeficientes [11] [14] [15] [16].

O vetor de parâmetros ou de características é utilizado no modelo acústico para reconhecimento dos padrões fonéticos. É possível otimizar esses parâmetros utilizando uma transformação linear que melhore a separabilidade entre os padrões a serem reconhecidos, o que produz um impacto positivo sobre a acurácia do sistema. Além disso, tal transformação descorrelaciona as dimensões do vetor de parâmetros e é possível reduzir a dimensão desse vetor (por exemplo, de 39 para 32) sem reduzir significativamente a acurácia, o que reduz o custo computacional do reconhecimento. Por isso, foi habilitada a criação de uma matriz de transformação do vetor de parâmetros utilizando duas transformações conhecidas que podem ser utilizadas em conjunto para aumentar a acurácia e reduzir o custo do reconhecimento: *Linear Discriminant Analysis* (LDA) e *Maximum Likelihood Linear Transform* (MLLT), com 32 dimensões [7] [11].

- **multithread**

A fim de reduzir o tempo necessário ao treinamento do modelo acústico, foi habilitado o processamento *multithread* (até 4 *threads* simultâneos, correspondendo ao limite do processador empregado).

- **forced alignment**

Foi habilitado também o alinhamento forçado (*forced alignment*), que não inclui no treinamento os arquivos de áudio que não puderem ser alinhados com as suas respectivas transcrições textuais.

O alinhamento das transcrições textuais com o áudio é realizado através de uma decodificação ou busca da sequência de palavras mais provável para a sequência de parâmetros extraída do sinal de voz. Idealmente, a busca deveria considerar todas as hipóteses possíveis para a sequência de estados. Como o cálculo de todos os caminhos possíveis dentro da árvore ou grafo do espaço de busca pode ser proibitivo pelo tamanho do vocabulário e pela complexidade dos modelos acústico e de linguagem, a busca pode ser otimizada computacionalmente se forem desconsiderados ("podados") os ramos mais improváveis. Note-se que há o risco de desconsiderar prematuramente um ramo pertencente ao caminho com maior probabilidade global, introduzindo erros de decodificação devido a essa "poda". Trata-se de uma decisão de compromisso entre custo computacional e acurácia. Há vários mecanismos de "poda" empregados na decodificação dos sistemas de RAV. O mais frequente é o controle de feixe (*beam*), em que são considerados apenas os ramos cuja probabilidade não caia abaixo de um limiar proporcional à probabilidade do ramo mais provável [11] [17].

O alinhamento forçado utilizou um controle de feixe (*beam*) com limiar muito baixo (10^{-100}), para evitar que um arquivo de áudio com transcrição correta pudesse ser indevidamente descartado.

- **modelo de conversão grafema-fonema**

Também foi habilitado o treinamento de um modelo de conversão grafema-fonema, que permite que sejam empregadas no treinamento do modelo acústico palavras que não constem no dicionário fonético utilizado (embora, para o treinamento de cada modelo acústico tenha sido utilizado um dicionário fonético específico com as 65.535 palavras mais frequentes do *corpus* de voz utilizado).

- **senones / gaussianas**

O modelo acústico do CMU Sphinx utiliza HMM (*Hidden Markov Model* – Modelo Oculto de Markov). Nesse modelo, assume-se que o processo segue uma sequência de estados que não são diretamente observáveis. A observação indireta, no caso, corresponde a um vetor de características extraído do sinal de voz. Os estados poderiam corresponder a fonemas mas, para levar em consideração as variações na realização de cada fonema devidas aos efeitos coarticulatórios, os fonemas são diferenciados pelo contexto de vizinhança fonética, utilizando "trifones" (um trifone representa um único fonema, dados o fonema anterior e o fonema posterior) e, para maior acurácia do sistema, divide-se cada trifone em três estados, denominados "senones", sendo o primeiro correspondente ao início do fonema, cuja observação é influenciada pelo efeito coarticulatório da transição a partir do fonema anterior, o segundo estado corresponde à parte intermediária e mais estável do fonema e o terceiro estado corresponde à parte final do fonema, cuja observação é influenciada pelo efeito coarticulatório da transição para o próximo fonema [11].

Um modelo HMM é caracterizado pelas probabilidades iniciais dos estados (representadas por um vetor), pelas probabilidades de transição entre os estados (representadas por uma matriz) e pelas probabilidades de observação. As probabilidades de observação são modeladas, normalmente, por uma mistura de gaussianas, pela flexibilidade que tal mistura oferece de aproximar qualquer distribuição de probabilidade. Uma mistura de gaussianas é representada pelo vetor das médias e matriz de covariância de cada componente e por um vetor com os pesos de cada componente [16].

Para reduzir a dimensionalidade do modelo pode-se associar as probabilidades dos estados (senones) semelhantes, tais como os estados intermediários dos trifones que representam o mesmo fonema e os estados inicial e final dos trifones que representam o mesmo fonema e que podem ser agrupadas por categorias fonéticas dos fonemas anteriores e posteriores, respectivamente [17].

Os dois parâmetros de configuração reconhecidamente mais críticos para a acurácia do sistema (devido ao impacto sobre a complexidade e a treinabilidade), cujos valores ótimos dependem do *corpus* de voz utilizado, foram ajustados de forma iterativa para cada modelo acústico: o número de senones (equivalente ao número de estados a serem treinados no Modelo Oculto de Markov, controlando o quanto os estados de trifones "semelhantes" são agrupados) e o número de gaussianas a serem treinadas para cada estado na modelagem das probabilidades de observação por mistura de gaussianas.

Avaliação do Desempenho de um Sistema de Reconhecimento Automático de Voz em Português do Brasil Baseado em Software Livre para Geração de Closed Caption
Luiz Fausto de Souza Brito, Flávio Luis de Mello

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

O modelo acústico inicial foi treinado com 250 senones e com número gaussianas inicial de 1 e final de 64 (gerando modelos com 1, 2, 4, 8, 16, 32 e 64 gaussianas). Nas etapas de adaptação de locutor e decodificação, detalhadas a seguir, foram utilizados os modelos a partir de 64 gaussianas, reduzindo o número de gaussianas até que não houvesse melhoria na acurácia do modelo adaptado. Em seguida, o modelos acústicos foram treinados novamente dobrando o número de senones e refazendo as etapas de adaptação de locutor e decodificação até que não houvesse melhoria na acurácia do modelo adaptado. Dessa forma, foram determinados o número ótimo (dentre os testados e quanto à acurácia) de senones e de gaussianas para cada modelo acústico. Por exemplo, para o modelo acústico treinado com apenas uma voz masculina, foi selecionada a configuração com 1000 senones e 16 gaussianas, como pode ser verificado pelos resultados de acurácia da Tabela 7.

TABELA 7
IMPACTO SOBRE A ACURÁCIA DO NÚMERO DE SENONES E GAUSSIANAS DO MODELO ACÚSTICO TREINADO COM UMA ÚNICA VOZ MASCULINA

		Senones			
		250	**500**	**1000**	**2000**
Gaussianas	**1**	-	-	-	69,50%
	2	-	-	-	69,60%
	4	-	-	-	69,51%
	8	-	-	69,52%	69,20%
	16	-	-	69,64%	69,15%
	32	66,94%	68,48%	69,43%	67,76%
	64	67,06%	68,67%	68,67%	62,52%

D. Adaptação do Modelo Acústico

Na adaptação do modelo acústico para a voz de um locutor específico (adaptação de locutor), é necessário um novo *corpus* de voz, com gravações da voz deste locutor. Para construção desse *corpus*, foram utilizados os mesmos arquivos de áudio gravados para a adaptação de locutor do IBM ViaVoice e os arquivos de áudio correspondentes às gravações das sentenças selecionadas para geração do *corpus* de texto específico de cada programa utilizado no teste, com 98 minutos e 23 segundos.

Na adaptação de locutor, duas técnicas são empregadas com frequência: MLLR (*Maximum Likelihood Linear Regression*) e MAP (*Maximum a Posteriori*). A técnica MLLR calcula matrizes de transformação dos vetores de médias das gaussianas do modelo acústico independente de locutor de forma a maximizar a probabilidade de observação dos vetores de parâmetros extraídos do *corpus* de adaptação. A técnica MAP ajusta todos os parâmetros do modelo acústico, interpolando o modelo original com um novo modelo treinado a partir do *corpus* de adaptação. O hiperparâmetro τ (tau) é usado para controlar o peso do modelo disponível *a priori*. A técnica MLLR possui custo computacional mais baixo e resulta em maior acurácia caso o *corpus* de adaptação seja muito pequeno. A técnica MAP possui custo computacional mais alto e resulta em maior acurácia caso o *corpus* de adaptação seja um pouco maior. A combinação das técnicas MLLR e MAP resulta na melhor acurácia para qualquer tamanho de *corpus* de adaptação [7] [11]. Assim, foi feita a adaptação de locutor através de MLLR e MAP, usando um valor fixo (100) para o hiperparâmetro τ (tau).

E. Acurácia

Para o teste de acurácia, foi utilizado o áudio de relocução das sentenças que não foram utilizadas na adaptação do modelo acústico. Para cada programa, foram empregados o dicionário e o modelo de linguagem específicos de cada programa gerados anteriormente. Foram usados os três modelos acústicos treinados e adaptados anteriormente.

Dada uma sequência de vetores de parâmetros X extraída de um sinal de voz através do processamento digital, a tarefa do Reconhecimento Automático de Voz é determinar a sequência correspondente de palavras W mais provável. Em termos probabilísticos, deseja-se descobrir a sequência de palavras W que maximiza a probabilidade condicional $P(W|X)$. Como essa probabilidade não é conhecida *a priori*, podemos utilizar o teorema de Bayes para inverter essa probabilidade condicional [15] [18]:

$$P(W|X) = P(X|W) \cdot P(W)/P(X) \quad (2)$$

Como a maximização de $P(W|X)$ é calculada com X fixo, isso equivale a maximizar:

$$P(X|W) \cdot P(W) \quad (3)$$

Isso permite a separação do problema do Reconhecimento Automático de Voz em dois problemas distintos: (1) um modelo acústico, $P(X|W)$ e; (2) um modelo de linguagem, $P(W)$.

O espaço de busca pode ser definido em um único modelo HMM integrando os modelos acústico, $P(X|W)$, e de linguagem, $P(W)$, em que, opcionalmente, pode-se atribuir empiricamente um peso diferenciado (LW) para o modelo de linguagem, $P(W)^{LW}$, para otimizar a acurácia [11].

O peso do modelo de linguagem (LW) foi ajustado de forma iterativa em cada programa, iniciando com LW=1 e somando 1 até que não houvesse melhoria na acurácia da decodificação. Um exemplo da variação de acurácia com o peso do modelo de linguagem, para o programa Bom Dia Brasil, utilizando o modelo acústico treinado com *corpus* maior, pode ser observado na Fig. 6.

Avaliação do Desempenho de um Sistema de Reconhecimento Automático de Voz em Português do Brasil Baseado em Software Livre para Geração de Closed Caption
Luiz Fausto de Souza Brito, Flávio Luis de Mello

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

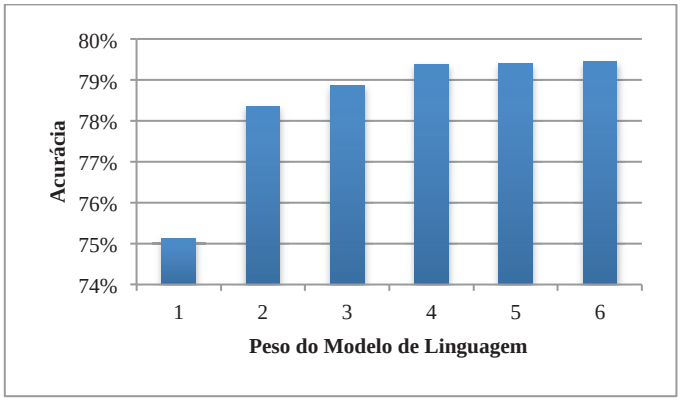

Fig. 6. Variação da acurácia com peso do modelo de linguagem, para o programa Bom Dia Brasil, utilizando o modelo acústico treinado com corpus maior

O modelo de linguagem também funciona como uma penalidade para inserir novas palavras durante a decodificação, isto é, com uma penalidade pequena o decodificador prefere utilizar mais palavras de menor duração e com uma penalidade alta o decodificador prefere utilizar menos palavras de maior duração. Ao modificar o peso do modelo de linguagem, essa penalidade é alterada. Uma penalidade para inserção de novas palavras (*IP*) pode ser incluída no modelo de linguagem, $P(W)^{LW} \cdot IP^{N(W)}$, onde $N(W)$ é o número de palavras, podendo também ser ajustada empiricamente para maximizar a acurácia [11]. Foi ajustada a mesma penalidade para inserção de novas palavras (*IP*) para todos os programas de forma iterativa, iniciando com *IP*=0,1 e somando 0,1 até que não houvesse melhoria na acurácia da decodificação, resultando no valor de *IP*=6,4.

A decodificação utilizou um controle de feixe (*beam*) com limiar muito baixo (10^{-100}), para evitar erros de decodificação devido à possível "poda" prematura de uma hipótese correta. Finalmente, a transcrição gerada foi comparada com a transcrição manual.

A execução de todos os processos de treinamento e teste do CMU Sphinx consumiu o equivalente a mais de 17 dias de processamento ininterrupto de um computador dedicado a essa atividade (utilizando um Apple Mac mini do final de 2012, com processador Intel Core i5 de 2,5 GHz, 16 GB de RAM e sistema operacional OS X 10.9.4), mesmo com a execução de *threads* paralelos quando possível.

Os resultados obtidos empregando o modelo acústico treinado com *corpus* mais reduzido, apenas com uma voz masculina, podem ser observados na Tabela 8.

TABELA 8
ACURÁCIA DO SISTEMA DE RAV BASEADO EM SOFTWARE LIVRE UTILIZANDO MODELO ACÚSTICO TREINADO COM *CORPUS* MAIS REDUZIDO, COM APENAS UMA VOZ MASCULINA

Telejornal (Bom Dia Brasil)			
Duração:	26m25s		
Número de Palavras:	3862		
Peso do Modelo de Linguagem:	4		
Acertos:	3129 (81,02%)		
Erros:	833 (21,57%)	*Substituições:*	544 (14,09%)
		Omissões:	189 (4,89%)
		Inserções:	100 (2,59%)
Acurácia:	78,43%		
Programa Jornalístico sobre Saúde (Bem Estar)			
Duração:	19m28s		
Número de Palavras:	3219		
Peso do Modelo de Linguagem:	5		
Acertos:	2218 (68,90%)		
Erros:	1077 (33,46%)	*Substituições:*	738 (22,93%)
		Omissões:	263 (8,17%)
		Inserções:	76 (2,36%)
Acurácia:	66,54%		
Programa de Auditório (Domingão do Faustão)			
Duração:	37m31s		
Número de Palavras:	6947		
Peso do Modelo de Linguagem:	4		
Acertos:	4785 (68,88%)		
Erros:	2349 (33,81%)	*Substituições:*	1586 (22,83%)
		Omissões:	576 (8,29%)
		Inserções:	187 (2,69%)
Acurácia:	66,19%		
TOTAL			
Duração:	01h38m24s		
Número de Palavras:	14028		
Senones:	1000		
Gaussianas:	16		
Acertos:	10132 (72,23%)		
Erros:	4259 (30,36%)	*Substituições:*	2868 (20,44%)
		Omissões:	1028 (7,33%)
		Inserções:	363 (2,59%)
Acurácia:	69,64%		

Os resultados obtidos com o modelo acústico treinado com *corpus* intermediário, com vozes masculinas apenas, podem ser observados na Tabela 9.

TABELA 9
ACURÁCIA DO SISTEMA DE RAV BASEADO EM SOFTWARE LIVRE UTILIZANDO MODELO ACÚSTICO TREINADO COM *CORPUS* INTERMEDIÁRIO, COM VOZES MASCULINAS APENAS

Telejornal (Bom Dia Brasil)			
Duração:	26m25s		
Número de Palavras:	3862		
Peso do Modelo de Linguagem:	3		
Acertos:	3184 (82,44%)		
Erros:	789 (20,43%)	*Substituições:*	511 (13,23%)
		Omissões:	167 (4,33%)
		Inserções:	111 (2,87%)
Acurácia:	79,57%		
Programa Jornalístico sobre Saúde (Bem Estar)			
Duração:	19m28s		
Número de Palavras:	3219		
Peso do Modelo de Linguagem:	3		
Acertos:	2251 (69,93%)		
Erros:	1072 (33,30%)	*Substituições:*	722 (22,43%)
		Omissões:	246 (7,64%)
		Inserções:	104 (3,23%)
Acurácia:	66,70%		
Programa de Auditório (Domingão do Faustão)			
Duração:	37m31s		
Número de Palavras:	6947		
Peso do Modelo de Linguagem:	4		
Acertos:	4915 (70,75%)		
Erros:	2225 (32,03%)	*Substituições:*	1458 (20,99%)
		Omissões:	574 (8,26%)
		Inserções:	193 (2,78%)
Acurácia:	67,97%		

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

TOTAL			
Duração:	01h38m24s		
Número de Palavras:	14028		
Senones:	2000		
Gaussianas:	4		
Acertos:	10350 (73,78%)		
Erros:	4086 (29,13%)	*Substituições:*	2691 (19,18%)
		Omissões:	987 (7,04%)
		Inserções:	408 (2,91%)
Acurácia:	70,87%		

Por fim, os resultados obtidos com o modelo acústico treinado com *corpus* maior, com vozes masculinas e femininas, podem ser observados na Tabela 10.

TABELA 10
ACURÁCIA DO SISTEMA DE RAV BASEADO EM SOFTWARE LIVRE UTILIZANDO MODELO ACÚSTICO TREINADO COM *CORPUS* MAIOR, COM VOZES MASCULINAS E FEMININAS

Telejornal (Bom Dia Brasil)			
Duração:	26m25s		
Número de Palavras:	3862		
Peso do Modelo de Linguagem:	6		
Acertos:	3160 (81,82%)		
Erros:	794 (20,56%)	*Substituições:*	515 (13,34%)
		Omissões:	187 (4,84%)
		Inserções:	92 (2,38%)
Acurácia:	79,44%		
Programa Jornalístico sobre Saúde (Bem Estar)			
Duração:	19m28s		
Número de Palavras:	3219		
Peso do Modelo de Linguagem:	3		
Acertos:	2292 (71,20%)		
Erros:	1031 (32,03%)	*Substituições:*	723 (22,46%)
		Omissões:	204 (6,34%)
		Inserções:	104 (3,23%)
Acurácia:	67,97%		

Programa de Auditório (Domingão do Faustão)			
Duração:	37m31s		
Número de Palavras:	6947		
Peso do Modelo de Linguagem:	4		
Acertos:	4915 (70,75%)		
Erros:	2219 (31,94%)	*Substituições:*	1482 (21,33%)
		Omissões:	550 (7,92%)
		Inserções:	187 (2,69%)
Acurácia:	68,06%		
TOTAL			
Duração:	01h38m24s		
Número de Palavras:	14028		
Senones:	2000		
Gaussianas:	16		
Acertos:	10367 (73,90%)		
Erros:	4044 (28,83%)	*Substituições:*	2720 (19,39%)
		Omissões:	941 (6,71%)
		Inserções:	383 (2,73%)
Acurácia:	71,17%		

F. Latência

A latência foi avaliada comparando os tempos de fim de três frases no áudio da relocução e na transcrição do CMU Sphinx. Nos testes, verificou-se a latência média de 652 ms.

G. Consumo de recursos computacionais

O CMU Sphinx, durante a operação de reconhecimento, ocupou cerca de 20% de todos os núcleos da CPU do computador (Intel Core i5 de 2,4 GHz) e cerca de 300 MB de RAM.

VIII. ANÁLISE DOS RESULTADOS

Um resumo dos resultados de acurácia obtidos nos testes está representado na Tabela 11.

TABELA 11
RESUMO DOS RESULTADOS DE ACURÁCIA OBTIDOS NOS TESTES

		Estenotipia		**Relocução**		**IBM ViaVoice**		**CMU Sphinx c/ *Corpus* Reduzido**		**CMU Sphinx c/ *Corpus* Intermediário**		**CMU Sphinx c/ *Corpus* Maior**	
Telejornal (Bom Dia Brasil)	*Acertos:*	84,90%		99,11%		83,91%		81,02%		82,44%		81,82%	
	Erros: *Substituições:*	17,97%	6,96%	1,10%	0,48%	20,93%	12,82%	21,57%	14,09%	20,43%	13,23%	20,56%	13,34%
	Erros: *Omissões:*		8,14%		0,40%		4,53%		4,89%		4,33%		4,84%
	Erros: *Inserções:*		2,87%		0,22%		3,58%		2,59%		2,87%		2,38%
	Acurácia:	82,03%		98,90%		79,07%		78,43%		79,57%		79,44%	
Programa Jornalístico sobre Saúde (Bem Estar)	*Acertos:*	66,95%		95,42%		71,49%		68,90%		69,93%		71,20%	
	Erros: *Substituições:*	36,52%	13,33%	5,01%	3,32%	32,18%	22,60%	33,46%	22,93%	33,30%	22,43%	32,03%	22,46%
	Erros: *Omissões:*		19,72%		1,26%		5,91%		8,17%		7,64%		6,34%
	Erros: *Inserções:*		3,47%		0,43%		3,67%		2,36%		3,23%		3,23%
	Acurácia:	63,48%		94,99%		67,82%		66,54%		66,70%		67,97%	
Programa de Auditório (Domingão do Faustão)	*Acertos:*	60,83%		94,69%		64,87%		68,88%		70,75%		70,75%	
	Erros: *Substituições:*	40,29%	9,02%	7,20%	3,14%	38,36%	27,36%	33,81%	22,83%	32,03%	20,99%	31,94%	21,33%
	Erros: *Omissões:*		30,15%		2,17%		7,77%		8,29%		8,26%		7,92%
	Erros: *Inserções:*		1,12%		1,89%		3,23%		2,69%		2,78%		2,69%
	Acurácia:	59,71%		92,80%		61,64%		66,19%		67,97%		68,06%	
TOTAL	*Acertos:*	68,75%		96,06%		71,52%		72,23%		73,78%		73,90%	
	Erros: *Substituições:*	33,37%	9,40%	5,05%	2,45%	32,24%	22,34%	30,36%	20,44%	29,13%	19,18%	28,83%	19,39%
	Erros: *Omissões:*		21,85%		1,49%		6,48%		7,33%		7,04%		6,71%
	Erros: *Inserções:*		2,12%		1,11%		3,42%		2,59%		2,91%		2,73%
	Acurácia:	66,63%		94,95%		67,76%		69,64%		70,87%		71,17%	

Na avaliação da Estenotipia, observa-se que o erro mais frequente é a omissão de palavras. Provavelmente, algumas palavras são suprimidas porque o estenotipista não consegue digitar rápido o suficiente o texto a ser transcrito. Entretanto, não foi feita nenhuma análise qualitativa sobre essas omissões

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

ou sobre os demais erros, quanto à perda de informação útil ou interferência na compreensão do texto.

Diferentemente da Estenotipia, a relocução utilizada na presente avaliação não foi realizada por um serviço profissional especializado e também não foi utilizado um estúdio de gravação adequado. Trata-se de uma gravação realizada pelo próprio autor deste trabalho. Portanto, os resultados obtidos a partir dessa relocução, incluindo os resultados de Reconhecimento Automático de Voz, provavelmente poderiam ser melhores se meios mais adequados fossem empregados. A relocução apresentou uma quantidade relativamente pequena de erros, que é somada aos erros dos sistemas de Reconhecimento Automático de Voz na aplicação proposta para esses sistemas.

Todos os sistemas de Reconhecimento Automático de Voz testados apresentaram acurácia superior à Estenotipia. Nota-se que o tipo de erro mais frequente nesses sistemas é a substituição. Na hipótese de se considerar que a substituição tem efeito mais negativo para a compreensão do texto do que a omissão de palavras, é possível modificar o sistema de RAV para exibir apenas as palavras reconhecidas com valor de confiança acima de um determinado limiar, o que reduziria o número de substituições e aumentaria o número de omissões.

Os testes utilizando o *software* livre CMU Sphinx apresentaram acurácia superior ao IBM ViaVoice, apesar de os *corpora* de voz utilizados no treinamento dos modelos acústicos serem muito pequenos (com durações totais variando de 8h50m a 11h45m) quando comparados com os modelos preconizados na literatura. Desta forma, vale notar que a documentação do sistema [7] recomenda que o treinamento do modelo acústico para a aplicação em telejornais utilize pelo menos 200 horas de áudio com a voz de, pelo menos 200 pessoas. Para o reconhecimento de conversação espontânea recomenda-se a utilização de 2000 horas de áudio. E trata-se de uma recomendação para a língua inglesa. A língua portuguesa, por ser muito flexiva e ter uma gramática relativamente livre em relação à ordem das palavras, requer dicionários ainda maiores e modelos de linguagem ainda mais complexos. Para obter a mesma acurácia que se obtém na língua inglesa, é necessário compensar aumentando a precisão do modelo acústico, o que requer um *corpus* de voz ainda maior. Neste sentido, os resultados obtidos podem ser considerados ainda melhores porque há espaço para incremento da acurácia, bastando que se proceda a construção de um *corpus* de voz maior e mais diversificado para treinamento do modelo acústico.

Os modelos acústicos treinados com *corpora* de voz um pouco maiores apresentaram acurácia maior. Aumentar o *corpus* de voz significativamente seria, portanto, muito benéfico para a acurácia do sistema. O *corpus* de voz maior (11h45m), mesmo sendo misto (com vozes masculinas e femininas) resultou em uma acurácia maior que o *corpus* de voz apenas masculino com duração um pouco menor (11h20m) no reconhecimento de uma voz masculina. Nesse caso, não houve vantagem em segmentar o *corpus* de voz por sexo, o que talvez seja vantajoso, como sugerido por diversos autores, apenas em um *corpus* de maior duração.

A adaptação de locutor, em uma aplicação real, poderia contar com um *corpus* cada vez maior, utilizando o reconhecimento revisado da relocução dos programas anteriores.

A utilização de dicionários e modelos de linguagem específicos de cada programa efetivamente reduziu a taxa de palavras fora do dicionário e a perplexidade, mesmo empregando apenas metade de um programa no treinamento. Em uma aplicação real também seria possível ter acesso a diversos programas anteriores e, talvez também a algum rascunho de roteiro do programa do dia, o que contribuiria para um incremento nos resultados.

Observa-se que, tanto na Estenotipia quanto na relocução e Reconhecimento Automático de Voz, o telejornal apresenta muito menos erros que o programa sobre saúde e que o programa de auditório. Dois motivos podem ser considerados: o vocabulário (especializado no programa sobre saúde, gírias no programa de auditório) e o estilo mais informal dos programas não jornalísticos, com mais diálogos espontâneos.

Em todas as alternativas de geração de *closed caption* ao vivo testadas, a taxa de acerto é muito inferior à mínima especificada na ABNT NBR 15290 [5] (98%). Pode-se considerar, portanto, que o desempenho especificado não é alcançável no atual estado da técnica.

Quanto à latência, a Estenotipia avaliada encontra-se no limite da norma (aproximadamente 4 segundos). A latência de aproximadamente um segundo observada na relocução é consistente com a relatada por Boulianne et al. [19] no Canadá. Para o IBM ViaVoice foi observada uma latência média de aproximadamente dois segundos, enquanto para o CMU Sphinx foi observada uma latência média de aproximadamente 650 ms. Em ambos os casos, essa medida de latência não inclui a latência da relocução, da codificação e da multiplexação do *closed caption*, processos que seriam necessários em uma aplicação real.

Quanto ao consumo de recursos computacionais, embora o CMU Sphinx consuma mais recursos que o IBM ViaVoice para o reconhecimento de voz, ambos podem ser executados facilmente por um computador de uso doméstico atual. Quanto ao treinamento do sistema de RAV, porém, pode ser necessário estudar otimizações para viabilizar uma atualização frequente, dado que no presente teste esse treinamento consumiu o equivalente a mais de 17 dias de processamento ininterrupto de um computador dedicado a essa atividade. Aparentemente, esta questão pode ser resolvida por um processo contínuo de treinamento, onde sistemas de RAV utilizam modelos acústicos construídos a partir do último treinamento disponível enquanto um novo treino é computado.

IX. Conclusão e Trabalhos Futuros

Observa-se claramente uma demanda por uma solução alternativa à Estenotipia na geração de legenda oculta de programas de televisão com fala espontânea ao vivo, tanto para reduzir os custos, como para conseguir a escala necessária para atender a todos os programas televisivos produzidos ao vivo no país, dada a escassez de profissionais de Estenotipia. Tal demanda é especialmente crítica nas

Avaliação do Desempenho de um Sistema de Reconhecimento Automático de Voz em Português do Brasil Baseado em Software Livre para Geração de Closed Caption
Luiz Fausto de Souza Brito, Flávio Luis de Mello

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

afiliadas das redes de televisão localizadas em cidades de pequeno e médio porte.

Além disso, há também o desejo de se obter maior qualidade, isto é, maior acurácia e menor latência. Como o limitador da qualidade da Estenotipia é o fator humano, é difícil acreditar que possa haver uma melhoria significativa na qualidade desse serviço no futuro próximo. O Reconhecimento Automático de Voz, que tem sua qualidade limitada por um sistema computacional, teria a oportunidade de melhoria mais incremental. Além disso, apresenta custo operacional baixo e é facilmente escalável.

Tanto o sistema comercial, já obsoleto, de RAV quanto o sistema baseado em *software* livre utilizados nesse teste apresentaram acurácia mais alta e latência mais baixa que a Estenotipia. O melhor desempenho foi do sistema baseado em *software* livre, que ainda apresenta muitas possibilidades de melhorias futuras.

Em todo caso, nem a Estenotipia, nem o Reconhecimento Automático de Voz puderam se aproximar da taxa de acerto mínima especificada pela NBR 15290 (98%), o que indica que se trata de um critério atualmente inalcançável, o que deveria ser reavaliado através de testes em sistemas reais de geração de legenda oculta ao vivo.

Para melhorar ainda mais o desempenho do sistema de RAV baseado em *software* livre para a geração de legenda oculta ao vivo, várias ações são possíveis, dentre as quais algumas ideias são destacadas a seguir.

O desenvolvimento que provavelmente seria mais significativo seria o de um *corpus* de voz suficientemente grande e diversificado. Isso pode ser feito a partir de outros *corpora* disponíveis publicamente, completando as transcrições e segmentações faltantes, e/ou utilizando *corpora* comerciais e/ou investindo na gravação de um novo *corpus*. Utilizar vozes com sotaque semelhante ao da localidade de aplicação do sistema também pode ser benéfico, embora a adaptação de locutor possa reduzir o problema da eventual diferença de sotaque das vozes utilizadas no treinamento do modelo acústico independente de locutor. Segmentar o *corpus* de voz em masculino e feminino pode ser útil se o *corpus*, para cada gênero, for suficientemente grande e diversificado.

Quanto à elaboração do *corpus* de texto é preciso desenvolver mecanismos completamente automáticos para extrair textos da *Internet*, tratando problemas de codificação de texto, erros de digitação, expansão de abreviações e números. Aparentemente este é um problema que pode ser resolvido com facilidade.

Quanto ao dicionário fonético, pode ser benéfico utilizar símbolos fonéticos diferentes para as vogais tônicas. Dicionários fonéticos diferenciados por sotaque também podem ser úteis, porém esta afirmação carece de um estudo mais específico.

Como o processo de treinamento do sistema de RAV utilizado nesse trabalho consumiu o equivalente a mais de 17 dias de processamento ininterrupto de um computador dedicado a essa atividade, seria necessário estudar otimizações para viabilizar uma atualização frequente (e.g. diária) dos modelos acústicos e de linguagem. Provavelmente, uma abordagem promissora é utilizar versionamento do treinamento enquanto um novo treino é computado.

Também seria interessante avaliar os vários *softwares* livres disponíveis para essa aplicação, bem como a possibilidade de customização deles (uma vez que possuem código-fonte aberto) ou mesmo de desenvolvimento de um novo sistema.

Para que o sistema alcançasse a maturidade necessária para uma aplicação real, seria necessária uma integração da solução completa em uma ferramenta simples de usar. Esta integração inclui a atualização automática (e.g. diária) dos dicionários e dos modelos de linguagem específicos de cada programa, do modelo acústico independente de locutor e da adaptação de locutor, a interface operacional para inserir pontuação e outros símbolos (e.g. nome do repórter ou apresentador) durante a relocução e para comutar para fontes de texto externas (e.g. *teleprompt*), as ferramentas de correção da transcrição automática em tempo real e/ou após a finalização do programa, a geração de relatórios de desempenho (acurácia e latência), a integração com ferramentas de codificação e multiplexação de *closed caption* no sinal a ser transmitido.

Por fim, seria interessante fazer uma avaliação qualitativa do desempenho obtido com o Reconhecimento Automático de Voz, preferencialmente com a participação de deficientes auditivos, para efetuar uma avaliação mais criteriosa sobre a identificação do conteúdo informativo das transcrições obtidas.

Referências

[1] L. F. S. Brito, E. Strauss e F. L. Mello, “Uso de Reconhecimento Automático de Voz em Português do Brasil na geração de Closed Caption,” *Revista de Radiodifusão - SET*, vol. 6, pp. 54-60, 2012.

[2] IBM, “IBM Desktop ViaVoice,” 2008. [Online]. Available: http:// www-01.ibm.com/ software/ pervasive/ viavoice.html. [Acesso em 13 05 2012].

[3] Voice Interaction, “Legendagem Automática - Audimus.Media,” 2014. [Online]. Available: http:// www.voiceinteraction.com.br/ ?page_id=376. [Acesso em 03 09 2014].

[4] Voice Interaction, “Sistema de ditado para MS Windows - VoxControl,” 2014. [Online]. Available: http:// www.voiceinteraction.com.br/ ?page_id=1086. [Acesso em 03 09 2014].

[5] Associação Brasileira de Normas Técnicas, “NBR 15290: Acessibilidade em comunicação na televisão,” Rio de Janeiro, 2005.

[6] L. F. S. Brito, “Sistema de Decisão Automático para Conversão de Áudio em Texto na Geração de Legenda Oculta,” Dissertação (mestrado profissional) - Universidade Estadual do Ceará, Centro de Ciências e Tecnologia, Mestrado Profissional em Computação Aplicada, Rio de Janeiro, 2015.

[7] Carnegie Mellon University, “CMU Sphinx,” 2012. [Online]. Available: http:// cmusphinx.sourceforge.net. [Acesso em 13 12 2012].

[8] I. Ahmer, “Automatic Speech Recognition for Closed Captioning of Television: Data and Issues,” Thesis (Master of Engineering) - University of South Australia, Adelaide, 2002.

[9] International Telecommunication Union, “Report ITU-R BT.2207-1: Accessibility to broadcasting services for persons with disabilities,” Geneva, 2011.

[10] C. F. Sanz e V. Quetschke, “CCExtractor v.0.64,” 2012. [Online]. Available: http:// ccextractor.sourceforge.net. [Acesso em 19 11 2012].

[11] X. Huang, A. Acero e H. W. Hon, Spoken Language Processing: A Guide to Theory, Algorithm, and System Development, Upper Saddle River: Prentice-Hall, 2001.

doi: 10.18580/setep.2015.1.11 Web Link: http://dx.doi.org/10.18580/setep.2015.1.11

[12] Universidade Federal do Pará, “FalaBrasil,” 2009. [Online]. Available: http:// www.laps.ufpa.br/ falabrasil. [Acesso em 14 05 2012].

[13] VoxForge, “Downloads - Portuguese,” 2014. [Online]. Available: http:// www.voxforge.org/ pt/ Downloads. [Acesso em 19 05 2014].

[14] N. A. Meseguer, “Speech Analysis for Automatic Speech Recognition,” Dissertation (Master of Science in Electronics) - Norwegian University of Science and Technology, Trondheim, 2009.

[15] T. Cincarek, “Selective Training for Cost-effective Development of Real-Environment Speech Recognition Applications,” Dissertation (Doctor of Engineering) - Nara Institute of Science and Technology, Ikoma, 2008.

[16] M. Segbroeck, “Robust Large Vocabulary Continuous Speech Recognition using Missing Data Techniques,” Dissertation (Doctor in Engineering) - Katholieke Universiteit Leuven, Leuven, 2010.

[17] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev e P. Woodland, The HTK Book, Cambridge: Cambridge University Engineering Department, 2009.

[18] C. P. A. Silva, “Um Software de Reconhecimento de Voz para Português Brasileiro,” Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal do Pará, Belém, 2010.

[19] G. Boulianne, J.-F. Beaumont, M. Boisvert, J. Brousseau, P. Cardinal, C. Chapdelaine, M. Comeau, P. Ouellet e F. Osterrath, “Computer-Assisted Closed-Captioning of Live TV Broadcasts in French,” em *In: Interspeech 2006: Proceedings of the International Conference of Spoken Language Processing*, Pittsburgh, 2006.

Luiz Fausto de Souza Brito possui Mestrado Profissional em Computação Aplicada pela UECE (2015), MBA Executivo em Tecnologia da Informação pela UFRJ (2011), curso de extensão em Redes e Vídeo sobre IP pela UFRJ (2009) e graduação em Engenharia Elétrica com ênfase em Eletrônica pela USU/UFRJ (2005). Atualmente é Engenheiro de Projetos de Telecomunicações da Rede Globo, membro da Delegação do Brasil no ITU-R (SG 6, CPM, JTG 4-5-6-7), membro do Grupo Técnico de Recepção (GT-Rx) do Grupo de Implantação do Processo de Redistribuição e Digitalização de Canais de TV e RTV (GIRED), membro do Grupo de Estudo de Espectro e do comitê da Diretoria de Tecnologia da Sociedade Brasileira de Engenharia de Televisão (SET) e membro da Comissão de Estudo de Acessibilidade em Comunicação na Televisão da ABNT.

Flávio Luis de Mello realizou seu DSc. em teoria da computação e processamento de imagens na Universidade Federal do Rio de Janeiro - UFRJ (2006), MSc. em computação gráfica pela Universidade Federal do Rio de Janeiro - UFRJ (2003), graduação em engenharia de sistemas e computação pelo Instituto Militar de Engenharia – IME (1998). Desenvolveu sistemas de comando e controle, implementou sistemas e protocolos de troca de mensagens em meio rádio HF durante doze anos no Exército Brasileiro e lecionou no Instituto Militar de Engenharia. Atua como líder do Grupo Witty e tem como responsabilidade o desenvolvimento de aplicações baseadas em prova de teoremas, sistemas baseados em conhecimento e representação do conhecimento. É professor adjunto do Departamento de Eletrônica e de Computação (DEL) da Escola Politécnica (Poli) na Universidade Federal do Rio de Janeiro (UFRJ) desde 2007.

Cite this article:
de Souza Brito,L.F., de Mello, F. L.; 2015. Avaliação do Desempenho de um Sistema de Reconhecimento Automático de Voz em Português do Brasil Baseado em Software Livre para Geração de Closed Caption. SET EXPO PROCEEDINGS. ISSN Print: 2447-0481.ISSN Online: 2447-049X. v.1. doi: 10.18580/setep.2015.1.11 Web-link: http://dx.doi.org/10.18580/ setep.2015.1.11