Validação de Corpus para Reconhecimento de Fala Contínua em Português Brasileiro

Fabiano Weimar dos Santos
Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil
fabiano.weimar@inf.ufrgs.br

Dante Augusto Couto Barone
Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brasil
barone@inf.ufrgs.br

André Gustavo Adami
Universidade de Caxias do Sul (UCS)
Rua Francisco Getúlio Vargas, 1130 – CEP 95070-560 – Caxias do Sul – RS – Brasil
agadami@ucs.br

ABSTRACT

The development of speech processing technologies requires the use of audio and text *corpus*. Despite these resources have been researched during years for several languages, there is not enough research made for Brazilian Portuguese language. This article describes the progress of the initiative of *corpus* creation and validation for Brazilian Portuguese, using Hidden Markov Models (HMM) based acustic models and statistical language models for large vocabulary continuous speech recognition.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Speech recognition and synthesis

General Terms

Experimentation

O desenvolvimento de tecnologias de processamento de fala requer o uso de *corpus* de áudio e suas transcrições textuais. Apesar desses recursos terem sido pesquisados durante anos para diversos idiomas, ainda não existem pesquisas suficientes para o idioma Português Brasileiro. Esse artigo descreve o progresso da iniciativa de criação e validação de *corpus* para o Português Brasileiro, usando modelos acústicos baseados em Modelos Ocultos de Markov (HMM) e modelos estatísticos de linguagem para o reconhecimento de fala contínua com grande vocabulário.

1. INTRODUÇÃO

As técnicas de processamento computacional da fala dependem da disponibilidade de *corpus* e de ferramentas adequadas. Diferentemente do que acontece com outros idiomas, especialmente o idioma Inglês (onde há bastante pesquisa desenvolvida e recursos disponíveis), é difícil encontrar *corpus* em Português Brasileiro.

Segundo [11] não existem reconhecedores de fala de domínio público específicos para o Português Brasileiro. As pesquisas sobre reconhecimento de fala para o idioma Português ainda carecem de ferramentas e essa carência prejudica os avanços nas pesquisas nessa área. Sabe-se que principalmente por falta de financiamento, não vigoram atualmente esforços para a criação de ferramentas para o Português Brasileiro visando sua disponibilização para a comunidade acadêmica. O mesmo artigo cita que há pesquisadores brasileiros que acabam voltando suas pesquisas para o idioma Inglês justamente pela falta de ferramentas computacionais específicas para o idioma Português.

A principal questão que dificulta a criação de um sistema de reconhecimento de fala é o seu custo [9]. Esse custo está associado principalmente com a criação e validação de *corpus*. Estima-se que a validação de uma base de áudio e de suas respectivas transcrições leve de 3 a 5 vezes o tempo de duração do áudio [16, p.5]. O processo de validação manual, além de ter um custo alto, leva um tempo que o torna impraticável em situações onde existe uma grande produção contínua de áudio.

Este artigo descreve o progresso atual das pesquisas sobre a validação de bases de áudio e suas respectivas transcrições textuais para a construção facilitada de sistemas de reconhecimento de fala contínua para grandes vocabulários (do inglês *Large Vocabulary Continuous Speech Recognition - LVCSR*) para o idioma Português Brasileiro.

O artigo está organizado da seguinte forma: na seção 2 é descrito um avanço obtido na generalização de um dicionário de pronúncia para o idioma Português Brasileiro, na seção 3 é apresentada uma análise da perplexidade [1] de alguns *corpora* e, por fim, são indicadas na seção 4 algumas conclusões preliminares a respeito dos avanços obtidos e sobre os aspectos que ainda demandarão pesquisas futuras.

2. DICIONÁRIO DE PRONÚNCIA

O primeiro desafio prático na construção de qualquer LVCSR é realizar de forma adequada a transcrição fonética de todas as palavras do vocabulário a ser reconhecido. Isso é necessário pois os modelos acústicos de LVCSR são baseados em sub-unidades menores que as palavras, na tentativa de reduzir a complexidade computacional do reconhecedor. Habitualmente são utilizados fonemas como sub-unidades, mas para que isso seja possível em um sistema de amplo vocabulário é necessário algum nível de automação do processo de transcrição fonética.

[1] A perplexidade pode ser interpretada como a média geométrica do valor de ramificação do texto quando apresentado ao modelo de linguagem.

Nesse sentido, um dos primeiros objetivos desse trabalho foi investigar a respeito de estratégias que facilitem o processo de transcrição fonética. As idéias predominantes na área são a adoção de dicionários de pronúncia (também chamados de *lexicon*) e o uso de algoritmos de conversão grafema-fonema (do inglês *grapheme-phoneme* - G2P). Este trabalho utiliza as duas estratégias de forma integrada, permitindo que dicionários de pronúncia sejam criados de forma automática e mantendo a possibilidade de um lingüista definir transcrições fonéticas explicitamente.

Existe pouco consenso sobre que formato de representação de fonemas, algoritmo G2P (e suas respectivas regras) ou lexicon simbolize o estado da arte no idioma Português Brasileiro. Isso em parte deve-se ao fato da não disponibilização de recursos de forma pública, o que permitiria que outros pesquisadores continuassem as pesquisas já realizadas utilizando os mesmos recursos. Há também outros fatores envolvidos, como a questão dos diferentes sotaques do idioma Português Brasileiro, o que dificulta a definição de regras de transcrição fonética que representem um sotaque neutro [12]. Há propostas de algoritmos publicados sobre essa questão [2, 17], mas a implementação desses algoritmos não é trivial. Esses algoritmos, que geralmente são baseados em um conjunto de regras definidas com apoio de um lingüista, muitas vezes dependem de informações de prosódia.

Um exemplo onde isso ocorre é na transcrição fonética da palavra "cama" /k'&~m&/[2]. O primeiro "a" possui um fonema diferente do segundo "a". Para que isso seja corretamente inferido, as regras utilizadas pelos algoritmos G2P dependem de conceitos que não estão explícitos na grafia original da palavra. No exemplo citado, a vogal estressada não possui nenhuma marca ortográfica (em sua grafia original) que a identifique como tal, mas essa informação a respeito do estresse da vogal seria necessária para a definição de uma regra adequada. Nesse exemplo, temos vogais "a" ortograficamente iguais mas foneticamente diferentes (de acordo com [2], Tabela 3, quinta regra).

Podemos assumir que, por mais elaboradas que sejam as regras de um algoritmo G2P, sempre poderão existir exceções a serem consideradas. A existência de exceções implica em uma forma de representação dessas exceções e, nesse sentido, dicionários de pronúncia ganham importância. O tratamento de exceções é importante pois bastam algumas palavras com transcrição fonética incorreta para que as taxas de reconhecimento dos modelos computacionais sejam prejudicadas. Por outro lado, manter manualmente um dicionário de pronúncia com dezenas de milhares de palavras pode ser simplesmente inviável, especialmente em domínios onde palavras novas devem ser consideradas freqüentemente (por exemplo, entidades nomeadas). Constatou-se que a forma mais adequada de tratar esse dilema é uma abordagem que utilize dicionários de pronúncia e ferramentas automatizadas que auxiliem na sua manutenção.

2.1 Criação de Dicionário de Pronúncia

Recentemente foi liberado publicamente o UFPADIC [13], um dicionário de pronúncia para o idioma Português Brasileiro.

[2]Transcrição fonética gerada pelo software de síntese de fala *eSpeak* (http://espeak.sourceforge.net), exibindo a localização do estresse com aspas simples antes do primeiro 'a'.)

Acredita-se que esse trabalho tenha sido realizado com objetivos semelhantes aos desse trabalho, ou seja, disponibilizar publicamente recursos e ferramentas computacionais que facilitem a construção de sistemas de reconhecimento de fala no idioma Português Brasileiro. Nesse sentido, o UFPADIC foi adotado como *lexicon* base da pesquisa e esperamos colaborar efetivamente com sua validação e expansão.

O UFPADIC é um dicionário relativamente grande, com cerca de 65000 palavras, transcritas foneticamente segundo o formato *Speech Assessment Methods Phonetic Alphabet* - SAMPA[20].

Existem diversos formatos para representação de fonemas. Apesar do formato oficial ser o *International Phonetic Alphabet* - IPA [1], o formato SAMPA é um formato bem aceito e amplamente discutido, com o objetivo de representar fonemas de forma simplificada, usando apenas caracteres ASCII. O formato IPA utiliza caracteres UNICODE, algo que torna difícil sua utilização em, por exemplo, linguagens de programação ou mesmo alguns editores de texto. Há outros padrões que também tem o mesmo objetivo do SAMPA, como X-SAMPA [19] (uma extensão padrão SAMPA), Worldbet [7] e o IPA ASCII [8] (também conhecido como formato *Kirshenbaum*).

Segundo [13], o dicionário de pronúncia UFPADIC foi desenvolvido em duas etapas distintas. Na primeira etapa foi criado um dicionário com 11827 palavras, adotando um processo de alinhamento manual entre letras e fonemas descrito em [6]. Esse processo é bastante custoso e, nesse sentido, limita o tamanho dos dicionários que podem ser construídos. O dicionário criado foi validado e posteriormente expandido na segunda etapa, onde um algoritmo de árvore de decisão foi aplicado para extrair regras e, utilizando-se dessas regras, produziu-se a transcrição fonética de um número maior de palavras, escolhidas dentre as palavras mais freqüentes do *corpus* CETEN-Folha [4].

Ainda segundo [13], existe um grau de confiança relativamente pequeno nesse dicionário de pronúncia e o mesmo precisa ser validado, construindo-se modelos computacionais e avaliando os resultados.

Com esse intuito, através do método descrito em [3] e da ferramenta *Sequitur G2P* [15], o dicionário UFPADIC foi modelado, validado e expandido com as palavras contidas nos textos da *Constituição da República Federativa do Brasil* e *Ato das Disposições Constitucionais Transitórias* - CF-ADCT [5]. Os resultados obtidos estão descritos na Tabela 1 e 2.

A Tabela 1 demonstra o potencial de generalização dos modelos criados com a ferramenta *Sequitur G2P* aplicados ao UFPADIC. Nesse teste foram utilizados 80% das palavras do UFPADIC para treinamento e 20% para validação. É possível observar que o modelo 1 possui taxas de reconhecimento baixas, mas isso deve-se ao fato desse modelo utilizar apenas unigramas. Como esperado, os modelos seguintes demonstraram maior eficiência. Planeja-se adotar uma estratégia de validação cruzada (do inglês, *n-fold cross validation*) após uma inspeção manual dos erros apontados pelo algoritmo G2P. A Tabela 2 demonstra os resultados obtidos

Table 1: G2P 'Teste' (strings: 13079, símbolos: 109174)

	E1	% E1	E2	% E2	Mem. Virtual	Mem. Res.	Tempo
modelo 1	13079	100%	33601	30.78%	96.3	54.1	23m (*)
modelo 2	1223	9.35%	1408	1.29%	134.0	91.5	130m (*)
modelo 3	478	3.65%	579	0.53%	189.5	135.9	56m
modelo 4	546	4.17%	670	0.61%	219.2	168.2	134m

E1 = Erro String, E2 = Erro Símbolo (*) Tempo de processamento possivelmente prejudicado por tarefas simultâneas.

Table 2: G2P Final (strings: 65191, símbolos: 545676)

	E1	% E1	E2	% E2	Mem. Virtual	Mem. Res.	Tempo
modelo 1	65191	100%	165181	30.27%	106.3	64.2	12m
modelo 2	6021	9.24%	6904	1.27%	162.7	119.6	49m
modelo 3	1635	2.51%	1786	0.33%	210.8	158.9	88m
modelo 4	1098	1.68%	1197	0.22%	282.6	219.3	155m

E1 = Erro String, E2 = Erro Símbolo

usando todo o UFPADIC para treinamento e validação. Em ambas as tabelas, os coeficientes E1 e E2 demonstram, segundo a terminologia utilizada pelo *Sequitur G2P*, os erros em strings (palavras) e em símbolos (fonemas). Percebe-se também que a quantidade de memória e tempo de processamento necessários para a criação de modelos é proporcional a ordem do modelo e o tamanho do vocabulário.

A validação efetiva desse dicionário expandido será obtida analisando os resultados de reconhecedores de fala construídos utilizando esse dicionário.

3. VALIDAÇÃO DE MODELOS

A eficiência de um sistema LVCSR depende da qualidade dos modelos acústicos e dos modelos de linguagem que podem ser criados com um *corpus*. Nesse sentido, a validação de *corpus* é um processo que analisa métricas de qualidade desses modelos.

Segundo [13], dois *corpora* públicos foram utilizados na tentativa de criar um sistema de referência para o processamento computacional do idioma Português Brasileiro: o OGI-22 e o Spoltech.

O *corpus Ogi 22 Language Telephone Speech Corpus* - OGI-22 [10] possui dados em 22 idiomas, incluindo o Português Brasileiro, adquiridos por telefone em 8KHz. É um esforço do *Center for Spoken Language Understanding* (CSLU), *OGI School of Science Engineering* (OHSU). Apesar de seu tamanho relativamente pequeno, essa base contém 2500 arquivos de áudio, com algumas transcrições ortográficas e nenhuma transcrição fonética.

O *corpus Spoltech* [14], criado em uma iniciativa conjunta da Universidade Federal do Rio Grande do Sul (UFRGS), Universidade de Caxias do Sul (UCS) e CSLU/OHSU, foi criado com melhor qualidade (44 KHz), mas possui também uma série de problemas de consistência, como transcrições ortográficas e fonéticas incorretas e um número elevado de fonemas.

As Tabelas 3 e 4 mostram a perplexidade dos *corpora* OGI-22 e Spoltech, utilizando-se o *toolkit* SRILM [18]. As colunas *ppl1, ppl2, ppl3, ppl4* identificam a perplexidade dos modelos de unigramas, bigramas, trigramas e 4-gramas respectivamente. As linhas SRILM representam o uso dos algoritmos padrão para desconto e retrocesso, SRILM-D representa o uso do algoritmo *Kneser-Ney* para desconto, SRILM-I representa o uso de interpolação e SRILM-ID, por analogia, representa o uso de ambos. SRILM-W representa o uso do algoritmo *Witten-Bell* para desconto e SRILM-IW representa o uso do mesmo com interpolação (essa notação também é válida para as Tabelas 5 e 6).

Analisando os dados das Tabelas 3 e 4 percebe-se que os *corpora* OGI-22 e Spoltech, além dos diversos problemas de consistência, são *corpora* muito reduzidos; afirmativa essa que vai de encontro com as conclusões expostas em [13].

Verificando-se as limitações expostas em [13], foram pesquisadas melhorias a serem adotadas na construção de modelos de linguagem usando o *corpus* CETEN-Folha. Os resultados obtidos são demonstrados na Tabela 5.

Percebe-se pelos dados da Tabela 5 que há possibilidade de obter melhorias significativas nos resultados se adotarmos modelos de linguagem com trigramas ou 4-gramas.

Por fim, analisou-se a perplexidade de um modelo de linguagem gerado com o *corpus* CF-ADCT. Os resultados obtidos estão descritos na Tabela 6.

Percebe-se que os modelos gerados são um pouco mais complexos que os encontrados no *corpus* OGI-22, mas um número maior de palavras. Acredita-se que esse novo *corpus* que está sendo criado e validado (CF-ADCT) terá complexidade semelhante ao OGI-22, mas será provavelmente mais robusto.

4. CONCLUSÕES

A abordagem de modelagem de dicionários de pronúncia permitiu expandir o FAPDIC com esforço reduzido, exibindo taxas de erro de transcrição de palavras de 1.68% e de fonemas de 0.22%. Esse índices são muito semelhantes aos lis-

Table 3: OGI 22 (sentenças: 1375, palavras: 15767)

	ppl1	ppl2	ppl3	ppl4
SRILM	244	26	20	20
SRILM-I	244	26	20	20
SRILM-D	298	36	25	21
SRILM-ID	298	31	22	18
SRILM-W	249	20	15	15
SRILM-IW	249	19	14	13

Table 4: Spoltech (sentenças: 1682, palavras: 9115)

	ppl1	ppl2	ppl3	ppl4
SRILM	91	12	9	8
SRILM-I	91	12	9	8
SRILM-D	139	20	14	9
SRILM-ID	139	18	11	8
SRILM-W	93	11	8	7
SRILM-IW	93	11	7	7

tados em [2], mas provavelmente mais representativos em relação aos dados de testes utilizados.

Os modelos de linguagem testados evidenciam os problemas dos *corpora* conhecidos e demonstram as potenciais melhorias que podem ser obtidas utilizando-se modelos de trigramas e 4-gramas, principalmente quando comparados aos modelos utilizados em [13] (onde são utilizados apenas bigramas).

Por fim, testes preliminares com o *corpus* CF-ADCT demonstram que ele possui complexidade semelhante ao OGI-22 e que poderá, em trabalhos futuros, servir como uma base de dados eficiente na produção de sistemas LVCSR.

Planeja-se utilizar o *corpus* CF-ADCT, assim como os modelos que generalizaram com eficiencia o *lexicon* UFPADIC, na validação de LVCSR em projetos futuros.

5. REFERÊNCIAS

[1] R. Albright. *The International Phonetic Alphabet: Its Backgrounds and Development*. Dept. of Speech and Drama, 1953.

[2] F. Barbosa, G. Pinto, F. Resende, C. Gonçalves, R. Monserrat, and M. Rosa. Grapheme-Phone Transcription Algorithm for a Brazilian Portuguese TTS. *PROPOR*, pages 23–30, 2003.

[3] M. Bisani and H. Ney. Joint-sequence models for grapheme-to-phoneme conversion. *Speech Communication*, 2008.

[4] CETEN-Folha. CETEN-Folha - Corpus de Extractos de Textos Electrónicos NILC/Folha de São Paulo, 2008. Publicado em http://acdc. linguateca. pt/cetenfolha/. Acesso em maio de 2008.

[5] CF-ADCT. Constituição da República Federativa do Brasil, 1988. Publicado em http://www2. camara. gov. br/acessibilidade/constituicaoaudio. html. Acesso em maio de 2008.

[6] R. Damper, Y. Marchand, J. Marseters, and A. Bazin. Aligning Letters and Phonemes for Speech Synthesis. In *Fifth ISCA Workshop on Speech Synthesis*. ISCA, 2004.

[7] J. Hieronymus. ASCII Phonetic Symbols for the World's Languages: Worldbet. *Journal of the International Phonetic Association*, 23, 1993.

[8] E. Kirshenbaum. Representing IPA phonetics in ASCII. *URL: http://www. kirshenbaum. net/IPA/ascii-ipa. pdf (unpublished), Hewlett-Packard Laboratories*, 2001.

[9] L. Lamel, F. Lefevre, J. Gauvain, and G. Adda. Portability issues for speech recognition technologies. *Proceedings of the first international conference on Human language technology research*, pages 1–7, 2000. Publicado em http://portal.acm.org/citation.cfm?id=1072133.1072211. Acesso em outubro de 2007.

[10] T. Lander, R. Cole, B. Oshika, and M. Noel. The ogi 22 language telephone speech corpus, 1995.

[11] N. S. Neto, E. Sousa, V. Macedo, A. G. Adami, and A. Klautau. Desenvolvimento de software livre usando reconhecimento e síntese de voz: O estado da arte para o português brasileiro. *6 Workshop Software Livre, 2005, Porto Alegre. Anais da Trilha Nacional do Workshop Software Livre*, 1, 2005.

[12] J. Ramos. Avaliação de dialetos brasileiros: o sotaque. *Revista de Estudos da Linguagem. Belo Horizonte: UFMG*, 6(5):103–125, 1997.

[13] N. Sampaio Neto, C. Patrick, A. G. Adami, and A. Klautau. Spoltech and ogi-22 baseline systems for speech recognition in brazilian portuguese (accepted). *PROPOR*, 2008.

[14] M. Schramm, L. Freitas, A. Zanuz, and D. Barone. A Brazilian Portuguese Language Corpus Development. In *Sixth International Conference on Spoken Language Processing*. ISCA, 2000.

[15] Sequitur G2P. Sequitur G2P - A trainable Grapheme-to-Phoneme converter, 2008. Publicado em http://www-i6. informatik. rwth-aachen.

Table 5: CETEN-Folha (sentenças: 1542881, palavras: 23414273)

	ppl1	ppl2	ppl3	ppl4
SRILM	1454	170	88	76
SRILM-I	1454	170	88	76
SRILM-D	2442	255	122	78
SRILM-ID	2242	231	109	71
SRILM-W	1456	166	84	69
SRILM-IW	1456	158	79	65

Table 6: CF-ADCT (sentenças: 2874, palavras: 68975)

	ppl1	ppl2	ppl3	ppl4
SRILM	467	27	17	16
SRILM-I	467	27	17	16
SRILM-D	696	48	28	18
SRILM-ID	696	41	24	15
SRILM-W	472	24	14	13
SRILM-IW	472	23	13	12

de/web/Software/g2p. html. Acesso em maio de 2008.

[16] A. J. Serralheiro, H. Meinedo, D. A. Caseiro, and I. Trancoso. Alinhamento de livros falados. *XVII Encontro Nacional da Associação Portuguesa de Linguística*, 2001. Publicado em http://www.inesc-id.pt/pt/indicadores/Ficheiros/737.pdf. Acesso em outubro de 2007.

[17] D. Silva, A. de Lima, R. Maia, D. Braga, J. de Moraes, J. de Moraes, and F. Resende. A rule-based grapheme-phone converter and stress determination for Brazilian Portuguese natural language processing. *Telecommunications Symposium, 2006 International*, pages 550–554, 2006.

[18] A. Stolcke. SRILM-an Extensible Language Modeling Toolkit. In *Seventh International Conference on Spoken Language Processing*. ISCA, 2002.

[19] J. Wells. Computer-coding the IPA: a proposed extension of SAMPA. *Revised draft*, 4(28):1995, 1995.

[20] J. Wells et al. SAMPA computer readable phonetic alphabet. *URL http://www. phon. ucl. ac. uk/home/sampa*, 2004.