Novos Recursos e Utilização de Adaptação de Locutor no Desenvolvimento de um Sistema de Reconhecimento de Voz para o Português Brasileiro

Patrick Silva, Nelson Neto e Aldebaro Klautau

***Resumo*— Este trabalho descreve o estágio atual do desenvolvimento de um sistema de reconhecimento de voz para o Português Brasileiro. Dentre os objetivos do trabalho tem-se a construção de um sistema de voz contínua para grandes vocabulários, apto a ser usado em aplicações em tempo-real. Os novos recursos produzidos consistem de corpora de voz digitalizada e texto. O corpus de texto vem sendo construído através da extração e formatação automática de textos de jornais na internet. Além disso, foram produzidos dois corpora de voz, um baseado em *audiobooks* e outro produzido especificamente para simular testes em tempo-real. O trabalho apresenta resultados experimentais usando-se tais recursos. O trabalho também propõe a utilização de técnicas de adaptação de locutor para resolução de ploblemas de descasamento acústico entre corpora de voz. Para construção dos modelos acústicos e de linguagem, fez-se uso das ferramentas HTK e SRILM, respectivamente. Todos os recursos desenvolvidos estão disponíveis no site do projeto *FalaBrasil*.**

***Palavras-Chave*— Corpus, reconhecimento de voz, sistemas em tempo-real.**

***Abstract*— This contribution describes an on-going work concerning the development of a speech recognition system for Brazilian Portuguese. The goal is to implement a large-vocabulary continuous speech recognition system, capable of operating in real-time. The new resources presented in this work are speech and text corpora. The text corpora was produced through a process of downloading and formatting the contents of some newspapers available on the internet. Also the work presents two new speech corpora: one based on audiobooks and another produced specifically for tests in real-time. The work presents experimental results using the new resources. This works also presents a environment adaptation procedure on corpora using speaker adaptation techniques. The HTK and SRILM toolkits were used to building acoustic and language models, respectively. The resources are publicly available and allow for reproducing results across different sites.**

***Keywords*— Speech corpora, speech recognition, real-time systems.**

I. Introdução

Universidades e centros de pesquisa vêm tentando encontrar soluções para problemas práticos na tarefa de reconhecimento automático de voz. Um desses problemas é a escassez de dados para treino de sistemas *large vocabulary continuous speech recognition* (LVCSR) para o Português Brasileiro (PB). Para línguas como a inglesa, já existem bases de dados (corpora) de referência como o TIMIT, WSJ (*Wall Street Journal*) e *Switchboard*. No Brasil, o corpus mais usado atualmente parece ser o Spoltech [1], entretanto o mesmo não é gratuito. Alguns grupos de pesquisa partiram para a construção de bases de dados proprietárias. O presente trabalho vai ao encontro de iniciativas anteriores [2], [3], complementando-as em alguns aspectos e diferenciando-se por disponibilizar todos os recursos. Uma motivação para a composição e uso de bases públicas é a recente comprovação de que pesquisas cujos resultados são passíveis de reprodução produzem maior impacto [4].

Em trabalhos anteriores [5], [6], buscou-se criar um sistema de referência utilizando os corpora Spoltech e OGI-22, porém o sistema se limitava a um reconhecimento "controlado", ou seja, treino e teste com características semelhantes (locutores, ambiente de gravação, etc). Já o trabalho atual, possui o objetivo de criar um sistema apto a trabalhar em condições de descasamento acústico entre os corpora de treino e teste.

Uma das contribuições deste trabalho é a construção e disponibilização dos seguintes recursos para o PB: um corpus de texto baseado em extração automática de textos de jornais na internet e dois corpora de voz: um baseado em *audiobooks* e outro criado para avaliação de desempenho de sistemas LVCSR, todos disponibilizados em domínio público. Outra contribuição é o uso de adaptação de locutor para diminuir o impacto da escassez de voz digitalizada e/ou uso de poucos locutores. Os resultados a serem apresentados mostram que técnicas de adaptação de locutor são eficazes para realizar a adaptação de ambiente e diminuir o impacto do *descasamento acústico*. Um terceiro aspecto é que muitos sistemas de reconhecimento de voz para o PB já foram produzidos [7]–[14], porém as informações da operação dos mesmos em tempo-real é escassa. O presente trabalho informa o tempo de decodificação.

Este trabalho está organizado da seguinte maneira. Na Seção II é descrito o dicionário fonético utilizado. Na Seção III são descritos os corpora construídos e utilizados na construção do sistema LVCSR. Nas Seções IV e V são mostrados os processos utilizados na construção dos modelos acústico e de linguagem, respectivamente. Na Seção VI são mostradas as técnicas de adaptação de locutor utilizadas. Na Seção VII são exibidos os resultados obtidos. Por fim, a Seção VIII apresenta as conclusões do trabalho e pesquisas futuras.

II. Dicionário Fonético

A conversão de uma seqüência de caracteres em seqüências de fones é um importante pré-requisito para serviços que

Laboratorio de Processamento de Sinais - LaPS, Faculdade de Engenharia da Computação, Universidade Federal do Pará, Belém - PA, Brazil, E-mails: {patrickalves, nelsonsampaio, aldebaro}@ufpa.br.

envolvem reconhecimento e/ou síntese de voz. Contudo, a tarefa não é trivial e diversas técnicas de conversão vêm sendo adotadas ao longo da última década. Existe um número bem menor de estudos na área dedicados ao PB quando comparado à língua inglesa, por exemplo.

Em [15], os autores disponibilizaram um algoritmo baseado em uma estrutura de regras descritas em [16] para conversão G2P (*Grapheme To Phoneme*) com determinação de vogal tônica para PB. Uma vantagem dos conversores baseados em regras é que o alinhamento lexical (dos grafemas) não se faz necessário, visto que o *software* não precisa ser treinado para gerar suas próprias regras. Ou seja, as propostas de conversão, baseadas em critérios fonológicos pré-estabelecidos, são fornecidas ao sistema de acordo com a língua a qual o aplicativo se destina. Sua arquitetura é *self-contained*, ou seja, não carece de estágios intermediários, nem depende de outros algoritmos, para realizar análises específicas, como divisão silábica ou identificação de pluralidade. Existe uma ordem obrigatória para aplicação das regras. Primeiro são analisadas as regras consideradas mais específicas e, por último, a regra, ou caso geral, que finaliza a análise. Nenhuma análise co-articulatória entre palavras foi realizada, já que este processo de conversão G2P lidou apenas com palavras isoladas.

Como produto do trabalho obteve-se um *software* para conversão G2P. O dicionário fonético resultante possui mais de 64 mil palavras e também se encontra disponível.

III. Corpora

A. LapsFolha

Os modelos de linguagem são tipicamente construídos utilizando-se de modelos interpolados de transcrições e textos de jornais, tais modelos são largamente utilizados em tarefas de reconhecimento de voz em tempo real. O corpus inicialmente utilizado foi o CETENFolha [17] (Corpus de Extratos de Textos Eletrônicos NILC/Folha de S. Paulo), um corpus com aproximadamente 24 milhões de palavras, criado pelo projeto Linguateca[1] com textos do jornal Folha de S. Paulo e compilado pelo centro de pesquisa NILC/São Carlos, Brasil.

O CETENFolha vem sendo complementado com textos recentes de outros jornais. Isso vem sendo feito através de um processo totalmente automatizado de coleta e formatação diária de jornais disponíveis na internet.

Aproximadamente dois meses de textos coletados e processados automaticamente encontram-se disponíveis em [18], os quais correspondem a aproximadamente 120 mil frases. O processo está sendo melhorado de forma a se obter 1 milhão de frases mensais. O corpus será expandido continuamente de forma a alcançar 5 milhões de frases.

B. LapsStory

É conhecido que um dois maiores problemas na construção de sistemas LVCSR é a carência de dados para treino. Para o Inglês existem corpora com mais de 240 horas de voz transcrita, como o *Switchboard* [19], por exemplo.

[1]www.linguateca.pt

Dentre as dificuldades encontradas em se produzir grandes corpora, tem-se a coleta de dados (voz) e transcrição ortográfica. Visando contornar tais problemas, foi construído um novo corpus de voz baseado em *audiobooks*. *Audiobooks* são livros falados disponíveis na internet. Com os arquivos de áudio e suas respectivas transcrições (livros) tem-se uma redução considerável na tarefa de produção de um corpus.

Inicialmente, foram obtidos 5 livros com aproximadamente 1 hora de duração cada. Os arquivos de áudio foram reamostrados para 22.050 Hz com 16 bits. Em seguida os mesmos foram segmentados em arquivos menores, com aproximadamente 30 segundos de duração cada, e por fim transcritos. Atualmente, o corpus é composto por 2 locutores do sexo masculino e 3 do sexo feminino. Os arquivos totalizam 5 horas e 19 minutos e o mesmo será expandido para 50 horas de voz com a mesma distribuição de gênero. Uma característica da utilização de *audiobooks* é que o ambiente de gravação utilizado é bastante controlado, sendo assim, os arquivos não possuem ruído audível, têm alta razão sinal/ruído, etc. Quando tais arquivos são usados para treinar um sistema que irá operar em ambiente ruidoso, tem-se um problema com o descasamento acústico. Esse problema pode ser aliviado com técnicas de adaptação de locutor, como será descrito na Seção VII.

C. LapsBenchmark

Com o intuito de obter uma boa avaliação de desempenho e possibilitar a comparação de resultados com outros grupos de pesquisas, vem sendo construído um corpus para o PB, atualmente composto por 500 frases. Busca-se aqui criar um corpus de referência com características mais próximas da operação de um sistema de reconhecimento de voz em ambiente de escritório, ou seja, com ruído do ambiente, condicionador de ar, etc. Isso distingue o LapsBenchmark do corpus LapsStory, previamente apresentado.

Para construção do corpus LapsBenchmark utilizou-se de frases retiradas do corpus CETENFolha. Mais especificamente foram utilizadas as frases descritas em [20]. Atualmente, o corpus possui 25 locutores com 20 frases cada, sendo 17 homens e 8 mulheres, o que corresponde a aproximadamente 42 minutos de áudio. Este corpus será expandido de forma a ter 50 locutores com a mesma distribuição, totalizando 1.000 frases. Todas as gravações foram realizadas em computadores utilizando microfones comuns. A taxa de amostragem utilizada foi de 22.050 Hz e cada amostra foi representada com 16 bits. Como mencionado, o ambiente não foi controlado, existindo a presença de ruído nas gravações.

O LapsBenchhmark precisa ter seu tamanho consideravelmente aumentado para ser utilizado plenamente na realização de experimentos considerados como LVCSR. Nesse trabalho, usa-se uma estratégia que busca imitar a operação de um sistema LVCSR: o modelo de linguagem possui mais de 60 mil palavras, e o decodificador precisa lidar com alta perplexidade e descasamento acústico. Obviamente, tal estratégia permite avaliar aspectos importantes mas possui limitações. Uma dessas limitações, inerente à pouca quantidade de dados para teste, é a robustez das estimativas de taxa de erro.

Diferentemente dos anteriores, o próximo corpus não foi desenvolvido pelos autores. O mesmo será brevemente de-

scrito por ter sido usado nos experimentos, após sofrer várias correções.

D. Spoltech

O Spoltech é um *corpus* para o português brasileiro desenvolvido pela Universidade do Rio Grande do Sul, Universidade de Caxias do Sul e OGI (*Oregon Graduate Institute of Science and Technology*)[2], em projeto subsidiado pelo CNPq/Brasil e pela NSF/Estados Unidos. O corpus é distribuído pelo LDC *Linguistic Data Consortium*[3] e OGI. O mesmo consiste de gravações via microfones de 477 locutores de várias regiões do Brasil com suas respectivas transcrições fonéticas e ortográficas. As gravações consistem tanto de leituras de frases curtas quanto de respostas a perguntas (no intuito de modelar a fala espontânea). No total, o *corpus* é composto de 8.080 arquivos de voz digitalizada (extensão wav), 2.540 arquivos com transcrições em nível de palavra (arquivos de texto sem alinhamento temporal, com extensão txt) e 5.479 arquivos com transcrições em nível de fonema (com alinhamento temporal e extensão phn).

O ambiente de gravação não foi sempre controlado. Sendo assim, algumas gravações foram feitas em estúdios e outras em ambientes ruidosos (feiras, escolas, etc). Os dados foram gravados a uma taxa de 44,1 KHz (mono, 16-bit). Para utilização do corpus foi necessário uma grande revisão e correção de vários arquivos (wavs e txts) como descrito em [5]. Além disso, para utilização do mesmo de forma compatível aos demais *corpora* citados, foi realizada uma reamostragem para 22.050 Hz. Nos testes foram utilizados 7.190 arquivos wav, os quais correspondem a aproximadamente 4 horas de voz.

IV. Modelos Acústicos

Para construção dos modelos acústicos (MA) foi utilizado o *software* HTK [21]. O HTK é um toolkit para manipulação de modelos ocultos de Markov [22] bastante utilizado em reconhecimento de voz. O *front-end* utilizado foram os conhecidos MFCC's (*Mel-Frequency Cepstral Coefficients*). Existem diversas outras alternativas ao MFCC que podem conduzir a melhores resultados, tais como coeficientes *perceptual linear predictive* (PLP) [23], mas os MFCC são tipicamente usados para comparações e daí terem sido adotados. Inicialmente obteve-se os 12 primeiros coeficientes ceptrais mais a energia do sinal de cada *frame* de voz, onde cada *frame* corresponde a uma janela de 25 ms e o deslocamento da janela de 10 ms. Em seguida foram extraídas a primeira e segunda derivada compondo um vetor com 39 parâmetros para cada *frame*. Por último os parâmetros MFCC's foram normalizados através da normalização da média cepstral (*cepstral mean subtraction*) [24].

Os modelos acústicos foram refinados de forma iterativa [25]. Iniciando com modelos baseados em monofones e com uma Gaussiana por mistura, as HMM's foram expandidas de forma a compor modelos com múltiplas Gaussianas por mistura e utilizando modelos trifones. Em todo o processo de treino das HMM's foi utilizado o algoritmo de Baum-Welch [26]. Foram utilizados 38 fones, retirados do dicionário fonético descrito na Seção II, mais o modelo de silêncio com HMM's com 3 estados utilizado a estrutura *left-to-right*. O modelo de pausa curta (*short-pause*) com apenas um estado emissor foi construído através da cópia do estado central do modelo do silêncio. Modelos trifones dependentes de contexto (*cross-word triphones*), foram criados a partir dos monofones. Modelos dependentes de contexto levam em conta os efeitos co-articulatórios presentes nas transições entre palavras. Foi realizado o vínculo (*tying*) das matrizes de transição do trifones derivados dos mesmos monofone.

Um problema clássico dos modelos dependentes do contexto é a ausência de dados de treino suficiente para suportar uma grande quantidade de trifones. Todavia, alguns efeitos coarticulatórios dos trifones são bastante similares. Nestes casos, o mais proveitoso, é representar tais trifones pelos mesmos modelos. Para isso, o método de vínculo de estados (*tied-state*) através de árvore de decisão fonética foi utilizado. Através de uma lista de questões linguísticas, foi realizado o vínculo dos trifones com características fonéticas similares.

Para ilustrar, seguem abaixo algumas regras de classificações de vogais e consoantes usadas na construção da árvore de decisão:

```
...
QS "R_V-Fechada" { *+i,*+e,*+o,*+u }
QS "R_V-Front"   { *+i,*+E,*+e }
QS "R_Palatais"  { *+S,*+Z,*+L,*+J }
QS "L_V-Back"    { u-*,o-*,O-* }
QS "L_V-Aberta"  { a-*,E-*,O-* }
...
```

Após o vínculo dos estados, finalizou-se o processo de treino do modelo com o incremento do número de Gaussianas nas misturas. Uma descrição mais detalhada do processo de treino das HMM's pode ser visto em [27].

V. Modelos de Linguagem

O modelamento da língua (ou de linguagem) foi composto a partir de N-gramas [14]. Para construção dos modelos foi utilizado o *SRI Language Modeling Toolkit* (SRILM). O SRILM [28] é uma ferramenta específica para construção e manipulação de modelos estatísticos de linguagem. O *software* também permite a implementação de várias técnicas de suavização (*smoothing*) nos modelos N-gramas.

O conjunto de treino foi composto por textos extraídos dos *corpora* CETENFolha, Spoltech, OGI-22 Language, West-Point, LapsStory e LapsFolha. No final obteve-se um corpus com 1,6 milhões de frases e 64.972 palavras distintas. O número total de palavras, contando-se as repetições, foi de 25,8 milhões. Nos testes realizados, foram utilizados modelos trigramas e 4-gramas. As características dos modelos são descritos na Tabela I. Para o cálculo da perplexidade foram utilizadas 10 mil frases do CETENFolha não vistas na fase de treino.

[2]http://cslu.cse.ogi.edu/corpora
[3]http://www.ldc.upenn.edu

TABELA I

CARACTERÍSTICAS DOS MODELOS DE LINGUAGEM UTILIZADOS.

Número de frases usadas no treino	1.645.327
Número de palavras usadas no treino	25.800.000
Número de palavras distintas	64.972
Perplexidade - Trigrama	169.0
Perplexidade - 4-grama	184.0
Técnica de suaviação	Kneser-Ney discounting

VI. ADAPTAÇÃO DE LOCUTOR

Alguns sistemas de reconhecimento de voz utilizam-se de modelos independentes de locutor. Nesses casos, o objetivo é que o sistema seja capaz de reconhecer a voz de qualquer locutor com uma boa precisão. Sistemas dependentes de locutor apresentam desempenho superior quando utilizados com o mesmo locutor usado na fase de adaptação (treino). Diante disso, busca-se através de técnicas de adaptação de locutor (*speaker adaptation*) aumentar a precisão do sistema desenvolvido.

Com o auxilio da ferramenta HTK, foram utilizadas duas técnicas de adaptação. A primeira técnica consiste na adaptação através de transformações lineares. Tal técnica é conhecida como MLLR [24] (*Maximum Likelihood Linear Regression*) e se utiliza de matrizes de transformação obtidas com o algoritmo de maximização e esperança (EM - *Expectation-Maximization*). A MLLR busca diminuir as diferenças entre o modelo inicial e os dados de adaptação. Isso é feito através da alteração das médias e variâncias das misturas de Gaussianas nas HMMs. A segunda técnica, conhecida como MAP (*Maximum a Posteriori*), utiliza o arcabouço do aprendizado Bayesiano, onde as distribuições iniciais dos modelos independente de locutor representam o conhecimento a priori e, são atualizadas através da regra de Bayes para se alcançar um novo modelo, dependente do locutor. Mais especificamente, na implementação do HTK, a média de cada Gaussiana é atualizada através de MAP usando-se a média da distribuição prior, dos pesos das Gaussianas e dos dados do locutor em questão, os quais são usados para a adaptação.

Ambas as técnicas foram utilizadas em treino supervisionado (*offline*), porém o HTK também possui suporte a treino não supervisionado (*online*). De forma a se obter uma melhor adaptação, é comum utilizar-se da combinação das duas técnicas, como será mostrado na Seção VII.

VII. RESULTADOS DOS EXPERIMENTOS

O modelo de linguagem trigrama citado na Seção V foi utilizado para testar o sistema. As medidas de desempenho utilizadas foram a WER (*word error rate*) e escala de tempo real RT (*real-time scale factor*) média. O RT é obtido dividindo-se o tempo que o sistema gasta para reconhecer uma frase, pela duração da mesma. Simulações foram realizadas utilizando o decodificador *HDecode* do HTK. Três modelos acústicos utilizando 14 Gaussianas por mistura com trifones dependentes de contexto foram criados a partir dos *corpora* Spoltech e LapsStory, dois utilizando os *corpora* individualmente e um combinando as bases. Os parâmetros utilizados na decodificação são mostrados na Tabela II. Para

TABELA II

PARÂMETROS UTILIZADOS PARA DECODIFICAÇÃO.

word insertion penalty	22
LM scale factor	20
pruning beam width	250
acoustic scale factor	1.0 e 2.0

TABELA III

RESULTADOS OBTIDOS COM OS MODELOS ACÚSTICOS DO LAPSSTORY E SPOLTECH. OS TESTES FORAM REALIZADOS COM O LAPSBENCHMARK.

Modelos Acústicos	**WER (%)**	**RT**
Spoltech	34.8	6
LapsStory	52.5	10
LapsStory+Spoltech	40	9
MA's com peso 2.0	**WER (%)**	**RT**
Spoltech	44.3	1.5
LapsStory	55.9	3
LapsStory+Spoltech	48.5	1.6

as simulações foi utilizado um computador Intel(R) Pentium Dual Core 1,8 GHz com 2 GB de memória RAM.

Todos os testes foram realizados com o LapsBenchmark. Com isso busca-se simular e avaliar como o sistema se comportaria em aplicações com descasamento acústico. A Tabela III mostra os resultados obtidos com os modelos acústicos criados a partir do LapsStory, Spoltech e com a combinação de ambos. Nos modelos treinados individualmente com o Spoltech e LapsStory, é visto que ambos obtiveram taxas de erro altas, principalmente no caso do LapsStory, já que o mesmo foi produzido em um ambiente acústico totalmente diferente do LapsBenchmark. Já a combinação dos *corpora*, mostrou-se ineficaz mesmo quando se realizou a normalização (normalização da média e variância dos parâmetros MFCC's), fato justificado pela grande diferença entre os mesmos. Foi observado que para todos os modelos acústicos utilizados, o sistema apresentou um alto valor de RT, o que impossibilita a utilização dos mesmos em aplicações de tempo-real. De forma a aumentar a velocidade do sistema, modificou-se o peso do modelo acústico para 2.0 no processo de decodificação (parâmetro *-a* no HDecode), o qual acelera o processo de busca aumentando os *scores* observados nas HMM's. Essa alteração torna o sistema até 4 vezes mais rápido, porém acarreta em perda de precisão, como mostrado na segunda parte da Tabela III.

Na segunda etapa de testes, foram adotadas as técnicas de adaptação de locutor, descritas na Seção VI, para melhorar o desempenho do sistema. Porém, diferente de uma adaptação de locutor convencional, realizou-se na verdade uma adaptação de ambiente (*environment adaptation*). Partindo do modelo acústico produzido com o LapsStory, fez-se uma adaptação utilizando o corpus Spoltech. Assim, os modelos acústicos do LapsStory, que tendem a ser homogêneos em função do pequeno número de locutores e alta razão sinal/ruído, são modificados de maneira a melhor modelar ambientes ruidosos e a fala de diversos locutores. Os resultados da adaptação são exibidos na Tabela IV.

A combinação MLLR+MAP apresentou melhores resultados. Trabalhos relacionados apontam vantagens do uso combi-

TABELA IV
RESULTADOS OBTIDOS COM AS TÉCNICAS DE ADAPTAÇÃO DE LOCUTOR.

Técnica de adaptação	**WER (%)**	**RT**
MLLR	30.2	9.5
MAP	29.3	7.0
MLLR+MAP	25.5	8.3
MA's com peso 2.0	**WER (%)**	**RT**
MLLR	34.2	2.2
MAP	35.6	2.0
MLLR+MAP	29.6	2.4

nados das duas técnicas [29]. De forma geral, a técnica MLLR trabalha com agrupamento de Gaussianas com características semelhantes via matrizes de transformação. Isso permite que a mesma seja efetiva mesmo para pequenos conjuntos de dados, mas não obtém desempenho ótimo quando o conjunto de dados é grande. Em contrate, a MAP utiliza as Gaussianas dos trifones individualmente. Isso permite aumento do desempenho ao custo da necessidade de um maior conjunto de dados para treino.

Para o problema proposto, o melhor resultado, levando em conta a WER e RT, foi com a técnica MLLR+MAP com peso 2.0 no modelo acústico a qual obteve 29.6% de WER com 2.4 de RT. Para adicionar alguma melhora ao sistema utilizou-se do procedimento de *rescoring* utilizando o modelo de linguagem 4-grama mencionado na Seção V. No procedimento de *rescoring* tem-se como saída do decodificador (HDecode) uma *lattice* (rede) para cada arquivo de entrada, cada uma com os 20 melhores candidatos de saída. Em seguida utiliza-se do modelo de linguagem 4-grama para recalcular os *scores* e encontrar o melhor candidato. Com o *rescoring* obteve-se uma melhora absoluta de 0.8%, alcançando-se 28.8% de WER. Obviamente, o tempo de processamento é bem maior e como a solução foi obtida via comandos (não automatizada), o fator RT não foi contabilizado nesse caso.

VIII. CONCLUSÕES

Este artigo apresentou novos recursos para pesquisa em sistemas LVCSR em PB. Resultados experimentais com o LapsBenchMark mostram a tentativa de criação de um sistema que possa ser utilizado em condições de descasamento acústico e em tempo real. O LapsStory mostrou-se relativamente eficiente para a construção de modelos acústicos quando combinado com técnicas de adaptação locutor, alcançando uma WER em torno de 29.6%. Apesar de alta, essa WER foi considerada satisfatória pois a proposta é validar todas as técnicas disponíveis para a construção do LVCSR e projetar que o desempenho irá melhorar com o aumento do tamanho do corpus. Ou seja, espera-se que com o aumento do corpus para 50 horas, várias dificuldades possam ser superadas. O valor de 2,4 para a RT também pode ser considerado razoável, mas não ideal para sistemas em tempo-real. Novas estratégias serão implementadas de forma a aumentar a velocidade do processo de decodificação, tal como o uso do decodificador Julius [30].

Deve-se registrar que todos os recursos citados encontram-se disponíveis na página do projeto *FalaBrasil* [18]. Espera-se que outros grupos possam utilizar os recursos e compartilhar resultados.

REFERÊNCIAS

[1] "Advancing human language technology in Brazil and the United states through collaborative research on portuguese spoken language systems," Federal University of Rio Grande do Sul, University of Caxias do Sul, Colorado University, and Oregon Graduate Institute, 2001.

[2] C. A. Ynoguti, P. A. Barbosa, and F. Violaro, "A large speech database for brazilian portuguese spoken language research," *Lecture Notes in Computer Science (LNCS)*, vol. 2721/9, pp. 193–196, 2003.

[3] C. A. Ynoguti and F. Violaro, "A brazilian portuguese speech database," *XXVI Simpósio Brasileiro de Telecomunicações - SBrt 2008*, 2008.

[4] P. Vandewalle, J. Kovacevic, and M. Vetterli, "Reproducible research in signal processing," *IEEE Signal Processing Magazine*, vol. 37, pp. 37–47, 2009.

[5] P. Silva, N. Neto, A. Klautau, A. Adami, and I. Trancoso, "Speech recognition for brazilian portuguese using the spoltech and OGI-22 corpora," *XXVI Simpósio Brasileiro de Telecomunicações - SBrt 2008*, 2008.

[6] N. Neto, P. Silva, A. Klautau, and A. Adami, "Spoltech and OGI-22 baseline systems for speech recognition in brazilian portuguese," *International Conference on Computational Processing of Portuguese Language - PROPOR*, 2008.

[7] R. Fagundes and I. Sanches, "Uma nova abordagem fonético-fonológica em sistemas de reconhecimento de fala espontânea," *Revista da Sociedade Brasileira de Telecomunicações*, vol. 95, 2003.

[8] L. Pessoa, F. Violaro, and P. Barbosa, "Modelo de língua baseado em gramática gerativa aplicado ao reconhecimento de fala contínua," in *XVII Simpósio Brasileiro de Telecomunicações*, 1999, pp. 455–458.

[9] S. Santos and A. Alcaim, "Um sistema de reconhecimento de voz contínua dependente da tarefa em língua portuguesa," *Revista da Sociedade Brasileira de Telecomunicações*, vol. 17, no. 2, pp. 135–147, 2002.

[10] I. Seara et al, "Geração automática de variantes de léxicos do português brasileiro para sistemas de reconhecimento de fala," in *XX Simpósio Brasileiro de Telecomunicações*, 2003, pp. v.1. p.1–6.

[11] M. Schramm, L. Freitas, A. Zanuz, and D. Barone, "A brazilian portuguese language corpus development," *ICSLP-2000, vol.2, 579-582*, 2000.

[12] C. A. Ynoguti and F. Violaro, "Influência da transcrição fonética no desempenho de sistemas de reconhecimento de fala contínua," in *XVII Simpósio Brasileiro de Telecomunicações*, 1999, pp. 449–454.

[13] R. Teruszkin and F. Vianna, "Implementation of a large vocabulary continuous speech recognition system for brazilian portuguese," *Journal of Communication and Information Systems, vol. 21, no. 3, pages 204-218*, 2006.

[14] E. Silva, M. Pantoja, J. C., and A. Klautau, "Modelos de linguagem n-grama para reconhecimento de voz com grande vocabulário," in *TIL2004 - III Workshop em Tecnologia da Informação e da Linguagem Humana*, 2004.

[15] A. C. Siravenha, N. Neto, V. Macedo, and A. Klautau, "Uso de regras fonológicas com determinação de vogal tônica para conversão grafema-fone em português brasileiro," *7th International Information and Telecommunication Technologies Symposium*, 2008.

[16] D. C. Silva, A. A. de Lima, R. Maia, J. F. d. M. Daniela Braga, J. A. de Moraes, and F. G. V. R. Jr, "A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing," *VI International Telecommunications Symposium*, 2006.

[17] "http://acdc.linguateca.pt/cetenfolha/," Visited in January, 2008.

[18] "http://www.laps.ufpa.br/falabrasil," Visited in April, 2008.

[19] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *Texas Instruments Inc. Dallas*.

[20] R. J. Cirigliano, C. Monteiro, F. L. de F. Barbosa, F. G. V. R. Jr, L. R. Couto, and J. A. de Moraes, "Um conjunto de 1000 frases foneticamente balanceadas para o português brasileiro obtido utilizando a abordagem de algoritmos genéticos," *XXII Simpósio Brasileiro de Telecomunicações*, 2005.

[21] "http://htk.eng.ac.uk," Visited in March, 2008.

[22] Huang, Ariki, and Jack, *Hidden Markov Models for Speech Recognition*, 1990.

[23] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," *Journal of the Acoustical Society of America*, vol. 87, no. 4, pp. 1738–52, Apr. 1990.

[24] S. e. Young, *The HTK Book*. Microsoft Corporation, Version 3.0, 2000.

[25] P. Woodland and S. Young, "The htk tied-state continuous speech recognizer," *In: Proc. Eurospeech'93, Berlim*, 1993.

[26] L. R. Welch, "Hidden markov models and the baum-welch algorithm," *IEEE Information Theory Society Newsletter.*, vol. 53, pp. 10–12, 2003.
[27] C. P. A. da Silva, "Sistemas de reconhecimento de voz para o português brasileiro utilizando os corpora spoltech e ogi-22," Tech. Rep., 2008.
[28] A. Stolcke, "Srilm - an extensible language modeling toolkit," *Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado*, 2002.
[29] S. G. Ralf and R. Kompe, "A combined MAP + MLLR approach for speaker adaptation," *Proc Sony Res Forum*, vol. 9th, pp. 9–14, 2000.
[30] "http://julius.sourceforge.jp/en/," Visited in may, 2009.