CC/SBTVD e Reconhecimento de Fala para Português Brasileiro: tentando preencher uma lacuna de Corpus de Fala

Rafael Martins Feitosa[1], Dante A. C. Barone[1], André G. Adami[2]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

[2]Universidade de Caxias do Sul
Rua Francisco Getúlio Vargas, 1130 – CEP 95070-560 – Caxias do Sul – RS – Brasil

{rafael.feitosa,barone}@inf.ufrgs.br, agadami@ucs.br

Abstract. *A Speech Corpus is required for acoustic modeling of Automatic Speech Recognition systems. However, the lack of such corpus for Brazilian Portuguese is a known problem already described in several works. This article describes, from a literature review, an alternative way consisting in the use of audio and Closed Caption from Brazilian Digital TV Broadcast programs as one potential source for the collection of spoken corpus.*

Resumo. *Um Corpus de Fala é necessário para modelagem acústica dos Sistemas de Reconhecimento de Fala. Entretanto a escassez deste tipo de corpus para o Português Brasileiro já é um problema descrito em vários trabalhos. Este artigo descreve, a partir de um levantamento bibliográfico, uma via alternativa, consistindo na utilização do áudio e Closed Caption dos programas transmitidos pelo Sistema Brasileiro de TV Digital (SBTVD) como fonte em potencial para coleta de corpus falado.*

1. Introdução

Os Sistemas de Reconhecimento de Fala (Automatic Speech Recognition – ASR, em inglês) atuais necessitam de uma grande quantidade de dados de treinamento. Para a tarefa de reconhecimento de fala contínua estes dados consistem em muitas horas de áudio e a transcrição de fala coletadas de dezenas a centenas de locutores. Estes sistemas ainda apresentam taxas de erro elevadas, sendo ainda um campo cheio de desafios e aberto à pesquisa. Uma das maneiras mais simples de se melhorar a acurácia dos sistemas de reconhecimento numa dada tarefa (por exemplo, reconhecimento em telefonia) é aumentar a quantidade de dados relevantes de treinamentos relacionados à tarefa, a partir do qual os modelos serão construidos [Baker 2009]. Entretanto vale ressaltar que a qualidade das transcrições tem papel importante, uma vez que a presença de falhas nas transcrições podem comprometer o modelo acústico [Jang 1999] e assim degradar a precisão do sistema no reconhecimento da fala.

Apesar de existirem alguns corpora para Português Brasileiro (PB), tais como West Point, Spoltech e OGI-22 estes ainda não são suficientes para o treinamento de sistemas de reconhecimento de fala contínua para PB [Neto et al. 2010]. Além da escassez de corpora, [Santos et al. 2010] aponta que há ainda problemas no que tange a qualidade dos que existem atualmente como erros de inconsistência, transcrições inexistentes, baixa frequência de alguns trifones e baixa qualidade de gravação.

Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology, pages 209–213,
Cuiabá, MT, Brazil, October 24–26, 2011.

Considerando algumas definições sobre o que é um corpus, [Sardinha 2004] destaca como um dos pontos importantes a necessidade de serem dados autênticos, que portanto não tenham sido produzidos com o propósito de serem alvo de pesquisa linguística – o que não se verifica com alguns dos corpora. Desta forma percebe-se que a falta de recursos para o PB, como corpora de fala é um obstáculo aos avanços de pesquisa na área. Devido aos altos custos envolvidos na produção de um corpora, o presente trabalho realizou um levantamento bibliográfico para encontrar soluções viáveis para este problema, apontando a utilização de legenda oculta (*Closed Caption – CC)* do SBTVD como uma possível fonte de coleta de corpus falado.

O restante do artigo organiza-se da seguinte forma: A Seção 2 descreve as características da legenda no Brasil. A Seção 3 apresenta a arquitetura proposta detalhando alguns passos, e a Seção 4 finaliza com as conclusões parciais e indicando trabalhos futuros.

2. Legendas Ocultas

Segundo [Araújo 2002], a legenda pode ser classificada do ponto de vista linguístico entre intralingual (destinada a telespectadores com deficiência auditiva, aprendizes estrangeiros) e interlingual (p.ex. a traduções de filmes de idioma estrangeiros); e do ponto de vista técnico entre aberta ou oculta (fechada). A legenda aberta é sobreposta ao vídeo antes da transmissão, sendo sempre exibida, enquanto que os códigos da legenda oculta são transmitidos separadamente – no intervalo vertical em branco (Vertical Blank Interval – VBI, em inglês) da transmissão analógica e em pacotes de fluxos distintos no caso do SBTVD. Desta forma a exibição da legenda oculta (quando disponível) depende da decisão do telespectador e do receptor ter a função CC.

No Brasil a legenda intralingual geralmente vem no formato de legenda fechada [Araújo 2002]. A legenda fechada intralingual tenta reproduzir toda a fala original aproximando-se de uma transcrição – o que a torna ideal para os objetivos de coleta de corpus de fala para treinamento de um ASR, diferentemente da legenda aberta interlingual em que muita condensação é feita.

Com relação à quantidade de horas de transmissão de programas contendo CC, há regulamentações governamentais e normatização ABNT que estabelecem metas de acessibilidade na transmissão de TV. As emissoras deverão transmitir no mínimo 6 horas de seu conteúdo diário acompanhado de CC com uma acurácia de 95%

Apesar de ser possível capturar o CC na transmissão analógica, optou-se pela TV digital por diversos motivos, tais como: a escassez de hardware com suporte à Closed Caption no sistema PAL-M; a transição gradual de TV analógica para a TV Digitial, menor presença de ruídos relativos ao canal de radiodifusão/conversão do sinal Analógico/Digital.

O método tradicionalmente utilizado para a produção dos CC é através de profissionais – estenotipistas que utilizam um estenótipo, um teclado computadorizado semelhante aos usados nos tribunais. Estes profissionais conseguem digitar até 160 palavras por minuto, sendo que uma das técnicas que os permite alcançar essa velocidade consiste em pressionar simultaneamente as teclas correspondentes à fonética aproximada da palavra [ARAÚJO 2002].

A partir de uma cooperação entre a emissora portuguesa RTP e o INESC-ID foi desenvolvido um sistema para produção em tempo-real das legendas do noticiário – o

sistema Audimus [Neto et al. 2008]. Como este sistema é voltado para o reconhecimento de Português Europeu (PE), houve a tentativa de portar este sistema para PB, porém não alcançou os resultados esperados [Abad et al. 2009]. Alternativamente a indústria apresentou um método que consiste em utilização de reconhecimento de voz, porém não aplicada diretamente ao áudio da programação da TV. Um locutor numa sala isolada ouve o que é dito na programação e repete num microfone. O sistema, dependente de locutor e sendo treinado por aproximadamente 40 horas, pode alcançar uma precisão de 95% [Panorama Audiovisual 2011]. Neste caso é interessante notar que as baixas taxas de erro destes sistemas dependentes de locutor com isolação de ruído ambiente, devido a sua menor complexidade, podem ser aproveitadas para produção de transcrições de alta qualidade de áudio proveniente de múltiplos locutores.

3. Arquitetura

A arquitetura geral proposta para contornar a escassez de Corpus de Fala é apresentada na Figura 1. Consiste em capturar, através de um receptor do sinal do SBTVD, extraindo seu áudio e CC, produzindo uma coletânea intermediária, que passará por um processo de alinhamento e verificação para que o corpus final contenha apenas as sentenças asseguradamente válidas.

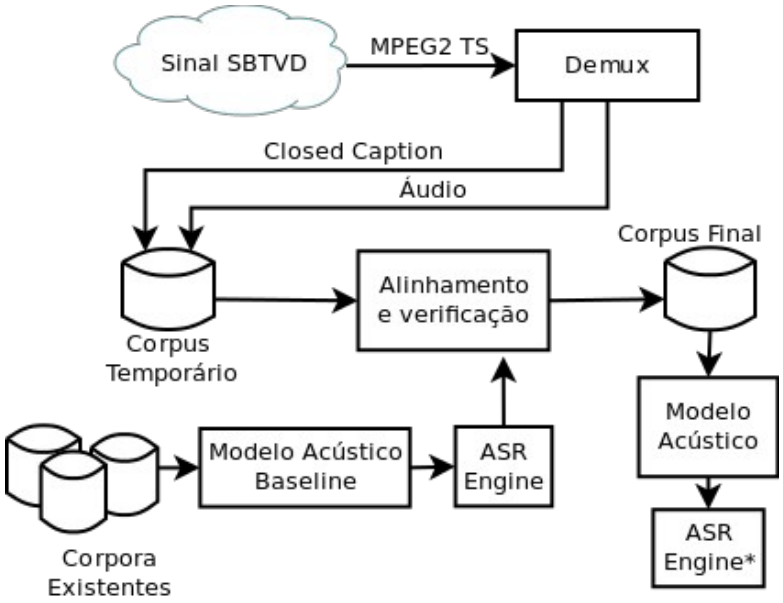

Figura 1: Arquitetura Geral

3.1 Demux

O Sinal SBTVD é transmitido por radiodifusão e captado por um receptor digital que pode ser *full seg* ou *1-seg*. No presente trabalho é empregado um receptor usb comercial *full-seg*, entregando ao computador ao qual está conectado um fluxo de bits MPEG2 TS, que contêm as trilhas de áudio (codificado em formato MPEG4 LC-AAC, encapsulado em LATM), vídeo (codificado em H.264), Closed Caption, entre outros dados multiplexados, como os referentes à interatividade provida pelo middleware Ginga. Estes formatos estão definidos nas normas da ABNT 15606-1 e ARIB STD B24.

Este fluxo MPEG2 TS é demultiplexado em *Elementary Streams* empacotados, extraindo-se apenas o canal de áudio, e o fluxo correspondente ao Closed Caption, que

pode também conter a informação de seus *timestamps* – marcações que situam temporalmente as legendas, e que facilitam o alinhamento.

3.2 Alinhamento e Verificação

A etapa de Alinhamento e Verificação tem como objetivo a produção de um corpus confiável, conforme exposto anteriormente, para evitar trechos onde o CC não corresponde ao áudio e que causariam efeitos contrário ao pretendido na modelagem acústica. Na literatura é possível encontrar trabalhos desenvolvidos sobre a produção de corpus de alta qualidade a partir de transcrições de CC [Jang 1999], [Lecouteux & Linares 2008], [Nguyen & Xiang 2004]. Para o alinhamento forçado, [Moreno 2009] descreve uma abordagem em casos onde há longos trechos de gravação de áudio e algumas ferramentas disponíveis para esta tarefa são apresentadas por [Knight 2010] e [Katsamanis 2011].

De acordo com a arquitetura proposta na Figura 1, a verificação automática da qualidade do CC (se realmente corresponde ao áudio) é realizada utilizando-se um modelo acústico inicial (*baseline*) como uma espécie de "semente". O áudio presente no Corpus temporário – que contém tanto as transcrições corretas como transcrições imprecisas, condensadas ou incompletas – é processado por este ASR treinado com o modelo acústico *baseline* e sua saída é comparada com o CC. Para esta tarefa é possível utilizar modelos acústicos pré-existentes, como o disponibilizados pelo projeto FalaBrasil [Neto et al. 2010].

3.3 Modelo Acústico Aprimorado

[Santos 2010] apresentou resultados que servem como *baseline* para sistemas de reconhecimento de fala para PB no qual foi utilizando o corpus West Point. [Jang 1999] demonstrou que mesmo para sistemas de reconhecimento já altamente otimizados, ainda é possível melhorar sua acurácia através da derivação de grandes quantidades de transcrições precisas. Desta forma espera-se que o novo corpus melhore a qualidade do modelo acústico e, consequentemente, melhore a acurácia em relação ao baseline apresentado por [Santos 2010] para o reconhecimento de PB.

4. Conclusões e Trabalhos Futuros

A partir do levantamento bibliográfico, a proposta de se utilizar o áudio e o CC, mostrou-se teoricamente viável, e empiricamente, através das coletas iniciais.

A etapa da demultiplexação e armazenamento do corpus temporário está sendo finalizada para que se avalie experimentalmente quais técnicas melhor se adequam na etapa de alinhamento e verificação. A coleta prosseguirá paralelamente a fim de produzir um corpus e experimentalmente verificar o impacto na taxa de erro dos sistemas de reconhecimento de fala. Pretende-se assim aprimorar o *baseline* de reconhecimento de fala contínua em PB desenvolvido e publicado por [Santos et al. 2010]

Espera-se como continuidade deste trabalho, desde que sanadas questões relativas a direitos autorais, seja possível criar um corpus de referência para PB. Esperamos também que seja de livre acesso, não apenas para a replicabilidade das pesquisas publicadas, como também motivar e dar continuidade à disseminação de tecnologias de fala em PB que vêm sendo desenvolvidas pelo Projeto FalaBrasil [Neto et al. 2010].

Referências

Abad, A., Trancoso, I., Neto, N. and Viana C. (2009). Porting an European Portuguese Broadcast News Recognition System to Brazilian Portuguese. *Interspeech 2009*, 92-95.

Araújo, V. L. S. (2002). O Processo de legendagem no Brasil. Revista do GELNE, Fortaleza, v. 1/2, n. 1, 156-159.

Baker, J. M., Deng, L., Glass, J., Khudanpur, S., Lee, C.-H., Morgan, N., and O'Shaughnessy, D. (2009). Research Developments and Directions in Speech Recognition and Understanding, Part 1. *IEEE Signal Processing Magazine*, (May), 75-80.

Jang, P. J. and Hauptmann A. G. (1999). Improving acoustic models with captioned multimedia speech. *Proceedings IEEE International Conference on Multimedia Computing and Systems*, 767-771. IEEE Comput. Soc.

Katsamanis, A., Black, M., Georgiou, P., Goldstein, L., and S. Narayanan, (2011) *SailAlign: Robust long speech-text alignment.* Proc. of Workshop on New Tools and Methods for Very-Large Scale Phonetics Research.

A. Knight and K. Almeroth, (2010) Fast Caption Alignment for Automatic Indexing of Audio. International Journal of Multimedia Data Engineering & Management (IJMDEM), vol. 1, no. 2, 1-17.

Lecouteux, B., and Linares, G. (2008). Using prompts to produce quality corpus for training automatic speech recognition systems. *MELECON 2008 - The 14th IEEE Mediterranean Electrotechnical Conference*, 841-846.

Neto, J. P. S., Meinedo H., Viveiros, M., Cassaca, R. M. F., Martins, C. A. D., and Caseiro, D. A., (2008) Broadcast News Subtitling System in Portuguese. ICASSP 2008 - Int. Conf. on Acoustics, Speech, and Signal Processing, IEEE, Las Vegas, USA.

Neto, N., Patrick, C., Klautau, A., and Trancoso, I. (2010). Free tools and resources for Brazilian Portuguese speech recognition. *Journal of the Brazilian Computer Society*, *17*(1), 53-68.

Nguyen, L. and Xiang, B. (2004). Light supervision in acoustic model training. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, I-185-8.

Moreno, P. J. and Alberti, C. (2009) *A factor automaton approach for the forced alignment of long speech recordings*. Proceedings of ICASSP. 2009, 4869-4872.

Panorama Audiovisual, Redação (2011). Acessibilidade na TV. Disponível em http://www.panoramaaudiovisual.com.br/2011/06/21/acessibilidade-na-tv/ . Último acesso em 27 de junho de 2011.

Sardinha T. B. (2004) Linguística de Corpus. Barueri/SP. Editora Manole.

Santos, F. W., Barone, D. A. C., and Adami, A. G. (2010). A Baseline System for Continuous Speech Recognition of Brazilian Portuguese Using the West Point Brazilian Portuguese Speech Corpus. *Computational Processing of the Portuguese Language, 6001*, 132–141. Springer.