Speech Recognition for Brazilian Portuguese using the Spoltech and OGI-22 Corpora

Patrick Silva, Nelson Neto, Aldebaro Klautau, Andre Adami and Isabel Trancoso

***Abstract*— Speech processing is a data-driven technology that relies on public corpora and associated resources. In contrast to languages such as English, there are few resources for Brazilian Portuguese (BP). This work describes efforts toward decreasing such gap and presents systems for speech recognition in BP using two public corpora: Spoltech and OGI-22. The following resources are made available: ATK and HTK scripts, pronunciation dictionary, language and acoustic models. The work discusses the *baseline* results obtained with these resources.**

Keywords— Speech recognition, Brazilian Portuguese, HMMs, pronunciation dictionary.

I. Introduction

Speech processing is a data-driven technology and researchers rely on public corpora and other speech-related resources to expand the state of the art. Three major factors that drive the speech processing community are: a) public corpora distributed by institutions such as the Linguistic Data Consortium (LDC) [1] and the Center for Spoken Language Understanding (CSLU) of the OGI School of Science and Engineering (OHSU) [2]; b) public and in some cases, free software with recipes for building baseline systems: HTK [3] (in C language), Sphinx 4 [4] (Java), ISIP [5] (C++), Festival [6], etc.; c) evaluation campaigns organized for specific tasks, such as the ones organized by National Institute of Standards and Technology [7] for speech and speaker recognition.

Automatic speech recognition (ASR) for BP has been investigated in several previous works [8]–[14]. To the best of the authors' knowledge, the resources are not publicly available. In fact, in contrast to languages such as English, there are very few public resources for ASR in BP. Recently, LDC released the catalog number LDC2008S04 [1], the West Point Brazilian Portuguese Speech, a read speech database of microphone digital recordings from native and non-native speakers. Besides, there are no publicly available scripts (or software *recipes*) to design BP baseline systems. These recipes considerably contribute towards shortening the development process.

This work discusses current efforts within the *FalaBrasil* initiative [15]. The overall goal is to develop and deploy resources and software for BP, aiming to establish baseline systems and allow for reproducing results across different sites. More specifically, the work presents resources and results for two baseline systems using the Spoltech and OGI-22 corpora. All corrected transcriptions and resources can be found in [15].

This paper is organized as follows. Section II describes the creation of a pronunciation dictionary for BP. Section III shows how the language model (LM) was built using the CETEN-Folha text corpus. Section IV describes the adopted front-end and HMM-based acoustic modeling. Section V discusses the speech corpora OGI-22 and Spoltech. Section VI presents the baseline results and Section VII concludes and suggests further research.

Laboratório de Processamento de Sinais - LaPS, Faculdade de Engenharia da Computação, Universidade Federal do Pará, Belém - PA, Brazil, Universidade de Caxias do Sul, Rua Francisco Getulio Vargas. 1180, 95070-560 Caxias do Sul, RS, Brazil and INESC, R. Alves Redol, 9, 1000 Lisbon, Portugal. E-mails: {patrickalves, nelsonsampaio, aldebaro}@ufpa.br, agadami@ucs.br, isabel.trancoso@inesc-id.pt.

II. UFPAdic: a pronunciation dictionary for BP

An important prerequisite for services involving ASR and/or speech synthesis is the information about the correspondence between the orthography and the pronunciation(s). For example, the development of a *large vocabulary continuous speech recognition* (LVCSR) system for BP, requires a *pronunciation* (or phonetic) dictionary, which maps each word in the lexicon to one or more phonetic transcriptions (pronunciation). Building a pronunciation dictionary for ASR is very similar to developing a grapheme-to-phoneme (G2P) module for text-to-speech (TTS) systems [16]–[18].

In [19], a two-step self learning approach that automatically derives algorithms for G2P conversion from training data was adopted. In the first step, corresponding grapheme and phoneme strings in the training data are aligned according to the method described in [20]. Lexicon alignment is an important and critical step of the whole training scheme of such G2P systems, as it gathers the data on which the learning methods extracts the transcription rules. This alignment can be done manually, but this is a time-consuming, error-prone task, and limits the size of datasets that can be used for training. In the second step, the Weka machine learning tool [21] was used to build a J4.8 decision tree classifier [22]. A sliding grapheme window moves over the word. The window takes into account a subsequence of the word including a focus (the central grapheme to be transcribed). For example, using a context of 1, means that a sliding window passes three graphemes (1 left + 1 focus + 1 right) to the classifier and obtains the phoneme (that could be a null symbol) corresponding to the focus grapheme.

In [19], a hand-labeled pronunciation dictionary *UFPAdic version 1* with 11,827 words in BP was released within the *FalaBrasil* initiative. The phonetic transcriptions adopted a modified version of SAMPA alphabet [15] and were validated by comparing results with other publicly available pronunciation dictionaries for other languages: NETtalk, 20,008 words,

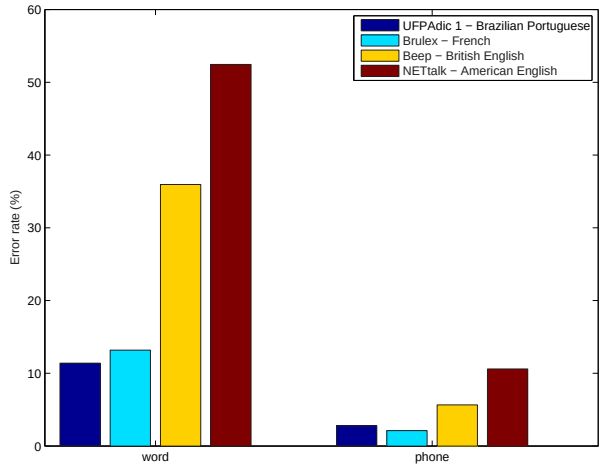

Fig. 1. A comparison of the performance of the dictionaries for context of 3 (total of 7 graphemes) and a decision tree learning algorithm [19]. The groups "word" and "phone" indicate the misclassification rate at the levels of words and phones, respectively.

American English [23]; Brulex, 27,473 words, French [24] and Beep 1.0, 256,980 words, British English [25]. UFPAdic 1 is the smallest among them, but its size is comparable to the dictionaries used in other recent studies (e.g., [18]). The validation results are summarized in Fig. 1, which shows the misclassification rate obtained for context of 3 (total of 7 graphemes) and a decision tree learning algorithm. The entries of a given dictionary were split into two disjoint sets for training and test. More details can be found in [19] and [26].

Using the whole UFPAdic 1 for training a decision tree and adopting the procedure described in [19], a new dictionary was built by selecting the most frequent words in CETENFolha corpus [27]. The new dictionary called UFPAdic 2, has approximately 60 thousand words. For the words that are not part of UFPAdic 1, there is no "ground-truth" because hand-labeling is too time-consuming. In this case, the goal is to validate the pronunciations obtained with the decision tree by analyzing the results of ASR experiments using UFPAdic 2.

III. Building language models from CETENFolha

The CETENFolha is a corpus of about 24 million words in BP, based on the texts of the newspaper *Folha de S. Paulo* and compiled by NILC/São Carlos, Brazil. The original corpus was adapted for ASR use with the HTK and ATK packages. Some examples of the formatting operations are:

- Removal of punctuation marks and tags ([ext], [t], [a] and others).
- Conversion to lowercase letters.
- Expansion of numbers and acronyms.
- Correction of grammatically incorrect words.

An example of the result of these operations is given below:

```
Before: O Senado tem uma <<caixa preta>>
de R\$ 2 milhões
After: o senado tem uma caixa preta de
dois milhões de reais
```

TABLE I

LM perplexities for different training set sizes.

	Number of sentences used to train the LM							
	10k	25k	50k	75k	100k	125k	150k	180k
Bigram	488	443	421	420	410	408	399	386
Trigram	446	403	373	367	355	353	343	326

O SRI Language Modeling Toolkit (SRILM) was used to build the *n*-gram ARPA format language models. The SRILM [28] is a toolkit for building and applying statistical language models. This software also enables the use of many *n*-gram smoothing implementations.

This experiment evaluates the LM perplexity against the number of sentences used to train it. The vocabulary used was kept constant during all the experiment and contains 3,285 words found on the 2,226 sentences from the OGI-22 corpus. The number of sentences used to train the LM ranged between 10,000 and the 180,000 from the CETENFolha corpus and the language models tested were the bigram and the trigram with Kneser-Ney smoothing.

The sentences used to measure the perplexity of each configuration were the 2,226 sentences from the OGI-22 corpus. As expected, the perplexity tends to diminish as the number of sentences used in the training increases [14]. This is related to the fact that the statistics of the trained models get improved as more occurrences of pairs and triples of words are registered in the CETENFolha corpus. The Table I shows the perplexities found in these experiments by counting all input tokens.

IV. Front-end and acoustic modeling

The preparation of the data is an essential stage of any developing speech recognizer project. Two different data sets are required: digitized voice (corpora) and transcribed at the level of words and/or the level of phonemes. This research is proposed to provide details for the development of resources that are specific to BP using the software HTK.

The front-end consists of the widely used 12 mel-frequency cepstral coefficients (MFCCs) using C0 as the energy component, appended with delta and acceleration coefficients, and computed every 10 milliseconds (i.e., 10 ms is the frame shift) for a frame of 20 ms. These static coefficients are augmented with their first and second derivatives to compose a 39-dimensional parameter vector per frame.

The acoustic models were iteratively refined [29]. Starting with continuous single mixture monophone models, the HMMs were gradually expanded to compose a multiple mixture output distributions and tied-state triphone system. The initial acoustic models for the 33 phones (32 monophones and a silence model) used 3-state left-to-right HMMs. The silence model was trained and then copied to create the tied short pause *tee* model with only one acoustic state [30]. Then, it was adopted the flat-start approach and re-estimation using the embedded Baum-Welch to train the monophones.

After that, triphone models were built from the monophone models. Both *word-internal*, where the context beyond the borders of the words are not considered, and *cross-word*, which takes into account the co-articulation effects between the

words boundary, were tested. Each triphone was cloned from the respective monophone that constitutes its base phoneme. The transition matrices of triphones that share the same base phoneme were tied and the triphones models were re-estimated using the embedded Baum-Welch algorithm.

A classic problem of the context-dependent models is insufficient training data to support a large quantity of triphones [31]. To circumvent this problem, tying (or sharing) parameters is crucial. Given a set of categories (also called *questions* (QS) [3]), a decision tree was designed for tying triphones with similar characteristics. An existing set of categories developed for the resource management task [3] was adapted to the used alphabet. For example, some categories used in the decision tree construction are shown below. The first QS command defines a question called *R_V-Fechada* which is true if the right context is either of the phones i, e, o, or u. This set of questions has also been made available.

```
...
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-* }
...
```

Notice that for a triphone system, it is necessary to include questions referring to both the right and left contexts of a phone. The questions should progress from wide, general classifications (such as consonant, vowel, nasal, diphthong, etc.) to specific instances of each phone. Ideally, the full set of questions loaded using the QS command would include every possible context which can influence the acoustic realization of a phone, and can include any linguistic or phonetic classification which may be relevant.

After tying, the triphone models were again reestimated using the Baum-Welch algorithm.

V. Speech Corpora

This section discusses the corpora OGI-22 and Spoltech. Also, it describes the training and test sets used for both corpora and the corrections made on their original orthographic transcriptions.

A. OGI-22 Corpus

The 22 Language Telephone Speech Corpus [32], which includes Brazilian Portuguese, is an effort of CSLU/OHSU [2]. This spontaneous speech corpus is very useful for the research of speech systems, despite its relatively small size. This telephone recordings database contains 2,500 files, but only about 100 files have orthographic transcriptions and there is no phonetic transcriptions.

A protocol was developed in English and then translated and recorded by native speakers for the other 21 languages. The protocol includes prompts which elicit a total of two to three minutes of speech from each caller. The responses fall into three categories: (a) Requests for specific information, such as age and gender; (b) Spontaneous speech on selected topics about the local climate, route to get to work, last meal, among others; (c) Extemporaneous speech with free topic, one necessarily in English and the other in the native language of the speaker. The orthographic transcriptions accuracy is not 100% guaranteed. Informal analysis made during this research, indicate a high level of inconsistency.

TABLE II
Examples of corrections in the OGI-22 transcriptions.

Before	After
aqui eu em inglês	português e inglês
uh se for como assim matson e	uh se for wisconsin matson e
sexta feira	terça feira
meu nome ursolandes gostaria	meu nome ursula landsy gostaria
em vinheiro caldas minas gerais	engenheiro caldas minas gerais

Initially, all the audio files and orthographic transcriptions of the corpus were verified. When necessary, the original orthographic transcriptions were corrected, and the nonexistent created. The pronunciation dictionary was updated to include all the words in OGI-22. The Table II shows some examples of transcriptions that had to be corrected.

For the experiments, the audio files in English and those which had poor recording quality were not used. So the OGI-22 training set was composed of 2,017 files, corresponding to 184.5 minutes, and the test set had 209 files with 14 minutes.

B. Spoltech Corpus

The Spoltech corpus [33] was created by the Universidade Federal do Rio Grande do Sul, Brazil, Universidade de Caxias do Sul, Brazil, and OHSU, USA, with funding from CNPq, Brazil and NSF, USA. The corpus has been distributed by LDC [1] (LDC2006S16) and CSLU/OHSU [2] (the version used in this work is the latter). It consists of waveform speech (WAV), orthographic (TXT) and phonetic transcriptions (PHN) files.

The utterances consist of both read speech (for phonetic coverage) and responses to questions (for spontaneous speech) from a variety of regions in Brazil. The acoustic environment was not controlled, in order to allow for background conditions that would occur in application environments. Although useful, Spoltech has several problems. Some WAV files do not have their corresponding TXT and PHN files, and vice-versa. Another problematic aspect is that both phonetic and orthographic transcriptions have many errors.

For this work, a pre-processing stage tried to find the files that have good quality (understandable utterances without much noise or artifacts) and 7,246 wav files were selected. These files were split into two disjoint sets, for training and test. Care was exercised to avoid having a given speaker participating in both sets.

In the Spoltech corpus, 183 different symbols can be found. These symbols are part of the Worldbet phonetic alphabet, which is adopted at OGI. Many of these symbols have only few occurrences and some of them are not valid Worldbet symbols (probably typos). For comparison purpose, it is interesting to note that the popular TIMIT corpus was annotated using a relatively narrow phonetic transcription. Hence, in

TABLE III
PARAMETERS USED FOR DESIGNING AND TESTING THE OGI-22 BASELINE SYSTEMS.

Parameter	Value	
	Bigram	Trigram
Pruning beam width	250	250
Max model pruning	150	150
Word end beam width	1000	1000
Decision tree outlier threshold	100	100
Decision tree termination threshold	430	430
Word insertion penalty	10	10
Language model scale factor	0	15
Grammar network scale factor	15	5
Number of tokens	1	20

TABLE IV
PARAMETERS USED FOR DESIGNING AND TESTING THE SPOLTECH BASELINE SYSTEMS.

Parameter	Value	
	Bigram	Trigram
Pruning beam width	250	250
Max model pruning	150	150
Word end beam width	1000	1000
Decision tree outlier threshold	100	100
Decision tree termination threshold	600	600
Word insertion penalty	25	25
Language model scale factor	0	0
Grammar network scale factor	21	21
Number of tokens	1	20

most ASR experiments the 61 TIMIT symbols are collapsed into 39 classes for scoring purposes [34]. This clearly indicates that the original number of symbols used in Spoltech is too large for ASR.

The phonetic alphabet used here was the same as the one used in the OGI-22 corpus, a total of 32 phones and a silence model. In the experiments, the Spoltech training set was composed by 5,246 files that corresponding to 180 minutes and the test set used the remaining 2,000 files corresponding to 40 minutes.

VI. BASELINE RESULTS

The Spoltech and OGI-22 baseline systems share the same front-end. In addition, the HMM-based acoustic models of both systems were estimated using the same procedure described in Section IV.

For the accomplishment of the tests and evaluate the acoustic modeling, it was used the AVite software, available in ATK package, version 1.6 [35]. The first experiments showed that AVite does not support energy normalisation. The variable ENORMALISE is by default true and performs energy normalisation on recorded audio files. It cannot be used with live audio and since the target system is for live audio, this variable should be set to false.

As described above, the training and test stages require selecting several parameters, such as word insertion penalty, pruning threshold and grammar scale factor, which have a significant impact on performance and computational complexity. Hence, it is necessary to properly tune these parameters to achieve the best results. The results for triphones obtained with word-internal were better than cross-word, this fact is observed by few data existing in the training set. Several tests were conducted and the best parameters for both OGI-22 and Spoltech tied-state word-internal triphone baseline systems are described in Table III and Table IV, respectively.

A. Results for n-gram LMs from corpora transcriptions

To evaluate the acoustic modeling, the first experiments used simplified *n*-gram language models, designed solely with the corpora transcriptions. For example, both the training and test orthographic transcriptions of the OGI-22 corpus were used to design a LM specific to this corpus. Similar procedure was adopted to created the Spoltech LM. Section VI-B describes results obtained with more general language models that include text from the CETENFolha corpus.

Initially, the Good-Turing and Kneser-Ney smoothing methods were evaluated. Both techniques resulted bad perplexities values when tested with simplified bigram language models, designed solely with the OGI-22 corpora transcriptions. This happened maybe because they rely on statistics called "count-of-counts", the number of words occurring *n* times [36]. The formulae for these methods become undefined if the counts-of-counts are zero, or not strictly decreasing. Some conditions are fatal (such as when the count of singleton words is zero), others lead to less smoothing.

To avoid these problems, since OGI-22 and Spoltech training corpus are small, the Witten-Bell interpolate discounting method was tested. The intuition is that the weight given to the lower order model should be proportional to the probability of observing an unseen word in the current context. This is the estimator where the first occurrence of each word is taken to be a sample for the "unseen" event. As expected, the Witten-Bell method worked better than the two other methods.

The first experiment used a OGI-22 bigram LM with Witten-Bell interpolate discounting and perplexity equal to 24. The same way, a OGI-22 trigram LM with Witten-Bell interpolate discounting and perplexity 18 was designed. The number of component mixture distributions was gradually increased from one to ten. The word error rate (WER) reduction can be observed in Fig. 2. The WER with 10-component Gaussian mixtures is 23.22% and 20.40% for bigram and trigram, respectively.

Similarly, *n*-gram LMs with 1,104 words and perplexity 5 for bigram and 4 for trigram, were designed using only the text of the selected Spoltech 7,246 phrases. The respective WER results are shown in Fig. 3, where the number of Gaussians per mixture was varied from 1 to 16. The WER with 14-component Gaussian mixtures is 6.13% and 5.27% for bigram and trigram, respectively. The experiments finished with 16-component Gaussian mixtures, because the WER stopped to decline.

B. Results with language models from CETENFolha

The language models mentioned in Section III were used to test the system, in order to evaluate the WER with respect to the LM. Simulations were performed setting the acoustic

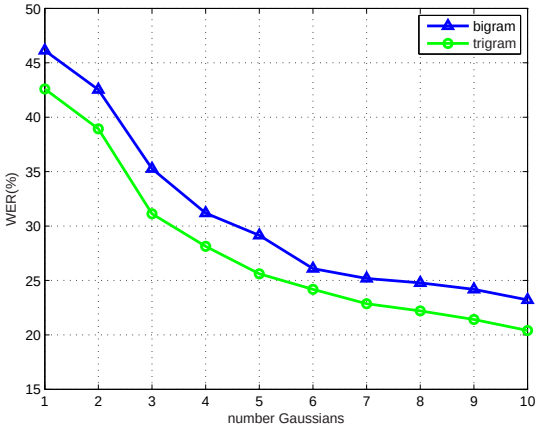

Fig. 2. Decrease in WER (%) with the number of Gaussians in each mixture for OGI-22 using simplified *n*-gram LMs.

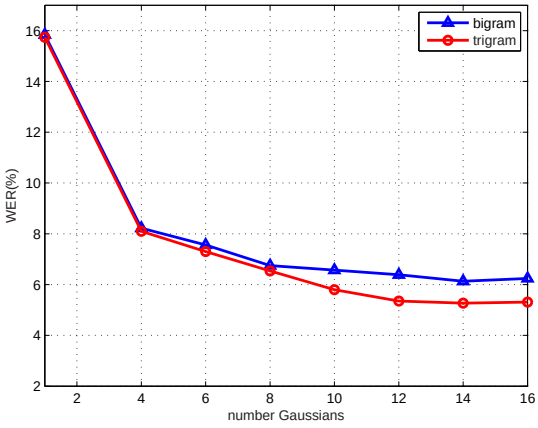

Fig. 3. Decrease in WER (%) for Spoltech using simplified *n*-gram LMs.

model created with the OGI-22 corpus and the number of Gaussians per mixture equal to ten. The results are shown in Fig. 4. Notice that the WER declines as the number of sentences used in the training increases.

On the executed experiments, the results remained nearly constant in some intervals. The reason for that could be related to a saturation of the language model, with almost all common *n*-gram sequences already appear in the language model, but rare ones are still unlikely to be seen in the training corpus. However, these uncommon *n*-grams are the ones whose probability is the hardest to estimate correctly, so adding small quantities of new data does not correspondingly improve the language model.

The best WER accuracy rate obtained with 180,000 sentences was 47.39% for trigram system. It is important to observe that there is no intercession between the OGI-22 and the CETENFolha corpus.

VII. Conclusions

This paper presented baseline results for ASR in BP. The resources were made publicly available and allow for reproducing results across different sites. It is clear that the OGI-22 and Spoltech corpora are too small for developing large-vocabulary ASR systems in BP. However, the strategy is to emphasize the creation of necessary resources even if they are not the ideal ones in terms of coverage, for example. This way the community can gradually improve aspects such as pronunciation dictionary and language model. Future work should concentrate efforts in collecting a larger corpus with broadcast news to put together a baseline system using a trigram language model, cross-word triphone models and an accurate pronunciation dictionary.

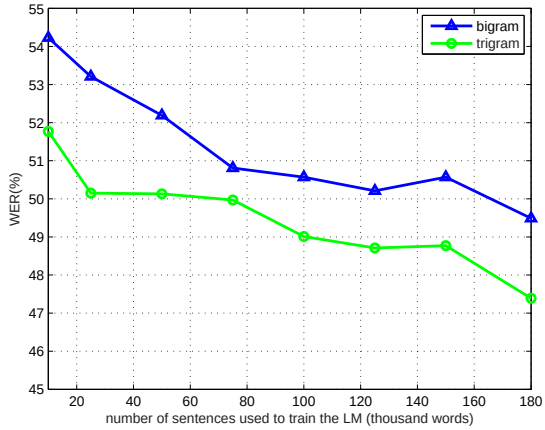

Fig. 4. WER (%) for OGI-22 using *n*-gram LMs from CETENFolha corpus.

Acknowledgements

This work was partially supported by CNPq, Brazil, project 478022/2006-9 *Reconhecimento de Voz com Suporte a Grandes Vocabulários para o Português Brasileiro: Desenvolvimento de Recursos e Sistemas de Referência*.

References

[1] "http://www.ldc.upenn.edu." Visited in March, 2008.
[2] "http://cslu.cse.ogi.edu/corpora," Visited in March, 2008.
[3] "http://htk.eng.ac.uk," Visited in March, 2008.
[4] "http://cmusphinx.sourceforge.net/sphinx4/," Visited in March, 2008.
[5] "http://www.isip.msstate.edu," Visited in March, 2008.
[6] "http://www.cstr.ed.ac.uk/projects/festival," Visited in March, 2008.
[7] "http://www.nist.gov/speech," Visited in March, 2008.
[8] 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.
[9] 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.
[10] 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.
[11] 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.
[12] M. Schramm, L. Freitas, A. Zanuz, and D. Barone, "A brazilian portuguese language corpus development," *ICSLP-2000, vol.2, 579-582*, 2000.
[13] 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.

[14] 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.
[15] "http://www.laps.ufpa.br/falabrasil," Visited in April, 2008.
[16] D. Caseiro, I. Trancoso, L. C. Oliveira, and M. do Céu Guerreiro Viana Ribeiro, "Grapheme-to-phone using finite-state transducers," in *In IEEE Workshop on Speech Synthesis*, 2002.
[17] P. Barbosa, F. Violaro, E. Albano, F. Simões, P. Aquino, S. Madureira, and E. Françozo, "Aiuruetê: a high-quality concatenative text-to-speech system for brazilian portuguese with demisyllabic analysis-based units and hierarchical model of rhythm production," in *Proceedings of the Eurospeech'99, Budapest, Hungary*, 1999, pp. 2059–2062.
[18] A. Teixeira, C. Oliveira, and L. Moutinho, "On the use of machine learning and syllable information in european portuguese grapheme-phone conversion," in *7th Workshop on Computational Processing of Written and Spoken Portuguese (to be presented) - Itatiaia, Brazil*, 2006.
[19] C. Hosn, L. A. N. Baptista, T. Imbiriba, and A. Klautau, "New resources for brazilian portuguese: Results for grapheme-to-phoneme and phone classification," *In VI International Telecommunications Symposium, Fortaleza*, 2006.
[20] R. I. Damper, Y. Marchand, J. Marsters, and A. Bazin, "Aligning letters and phonemes for speech synthesis," in *5th ISCA Speech Synthesis Workshop - Pittsburgh*, 2004, pp. 209–214.
[21] "http://www.cs.waikato.ac.nz/ml/weka/," Visited in March, 2008.
[22] I. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*. Morgan Kaufmann, 1999.
[23] T. J. Sejnowski and C. R. Rosenberg, "Parallel networks that learn to pronounce english text," *Complex Systems*, vol. vol. 1, pp. 145–168, 1987.
[24] A. Content, P. Mousty, and M. Radeau, "Brulex: Une base de données lexicales informatisée pour le français éecrit et parlé," *L'Année Psychologique*, pp. 551–566, 1990.
[25] "ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries," Visited in March, 2008.
[26] C. Hosn, "Conversão grafema-fone para um sistema de reconhecimento de voz com suporte a grandes vocabulários para o português brasileiro," Master's thesis, Universidade Federal do Pará, Centro Tecnológico, 2006.
[27] "http://acdc.linguateca.pt/cetenfolha/," Visited in January, 2008.
[28] A. Stolcke, "Srilm - an extensible language modeling toolkit," *Proc. Intl. Conf. Spoken Language Processing, Denver, Colorado*, 2002.
[29] P. Woodland and S. Young, "The htk tied-state continuous speech recognizer," *In: Proc. Eurospeech'93, Berlim*, 1993.
[30] S. Young, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department, 2006.
[31] P. Woodland, J. Odell, V. Valtchev, and S. Young, "Large vocabulary continuous speech recognition using htk," *IEEE Int'l Conf. Acoustics, Speech, and Signal Processing, vol. 2, pp. 125–128, Adelaide*, 1994.
[32] T. Lander, R. Cole, B. Oshika, and M. Noel, "The ogi 22 language telephone speech corpus," *In: Proc. Eurospeech'95, Madrid*, 1995.
[33] "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.
[34] K.-F. Lee and H.-W. Hon, "Speaker-independent phone recognition using hidden Markov models," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 37, no. 11, pp. 1641–8, Nov. 1989.
[35] S. Young, *ATK - An Application Toolkit for HTK (Version 1.6)*. Cambridge University Engineering Department, 2007.
[36] S. F. Chen and J. Goodman, *An Empirical Study of Smoothing Techniques for Language Modeling*. Computer Science Group, Harvard University, 1998.