INTERSPEECH 2009 BRIGHTON

Porting an European Portuguese Broadcast News Recognition System to Brazilian Portuguese

Alberto Abad[1], *Isabel Trancoso*[2], *Nelson Neto*[3], *M.Céu Viana*[4]

[1]INESC-ID Lisboa, Portugal
[2]IST / INESC-ID Lisboa, Portugal
[3]Federal University of Pará, Belém, Brazil
[4]Center of Linguistics of the University of Lisbon, Portugal
Alberto.Abad@inesc-id.pt

Abstract

This paper reports on recent work in the context of the activities of the PoSTPort project aimed at porting a Broadcast News recognition system originally developed for European Portuguese to other varieties. Concretely, in this paper we have focused on porting to Brazilian Portuguese. The impact of some of the main sources of variability has been assessed, besides proposing solutions at the lexical, acoustic and syntactic levels. The ported Brazilian Portuguese Broadcast News system allowed a drastic performance improvement from 56.6% WER (obtained with the European Portuguese system) to 25.5%.
Index Terms: speech recognition, accent

1. Introduction

The goal of the PoSTPort project is porting spoken language technologies originally developed for European Portuguese to other varieties of Portuguese. Although ideally we would like to cover all the varieties spoken in CPLP countries (Community of Portuguese-speaking Countries), we had to exclude East Timor from our list, because of the difficulties in collecting corpora from this variety. The project therefore covers only 3 broad varieties:

- European Portuguese (henceforth denoted as EP), the variety spoken in Portugal, for which all the speech technologies available at the Spoken Language Systems Group of INESC-ID were originally developed.
- Brazilian Portuguese (henceforth denoted as BP), the variety spoken in South America, with the largest number of speakers.
- African Portuguese (henceforth denoted as AP), the generic name that covers all the varieties spoken in African countries that have Portuguese as official language (PALOP countries): Angola (AN), Cape Verde (CV), Guinea-Bissau (GB), Mozambique (MO) and São Tomé and Príncipe (ST).

The current paper describes our porting efforts for Broadcast News (BN) recognition. Although these efforts were conducted in parallel for BP and AP, for the sake of space, we shall describe only the BP work. The paper is structured into 3 main sections. Section 2 describes the corpora used in these experiments. Section 3 summarizes our porting efforts in terms of the grapheme-to-phone conversion module (GtoP), which is essential not only for the generation of the lexical model of recognition systems, but also for synthesis systems. Section 4 describes our porting efforts in terms of automatic speech recognition (ASR). For the sake of completeness, we shall briefly describe our baseline modules developed for EP.

2. Corpora

The amount of training data can drastically influence the performance of an automatic speech recognition system. This justifies the inclusion of a brief description of our EP and BP Broadcast News corpora.

2.1. European Portuguese

The original EP corpus involves different types of news shows, national and regional, from morning to late evening, including both normal broadcasts and specific ones dedicated to sports and financial news. The corpus was collected in 2000/2001 and is divided into 2 main subsets:

- SR (Speech Recognition) - The SR corpus contains around 61h of manually transcribed news shows, collected during a period of 3 months, with the primary goal of training acoustic models and adapting the language models of our large vocabulary speech recognition component of our system. The corpus is subdivided into training (51h), development (6h), and evaluation sets (4h).
- JE (Joint Evaluation) - The JE corpus contains around 13h, corresponding to the last two weeks of the collection period. It was fully manually transcribed, both in terms of orthographic and topic labels.

2.2. Brazilian Portuguese

Three corpora were identified for Brazilian Portuguese: OGI-22 [1], Spoltech [2], and West Point. The last two were available from LDC[1]. The OGI-22 corpus consists of telephone speech from 22 languages, amounting to a total of around 200 minutes of speech, for Portuguese. The Spoltech corpus contains microphone recordings (read speech and spontaneous answers to questions) from 477 speakers. A pre-selection of the usable files resulted in around 3h 40min [3]. The West Point Brazilian Portuguese Speech corpus consists of read speech from 128 native and non-native speakers. A significant part of this subset, however, was distributed with empty audio files, amounting to 8h of usable audio [4].

[1]http://www.ldc.upenn.edu

The fact that none of these corpora were specific for Broadcast News prompted the collection of a BN corpus recorded from the Record channel transmitted by cable TV in Portugal. We recorded several broadcast news and debates. The corpus was divided into training (851.4 min.), development (102.4 min.), and test (106.6 min.) subsets. Manual annotation at the orthographical level was done using the Transcriber tool[2], using standard guidelines. The manual transcriptions of the BP TV corpus included around 131k words for training, 15k words for development, and 18k words for testing. All the experiments described here used only this BN corpus, for the time being.

2.3. Newspaper text corpora

The language models of BN recognition systems are typically built using interpolated models of transcriptions and newspaper texts. For EP, we used a newspaper corpus of approximately 750 million words. For BP, our initial newspaper corpus was CETENFolha (Corpus de Extractos de Textos Electrónicos NILC/Folha de S. Paulo), a corpus with about 24 million words, created by the Linguateca project [3] with texts from the newspaper *Folha de S. Paulo* (from 1994) that were compiled by the NILC research center in Brazil.

Since the CETENFOLHA corpus is not so recent, its use as a source for language modeling for a task like Broadcast News subtitling was complemented with recent newspaper corpora. This was done by fully automatizing the collection of three daily newspapers, available from the Internet: Diário do Nordeste, O Globo, and Zero Hora. Approximately four months of automatically collected newspaper data were available at the time of building the first language model for BN speech recognition, which was insufficient for training purposes. In order to increment the amount of text data available, we downloaded and processed the archives of the Diário do Nordeste available from the Internet since September 2002.

3. Grapheme-to-phone conversion

3.1. Baseline GtoP

Although several GtoP modules were available for EP, we choose to port the one that was based on the Weighted Finite State Transducers (WFSTs) framework [5], because of its flexibility. In this baseline system, rules are organized in various phases [4], each represented by transducers that can be composed to build the full system.

The first phase is the stress phase, which consists of 27 rules that mark the stressed vowel of the word. The second phase is just the simple rule that inserts an empty phone placeholder after each grapheme. The third phase consists of pronunciation rules for compound words, namely with roots of Greek or Latin origin such as ”tele” or ”aero”. It includes 92 rules. The fourth and main phase consists of 340 rules, that convert the graphemes (including graphically stressed versions of vowels) to phones. The fifth phase implements word co-articulation rules across word boundaries. Finally, the last phase removes the graphemes in order to produce a sequence of phones.

The performance of this WFST-based GtoP module was compared to the one of our original rule-based approach on a pronunciation lexicon that was randomly selected from PF (“Português Fundamental”), amounting to about 6k words. The performance was slightly inferior, which we attributed to the fact that the exception lexicon was not originally included in the WFST approach. Once included, the performance was the same: 3.25% error at word level, and 0.50% at phone level.

3.2. Porting to BP

Several GtoP rule sets for BP have been described in the literature [6] [7] [8], including the generation of pronunciation variants for speech recognition [9].

Our efforts in terms of porting the WFST-based GtoP module were concentrated on the fourth phase, as expected. Prior to that, the grapheme set had to be augmented with the symbol ü, and the phoneme set with three extra symbols. The two affricate symbols for [tʃ] and [dʒ] could have been considered as a sequence of two phonetic symbols as well, but the chosen alternative was more convenient from programming and linguistic points of view. An additional SAMPA symbol “x” was also adopted to take into account the different realization of “r’s” in coda position most commonly found in our BP data.

The number of rules derived for BP was smaller than for EP (around 250). Over 30 EP rules are not present in the BP set because of the adoption of a different orthography. The remaining differences could be mostly found in the extra rules for pronouncing graphemes *a, e, o* in EP, which were not yet counterweighted by a full implementation of vowel harmony rules in BP. It was very interesting to notice that the type of errors, however, was very similar.

Additionally, multiple pronunciations were automatically added in order to take into account some of the different variations that can be obtained as a result of word co-articulation rules. Some examples are shown below. Other pronunciation variants could have been added, accounting for instance for the deletion of the [r] in coda position. This step was never performed for EP, where multiple pronunciations were only added manually to words as pronounced in isolation.

```
carros   k a R u s    casar   k a z a r
carros   k a R u z    casar   k a z a R
carros   k a R u Z    casar   k a z a x
```

In order to formally evaluate the GtoP module, a manual pronunciation lexicon is needed. For this purpose, we investigated the use of the lexicon included in the West Point Brazilian corpus, but it used a smaller vowel set than the one adopted in this project, which precluded the comparison with the automatically produced pronunciations. We are currently looking for a manually corrected pronunciation lexicon in order to formally evaluate the GtoP module.

4. Automatic speech recognition

4.1. Baseline ASR

Our baseline recognizer (AUDIMUS) is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of hidden Markov models (HMMs) with the pattern discriminative classification capabilities of MLPs. The acoustic modeling combines monophone probabilities generated by several MLPs trained on distinct feature sets: PLP (perceptual linear prediction), Log-RASTA (log-RelAtive SpecTrAl), and MSG (Modulation SpectroGram). Each MLP classifier incorporates local acoustic context via an input window of 13 frames (MSG uses 15 frames). The resulting networks have two fully connected non-linear hidden layers with 2,000 units each and 39 soft-max output units (38 phones plus silence).

[2] http://trans.sourceforge.net/

[3] www.linguateca.pt

[4] Text normalization is applied prior to the GtoP module.

The baseline vocabulary includes around 57 k words. The lexicon includes multiple pronunciations, totaling 65 k entries. The corresponding out-of-vocabulary (OOV) rate is 1.4%. The language model which is a 4-gram back-off model was created by interpolating a 4-gram newspaper text language model built from over 604 M words with a 3-gram model based on the transcriptions of the SR training set with 532 k words. The language models were smoothed using Knesser-Ney discounting and entropy pruning. The perplexity obtained in a development set is 112.9. Our decoder is based on the weighted finite-state transducer (WFST) approach to large vocabulary speech recognition. In this approach, the search space is a large WFST that maps HMMs (or in some cases, observations) to words. This WFST is built by composing various components of the systems represented as WFSTs. In our case, the search space integrates the HMM/MLP topology transducer (one single state HMM per monophone with a fixed minimum duration), the lexicon transducer, and the language model one. For BN in European Portuguese, this hybrid recognizer achieves a word error rate (WER) of 23.5% for all conditions, and 11.3% for read clean speech [10] on the European Portuguese JE corpus described in section 2.1. These results and the next ones in this paper are based on manual segmentation.

An improved version of this recognizer was developed, targeted at a vocabulary of 100k words [11]. In this version, the acoustic model which had been initially trained with manually annotated BN data (SR training corpus), was improved using unsupervised training. Recognized words that have a confidence measure above 91.5% are chosen as new training data. This is an iterative and never ending process while we get better performance with more data. We are currently using 378 hours of training data, 332 of which were automatically annotated using word confidence measures. This retraining yielded a significant 8.5% relative improvement in Word Error Rate. The new error rates on the JE corpus are 21.5% for all conditions, and 10.5% for read clean speech. The corresponding OOV reduced to 0.71%.

4.2. Porting to BP

The need for porting all the key modules of the speech recognizer was first stated by applying the EP recognition system to the BP test data (see section 2.2). The result of this preliminary experience was an expected low performance of 56.6% word error rate (WER), using the 100k vocabulary version.

In order to overcome the confirmed mismatch between European Portuguese and Brazilian Portuguese, several adaptation/development steps were mandatory in the baseline speech recognizer. More concretely: the use of the grapheme-to-phoneme conversion module for Brazilian Portuguese in order to build new lexical models, the development of new acoustic models based on BP data, and building new language models that could model the syntactic differences.

4.2.1. Lexical model

The integration of the developed GtoP module for BP was not a straight-forward task, since there are some phonetic symbols that only exist in BP. Consequently, there was not an acoustic model trained for these phonetic units in the EP baseline system, and it was not possible to directly replace the EP GtoP module by the BP GtoP module.

Our first experiments were made using the EP 100k list of words, for which the BP GtoP generated pronunciations. At this early stage, we mapped the phonetic symbols that only exist in BP to the most similar symbols (or sequence of symbols) in EP. In this way, we could approximately assess the influence of the pronunciation differences without modifying the acoustic and language models as well.

The performance achieved by the recognition system with the integration of the BP GtoP module was 46.2%, which is still far from the performance achieved for EP, but represents a significant improvement if one takes into account that only lexicon adaptation was performed with respect to the previous recognition experiment with BP data. Moreover, the construction of the pronunciation lexicon was totally automatic.

It is likely that part of the WER can be attributed to the different orthography of BP and EP words, but this percentage was not yet evaluated.

The development of this system with an adapted Brazilian GtoP module is a necessary step for boot-strapping the alignment of the BP data that is needed in the next porting step.

4.2.2. Acoustic model

The previous system was used to generate forced alignments of the BP BN data with the orthographic transcriptions available. Once the phonetic level alignments were obtained, it was possible to convert back the equivalences in EP of the phonetic symbols only existent in BP to the correct target symbols. That is, the new phonetic networks that are trained for BP already include the specific phonetic symbols. In this case, the Brazilian GtoP module can directly be used without modifications in the following alignments (generation of the lexicon for all the words in the training data) or recognition tests (generation of the lexicon for the active vocabulary).

Like in the EP system, MLP networks were trained on distinct feature sets (PLP, Log-RASTA and MSG). Each MLP classifier incorporates local acoustic context via an input window of 13 frames (15 in the case of MSG features). Due to the reduced amount of data available for training (compared to EP), the size of the two nonlinear hidden layers is only 600 units.

The process for the estimation of the monophone classification MLP networks consisted of several iterations of realignment and re-training until a stable phone classification rate is achieved in the development data set. The usefulness of the new acoustic models together with the Brazilian Portuguese GtoP was validated in the evaluation test set, achieving a WER performance of 31.6%. Despite the relative small amount of data used for training purposes, it is clear that training specific acoustic models for BP allows great performance improvements.

4.2.3. Language model

The last stage for porting the EP recognition system to a first complete BP version consists of the adaptation of the language model. In fact, in addition to better modeling the already mentioned syntactic differences among the two varieties, the redefinition and selection of a new active recognition vocabulary is crucial, since there are words (for instance proper nouns of toponyms or local personalities) that can be much more frequent in one variety that in the other or even have a different orthographic representation. In fact, the OOV rate in the BP development set obtained with the 100k EP vocabulary of previous experiments is 3.3%. This is a considerable increase with respect to EP Broadcast News data.

Hence, vocabulary selection was the first step of this adaptation stage. We built an initial vocabulary with all the words of the transcriptions of the training corpus, and completed it with

the most frequent words of the newspaper corpus, in order to achieve 64k different word forms. Given the scarcity of text material, we did not aim at this stage at building a 100k vocabulary. Despite using a smaller vocabulary we obtained an OOV rate of 1.3% with the new BP active vocabulary.

The language model trained was a 3-gram back-off model created by interpolating three individual 3-gram language models built from three different sources: the CETENFolha corpus, the recent newspapers corpora automatically obtained from the Internet and the manual transcriptions of the training set. The language models were smoothed using Knesser-Ney discounting and entropy pruning. The perplexity obtained in the development set is 197.

The use of this new language model and new vocabulary together with the BP GtoP conversion module and the acoustic models previously described resulted in a 26.9% WER performance. The specific language model allows additional performance improvements even using a relatively small amount of text data compared to the amount used for the EP models.

These results were obtained with a lexicon of 99.5k entries, due to the above mentioned use of multiple pronunciations. When single pronunciations were used instead in all the porting steps, the word error rate increased by 1.9%, with a corresponding 5% reduction in processing time.

4.2.4. Acoustical modeling of phone transitions

Our speech recognition experiments for English, using a vocabulary of 5k, have shown us the potential advantages of using, in addition to the monophone units modeled by a single state, multiple-state monophone units and a fixed set of diphone units aimed at specifically modeling phonetic transitions [12]. Using this strategy for BP, we incorporated the sub-phonetic units (multiple-state phones and diphones), which caused the trained networks to have 320 soft-max output units. Preliminary results already show a small improvement in WER: 25.5%.

5. Conclusions

Our European Portuguese Broadcast News system dramatically fails when it is used for transcription of Brazilian Portuguese data. By means of several stages of adaptation, including lexical conversion, acoustic modeling and language modeling, we are able to port our EP speech recognition system to the specific characteristics of the BP achieving reasonable good recognition results even with a considerable limitation of available corpora, both audio and textual. A potential cause is the fact that EP recognition systems have to deal with much more pronounced vowel reduction phenomena that create very large consonant clusters and render EP much more difficult than BP. Other potential causes are the smaller speaker diversity and lower percentage of spontaneous speech in the BP test set, compared to the EP one.

We expect that the future use of additional resources will allow remarkable improvements in the BP speech recognizer. We are planning to update the language model using the newspaper data that is continuously being collected from the Internet and include a larger active vocabulary. Concerning acoustic modeling, we intend to integrate both manually transcribed data and also automatically transcribed data, for re-training the MLP networks with a larger variety of data sources (in fact, the Spoltech and West Point corpora have already been forced aligned).

Our target is the development of a fully automatic subtitling system for BP, similar to the one we developed for EP and which has been deployed at the public Portuguese TV, since March 2008, integrating dynamic lexical and language models. It is interesting to notice that the punctuation and capitalization modules which were originally developed for EP [13], seem to be robust enough to be used for another variety, although the formal evaluation was not yet conducted.

6. Acknowledgments

This work was funded by FCT project PTDC/PLP/72404/2006. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III".

7. References

[1] T. Lander, R. Cole, B. Oshika, and M. Noel, "The OGI 22 language telephone speech corpus," in *Proc. Eurospeech '1995*, Madrid, Spain, Sep. 1995.

[2] "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, Tech. Rep., 2001.

[3] P. Silva, N. Neto, A. Klautau, A. Adami, and I. Trancoso, "Speech recognition for Brazilian Portuguese using the Spoltech and OGI-22 corpora," in *XXVI Simpósio Brasileiro de Telecomunicações*, Rio de Janeiro, Brazil, Sep. 2008.

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

[5] D. Caseiro, I. Trancoso, L. Oliveira, and C. Viana, "Grapheme-to-phone using finite state transducers," in *Proc. 2002 IEEE Workshop on Speech Synthesis*, Santa Monica, CA, USA, Sep. 2002.

[6] D. Silva, A. de Lima, R. Maia, D. Braga, J. de Moraes, J. de Moraes, and F. R. Jr., "A rule-based grapheme-phone converter and stress determination for brazilian portuguese natural language processing," in *IEEE Int. Telecomm. Symposium (ITS 2006)*, Fortaleza, Brazil, Sep. 2006.

[7] E. Albano and A. Moreira, "Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese," in *Proc. of the Int. Conf. on Spoken Language Processing (ICSLP)*, Philadelphia, USA, Oct. 2006.

[8] F. Barbosa, G. Pinto, F. Resende, C. Gonçalves, R. Monserrat, and R. Rosa, "Grapheme-phone transcription algorithm for a Brazilian Portuguese TTS," in *Proc. of 6th. Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR 2003)*, Faro, Portugal, Jun. 2003.

[9] I. Seara, F. Pacheco, R. S. Jr., S. Kafka, and S. Seara, "Geração automática de variantes de léxicos do Português Brasileiro para sistemas de reconhecimento de fala," in *XXVI Simpósio Brasileiro de Telecomunicações*, Rio de Janeiro, Brazil, Oct. 2003.

[10] R. Amaral, H. Meinedo, D. Caseiro, I. Trancoso, and J. Neto, "A prototype system for selective dissemination of broadcast news in European Portuguese," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 37507, May 2007.

[11] H. Meinedo, M. Viveiros, and J. Neto, "Audimus.media: a broadcast news speech recognition system for the European Portuguese language," in *Proc. PROPOR '2003*, Faro, Portugal, Jun. 2003.

[12] A. Abad and J. Neto, "Incorporating acoustical modelling of phone transitions in an hybrid ANN/HMM speech recognizer," in *Proc. Interspeech 2008*, Brisbane, Australia, Sep. 2008.

[13] F. Batista, D. Caseiro, N. Mamede, and I. Trancoso, "Recovering capitalization and punctuation marks for automatic speech recognition: Case study for the Portuguese broadcast news," *Speech Communication*, vol. 50, no. 10, pp. 847–862, Oct. 20078.