The L2F Broadcast News Speech Recognition System

Hugo Meinedo[1], *Alberto Abad*[1], *Thomas Pellegrini*[1], *João Neto*[1,2], *Isabel Trancoso*[1,2]

[1]L2F - Spoken Language Systems Lab, INESC-ID Lisboa, Portugal
[2]Instituto Superior Técnico, Lisboa, Portugal
{hugo.meinedo, alberto.abad, thomas.pellegrini, joao.neto, isabel.trancoso}@l2f.inesc-id.pt

Abstract

Broadcast news play an important role in our lives providing access to news, information and entertainment. The existence of an automatic transcription is an important medium that not only can provide subtitles for inclusion of people with special needs or be an advantage on noisy and populated environments, but also because it enables data search and retrieve capabilities over the multimedia streams. In this work we will describe and evaluate the automatic speech recognition systems developed for two Iberian languages, European Portuguese and Spanish and also for Brazilian Portuguese, African Portuguese and English. The developed systems are fully automatic and capable to subtitling in real-time Broadcast News stream with a very small delay.

Index Terms: Speech Recognition, Broadcast News, Iberian languages, Accent, Online processing

1. Introduction

The Broadcast News (BN) processing system developed at the Spoken Language Systems Lab of INESC-ID integrates several core technologies, in a pipeline architecture: jingle detection, audio segmentation, automatic speech recognition, punctuation, capitalization, topic segmentation/indexation, summarization, and translation. The first modules of this system were optimized for on-line performance, given their deployment in the fully automatic speech recognition subtitling system that is running on the main news shows of the public TV channel in Portugal (RTP), since March 2008.

To our knowledge, the majority of subtitling systems described in the literature rely on speech-to-text alignment rather than full automatic speech recognition [1]. Re-speakers also are commonly used to simplify the original speech, and speech recognition engines are adapted to the captioner voice [2].

This paper concerns the third module in the pipeline - speech recognition, emphasizing the most recent improvements, and our efforts to port it to other languages (English and Spanish), and to other varieties of Portuguese, namely those spoken in the South American and African continents.

The development of a system for a new language is a challenging task due to the need of new acoustic training data, vocabulary definition, lexicon generation and language model estimation [3].

The paper starts with a description of the main modules of our recognition engine, emphasizing the two language independent components - feature extraction and decoder. The next three sections are devoted to the three varieties of Portuguese covered by our system: the original one (European Portuguese, henceforth designated as EP), Brazilian Portuguese (BP), and African Portuguese (AP). The porting efforts for the other two languages (European Spanish and American English) are described in Sections 6 and 7, respectively. For each of these sections, we shall detail the corpora, vocabulary, and lexical and language model generation, ending with performance results. The final section discusses the main advantages and shortcomings of these systems, namely in what concerns real time close captioning applications.

2. Automatic Speech Recognition

Our Broadcast News automatic speech recognition engine named Audimus [4, 5] is a hybrid automatic speech recognizer that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of Multi-Layer Perceptrons (MLPs). A block diagram is shown in Figure 1.

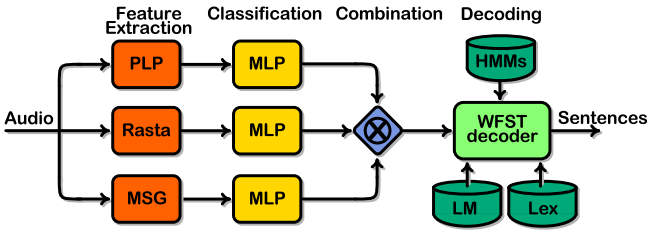

Figure 1: *Audimus block diagram.*

2.1. Feature extraction

The MLP/HMM acoustic model combines posterior phone probabilities generated by three phonetic classification branches. Different feature extraction and classification branches effectively perform a better modeling of the acoustic diversity, in terms of speakers and environments, present in BN data. The first branch extracts 26 PLP (Perceptual Linear Prediction) features, the second 26 Log-RASTA (log-RelAtive SpecTrAl) features and the 3rd uses 28 MSG (Modulation Spectrogram) coefficients for each audio frame. Each MLP classifier incorporates local acoustic temporal context via an input window of 13 frames (the MSG branch uses 15 frames) and two fully connected non-linear hidden layers. The number of units of each hidden layer as well as the number of softmax outputs of the MLP networks differs for every language. Usually, the hidden layer size depends on the amount of training data available, while the number of MLP outputs depends on the characterisitc phonetic set of each language.

2.2. Decoding process

The Audimus decoder is based on the Weighted Finite-State Transducer (WFST) approach, where the search space is a large

WFST that results from the integration of the HMM/MLP topology transducer, the lexicon transducer and the language model one [6]. This decoder uses a specialized WFST composition algorithm of the lexicon and language model components in a single step. Furthermore it supports lazy implementations, where only the fragment of the search space required in runtime is computed. Besides the recognized words, the decoder outputs a series of values describing the recognition process. In order to generate a word confidence measure these features are combined through a maximum entropy classifier, whose output represents the probability of each word being correct [6]. Confidence measures for the recognized text are fundamental not only to select new acoustic training data but also to filter the output text in the subtitling composition stage.

3. European Portuguese

The initial EP acoustic model (EP baseline) was trained with 46 hours of manually annotated BN data collected from the public Portuguese TV. Currently automatically collected and transcribed data is being reused to perform unsupervised training. Recognized words that have a confidence measure above 91.5% are chosen for new training data. This is an iterative and never ending process while we get better performance with more data. The first iteration (EP iteration 1) used 378 hours of useful training speech data, 332 of which were automatically annotated using word confidence measures. The current iteration (EP iteration 2) used a total of 1000 hours of data mostly news shows from several EP TV channels. The EP MLPs are formed by 2 hidden layers with 2000 units each and have 500 softmax output units that correspond to 38 three state monophones of the EP language plus a single-state non-speech model (silence) and 385 phone transition units which were chosen to cover a very significant part of all the transition units present in the training data.

The Language Model (LM) is a statistically 4-gram model and results from the interpolation of three specific LMs. The first is a backoff 4-gram LM, trained on a 700M word corpus of newspaper texts, collected from the Web from 1991 to 2005 (out-of-domain corpus). The second LM is a backoff 3-gram LM estimated on a 531k word corpus of broadcast news transcripts (in-domain corpus). The third model is a backoff 4-gram LM estimated on the EP web newspapers texts collected from the previous seven days. These three LMs were linearly interpolated. For weight optimization we have used the automatically transcribed texts from the last twenty one days of news shows from RTP channels 1 and 2. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing, 100k words (or 1-gram), 7.5M 2-gram, 14M 3-gram and 7.9M 4-gram.

The EP engine uses a 100k word vocabulary adapted on a daily basis to reflect the new words that appear in web newspaper texts [7]. This daily modification of the vocabulary implies a re-estimation of the language model and retraining of the word confidence measures classifier. In order to validate the new vocabulary and language model generated, a benchmark test with one hour long news show was created, running after the daily adaptation process. This validation data is then used to retrain the confidence measure classifier in order to linearize the confidence threshold.

After the 100k word vocabulary adaptation, the pronunciation lexicon is built automatically by dividing the words into two categories. The "known" ones, for which we are able to produce a correct pronunciation, and the "unknown" ones. The correct pronunciation is either retrieved from an in house lexicon, or generated by our rule-based grapheme-to-phone (GtoP) conversion module [8]. This module can only process words which follow the Portuguese pronunciation rules, so spelled acronyms and foreign words have to be filtered out. These unknown words are then automatically split into spelled acronyms and foreign words. For spelled acronyms, rule-based pronunciations are generated. For the foreign words, a further subdivision is made, in order to identify the ones that exist in the public domain lexicon provided by CMU, for which a nativized version was produced. For the words not included in the CMU lexicon, grapheme nativization rules are applied prior to using the EP GtoP module to generate the pronunciation. The final multiple-pronunciation EP lexicon generally includes 114k entries.

Table 1 summarizes the Word Error Rate (WER) obtained in one of our BN evaluation test sets, RTP07, which is composed by six one hour long news shows from 2007. EP iteration 1 system reduced the WER by using more training data and switching to multi-state-monophones and transition units. Our current ASR, denoted EP iteration 2 significantly reduces the WER by using an extended training set with 1000 hours, larger MLPs and also the daily adaptations of the vocabulary, lexicon and language models.

Training data	Train	WER (%)
EP baseline	46 h	23.5
EP iteration 1	378 h	21.5
EP iteration 2	1000 h	18.4

Table 1: *Word Error Rates (WERs) achieved on RTP07 evaluation test set for our European Portuguese BN Recognition systems.*

4. Brazilian Portuguese

The need for porting all the key modules of the speech recognizer was first stated by using the EP BN transcription system for testing Brazilian Portuguese (BP) BN data. 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 EP and BP, several adaptation/development steps were mandatory in the baseline speech recognizer. More concretely: the use of a GtoP module for BP in order to build new lexicon models, the development of new acoustic models based on BP data, and building new language models that could model the syntactic differences.

Details on the G2P developed for BP can be found in [9]. 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 already represents a significant improvement.

Due to the reduced amount of data available for training compared to EP, the size of the two nonlinear hidden layers of the MLPs is of 600 units. The process for the estimation of the

monophone classification MLP networks, which corresponds to 40 outputs, consisted of several iterations of re-alignment 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 same evaluation data set of the previous experiments, achieving a WER performance of 31.6 %.

The last stage for porting the EP recognition system to a first complete BP version consists of the adaptation of the language model. 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 100k different word forms. The next step was the automatic addition of multiple pronunciations in order to take into account some of the different variations that can be obtained as a result of word co-articulation rules.

The language model is a 4-gram backoff model created by interpolating three individual LMs built from three different sources: the CETENFolha corpus which has around 24M words, the recent newspapers corpora automatically obtained from the Internet which amounts to 18M words 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 by incorporating multiple-state phones and transitions, which caused the MLPs to have 320 outputs obtained an improvement in WER of 25.5 %. Afterwards we increased the training data with more 33 hours of automatically transcribed material and we were able to achieve 22.4 % of WER. Finally we built a new 100k vocab and LM using more 44M words of web newspaper texts which resulted in a 21.6 % WER performance.

5. African Portuguese

The EP trained speech recognizer was tested on the AP data, using the 100k vocabulary version, yielding a WER of 29.7 %, which is as expected worse than the one obtained for the EP test data, but is significantly better than the one obtained for the BP test data. Given that the GtoP module was not yet ported to AP, and no pronunciation lexicon was available for AP, our first efforts concerning porting the ASR to AP were thus restricted to porting the acoustic and the language models, starting with the latter.

The language model for AP is a 3-gram model created by interpolating three individual language models. The first is the same 4-gram LM from the EP, the second LM was built from recent AP newspapers automatically obtained from the Internet with 1.6M words and the third LM was built from the manual AP transcriptions of the training set which amount to 86k words. The language models were smoothed using Knesser-Ney discounting and entropy pruning. The perplexity obtained in the test set with only the EP model were 165.4. The use of the final AP interpolated LM resulted in perplexity of 150.6.

In terms of acoustic model training two distinct approaches were followed. The first one consisted of training new acoustic models using only the small amount of AP training data available which is around 7.5 hours and using around 17 hours of automatically detected and transcribed AP data [10]. The automatic transcription was done using the EP recognizer and selecting only words with a high confidence score. Nevertheless this very short amount of training material was the cause that we attributed for the worse performance, 27.9 % WER.

The second approach consisted of adapting the EP acoustic models using only the 7.5 hours of manually transcribed AP data. This resulted in 23.7 % WER which is a very significant improvement relative to the value obtained using only the EP acoustic models that had obtained 29.7 % WER.

6. European Spanish

In order to create an acoustic model for European Spanish (ES), an initial audio/phone alignment was necessary. We decided to transform the EP BN acoustic model output to bootstrap the ES phone alignment. A mapping was created between the ES phone set (29 phones plus silence) and EP phone set (39 plus silence) by choosing the EP phone with the most similar sound to the ES phone. With this transformed phone set, an initial ES alignment was created for the training data which consisted of 10 hours of manually annotated news shows from the national Spanish TV station (TVE). We made 4 realignment / MLP training iterations, until there was no significant improvement in the recognition results. With this ES acoustic model it was possible to automatically transcribe more data, and reuse it for training. Currently we have available an additional 148 hours of BN automatically recognized data. To improve phone classification, the 29 ES monophones where extended to multi-state-monophones (87 units), plus 112 phone transition units plus silence, totalizing 200 output units. To cope with the increased number of output units and the larger training set, the hidden layers of the MLP classifiers were enlarged to 1500 units each.

Similar to the EP the ES LM is also a statistically 4-gram model but results from the interpolation of four specific LM. The first is a 4-gram LM trained on a 1.2G word corpus of newspapers texts named Spanish Gigaword from LDC catalog. The second is also a 4-gram LM composed by data from several online ES newspapers text raging from 2001 to 2008 totalizing 72M words. This data is more recent than the Gigaword corpus and its content is similar to the BN news shows. The third is a 3-gram LM estimated on the BN training transcriptions which has 466k words. The fourth model is a 4-gram LM estimated on the ES web newspapers texts collected from the previous seven days to better cover and reflect the vocabulary adaptation. These four LMs were linearly interpolated with optimization of the weights on the automatic transcription texts from the last twenty one days of news shows from TVE channel. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing, with 100k words (1-gram), 6.6M 2-gram, 12.5M 3-gram and 10M 4-gram.

Audimus ES ASR system also uses the vocabulary adaptation process developed for the EP system. A new vocabulary containing 100k words is built in a daily base to reflect the new words that appear in web newspaper texts [7]. After the 100k word vocabulary adaptation the pronunciations lexicon is built automatically by dividing the words according to three categories. The acronyms, foreign words and “normal” words. For the “normal” words a correct pronunciation is built using in

house lexica and our rule-based GtoP conversion system. For the acronyms, rule base pronunciations are generated. For the foreign words grapheme transformation rules are applyed after our ES GtoP system is used to generate the pronunciation. The final multiple-pronunciation EP lexicon includes 100k entries.

Training data	Train	WER (%)
ES baseline	10 h	22.6
ES current	158 h	15.7

Table 2: *Word Error Rates (WERs) achieved on TVE evaluation test set for our European Spanish BN recognition systems.*

Table 2 presents the evaluation results for the ES recognition system. These results were obtained in a test set with nine BN news shows from TVE station totalizing aproximately 9 hours of audio. The first line refers to the ES baseline which was trained using just the 10 hours of manually anotated audio and used only monophones. The second line shows a significant improvement by using much more acoustic training data (158 h), larger MLPs and multi-state-monophones plus phone transition units. Both systems used the same 100k vocabulary, lexicon and LM.

7. American English

The HUB-4 1996 (LDC97S44) and 1997 (LDC98S71) data sets were used to train MLP networks. Both data sets are distributed by LDC, and contain respectively 73 and 67 hours of manually transcribed speech, coming from ABC, CNN and CSPAN television networks and NPR and PRI radio networks. A set of 39 multiple-state monophones plus two single-state non-speech models (one for silence and one for breath) and 336 phone transition units (chosen to cover more than 90% of all the transition units present in the training data), was trained. The number of final recognition units (output layer size) totalizes 455 units. The use of context dependent units led to gains larger than 20% relative, when compared to context independent ("monophones") units.

For language modeling, we built one LM per source, including speech transcripts and written text sources. Nine LMs were linearly interpolated with optimization of the weights on a subset of the HUB-4 1997 training corpus used as development corpus. The final interpolated language model is a 4-gram LM, with Kneser-Ney modified smoothing, comprised of 64k words (or 1-gram), 12 M 2-gram, 5.8M 3-gram and 4.5M 4-gram.

The 64k word vocabulary consists of all the words contained in the HUB-4 training set plus the most frequent words in the broadcast news texts and Newspapers texts. The pronunciations were extracted from the public domain lexicon provided by CMU. For words not included in this lexicon, a rule-based GtoP conversion system was used. The multiple-pronunciation lexicon included 70k entries.

Table 3 shows the best performances achieved on four official NIST test sets. The use of multi-state monophones and transition units greatly improved the performance with 15 % relative gain on average.

Corpus	Eval'97	Eval'98	Eval'99	Eval'03
WER (%)	22.0	20.4	23.3	20.6

Table 3: *Word Error Rates (WERs) achieved on four NIST evaluation test sets for our American English BN system.*

8. Conclusions

These ASR systems are the result of several years of research and development in the BN area for the Portuguese language, which were recently expanded to encompass other languages and varieties.

The main characteristic of these systems is their low latency requirements, which makes them suitable for BN subtitling. The evaluation results presented on this paper are strong and the systems have several and innovative differences. We will continue in the near future to explore these differences, mainly introducing language and variety identification, new speaker adaptation techniques, gender and bandwidth dependent acoustic models and improving language modeling and reducing vocabulary size to better accommodate the word types generated each day, in order to further improve the system performance.

9. Acknowledgements

This work was funded by FCT projects PTDC/PLP/72404/2006 and CMU-PT/HuMach/0039/2008. INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitrio de Apoio III".

10. References

[1] A. Ortega, J. Garcia, A. Miguel, and E. Lleida, "Real-time live broadcast news subtitling system for spanish," in *Proc. Interspeech 2009*, Brighton, September 2009.

[2] "Speech Recognition in Assisted and Live Subtitling for Television, BBC R&D White Paper WHP 065," September 2003.

[3] T. Schultz, A. Black, S. Badaskar, M. Hornyak, and J. Kominek, "SPICE: Web-based tools for rapid language adaptation in speech processing systems," in *Proc. Interspeech 2007*, Antwerp, Belgium, 2007.

[4] J. Neto, H. Meinedo, M. Viveiros, R. Cassaca, C. Martins, and D. Caseiro, "Broadcast news subtitling system in portuguese," in *Proc. ICASSP 2008*, Las Vegas, USA, 2008.

[5] H. Meinedo, "Audio pre-processing and speech recognition for broadcast news," Ph.D. dissertation, IST, Lisbon, Portugal, 2008.

[6] D. Caseiro and I. Trancoso, "A specialized on-the-fly algorithm for lexicon and language model composition," *IEEE Transactions on Audio, Speech and Lang. Proc.*, vol. 14, no. 4, Jul. 2005.

[7] C. Martins, A. Teixeira, and J. Neto, "Dynamic language modeling for a daily broadcast news transcription system," in *Proc. ASRU 2007*, Kyoto, Japan, 2007.

[8] 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.

[9] A. Abad, I. Trancoso, N. Neto, and C. Viana, "Porting an european portuguese broadcast news recognition system to brazilian portuguese," in *Proc. Interspeech 2009*, Brighton, UK, 2009.

[10] O. Koller, A. Abad, I. Trancoso, and C. Viana, "Exploiting variety-dependent phones in portuguese variety identification applied to broadcast news transcription," in *Proc. Interspeech 2010*, Makuhari, Japan, 2010.