AUDIMUS.MEDIA: A Broadcast News Speech Recognition System for the European Portuguese Language

Hugo Meinedo, Diamantino Caseiro, João Neto, and Isabel Trancoso

L^{2}F – Spoken Language Systems Lab
INESC-ID / IST, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
{Hugo.Meinedo,Diamantino.Caseiro,Joao.Neto,Isabel.Trancoso}@l2.inesc-id.pt
http://l2f.inesc-id.pt

Abstract. Many applications such as media monitoring are experiencing a large expansion as a consequence of the different emerging media sources and can benefit dramatically by using automatic transcription of audio data. In this paper, we describe the development of a speech recognition engine, AUDIMUS.MEDIA used in the Broadcast News domain. Additionally we describe recent improvements that permitted a relative recognition error decrease of more than 20% and a 4x speed-up.

1 Introduction

The development of speech recognition systems associated to Broadcast News (BN) tasks open the possibility for novel applications where the use of automatic transcriptions is a major attribute.

We have been developing a system for selective dissemination of multimedia information in the scope of an IST-HLT European programme project. To accomplish that goal we have been working in the development of a broadcast news speech recognition system associated with automatic topic detection algorithms [1]. The idea was to build a system capable of identifying specific information in multimedia data consisting of audio-visual streams, using continuous speech recognition, audio segmentation and topic detection techniques. The automatic transcriptions produced by our speech recognition system are used for a topic detection module that outputs a set of topics to be sent to end users.

This paper describes in detail our BN speech recognition system engine, AUDIMUS.MEDIA. Section 2 introduces the BN corpus used for training the system. Section 3 describes the acoustic models. In Sects. 4 and 5 we present the vocabulary and lexicon building and the language model. The decoder algorithm is described in Sect. 6, followed by our latest BN recognition results in Sect. 7. Finally some concluding remarks are presented in Sect. 8.

N.J. Mamede et al. (Eds.): PROPOR 2003, LNAI 2721, pp. 9–17, 2003.

2 BN Corpus

To support the research and developments associated with this task it was necessary to collect a representative Portuguese BN corpus in terms of amount, characteristics and diversity. We started by defining the type of programs to monitor, in close cooperation with RTP the Portuguese public broadcast company, being selected as primary goals all the news programs, national and regional, from morning to late evening, including both normal broadcasts and specific ones dedicated to sports and financial news. Given its broader scope and larger audience, the 8 PM news program was selected as the prime target. This corpus serves two main tasks: the development of a BN speech recognition system and a system for topic segmentation and indexing. In that sense we divided our corpus in two parts: the Speech Recognition Corpus and the Topic Detection Corpus. Since each of these parts serves different purposes it will also have different features. Prior to the collection of these corpora we started with a relative small Pilot Corpus of approximately 5 hours, including both audio and video, which was used to setup the collection process, and discuss the most appropriate kind of programs to collect. The Speech Recognition Corpus was collected next, including 122 programs of different types and schedules and amounting to 76h of audio data. The main goal of this corpus was the training of the acoustic models and the adaptation of the language models used in the large vocabulary speech recognition component of our system. The last part of our collection effort was the Topic Detection Corpus, containing around 300 hours of audio data related to 133 TV broadcast of the 8 PM news program. The purpose of this corpus was to have a broader coverage of topics and associated topic classification for training our topic indexation module. RTP as data provider was responsible to collect the data in their installations. The transcription process was jointly done by RTP and our institution, and made through the Transcriber tool following the LDC Hub4 (Broadcast Speech) transcription conventions. Most of the audio data was first automatically transcribed. The orthographic transcriptions of the Pilot Corpus and the Speech Recognition Corpus were manually verified. For the Topic Detection Corpus, we have only the automatic transcriptions and the manual segmentation and indexation of the stories made by RTP staff in charge of their daily program indexing. Our institution was responsible for the validation process, training the annotators and packaging the data.

2.1 Pilot Corpus

The Pilot Corpus was collected in April 2000 and served as a test bed for the capture and transcription processes. We selected one of each type of program resulting in a total of 11 programs and a duration of 5h 33m. After removing the jingles and commercial breaks we ended up with a net duration of 4h 47m. The corpus was manually transcribed at RTP. The corpus includes the MPEG-1 files (.mpeg) where the audio stream was recorded at 44.1 kHz at 16 bits/sample, a separated audio file (.wav) extracted from the MPEG-1 data, a transcription file (.trs) resulting from the manual annotation process in the Transcriber tool, and a

Table 1. Pilot Corpus programs

Program	Duration	Type
Notícias	0:08:02	Morning news
Jornal da Tarde	0:57:36	Lunch time news
País Regiões	0:16:05	Afternoon news
País Regiões Lisboa	0:24:21	Local news
Telejornal	0:45:31	Evening news
Remate	0:07:30	Daily sports news
24 Horas	0:24:23	Late night news
RTP Economia	0:09:43	Financial news
Acontece	0:20:39	Cultural news
Jornal 2	0:49:34	Evening news
Grande Entrevista	1:09:38	Political / Economic interview (weekly)
Total	5:33:00	

.xls (Excel format) file including the division into stories and their corresponding summary that results from the daily process of indexation at RTP. The final contents of the Pilot Corpus in terms of programs, duration and type is presented in Table 1.

2.2 Speech Recognition Corpus

The Speech Recognition Corpus was collected from October 2000 to January 2001, with small changes in the programs when compared with the Pilot Corpus. This corpus was divided in three sets: training, development and evaluation. A complete schedule for the recordings was elaborated leaving some time intervals between the different sets. We got 122 programs in a total of 75h 43m. We adopted the same base configuration as for the Pilot Corpus, except that we did not collected the video stream. Also the audio was recorded at 32 kHz, due to restrictions of the hardware, and later downsampled to 16 kHz which was appropriate to the intended processing. This corpus was automatically transcribed and manually verified. As a result it includes only a audio stream file (.wav) and a transcription file (.trs). The final contents of the Speech Recognition Corpus in terms of programs, duration and type is presented in Table 2.

3 AUDIMUS.media Recognition System

AUDIMUS.media is a hybrid speech recognition system that combines the temporal modeling capabilities of Hidden Markov Models (HMMs) with the pattern discriminative classification capabilities of multilayer perceptrons (MLPs). In this hybrid HMM/MLP system a Markov process is used to model the basic temporal nature of the speech signal. The MLP is used as the acoustic model estimating context-independent posterior phone probabilities given the acoustic data at each frame. The acoustic modeling of AUDIMUS.media combines phone

Table 2. Speech Recognition Corpus programs

	Training		Development		Evaluation		
Programs	Number	Duration	Number	Duration	Number	Duration	Type
Notícias	7	0:43:52	1	0:10:38	1	0:10:41	Morning news
Jornal da Tarde	8	7:55:46	1	1:13:10	1	1:02:59	Lunch time news
País Regiões	12	6:43:47	1	0:32:49	1	0:33:46	Afternoon news
País Regiões Lx	7	2:16:47	1	0:20:32	1	0:20:16	Local news
Telejornal	30	32:41:34	3	3:37:13	2	1:54:18	Evening news
24 Horas	4	1:18:53	2	1:02:43	2	0:38:35	Late night news
RTP Economia	13	1:53:02	2	0:10:38	2	0:19:58	Financial news
Acontece	9	3:05:38	1	0:19:47	1	0:17:50	Cultural news
Jornal 2	7	4:53:39	1	0:46:04	1	0:38:26	Evening news
Total	97	61:32:58	13	8:13:34	12	5:56:49	

probabilities generated by several MLPs trained on distinct feature sets resulting from different feature extraction processes. These probabilities are taken at the output of each MLP classifier and combined using an average in the log-probability domain [2]. All MLPs use the same phone set constituted by 38 phones for the Portuguese language plus silence and breath noises. The combination algorithm merges together the probabilities associated to the same phone.

We are using three different feature extraction methods and MLPs with the same basic structure, that is, an input layer with 9 context frames, a non-linear hidden layer with over 1000 sigmoidal units and 40 softmax outputs. The feature extraction methods are PLP, Log-RASTA and MSG.

4 Language Modeling

During the few last years we have been collecting Portuguese newspapers from the web, which allowed us to build a considerably large text corpus. Until the end of 2001 we have texts amounting to a total of 24.0M sentences with 434.4M words.

A language model generated only from newspaper texts becomes too much adapted to the type of language used in those texts. When this language model is used in a continuous speech recognition system applied to a Broadcast News task it will not perform as well as one would expect because the sentences spoken in Broadcast news do not match the style of the sentences written in the newspaper. A language model from Broadcast News transcriptions would probably be more adequate for this kind of speech recognition task. The problem is that we do not have enough BN transcriptions to generate a satisfactory language model. However we can adapt a newspaper text language model to the BN task by combining it with a model created from BN transcriptions using linear interpolation, and thus improve performance.

One of the models is always generated from the newspaper text corpus while the other is a backoff trigram model using absolute discounting and based on the training set transcriptions of our BN database. The optimal weights used in

the interpolation are computed using the transcriptions from the development set of our BN database.

The final interpolated model has a perplexity of 139.5 and the newspapers model has 148.0. It is clear that even using a very small model based in BN transcriptions we can obtain some improvement in the perplexity of the interpolated model.

5 Vocabulary and Pronunciation Lexicon

From the text corpus with 335 million words created from all the newspaper editions collected until the end of 2000, we extracted 427k different words. About 100k of these words occur more than 50 times in the text corpus. Using this smaller set, all the words were classified according to syntactic classes. Different weights were given to each class and a subset with 56k words was created based on the weighted frequencies of occurrence of the words. To this set we added basically all the new words present in the transcripts of the training data of our Broadcast News database then being developed, giving a total of 57,564 words. Currently the transcripts contain 12,812 different words from a total of 142,547.

From the vocabulary we were able to build the pronunciation lexicon. To obtain the pronunciations we used different lexica available in our institution. For the words not present in those lexica (mostly proper names, foreign names and some verbal forms) we used an automatic grapheme-phone system to generate corresponding pronunciations. Our final lexicon has a total of 65,895 different pronunciations.

For the development test set corpus which has 5,426 different word in a total of 32,319 words, the number of out of vocabulary words (OOVs) using the 57k word vocabulary was 444 words representing a OOV word rate of 1.4%.

6 Weighted Finite-State Dynamic Decoder

The decoder underlying the AUDIMUS.media system is based the weighted finite-state transducer (WFST) approach to large vocabulary speech recognition [3]. In this approach, the search space used by the decoder is a large WFST that maps observation distributions to words. This WFST consists of the composition of various transducers representing components such as: the acoustic model topology H; context dependency C; the lexicon L and the language model G. The search space is thus $H \circ C \circ L \circ G$[1], and is traditionally compiled outside of the decoder, which uses it statically.

Our approach differs in that our decoder is dynamic and builds the search space "on-the-fly" as required by the particular utterance being decoded [4]. Among the advantages provided by the dynamic construction of the search space are: a better scalability of the technique to large language models; reduction of

[1] We use the matrix notation for composition.

the memory required in runtime; and easier adaptation of the components in runtime.

The key to the dynamic construction of the search space is our WFST composition algorithm [5] specially tailored for the integration of the lexicon with the language model (represented as WFSTs). Our algorithm performs simultaneous the composition and determinization of the lexicon and the language model while also approximating other optimizing operations such as weight pushing and minimization [6].

The goal of the determinization operation is to reduce lexical ambiguity in a more general way than what is achieved with the use of a tree-organized lexicon. Weight pushing allows the early use of language model information, which allows the use of tighter beams thus improving performance. Minimization essentially reduces the memory required for search while also giving a small speed improvement.

6.1 Alignment

In order to allow the registration of time boundaries the decoder allows the use of a special label *EOW* in the input side of the search space transducer. Whenever that label is crossed while searching, the decoder records the time of the crossing. That label is thus used to mark word or phone boundaries.

Using the *EOW* labels the decoder can be used for alignment, in alignment mode the search space is usually build as $H \circ L \circ S$ where S is the orthographic transcription of the utterance being recognized. The fact that the decoder imposes no a priori restrictions on the search space structure gives us great flexibility, for example, alternative pronunciation rules can be used by compiling then in a finite-state transducer R, and building the search space as $H \circ R \circ L \circ S$[7].

The *EOW* label is also given other uses in the decoder, for example, it can be used as an aid in the construction of word lattices wherein the labels mark the end of segments corresponding to arcs in the lattice. One other use of the label is to collect word-level confidence features that can be used to compute confidence scores.

6.2 Pruning

Pruning is fundamental to control the search process in large vocabulary recognition. Our decoder uses 3 forms of pruning: beam pruning; histogram pruning; and phone deactivation pruning. Each form of pruning deals with a different aspect of the search process.

Beam pruning is probably the most important, and consists of pruning the hypotheses with a score worse than a given amount (the beam) from the best one among the hypotheses ending at a particular frame. This form of pruning is used by most large vocabulary decoders.

Our form of beam pruning differs from most in its eagerness, which allows the pruning of hypotheses while they are being generated, by using the cost of the best hypothesis so far as a reference. When an hypothesis with cost c_t at time t is

propagated though an edge with input label d and weight w, its cost in the next frame is updated with two components: a transition weight w that incorporates linguistic constraints; and an acoustic weight $distr(d, t+1)$ obtained from the speech signal.

Because the transition weight is often of the same order of magnitude as the beam, we obtain significant improvements by performing the pruning test twice: first the cost $c_t + w$ is tested and then $c_t + w + distr(d, t+1)$. If the first test fails, we avoid the expensive computation of both $distr(d, t+1)$ and the bookkeeping associated with the expansion of the hypothesis.

The function of *histogram pruning* is to reduce peak resource usage, time and memory, to a reasonable limit. It consists of establishing the maximum number m of hypotheses that are expanded at each frame. Whenever their number is over the limit, only the m best are kept and the other are pruned. If the value of m is set to a reasonable value (such as 100000) then it was virtually no negative effect on the accuracy of the decoder while preventing it from staling when there is a severe acoustic mismatch relative to the training conditions.

Phone deactivation pruning [8] takes advantage of the fact that the MLP directly estimates the posterior probability of each phone. This form of pruning consists of flooring the posterior probability of a given phone to a very low value when it is below a given threshold. This has the effect of allowing the MLP to deactivate some unlikely phones. There is usually an optimal value for the threshold, if too large then the search will be faster but more error prone, if it too low, then the search will be slower with no advantage regarding the accuracy (sometimes the accuracy is even worse due to the MLP having difficulty modeling low probabilities).

7 Speech Recognition Results

Speech recognition results were conducted in the development test set which has over 6 hours of BN data. The experiments were conducted in a Pentium III 1GHz computer running Linux with 1Gb RAM. Table 3 summarizes the word error rate (% WER) evaluation obtained by AUDIMUS.media. The lines in Table 3 show the increase in performance by each successive improvement to the recognition system.

The first column of results refers to the F0 focus condition where the sentences contain only prepared speech with low background noise and good quality audio. The second results column refers to the WER obtained in all test sentences, including noise, music, spontaneous speech, telephone speech, non-native accents and also including the F0 focus condition sentences.

The first line of results in Table 3 were obtained using MLPs with 1000 hidden units and a stack decoder. Compared with the second line of results we see that there was a significant increase in performance obtained when we switched to the new WFST dynamic decoder, especially in decoding time, expressed in the last column as real-time speed. The third line of results shows the improvement obtained by substituting the determinized lexicon transducer by one that was

Table 3. BN speech recognition evaluation using the development test set

		% WER		
MLPs	Decoder	F0	All	xRT
1000	stack	18.3	33.6	30.0
1000	WFST	18.8	31.6	4.8
1000	+ min det L	18.0	30.7	4.3
4000	"	16.9	29.1	3.7
4000	+ eager pruning	16.7	28.9	3.9
4000	+ shorter HMMs	14.8	26.5	7.6

also minimized. The fourth line shows 8% relative improvements obtained from increasing the hidden layers to 4000 units. This increase was necessary because the acoustic models MLPs were no longer coping with all the variability present in the Speech Recognition and Pilot corpus that were used as training data. The fifth line shows the positive effect of the eager pruning mechanism described in Sect. 6.2. Our current system, shown in line six, achieves another 8% relative improvement by decreasing the minimum duration of phone models by one frame.

8 Concluding Remarks

Broadcast News speech recognition is a very difficult and resource demanding task. Our recognition engine evolved substantially through the accumulation of relatively small improvements. We are still far from perfect recognition, the ultimate goal, nevertheless our current technology is able to drive a number of very useful applications, including audio archive indexing and topic retrieval.

In this paper we have described a number of improvements that permitted a relative recognition error decrease of more than 20% and speed-up from 30x real-time to as little as 7.6 xRT.

References

1. Amaral, R., Langlois, T., Meinedo, H., Neto, J., Souto, N., Trancoso, I.: The development of a portuguese version of a media watch system. In: Proceedings EUROSPEECH 2001, Aalborg, Denmark (2001)
2. Meinedo, H., Neto, J.: Combination of acoustic models in continuous speech recognition. In: Proceedings ICSLP 2000, Beijing, China (2000)
3. Mohri, M., Pereira, F., Riley, M.: Weighted finite-state transducers in speech recognition. In: ASR 2000 Workshop. (2000)
4. Caseiro, D., Trancoso, I.: Using dynamic wfst composition for recognizing broadcast news. In: Proc. ICSLP '2002, Denver, Colorado, USA (2002)
5. Caseiro, D., Trancoso, I.: On integrating the lexicon with the language model. In: Proc. Eurospeech '2001, Aalborg, Denmark (2001)
6. Caseiro, D., Trancoso, I.: Transducer composition for "on-the-fly" lexicon and language model integration. In: Proc. ICASSP '2003, Hong Kong, China (2003)

7. Caseiro, D., Silva, F.M., Trancoso, I., Viana, C.: Automatic alignment of map task dialogs using wfsts. In: Proc. PMLA, ISCA Tutorial and Research Workshop on Pronunciation Modelling and Lexicon Adaptation, Aspen, Colorado, USA (2002)
8. Renals, S., Hochberg, M.: Efficient search using posterior phone probability estimates. In: Proc. ICASSP '95, Detroit, MI (1995) 596–599