Transcrigal: A Bilingual System for Automatic Indexing of Broadcast News

**Carmen Garcia-Mateo, Javier Dieguez-Tirado,
Laura Docio-Fernandez, Antonio Cardenal-Lopez**

Dpto. Teoria de la Señal y Comunicaciones
ETSI Telecomunicacion – University of Vigo
VIGO (SPAIN)
carmen,jdieguez,ldocio,cardenal@gts.tsc.uvigo.es

Abstract

This paper describes a Broadcast News (BN) database called Transcrigal-DB. The news shows are mainly in Galician language, although around 11% of data is in Spanish. This database has been constructed for automatic speech recognition (ASR) purposes. A BN-ASR reference system is also described and evaluated on the test partition of Transcrigal-DB. The reference system has been designed having in mind that both languages, Spanish and Galician, may be used. Performance of the reference is improved when language adaptation techniques are taken into consideration.

1. Introduction

Automatic transcription of broadcast news (BN) is still a challenging recognition problem due to many unresolved issues: frequent and unpredictable changes that occur in speaker, speaking style, topic, channel and background conditions, vocabulary, etc. In order to conduct research in this area, high-quality language resources (LRs) are required. Although BN language resources for languages such as English do exist, this is not the case for many others. This is specially true for a minotity language like Galician, a romance language spoken in Galicia (Spain) similar to Spanish and Portuguese.

The bilingual aspect of this work derives from the fact that both Spanish and Galician languages coexist in Galicia. While anchorpersons and reporters in BN programs use only Galician, the rest of the speakers may use any of these two languages indistinctly and even change between them in the course of the same conversation. This latter aspect makes speech recognition difficult if no language detection algorithm is envisaged. Our approach is to develop a "bilingual" ASR system capable of handling both languages at the decoding stage.

Our reference ASR system uses a bilingual language model as well as bilingual acoustic models as baseline LRs. To improve upon the results of the baseline system, we adapt the language models to language, topic and style, and also perform acoustic adaptation for the reporters.

The rest of the paper is organized as follows. In Section 2 the BN database called Transcrigal-DB is described. Section 3 gives a brief description of other Language Resources (LRs) required to build the bilingual BN-ASR system. Section 4 is devoted to the Speaker Segmentation algorithm. In Section 5, the BN-ASR system is described while its performance on several configurations is reviewed in Section 6. Finally, some conclusions and guidelines for future work are given in section 7.

2. The Transcrigal-DB database

The collection and annotation of the data is still on progress, but 14 BN shows have already been selected to become the first release of Transcrigal-DB. These recordings have been made in AVI format and appended with the corresponding transcriptions generated using Transcriber software (Barras et al., 2000). These broadcasts were captured from the afternoon edition of the TV news show "Telexornal" of public Galician Television (TVG), during October 2002, and consist of approximately 15 hours of material. Part of this material has been incorporated to the COST278 pan-European Broadcast News Database (Vandecatseye et al., 2004). Each show consists of 3 differentiated, sequential blocks with its corresponding anchorperson (Table 1). Blocks 1 and 2 include "Speakers" in addition to "Reporters". Reporters speak only in Galician while Speakers may use Spanish, Galician or a personal combination of both. Approximately 11% of each BN show is in Spanish.

Block	Content	Anchor	Approx. length
1	General news	female	40 mins
2	Sports	female	15 mins
3	Weather forecast	male	5 mins

Table 1: Data distribution within each news show of Transcrigal-DB

In order to be able to exploit this limited database in a more efficient way, we have established 2 different partitions, each consisting of a set of training, testing and validation data (Table 2). BN shows for each partition were chosen at random. Additionally, the testing material within each partition (9 shows each) is divided into three 3-show subpartitions. The reason for doing that is to increase the amount of available material for language model adaptation. Therefore, testing is performed in series of six experiments using 3 shows each.

part.	train	valid.	test
A	1,2,3	4,5	6,7,8 / 9,10,11 / 12,13,14
B	4,7,14	2,13	3,1,5 / 6,8,9 / 10,11,12

Table 2: Partitions of the Transcrigal-DB database

Table 3 lists the percentage of data in the test-set by class of speakers. We have grouped speakers into four

Block	Group	Partition A		Partition B	
		Male	Female	Male	Female
1	Anchor	–	22.66	–	23.15
	Reporters	16.87	17.02	16.72	17.78
	SpeakersGA	4.70	1.59	4.30	1.59
	SpeakersSP	6.45	0.63	5.63	0.90
2	Anchor	–	5.02	–	5.05
	Reporters	7.43	5.71	8.53	4.82
	SpeakersGA	1.13	–	0.92	–
	SpeakersSP	4.13	–	4.39	–
3	Anchor	6.65	–	6.20	–
		47.36	52.63	46.69	53.29

Table 3: Test-set speaker classification (% of testing data)

classes: anchorpersons, reporters, speakersGA (all the other Galician-speaking persons) and speakersSP (all the other Spanish-speaking persons).

3. Other Spoken and Textual LRs

In addition to the Transcrigal-DB database, we have collected two other LRs required to design the bilingual ASR system:

- Approximately 40 hours of speech (15 hours in Galician and 25 hours in Spanish), taken from the corresponding SpeechDAT databases have been used for training the acoustic models.
- A corpus of journalistic text (Table 4) to train the statistical language models. ECG stands for "El Correo Gallego" and "GH" stands for "Galicia Hoxe". These as well as "Vieiros" are Galician journals captured on the Internet. The text named "Escaletas" corresponds to news prompts provided by TVG TV station. Finally, "TRS' ' are the actual transcriptions of the BN shows.

Name	Lang	Time frame (mm/yy)	MBytes
ECG	SP	12/00 - 05/03	270
ECG	GA	12/00 - 12/03	116
GH	GA	05/03 - 12/03	45
Vieiros	GA	03/01 - 12/03	10
Escaletas	GA	10/02 - 03/03	12
TRS	GA-SP	10/02	1

Table 4: Text corpus collected for the developing of Language Models

4. Speaker segmentation module

In multimedia recordings like these, we have two related information sources, namely audio and video streams. Video and audio events are often synchronized: acoustic changes are more likely to occur in the neighborhood of video shot boundaries. With this consideration in mind, a multimedia approach (audio + video processing) for audio segmentation was designed. Such an approach is based primarily in the Bayesian Information Criterion (BIC), and

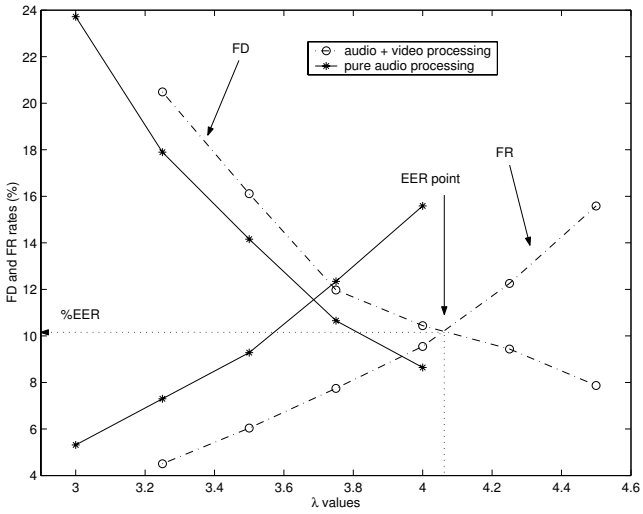

Figure 1: Performance of the speaker segmentation module with and without considering video information)

in addition takes into account useful information extracted from the video stream to improve performance. Details about the algorithm are described in (Perez-Freire and Garcia-Mateo, 2004). Results obtained on Transcrigal-DB are shown in Figure 1. As can be seen, inclusion of video information leads to a remarkable reduction of the overall error rates (for example, Equal Error Rate (EER) goes down from 12% to 10%). Identification of anchors and gender classification is performed using a GMM-based speaker recognition framework.

5. Automatic Speech Recognizer

The recognition engine is a two-pass recognizer: (i) a Viterbi algorithm which works in a synchronous way with a beam search; and (ii) an A^* algorithm. This recognizer was developed for large vocabulary continuous speech recognition applications (Cardenal-Lopez et al., 2002).

5.1. Acoustic Modeling

We should cope with two problems: (i) the lack of large speech databases to train the acoustic models, and (ii) to cover both Galician and Spanish languages.

To train the acoustic models we start from a set of seed models built from the Galician and Spanish SpeechDAT databases (Docio-Fernandez and Garcia-Mateo, 2004). As training data we have used 15 hours in Galician and 25 hours in Spanish. These speech corpora were recorded through the public fixed telephone network, sampled at 8 KHz and codified by the A-law using 8 bits per sample.

The recognition engine makes use of continuous density hidden Markov models (CDHMM). As acoustic units we used demiphones. We used 627 demiphones. Each demiphone consists of a 2-state HMM. Each HMM-state is modeled by a mixture of 4 or 6 Gaussian distributions with a 39-dimensional feature space: 12 mel-frequency cepstrum coefficients (MFCC), normalized log-energy, and their first- and second-order time derivatives.

To compensate for the acoustic mismatch between training models and test data, and also to adapt speaker inde-

type	LM	#n-grams			validation		test	
		1	2	3	PPL	%OOV	PPL	%OOV
component	JournalSP	20.1K	3.08M	3.25M	505.9	12.97	523.7	12.88
	JournalGA	20.1K	2.79M	2.48M	174.5	6.67	196.2	7.37
	escaletas	20.8K	0.47M	0.25M	218.1	6.55	253.3	7.22
	TRS	12.8K	0.05M	0.06M	276.8	8.95	293.3	9.66
baseline candidates	1	20K	2.49M	2.36M	149.3	5.14	168.7	5.55
	2	20K	4.19M	4.82M	140.8	4.88	155.3	5.20
	3	20K	4.42M	5.00M	138.4	4.94	154.1	5.21
	4	**20K**	**4.32M**	**4.93M**	**130.7**	**4.50**	**144.8**	**4.81**

Table 5: Reference language models

pendent system to individual speakers, we have used supervised acoustic adaptation based on MLLR (Maximum Likelihood Linear Regression) and MAP (Maximum a Posteriori) techniques. Only material from F0 focus condition (studio, planed, native, clean) was used to adapt our seed model set. A total of 22 and 24 minutes of speech have been used for adaptation purposes in parttion A and B respectively. The adaptation process is done in three passes. On the first pass a global speech MLLR adaptation is performed. The second pass uses the global transformation on the model set, producing better frame/state alignments. This information is used to estimate a set of more specific transforms, using a regression class tree. Finally, the previous MLLR-adapted models are further improved using the MAP technique.

5.2. Language Modeling

In order to build the bilingual language models (LM), first we have trained separate language models for different kinds of text sources, namely *JournalSP* (text in Spanish from ECG), *JournalGA* (text in Galician from ECG and Vieiros), *escaletas*, and *TRS*. For this latter case, 3 LM are trained for each partition using 9 news shows each. Upper half of Table 5 shows the size of the LM in n-grams, and perplexity (PPL) and out-of-vocabulary rate (OOV) on the validation and test sets, averaged for the 2 partitions.

Secondly, these individual LMs are combined in different configurations using linear interpolation in order to find a baseline LM. Last, this baseline LM is further improved by adapting to block, language and style.

5.2.1. Baseline Language Model

We have combined the component LMs in 4 different ways in order to choose the best-performing baseline LM.

The weights that minimize perplexity of the validation set are shown in Table 6 across several combination configurations. It can be seen that JournalSP has got a low weight compared to the rest of the components, even though this LM has been trained with more text. This is because BN programs have got a very limited amount of Spanish. Similarly, TRS LMs have been trained with little material but have got a high weight in the interpolated LMs, because they are the LMs that better model the spontanous speech present in the BN shows.

The perplexities and OOV rate of these language models are shown in the second half of Table 5. It can be seen that they show better performance than the component LMs. We have chosen *LM number 4* as our baseline model.

LM id	component weights			
	JournalSP	JournalGA	escaletas	TRS
1	–	0.6834	–	0.3166
2	0.1258	0.6212	–	0.2530
3	0.1453	0.4653	0.3894	–
4	**0.1147**	**0.4215**	**0.2919**	**0.1719**

Table 6: Components of the baseline LM candidates

5.2.2. Block-Adapted LMs for planned Galician

An immediate improvement is achieved by creating a different LM for each block of a BN show. (Tab.1). As the structure of a BN show is fixed, the appropriate LM can always be applied without the need for a classifier.

The LMs were constructed by using the 4 components of baseline LM, separating the validation text in three different blocks and recalculating the weights for each validation block.

However, this time we chose only planned (F0 focus) Galician speech in the validation material, as we found that introducing spontaneous text degrades the overall WER.

Having built a Galician planned-speech LM for each block, the next step will be to train a LM for spontaneous speech, both for Galician and Spanish speakers.

5.2.3. Spontaneous speech LMs

Spontaneous speech LMs were created for each block. This was again accomplished by classifying the corresponding validation text into these categories and choosing the optimal mixture weights in each case.

In block 1 there is enough material for both Spanish and Galician speakers (Tab. 1) so a separate LM was trained for each. For block 2 (Sport news) we found that most of the speakers are non-native so we opted for creating a bilingual LM for spontaneous speech merging both kinds of speakers. Block 3 (Weather) consists solely of a single anchorperson turn, so no spontaneous models were needed.

6. Experimental Results

Table 7 shows the Word Recognition Rate (%Corr) across the different classes of speakers for the experiments

Block	Group	Reference: Baseline LM #4		Experiment 1: Block-Adapted LM		Experiment 2: +LM spont.		Experiment 3: +Ac anch & rep.	
		M	F	M	F	M	F	M	F
1	Anchor	–	79.72	–	**80.75**	–	80.75	–	**83.32**
	Reporters	50.58	70.83	**52.08**	**72.24**	52.08	72.24	**64.48**	**70.13**
	SpeakersGA	28.56	49.19	28.85	47.67	**29.83**	**49.97**	29.83	49.97
	SpeakersSP	22.07	40.28	17.66	30.95	**30.60**	**54.18**	30.60	54.18
2	Anchor	–	70.63	–	**74.70**	–	74.70	–	**76.59**
	Reporters	47.98	56.66	**52.01**	**62.15**	52.01	62.15	**58.86**	**60.04**
	SpeakersGA	16.28	–	14.41	–	**25.84**	–	25.84	–
	SpeakersSP	15.05	–	12.52	–	**23.80**	–	23.80	–
3	Anchor	67.02	–	**74.23**	–	74.23	–	**78.34**	–
		42.58	**72.16**	**43.96**	**73.82**	**47.02**	**74.22**	**53.15**	**74.61**

Table 7: Recognition results (Word Recognition Rate, average 2 partitions)

we have conducted. Fist of all we extracted reference results using our baseline acoustic and language models (Sections 5.1. and 5.2.1. respectively), to be compared with another 3 experiments performed.

Experiment 1 uses block-adapted LMs (Sec. 5.2.2.). As these are planned speech LMs, results for "Speakers" are worse than the reference, but the %Corr rate for reporters and anchorpersons is better. Blocks 2 and 3 are the most improved because the baseline LM was more biased to optimize block 1 as it had the most material (Tab. 1).

The second experiment was performed using spontaneous speech LMs for SpeakersGA and SpeakersSP (Sec. 5.2.3.), which increased their recognition rate.

The last experiment shows the effect of using adapted acoustic models for each of the anchorpersons, male reporters (excluding anchors) and female reporters (excluding anchors). We find that anchorpersons and male speakers improve significantly, however for female reporters the adapted models perform worse than the baseline ones. Therefore, leaving out the training material of the female anchorpersons for adaptation seems counterproductive. If the baseline acoustic models would have been used for female reporters, the recognition rate for female speakers would be of 75.51% instead of 74.61%.

7. Discussion and Further Work

In this paper a system for automatic indexing of bilingual BN shows has been presented. Its development has entailed the compilation of a new database for Galician BN, and the building of a specialized ASR, both described on the beginning sections.

We have also presented some results obtained with our baseline experiment. Several improvements have been achieved by adapting the language models to topic, language ans style, and the acoustic models to reporters and anchorpersons. This has improved the Word Recognition Rate for each of the speaker classes.

However, in spite of our efforts to build a bilingual system, we find that the recognition rate for Spanish is still very low. The explanation can be found in the fact that all the Spanish speakers present in the BN shows use spontaneous speech, and the language models we have used are based mainly on journalistic sources. The small amount of manual transcriptions helps to build specific spontaneous speech LMs, but this material is still very limited to achieve a good performance. We aim to overcome this limitation by incorporating text from novels, and by transcribing more BN programs (Transcrigal-DB2 database).

In addition to these techniques, there are other well known strategies to improve language modeling. We are working on incorporating cache-based language modeling in our decoder (Clarkson and Robinson, 1997) as well as LM adaptation using information retrieval techniques (Mahajan et al., 1999).

8. Acknowledgements

This project has been partially supported by Spanish MCyT under the projects TIC2000-1104-C02-01 and TIC2002-02208, and Xunta de Galicia under the projects PGIDT01PX132201PN and PGIDT03PXIC32201PN.

9. References

Barras, C., E. Geoffrois, Z. Wu, and M. Liberman, 2000. Transcriber: development and use of a tool for assisting speech corpora production. *Speech Communication, Vol 33, No 1-2*.

Cardenal-Lopez, A., F.J. Dieguez-Tirado, and C. Garcia-Mateo, 2002. Fast LM look-ahead for large vocabulary continuous speech recognition using perfect hashing. In *Proc. ICASSP*.

Clarkson, P. and A. Robinson, 1997. Language Model Adaptation Using Mixtures and an Exponentially Decaying Cache. In *Proc. ICASSP*.

Docio-Fernandez, L. and C. Garcia-Mateo, 2004. Acoustic Modeling and Training of a Bilingual ASR System when a Minority Language is Involved. In *Proc. of LREC 2002, Gran Canaria (Spain)*.

Mahajan, M., D. Beeferman, and X.D. Huang, 1999. Improved Topic-Dependent Language Modeling Using Information Retrieval Techniques. In *Proc. ICASSP*.

Perez-Freire, L. and C. Garcia-Mateo, 2004. A multimedia approach for audio segmentation in tv broadcast news. *Accepted for publication at Proc. of ICASSP 2004, Montreal (Canada)*.

Vandecatseye, A., J.P. Martens, J. Neto, H. Meinedo, C. Garcia-Mateo, F.J. Dieguez, F. Mihelic, J. Zibert, J. Nouza, P. David, M. Pleva, A. Cizmar, H. Papageorgiou, and C. Alexandris, 2004. The COST78 pan-European Broadcast News Database. In *Proc. of LREC 2004, Lisbon (Portugal)*.