ADAPTATION STRATEGIES FOR THE ACOUSTIC AND LANGUAGE MODELS IN BILINGUAL SPEECH TRANSCRIPTION

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

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

ABSTRACT

This paper describes our current work on speech-to-text transcription for recordings in two languages. The experimental framework consists of Television News shows in Galician and at some extent Spanish language. A priori language detection is avoided so a bilingual speech recognition system has been developed and its performance is presented. Better results are obtained when speaker and speech style is taken into account through adaptation of both acoustic and language models. Special attention must be paid to the limited resources available in the experimental framework.

1. INTRODUCTION

Speech transcription in multiple languages has drawn considerable attention in the recent years. Usual tasks in this field include conversational telephone speech (CTS) and broadcast news (BN), where several systems for different languages are being developed and compared [1, 2].

Our novel approach is to develop a system in a bilingual scenario for the BN task. Thus, in addition to the usual problems of changing acoustic conditions, speakers, styles and topics, we add the problem of changing language. For this purpose we have captured and annotated a bilingual database of news shows broadcasted in the Galician region of Spain (Transcrigal-DB). These news shows consist mostly of speech in Galician language, but some non-reporter speakers may also use Spanish language, and even a speaker-dependent mixture of both languages.

Our aim is to provide a rich transcription of these news shows, where the textual representation of speech is also tagged with information such as topic, language or speaker type. At the same time, we want to avoid a priori detection of language. For this reason, we have relied upon the significant overlap between Galician and Spanish [3], and have designed a bilingual transcription system that takes into account both languages simultaneously. This system, "Transcrigal" [4], has been developed under very limited resources, and is based on multiple passes. The first pass relies only on the audio signal to extract a transcription, while subsequent passes may use intermediate transcriptions to dynamically adapt acoustic and language models (LMs).

In this paper we concentrate solely on the first pass of Transcrigal, and we compare two different approaches:

- "universal", using a unique set of universal acoustic and language models to cover all speech conditions at once.
- "adapted", training several sets of acoustic and language models constrained to different conditions, whilst providing a way to choose the correct models for each speaker turn prior to recognition.

In order to attain an effective adaptation, an emphasis has been put to exploit the limited training material in the best possible way. Adapted acoustic models were trained for the anchorpersons, and for male and female speakers. Also, six different adapted LMs were trained to take into account different topic, style and language combinations.

We find that using adapted models considerably improves the performance of the system. In particular, the adapted LMs enhance the bilingual capabilities of the system by improving the transcription of the small amount of Spanish speech.

The rest of the paper is organized as follows: first we present an overview of the Transcrigal framework (Sec. 2), we then explain the acoustic and language modeling in detail (Secs. 3 and 4), we continue showing our experimental results (Sec. 5), and finally we give our conclusions and outline our future work (Sec. 6).

2. EXPERIMENTAL FRAMEWORK

Our framework to investigate bilingual transcription is the Transcrigal Galician-Spanish BN system [4], which consists of a BN database (Transcrigal-DB) collected in our laboratory, a set of language resources (LRs) and an automatic speech recognition (ASR) engine.

Transcrigal-DB is a collection of 14 news shows collected from Galician Television broadcasts during October 2002, and is currently being augmented with 14 new shows recorded in 2003–2004. The structure of the news shows is presented in Table 1. Each show is approximately 60 minutes long, and consists of three well separated blocks with a corresponding anchorperson: "news", "sports" and "weather". Speakers may be classified in three main

Block	GA reporters	GA non-rep.	ES non-rep.	Total
News	58.01	6.58	6.57	71.17
Sports	17.71	0.81	4.27	22.78
Weather	6.05	–	–	6.05
Total	81.77	7.39	**10.84**	100.0

Table 1. Structure of the news shows (%words, average 14 shows)

part.	train		valid.	test
	acoustic	LM		
A1	1,2,3	1,2,3 + test A2,A3	4,5	6,7,8
A2	1,2,3	1,2,3 + test A1,A3	4,5	9,10,11
A3	1,2,3	1,2,3 + test A1,A2	4,5	12,13,14
B1	4,7,14	4,7,14 + test B2,B3	2,13	3,1,5
B2	4,7,14	4,7,14 + test B1,B3	2,13	6,8,9
B3	4,7,14	4,7,14 + test B2,B3	2,13	10,11,12

Table 2. Transcrigal-DB partitions

Set	Type	Lang.	Description	Size
Transcrigal-DB	Speech	bil.	Audio from BN shows	14h
	Text	bil.	Manual transcriptions	1MB
Other	Speech	ES	SpeechDAT	15 h
		GA	SpeechDAT	25 h
	Text	GA	Closed-caps. 10/02-03/03	12 MB
			Journals 12/00-12/03	271 MB
		ES	Journals 12/00-05/03	270 MB

Table 3. Language resources used in Transcrigal

groups: Galician reporters, Galician non-reporters and Spanish non-reporters. There are an average of 152 speaker turns for each show.

Both audio and video have been captured for each show. The audio portion of the recordings has been annotated by means of the Transcriber tool [5]. The video has proven useful in improving acoustic segmentation [6]. The Transcrigal database has been partitioned in three subsets: train (9 shows), validation (2 shows) and test (3 shows). Due to the limited amount of manually annotated material, six different partitions have been made in order to rotate training and test data. Table 2 shows the partitions, where each show has been chronologically numbered from 1 to 14.

The Transcrigal-DB is complemented by a number of LRs, in order to build the acoustic and language models. In Table 3 we summarize the LRs collected for Galician, Spanish and bilingual.

The ASR 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 [7].

We can compare Transcrigal to state of the art BN systems for different languages [1], taking into account the amount of resources available. We find that our system is similar in both training resources and word error rate (WER) to the Portuguese LIMSI system.

3. 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 [3]. 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 to 8 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.

Type	# MLRR regr. classes	# minutes F0 (avg A,B)
Universal	16	27.0
News anchor	12	7.1
Sports anchor	6	4.6
Weather anchor	6	3.9
Male reporters (non-anchor)	6	5.2
Female reporters (non-anchor)	6	6.2

Table 4. Universal and particular acoustic models

To compensate for the acoustic mismatch between training models and test data, and also to adapt speaker independent 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, planned, native, clean) was used to adapt our seed model set. 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.

Using this adaptation framework, an universal acoustic model set has been created by using F0 material of each available speaker, and 5 particular model sets have been trained by selecting F0 material only from male reporters, female reporters, and each anchorperson, respectively. Their characteristics are shown in Table 4.

4. LANGUAGE MODELING

Our problems for language modeling are the same as in the acoustic case: (i) the available text databases are very limited and (ii) we have to cover both languages.

In order to combine all available text sources in an efficient manner, we have used mixture based n-gram models [8]. The mixture process consists of several steps. First, separate trigram language models are trained for each of the four text sources: bilingual training-set manual transcriptions (TRS), Spanish Journals (jour-ES), Galician journals (jour-GA), and closed captions (cap-GA). We have used Good-Turing discounting and backoffs. Next, the component LMs are linearly interpolated. The mixing weights are chosen by minimizing the perplexity of the validation set transcriptions, by means of the EM algorithm. Lastly, the vocabulary size is limited to 20K words, and entropy-based pruning [9] is performed with a $2,5 \cdot 10^{-8}$ threshold, to limit the LM size to 1,2M bigrams and 0,8M trigrams on average with only a small perplexity increase.

In an analogous manner to the acoustic models, both universal and particular LMs can be obtained by the same training procedure (Fig. 1): for universal models weights are chosen by minimizing the perplexity of the entire validation transcriptions. For the particular models only a subset of the validation TRS is used. In this last case, to improve matching between validation and training data, we have performed additional filtering of some LM sources, and

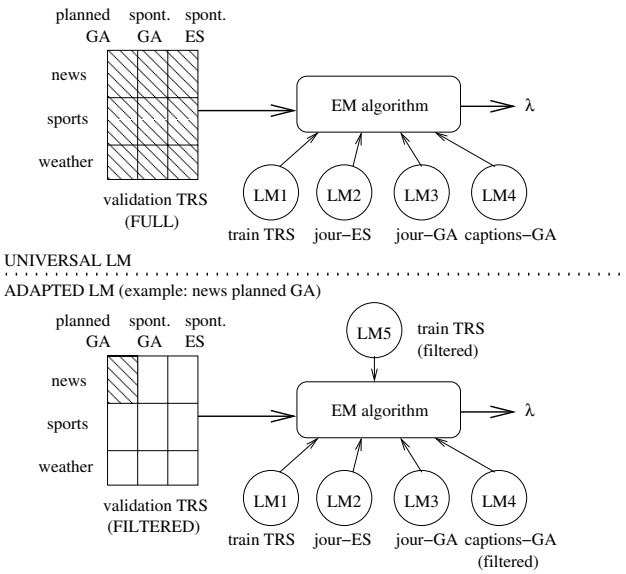

Fig. 1. Mixture LM training: universal and adapted

Block	Type	Lang.	λ_1 trs	λ_2 j-GA	λ_3 j-ES	λ_4 cap	λ_5 trs-f
universal			0.17	**0.42**	0.11	0.29	–
news	planned	GA	0.01	**0.55**	0.02	0.38	0.04
	spont.	GA	0.06	**0.66**	0.08	0.05	0.15
	spont.	ES	0.01	0.03	**0.80**	0.00	0.15
sports	planned	GA	0.02	0.29	0.02	**0.55**	0.12
	spont.	GA+ES	0.01	0.03	**0.61**	0.02	0.33
weather	planned	GA	0.04	0.09	0.00	0.07	**0.79**

Table 5. Mixture LM weights (average 6 partitions)

Block	Speaker Class	#words	% OOV		Perplexity	
			univ.	adap.	univ.	adap.
news	anchor + reps.	106304	4.68	4.33	118.4	110.5
	non-rep. GA	11346	5.62	5.54	231.7	207.8
	non-rep. ES	12656	6.89	**4.20**	381.5	**140.1**
sports	anchor + reps.	34036	5.36	3.43	198.5	139.3
	non-rep.	9795	5.17	**3.18**	403.8	**129.3**
weather	anchor	11965	1.19	**0.87**	77.7	**27.3**
		186102	4.81	**3.95**	151.8	**111.8**

Table 6. Lexicon coverage and perplexity (avg. 6 partitions)

we have added one additional source consisting of style-matched manual transcriptions.

In order to choose the conditions on which to adapt the particular LMs, we have taken advantage of the structure of the BN programs (Table 1), where there are three consecutive blocks clearly separated by jingles and commercials: news, sports and weather sections. A straightforward improvement over the universal model would be to train three different LMs adapted for each block, as the correct LM (current block) may always be known. This would constitute a rudimentary topic adaptation.

However, larger improvements may be expected by also adapting to language and style. Therefore, we have created three different language models for the "news" block: planned GA, spontaneous GA and spontaneous ES. In this case, a mechanism to select among them before recognition must be chosen. We have tested three methods: (i) to use always the planned-GA LM, which will benefit the majority of the speakers (ii) to run three decoders in parallel using the different LMs, and selecting the better scored execution (iii) to follow an oracle-based approach, simulating a perfect detector to serve as a performance ceiling.

In the case of the "sports" block, we found that there was very little amount of spontaneous material. So, to improve training and to reduce decision confusability, we decided to merge both spontaneous ES and GA in a single LM. The "weather" block always consists of a single anchor turn, so only one LM is needed for it.

Table 5 shows the weights of the universal and the six particular language models. The universal language model shares similar weights with the planned-GA "news" LM, as this is the most common style in the BN shows. The remaining particular LMs deviate from this behavior, and more importance is given to different sources.

5. EXPERIMENTAL RESULTS

We have performed language model experiments and recognition experiments using our test set.

Our LM experiments measure the potential improvement of the adapted language models by means of lexicon coverage and perplexity. For this purpose, we have divided the test data into six groups, and have applied both the universal and the adequate (oracle) adapted language model to each group. In order for perplexities to be comparable, each LM pair was constrained to their intersected vocabulary before measuring perplexity. Results are presented in Table 6, showing a 17.9% relative out of vocabulary (OOV) rate decrease and a 26.3% relative perplexity reduction in the overall test. We find that some speaker groups would benefit more than others from LM adaptation, particularly Spanish speakers. The weather section also shows a considerable improvement, as its very homogeneous language style becomes better modeled by increasing the weight of the small amount of available manual transcriptions (Table 5).

For the recognition experiments, we have made a number of assumptions, as we wanted to concentrate on the problems derived from bilingualism. First, we have used manual turn segmentation, avoiding problems introduced by segmentation errors. Secondly, for the use of adapted acoustic models, we also assume that a reasonably good acoustic detector may be implemented that classifies the speaker turn into "male", "female" or "anchorperson" using Gaussian mixture models (GMMs). As we have still not implemented this detector, for these experiments we assume an oracle-based approach. The pruning parameters of the ASR engine have been tuned to achieve approximately 3x real-time (RT) decoding on current x86 servers (2,4–3,06 GHz).

The recognition experiments and results are presented in Table 7, where we have separated results for Galician (GA) and Spanish (ES). We have performed five experiments, using both universal and adapted acoustic language models (experiment types "a" and "b", respectively). The first recognition experiment uses universal language models. The second experiment uses a straightforward topic or "block" adaptation, and experiments 3–5 analyze three different ways of choosing the adequate adapted language model for each speaker turn, as proposed in Section 4.

The results indicate that LM and acoustic improvements are approximately additive. The topic adaptation proves itself as a

Acoustic Models	exp.	LM	% Word error rate		
			GA	ES	Total
Universal	1-a	universal	36.98	79.35	41.72
	2-a	topic adapted	35.44	79.15	40.32
	3-a	fully adapted planned-GA	34.73	83.39	40.17
	4-a	fully adapted parallel	**34.60**	**72.93**	**38.88**
	5-a	fully adapted oracle	34.50	70.51	38.53
Adapted	1-b	universal	32.14	72.17	36.62
	2-b	topic adapted universal	30.70	71.91	35.31
	3-b	fully adapted planned-GA	30.15	77.49	35.44
	4-b	fully adapted parallel	**29.96**	**62.33**	**33.57**
	5-b	fully adapted oracle	29.90	60.95	33.37

Table 7. Recognition results for each language (avg. 6 partitions)

simple way to improve results for both languages over the baseline case. The fully adapted models yield further improvements. Using always the planned-GA model may benefit the total result but degrade Spanish turns. Using a parallel approach proves to be almost as good as using the ideal decoder, at the expense of more computational time. The realistic case of using an acoustic detector with parallel adapted language models (exp. 4-b) provides a 19.5% relative improvement in word error rate for the overall test, and a 23.2% improvement for Spanish speakers, enhancing the bilingual capabilities of the system.

6. CONCLUSIONS AND FUTURE WORK

In this paper we have presented our current work regarding bilingual speech transcription. We have used the Transcrigal Galician-Spanish BN system as a our framework, and have investigated the handling of changing conditions, in particular multiple speakers, speaking styles, topics and languages. We have shown adaptation methods for both the acoustic and language models, and have successfully applied them to provide a small performance increase for the overall test, and a considerable improvement for the minor portion of spontaneous, Spanish speakers.

However, we find that the performance for Galician and Spanish in our system is still unbalanced. This is because Galician corresponds mainly to planned speech, while the Spanish turns are mainly spontaneous. As our LRs do not contain spontaneous text databases, the training for spontaneous speech relies mainly on the limited number of manual transcriptions. In addition to this problem, the language modeling method for this kind of speech needs to be questioned, as we are not taking into account speech disfluencies that normally occur. In the near future, we will double the amount of manual transcriptions in order to improve training, and are working towards acquiring spontaneous Galician and Spanish transcriptions from films and novels.

Our future work will be centered in the second pass of the system, where we aim to introduce further adaptation. Using the first-pass transcription, we will perform speaker turn-dependent acoustic model adaptation, and news-story based topic adaptation using information retrieval techniques [10, 11], in order to improve the transcription of both languages.

7. ACKNOWLEDGEMENTS

This project has been partially supported by Spanish MCyT under the project TIC2002-02208, and Xunta de Galicia under the project PGIDT03PXIC32201PN.

8. REFERENCES

[1] L. Lamel, J-L. Gauvain, G. Adda, M. Adda-Decker, L. Canseco, L. Chen, O. Galibert, A. Messaoudi, and H. Schwenk, "Speech transcription in multiple languages," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, vol. 3, pp. 757–760.

[2] R. Schwartz, T. Colthurst, N. Duta, H. Gish, R. Iyer, C-L. Kao, D. Liu, O. Kimball, J. Ma, J. Makhoul, S. Matsoukas, L. Nguyen, M. Noamany, R. Prasad, B. Xiang, D-X. Xu, J-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and L. Chen, "Speech recognition in multiple languages and domains: The 2003 BBN/LIMSI EARS system," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, vol. 3, pp. 753–756.

[3] L. Docio-Fernandez and C. Garcia-Mateo, "Acoustic modeling and training of a bilingual ASR system when a minority language is involved," in *Proc. Int. Conf. on Language Resources and Evaluation*, Gran Canaria, Spain, May 2002, vol. 3, pp. 873–876.

[4] C. Garcia-Mateo, J. Dieguez-Tirado, A. Cardenal-Lopez, and L. Docio-Fernandez, "Transcrigal: A bilingual system for automatic indexing of broadcast news," in *Proc. Int. Conf. on Language Resources and Evaluation*, Lisbon, Portugal, May 2004.

[5] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: development and use of a tool for assisting speech corpora production," *Speech Communication*, vol. 33, no. 1–2, pp. 5–22, January 2001.

[6] L. Perez-Freire and C. Garcia-Mateo, "A multimedia approach for audio segmentation in TV broadcast news," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Montreal, Canada, May 2004, vol. 1, pp. 369–372.

[7] A. Cardenal-Lopez, F. J. Dieguez-Tirado, and C. Garcia-Mateo, "Fast LM look-ahead for large vocabulary continuous speech recognition using perfect hashing," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Orlando, FL, May 2002, vol. 1, pp. 705–708.

[8] P. Clarkson and A. J. Robinson, "Language model adaptation using mixtures and an exponentially decaying cache," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Munich, Germany, April 1997, vol. 2, pp. 799–802.

[9] A. Stolcke, "Entropy-based pruning of backoff language models," in *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, 1998, pp. 270–274.

[10] L. Chen, J. L. Gauvain, L. Lamel, and G. Adda, "Unsupervised language model adaptation for broadcast news," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Hong Kong, China, April 2003, vol. 1, pp. 220–223.

[11] M. Mahajan, D. Beeferman, and X. D. Huang, "Improved topic-dependent language modeling using information retrieval techniques," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, Phoenix, AZ, March 1999, vol. 1, pp. 541–544.