

TC-STAR 2006 Automatic Speech Recognition Evaluation: The UVIGO System

Laura Docio-Fernandez, Antonio Cardenal-Lopez, Carmen Garcia-Mateo

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

Abstract

This paper describes the ongoing development of the University of Vigo's Automatic Speech Recognition system (UVIGO) for the automatic transcription of Spanish European Parliamentary Plenary sessions and Spanish Parliamentary sessions. The system was developed to partake in the 2006 TC-STAR Automatic Speech Recognition evaluation campaign in the Spanish language section. The UVIGO system was derived from the University of Vigo's Galician Broadcast News (BN) Transcription system by adapting the BN acoustic and language models to the TC-STAR domain. A detailed discussion of the front-end processing, acoustic and language modelling, and decoding process are presented. The proposed decoding strategy was developed to make the best possible use of gender- and speaker-dependent acoustic models without a prior gender or speaker identification process. In addition to describing the system architecture and reporting the evaluation results, we also highlight further improvements that we are planning to make to the overall system.

1. Introduction

The 2006 TC-STAR Automatic Speech Recognition (ASR) evaluation campaign assessed speech recognition performance for three languages (English, Spanish and Mandarin) and two domains (Parliament Plenary Sessions and Broadcast News). Specifically, the following language/task combinations were evaluated: European Parliament Plenary Sessions (EPPS) for European English and Spanish; Spanish Parliament Plenary Sessions (PARL) for Spanish; and Broadcast News (BN) from VOA for Mandarin Chinese.

Participants were able to use data made available by the TC-STAR project and any publicly available data (essentially from LDC and ELDA) that predated the training cut-off date for system development. They were also allowed to use any data they considered would be of use, as long as the data predated the established cut-off date. Three training conditions were proposed for the Parliament tasks:

- A *restricted* training condition, under which systems could only be trained on data supplied as part of the TC-STAR project.
- A *public* data condition, under which systems could be trained on any publicly available data.
- An *open* condition, where the only constraint was a pre-established cut-off date for the training data.

Participants had to submit a system that had been trained under at least the restricted or the public condition.

The following three test conditions will be used for the TC-STAR'06 evaluation:

- European Parliament Plenary Session in English (EPPS_ENG), which should preferably include original speech (as opposed to interpreters' speech). 3-4 hours of test material from September, October and November 2005.
- European Parliament Plenary Session in Spanish (EPPS_SP), which should also preferably include original speeches (as opposed to interpreters' speech). 3-4

hours of test material from September, October and November 2005. Spanish Parliament Plenary Session (PARL_SP). 3-4 hours of test material from November 2005.

- Broadcast News in Mandarin (BN_MAN).

Our research group (the Signal Processing Group at the University of Vigo) has extensive experience in the field of natural language processing, having covered areas ranging from automatic speech and speaker recognition to speech synthesis and voice conversion. We began developing our own state-of-the-art systems several years ago, and as far as ASR systems are concerned, we have developed a Galician BN transcription system called 'Transcrigal' (Garcia-Mateo et al., 2004)(Dieguez-Tirado et al., 2005). Transcrigal works in a bilingual (Galician and Spanish) environment and was developed using very limited linguistic and acoustic resources. To develop the system we captured and annotated a bilingual database of news shows broadcast in the Galician region of Spain. These news shows consisted mostly of speech in the Galician language although they also included Spanish speech by non-reporter speakers and a speaker-dependent mixture of both Galician and Spanish. As mentioned above, the highly limited amount of BN-specific training data available substantially limited the performance of Transcrigal in a number of scenarios such as spontaneous speech. The acoustic and linguistic resources available for the TC-STAR project were considerably larger, however, and we therefore did not expect to run into problems caused by data limitations. BN-type speech is also significantly different from parliament plenary session-type speech, both acoustically (due to the environment) and linguistically (due to the way in which politicians speak).

This paper describes the speech recognition system developed by the University of Vigo (UVIGO) and its performance within the framework of the 2006 TC-STAR ASR evaluation project. As the system developed was basically a slightly modified version of our existing BN recognition system, we expected our results to provide little more than

a baseline for future work. The UVIGO system was submitted to the Spanish ‘restricted’ condition only.

The rest of the paper is organized as follows: we will summarize the audio and text corpora used for training and testing the system (Sec. 2.) before describing the recognition system in detail (Sec. 3.); we will then present the results of our tests on different data sets, (Secs. 5. and 6.) and finally discuss the system (Sec. 7.) and outline our future work (Sec. 8.).

2. Training and test data sets

We investigated both of the Spanish TC-STAR tasks - the European Parliament Plenary Sessions (EPPS) and the Spanish Parliament Sessions (PARL) separately, and focused only on the restricted condition (TC-STAR data only). Our acoustic data therefore consisted of manually transcribed EPPS data (totalling about 57 hours from May, July, September, October and December 2004 and January 2005) and manually transcribed PARL data (totalling about 36 hours from July, September, October and December 2004). Table 1 shows the number of hours of training data classified by gender and by both original speech and interpreters’ speech in the case of EPPS data.

	EPPS	PARL
male	30.44 h (6.75 h speeches)	26.9 h
female	26.61 h (1.2 h speeches)	9.37 h
Total	57.05 h (7.95 h speeches)	36.27 h

Table 1: Selected Spanish TC-STAR acoustic training data sets.

Table 2 shows the sources and text sizes that were made available to generate the language models for the restricted condition. The EPPS transcripts were extracted from EPPS parallel texts supplied by RWTH.

Source	Period	Size (MW)
SP EPPS transcr.	Apr. 1996 to Jun. 1999	12
SP EPPS transcr.	Jul. 1999 to Sep. 2004	18.2
SP EPPS transcr.	Dec. 2004 to May 2005	1.7
SP PARL transcr.	Jan. 2000 to Oct.2004	43.2

Table 2: Description of the Language Resources.

We used the development and evaluation data sets defined for the 2006 TC-STAR evaluation project to assess the performance of our system. The development data for EPPS and PARL were recorded during the first half of June and July 2005 and on December 1st and 2nd 2004 respectively, while the evaluation data were recorded from September to November 2005 (EPPS) and in November 2005 (PARL). Tables 3 and 4 show the total number of minutes that make up each set of data.

3. UVigo system

This section describes the UVIGO speech recognition system, which consists of two main modules: feature extraction and decoding. A description of the acoustic and language model training processes is also provided.

	EPPS	PARL
male	112 min	169 min
female	32 min	47 min
Total	144 min	216 min

Table 3: Selected Spanish TC-STAR development data set: Dev06es.

	EPPS	PARL
male	149 min	143.6 min
female	24 min	50.3 min
Total	173 min	193.9 min

Table 4: Selected Spanish TC-STAR evaluation data set: Eval06es.

3.1. Feature extraction

The acoustic front-end of the UVIGO speech recognition system used 39-dimensional cepstral feature vectors obtained from a Mel frequency spectrum estimated on the 0-8 KHz band every 10 ms. The Mel scale magnitude spectrum was computed using a Hamming window of 25 ms length. This spectrum was then filtered with a filter bank consisting of 27 triangular overlapping filters positioned at equidistant points on the Mel frequency axis. The logarithms of the filter outputs were cepstrally decorrelated (discrete cosine transform) to produce 12 dimensional vectors. The Mel-Frequency Cepstral Coefficients were normalized using utterance-based cepstral mean removal. The normalized log-energy was appended to the 12-dimensional cepstral vector, and the first- and second-order time derivatives of the 13-dimensional feature vector were also joined, resulting in the above-mentioned 39-dimensional feature vector.

3.2. Decoding

The decoding process involved two separate stages.

In the first stage, which followed the feature extraction of the whole file, a log-energy-based VAD was used to provide the recognition engine with an initial segmentation of the input data. Each VAD segment was input to the recognition engine, which performed recognition using a two-pass search strategy (Cardenal-Lopez et al., 2002). Standard synchronous Viterbi alignment involving the application of trigrams and demiphones was performed in the first pass. Several null nodes connected to the edges of the vocabulary tree were used to collect a subset corresponding to the most likely hypothesis on each frame. Of this set, only a few paths were further propagated to the tree, but the remainder were stored for later use. In the second recognition pass, all the previously stored paths were used to build a word lattice, which was then rescored using the same language and acoustic models as in the first pass but with the additional application of an A* algorithm. Second pass recognition usually provides an improvement of approximately 5 to 10 points.

Although we did not use gender detection or speaker segmentation, we did apply male, female and speaker-

independent models to the entire data set. By running the recognition engine through three (and sometimes more) different sets of acoustic models, we obtained three (or more) recognition hypotheses that were combined in the second stage to yield the final recognition hypothesis. The hypotheses were combined using the segmentation provided by the VAD, and only acoustic probabilities were taken into account. In other words, for each VAD segment, the recognition hypothesis with the highest overall acoustic scoring was selected as the final result.

It is worth noting that only one recognition pass was performed in the evaluation process, i.e. the system did not use a second recognition pass involving supervised adaptation from the first recognition pass. Recognition improvements could obviously be expected in cases where the output of the first decoding process is used to supervise MLLR adaptation of available recognition models.

3.3. Acoustic modelling

The recognition engine makes use of continuous density hidden Markov models (CDHMM). As acoustic units we used demiphones (Mario et al., 1997). We used 627 demiphones derived from a total set of 25 phones. Table 5 clearly illustrates the SAMPA/DARPA code of the set of phones used.

unit	SAMPA	unit	SAMPA
#1	p	#14	l
#2	b, B	#15	Z
#3	t	#16	rr
#4	d, D	#17	S
#5	k	#18	s
#6	g, G	#19	N
#7	m	#20	r
#8	n	#21	a
#9	J	#22	e
#10	tS	#23	i
#11	f	#24	o
#12	T	#25	u
#13	x		

Table 5: Set of used phones. The allophonic variations b and B are merged in an unique unit. The same is made for d and D, and for g and G.

Each demiphone consisted of a 2-state HMM, and each HMM-state was modelled using a mixture of 4 to 8 Gaussian distributions.

To obtain the acoustic models we used a scheme for rapidly porting/adapting a source system to a new target task. In previous work (Gales et al., 2003) a number of porting techniques were described. We started from a source set of models built from the Galician and Spanish SpeechDAT databases (Docio-Fernandez and Garcia-Mateo, 2002). 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. To compensate for the acoustic mismatch between these seed models and the TC-STAR data, we then used supervised acoustic adaptation based on MLLR (Maximum Likelihood Linear

Regression) and MAP (Maximum a Posteriori) techniques. The adaptation process was performed in three passes: a first pass involving global speech MLLR adaptation; a second pass involving global transformation of the model set to produce better frame/state alignments (this information was used to estimate a set of more specific transforms using a regression class tree); and a third and final pass involving further improvement of the previous MLLR-adapted models using the MAP technique (Dieguez-Tirado et al., 2005). When building the acoustic models we distinguished between EPPS speech and PARL speech. We used the above-described adaptation framework to create two universal speaker-independent acoustic model sets: one which was trained on speech data from the European Parliamentary Plenary Sessions another which was trained on speech data from the Spanish Parliament Sessions. In addition to these speaker-independent acoustic model sets, we also built gender-dependent (male and female) acoustic models. The gender-dependent models were built using MLLR and MAP adaptation of the speaker-independent models. We also trained a set of specific EPPS acoustic models corresponding to Josep Borrell, President of the European Parliament. The amount of speaker-specific training data for adapting the speaker independent models to this particular speaker was 4 hours, approximately.

4. Language modelling

We created two trigram language models using the training text data supplied by RWTH for the TC-STAR project:

1. The first model was trained on the Spanish EPPS final transcriptions (about 31.7 million words) extracted from parallel texts (April 1996 to June 1999, July 1999 to September 2004, December 2004 to May 2005). The perplexity value and out-of-vocabulary rate measured on the manual transcriptions of the EPPS development set were 71 and 0.64% respectively.
2. The second model was estimated from the Spanish Parliament texts provided by UPC (about 43.0 million words corresponding to the version dated 23 November 2004).

The characteristics of the language models are summarized in Table 6. These values were obtained using the SRI language model toolkit and the Dev06es data set. End-of-sentence tags were excluded when computing perplexity.

	EPPS		PARL	
	25k	40k	25k	40k
#words	25k	40k	25k	40k
%OOV	1.34	0.64	1.33	0.68
perplex.	67.94	70.96	74.87	78.14

Table 6: Summary of language models characteristics.

The SRI language modelling toolkit (Stolcke, 2002) (with standard options) was used for language model estimation, and entropy-based pruning (Stolcke, 1998) was performed with a $2, 5 \cdot 10^{-8}$ threshold to limit the LM size yet increase perplexity only marginally.

We investigated two different-sized vocabularies. The first had a recognition lexicon of 40k words and the second a lexicon of 25k words. Specifically, the Spanish Parliament lexicon contained 40284 words, which were expanded to 44114 phonetic transcriptions, and the European Parliament lexicon 40611 words and 44761 phonetic transcriptions. In the case of the smaller vocabulary, the Spanish Parliament lexicon contained 25336 words and 27746 phonetic transcriptions, and the European Parliament lexicon 25084 words and 27618 phonetic transcriptions.

In addition to the above-mentioned lexicon we also used a silence model, a short-pause model, and a set of specific models representing some frequent spontaneous events. These specific-word models are three filler word models representing hesitations like ehh, ahh, and mmm. And also four models for speaker noise such as breath, throat, applause and rustle.

The phonetic transcriptions were based on a set of 25 phones, and were automatically generated using the phonetic transcriber that forms part of the UVIGO Text-to-Speech system (Campillo and Banga, 2002).

It is worth noting here that the abbreviations, acronyms and foreign names contained in the source texts posed a problem for the automatic phonetic-transcription tool. This tool works only with Spanish and Galician, and it uses a set of dictionaries for acronyms and abbreviations. These dictionaries, however, did not contain all the acronyms and abbreviations required to transcribe the texts. Furthermore, the manner in which the words were written varied with the person who had written the text. In order to address this problem we decided to transcribe this set of words manually using the more appropriate ‘Spanish-based’ phonetic transcription.

5. Results on the Dev06es test set

This section presents the results of our testing of the UVIGO system using the development test set (Dev06es). Table 7 shows the word error rates (WER) for the recognition system using both decoding approaches. The results are classified by European Parliamentary Plenary Session (EPPS) data and Spanish Parliamentary Session (PARL) data. As can be seen, performance was significantly worse in the case of the PARL data due to its greater complexity. Specifically, PARL speech is more spontaneous than EPPS speech; it contains a large number of spontaneous effects such as disfluencies and hesitations, and speakers also interrupt each other frequently. The difficulty of obtaining good performance from speech recognition systems using this kind of speech is well known as spontaneous speech is more poorly articulated, grammatically ill-formed, and garbled by noise.

As described in Section 3.3. we trained a number of different sets of acoustic models (AM), including speaker-independent (SI) models, gender-dependent (male and female) models, and speaker-dependent (president) models, and used these during the decoding process (Sec. 3.2.). In order to evaluate the proposed decoding approach, we also applied a recognizer that uses the SI acoustic models only. For the worst-case (40k vocabulary and the decoding approach presented in Sec. 3.2.) the system works at about

6.75 times real-time on a 3GHz Pentium 4.

5.1. Discussion

Table 7 shows how consistent gains were obtained for both the EPPS and PARL data when male, female and speaker-independent AMs were all used in the decoding process. When compared to the use of an SI model only, the combined AMs produced a relative WER reduction of 11.26% and 5.58% for EPPS and PARL data respectively.

It is also interesting to compare the system’s performance in terms of male and female speakers. The SI acoustic models performed better for the male speakers than for the female speakers, which is due to the fact that there was more training data available for the male speakers (Tables 1, 3). When we added the gender-dependent AM to the decoding process, we saw a considerable improvement in terms of recognition performance for the female speakers. The WER for PARL data decreased from 33.32% to 26.00%, which represents a relative error reduction of 21.96%.

We also investigated the influence of vocabulary size on the LM, and studied two vocabulary sizes: 25k words and 40k words. Using the same decoding approach, the 40k vocabulary outperformed the 25k vocabulary. The reduction in WER is attributable to the reduction in the out-of-vocabulary (OOV) rate. We observed a 4.6% increase in relative WER reduction for EPPS data for the 40k vocabulary with respect to the 25k vocabulary. The corresponding OOV relative reduction was 52.2%.

When we analysed the alignment between the true transcriptions of the data and the recognition hypothesis obtained, we found that the large number of hesitations and broken words were a problem for our system. A number of hesitations had remained undetected and the decoder does not detect broken words. These events, which are very common in spontaneous speech, normally produce basic insertion and substitution errors.

	LM: 25k			
	AM: SI		AM: All	
	EPPS	PARL	EPPS	PARL
male	20.87	30.76	18.94	30.71
female	21.58	33.32	17.79	26.00
president	20.88	–	16.28	–
total	21.04	31.36	18.67	29.61
	LM: 40k			
	AM: SI		AM: All	
	EPPS	PARL	EPPS	PARL
male	20.24	30.01	18.06	27.18
female	21.11	32.73	16.98	25.19
president	19.99	–	15.15	–
total	20.44	30.65	17.81	26.71

Table 7: %WER on the development test set Dev06es. AM:SI stands for speaker-independent acoustic models and AM:All stands for gender-dependent plus speaker-dependent acoustic models. The row ‘president’ refers to the President of the European Parliament.

6. Results on the Eval06es test set

The results (%WER) on the TC-STAR 2006 evaluation set Eval06es are shown in Table 8. The results are also classified by European Parliamentary Plenary Session (EPPS) data and Spanish Parliamentary Session (PARL) data. The evaluation results were similar to those obtained for the Dev06es data in that recognition performance was worse in the case of PARL data than in the case of EPPS data. This, once again, may indicate the more ‘complex’ nature of the Spanish parliamentary-style speech.

	LM: 25k			
	AM: SI		AM: All	
	EPPS	PARL	EPPS	PARL
male	25.1	35.8	20.7	34.2
female	28.6	48.6	24.0	41.3
president	21.6	–	20.1	–
total	22.5	39.3	21.1	35.8
	LM: 40k			
	AM: SI		AM: All	
	EPPS	PARL	EPPS	PARL
male	20.5	35.0	19.8	33.4
female	27.3	48.3	23.1	40.8
president	20.7	–	18.9	–
total	21.4	38.6	20.2	35.0

Table 8: %WER on the evaluation test set Eval06es. AM:SI stands for speaker independent acoustic models and AM:All stands for gender-dependent plus speaker-dependent acoustic models. The row ‘president’ refers to the President of the European Parliament.

6.1. Discussion

The UVIGO system performed only slightly worse on the Eval06es data set than on the Dev06es data set.

Table 8 shows how the WER decreased when we used speaker-dependent, and male/female speaker-dependent acoustic models in the decoding process. The error rate for EPPS data and a vocabulary of 40k words decreased from 21.4% when the speaker-independent acoustic model alone was used to 20.2% when all the available acoustic models were used. This reduction represents an overall relative WER reduction of 5.6%.

Recognition performance was also better for male speakers than for female speakers with this data set, and once again, a considerable improvement was seen when a combination of all available acoustic models were used. The EPPS data WER decreased from 27.3% to 23.1%, which represents an overall relative reduction of 15.38%.

7. Discussion

In this paper, we have presented the tests we carried out within the framework of the TC-STAR 2006 ASR evaluation project, in which we applied our automatic speech recognition system (UVIGO) to the Spanish Parliament Sessions task under the restricted condition. The above sections summarize the results of system performance using both the Dev06es and Eval06es data sets.

The UVIGO system evaluated used a standard front-end for feature extraction but did not use speaker normalization/adaptation techniques of any kind to obtain more robust features related to speaker and environment variability. Furthermore, the decoding process involved one pass only, i.e. it did not perform supervised adaptation on the original models to implement a new decoding pass.

We carried out several tests that compared the effect of two different vocabulary sizes on LM generation, and assessed the performance of the decoding strategy proposed in Section 3.2.. This decoding strategy consistently yielded a lower error rate, but was more costly in terms of computation time than the strategy involving only the set of speaker-independent acoustic models.

We also compared the error rates obtained for the European Parliamentary Plenary Session (EPPS) data and the Spanish Parliament Session (PARL) data, and found major differences in the style of speech used in both domains. We concluded that PARL speech is more spontaneous than EPPS speech, and is therefore more difficult for the recognition system to process.

8. Future Work

This section discusses the main improvements that we are planning to include in our system in the near future.

Although we believe that the UVIGO recognition system has a solid foundation, we are aware of the amount of work that is required in order to deliver the performance expected of current state-of-the-art recognizers. The majority of the improvements required are related to language and acoustic modelling. We also need to address several problems related to the origin and development of our system, which is closely related to Transcrigal, a Broadcast News transcription project that works in the Galician language environment.

In order to further improve recognition performance, we are therefore planning to focus our work on two particular areas: feature extraction and decoding.

As far as feature extraction is concerned, it is widely recognized that one of the major flaws of current ASR technology is an excessive sensitivity to speech signal characteristics that are unrelated to content, such as changes in the acoustic environment (background noises, microphone and channel distortions, etc). We are planning to address this problem with normalization techniques designed to reduce this variability by modifying the representation utterance. This may involve, for example, including a combination of different speaker and channel normalization techniques in order to reduce inter-speaker and channel acoustic variability (Giuliani et al., 2006), and acoustic parameters based on auditory models that are less sensitive to irrelevant aspects of the signal.

As far as decoding is concerned, the simple approach of using a single pass rather than a standard approach involving several recognition passes with increasingly specialized models has many obvious limitations. Of the many possible improvements in this area, we plan to include on-line speaker acoustic model adaptation techniques featuring a number of well known adaptation algorithms. The study of the use of on-line language model adaptation (by means

of topic detection and information retrieval mechanisms) within the framework of the Transcrigal project has already produced promising results, but we believe that its usefulness may be much more limited in a domain like TC-STAR, in which vocabulary, language and topic variability is much smaller.

One of the problems associated with a minority language like Galician is its relative lack of resources. Several design choices that were adopted with this problem in mind might be suboptimal in the environment defined by the TC-STAR evaluation project. Demiphones are a good choice for acoustic modelling if the amount of training data is small but they could be out-performed by triphones if enough training material was available. As far as language modelling is concerned, in Transcrigal we tested the use of 4-grams in the lattice rescoring pass of the recognizer but saw little or no improvement. This is probably due to the poor training of the language model as a result of the few text resources available, although this needs to be clarified with further research.

Finally, we believe that considerable improvements could be achieved by using a good pronunciation model. We found that by using a careful selection of alternative transcriptions, we improved recognition results in several areas for a number of speakers with strong regional accents. Unfortunately, we also found that the recognition rate for other speakers with different accents decreased, which resulted in an overall reduction in accuracy. The selection of an adequate pronunciation model for each speaker using confidence measures, LM and acoustic scores, or a probabilistic pronunciation model, is another area that we believe deserves further investigation.

9. Acknowledgements

This project has been partially supported by the Xunta de Galicia under project number PG10IT05TIC32202PR.

10. References

- F. Campillo and E. Rodriguez. Banga. 2002. Combined prosody and candidate unit selections for corpus-based text-to-speech systems. In *Proc. Int. Conf. Spoken Language Processing*, pages 141–144.
- A. Cardenal-Lopez, F. J. Dieguez-Tirado, and C. Garcia-Mateo. 2002. Fast LM look-ahead for large vocabulary continuous speech recognition using perfect hashing. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, volume 1, pages 705–708, Orlando, FL, May.
- J. Dieguez-Tirado, C. Garcia-Mateo, A. Cardenal-Lopez, and L. Docio-Fernandez. 2005. Adaptation strategies for the acoustic and language models in bilingual speech transcription. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, volume 1, pages 833–836, Philadelphia, PA, March.
- L. Docio-Fernandez and C. Garcia-Mateo. 2002. Acoustic modeling and training of a bilingual ASR system when a minority language is involved. In *Proc. Int. Conf. on Language Resources and Evaluation*, volume 3, pages 873–876, Gran Canaria, Spain, May.
- M.J.R. Gales, Y. Dong, D. Povey, and P.C. Woodland. 2003. Porting: Switchboard to the voicemail task. In *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, volume 1, pages 536–540, Hong Kong, April.
- C. Garcia-Mateo, J. Dieguez-Tirado, A. Cardenal-Lopez, and L. Docio-Fernandez. 2004. Transcrigal: A bilingual system for automatic indexing of broadcast news. In *Proc. Int. Conf. on Language Resources and Evaluation*, volume 6, pages 2061–2064, Lisbon, Portugal, May.
- D. Giuliani, M. Gerosa, and F. Brugnara. 2006. Improved automatic speech recognition through speaker normalization. *Computer Speech and Language*, 20(1):107–123, January.
- J. Mario, A. Nogueiras, and A. Bonafonte. 1997. The demiphone: an efficient subword unit for continuous speech recognition. In *Proc. Eurospeech*, pages 1215–1218, Rhodes, Greece, September.
- A. Stolcke. 1998. Entropy-based pruning of backoff language models. In *Proc. DARPA Broadcast News Transcription and Understanding Workshop*, pages 270–274, Lansdowne, VA.
- A. Stolcke. 2002. SRILM – an extensible language modeling toolkit. In *Proc. Int. Conf. Spoken Language Processing*, volume 2, pages 901–904, Denver, CO, September.