The ITC-irst transcription systems for the TC-STAR-06 evaluation campaign

F. Brugnara, D. Falavigna, D. Giuliani, R. Gretter, D. Pineda[1], D. Seppi and G. Stemmer[2]

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica, 38050 Panté di Povo, Trento, Italy
(brugnara,falavi,giuliani,gretter,seppi)@itc.it

Abstract

This paper describes the ITC-irst systems used in the TC-STAR'06 evaluation campaign for transcribing parliamentary speeches delivered in both English and Spanish languages. Systems use a three pass decoding strategy with cluster-based unsupervised acoustic models adaptation. Both first and second decoding passes use a trigram language model, while the third decoding pass employs a fourgram language model. Acoustic and language models of both English and Spanish transcription systems were trained exploiting the language resources released for the TC-STAR evaluation campaign of year 2006. An additional language resource, i.e. a 200M word text corpus distributed by the Linguistic Data Consortium (LDC), was also utilized to train language models for English. The Word Error Rates (WERs) of the primary English transcription system were 13.0% and 11.0% on the EPPS English development and evaluation data sets, respectively. On both the Spanish development and evaluation data sets, which include both EPPS and Spanish Parliament speech data, the transcription system provided a WER of 13.3%.

1. Introduction

In this paper, the main features of the ITC-irst transcription systems used in the TC-STAR'06 evaluation campaign are described.

ITC-irst submitted results for both '06 English and Spanish evaluation data sets. While the '06 English evaluation data set includes only EPPS (European Parliament Plenary Sessions) data, the '06 Spanish evaluation set includes data stemming from both EPPS and Spanish Parliament. Furthermore, a direct comparison of the capabilities of the transcription systems adopted in TC-STAR 2005 and 2006 evaluation campaigns will be given.

With respect to the transcription systems developed for the '05 evaluation campaign, the systems used in the '06 campaign feature: an audio partitioner for identifying and clustering speech segments, Heteroscedastic Linear Discriminant Analysis (HLDA) for augmenting acoustic features discrimination capabilities, and a three pass decoding strategy. The third decoding pass is carried out by exploiting a fourgram language model (LM), on a restricted search space given by word lattices generated in the second pass.

The paper is organized as follows. Main features common to all the ITC-irst transcription systems are presented in Section 2. The transcription system developed for the EPPS English task is described in Section 3, while the system developed for the EPPS Spanish task is described in Section 4. Experiments and results are reported in Section 5. Some conclusions are presented in Section 6.

2. Transcription system overview

The ITC-irst transcription system consists of two main components: the audio partitioner and the speech recognizer.

The aim of the audio partitioner is to divide the continuous audio stream into homogeneous non-overlapping segments and to cluster these segments into homogeneous groups.

The partitioner consists of three main modules (Brugnara et al., 2002): the segmenter, the classifier and the clustering module.

Audio Segmentation. Usually segmenting an audio stream means detecting the time indexes corresponding to changes in the nature of audio, in order to isolate segments that are homogeneous in terms of bandwidth and speaker. However, in the current version of the ITC-irst partitioner, the segmenter just identifies the region of the audio stream with high energy through the application of a start-end point activity detector. The identification of acoustically homogeneous segments within these regions is embedded into the classification process.

Segment Classification. An intermediate goal of the partitioning stage is to identify each acoustically homogeneous segment and to classify it in terms of broad acoustic classes. For this purpose, acoustic classes are modeled by a set of Gaussian Mixture Models (GMMs) and the classification is done applying the Viterbi algorithm to a search space in which the activation of a new class is possible at any time: this is accomplished through a network with loop topology. This process induces a refinement of the raw segmentation made by the segmenter, since the time indexes of class changes correspond to new segment boundaries.

Segment Clustering. Identified acoustically homogeneous segments are finally clustered employing a method (Cettolo, 2002) based on the Bayesian Information Criterion (BIC).

The partitioner is applied to each audio file and the speech recognizer, which uses continuous density hidden Markov Models (HMMs), generates a word transcription for each speech segment. In addition, initial and final temporal instants of each word are supplied.

Word transcription is generated in three passes.

First pass: Preliminary Decoding Step. This step generates an initial word transcription which is used as a supervision for performing cluster-based normalization of acoustic features and acoustic model (AM) adaptation. Continuous

(1) Visiting student - Universitat Politecnica de Catalunya, Barcelona, Spain.

(2) G. Stemmer is now with Siemens AG, Corporate Technology, Munich, Germany.

density triphone HMMs and a trigram language model are used in this step.

Second pass: Word Lattice Generation. The second decoding step, exploiting normalized acoustic features and adapted acoustic models, generates the best word hypothesis, as well as word lattices.

Third pass: Final Decoding Step. In the third decoding step a word graph is generated for each word lattice produced in the second step. First, a bigram constrained word graph is computed for each given word lattice and a pruning procedure, based on estimates of posterior probabilities, is applied to it (this procedure allows to reduce the average word graph size of about 90% without significant increase in graph error rate). Then, an expansion algorithm, similar to the one reported in (Weng et al., 1998), allows to introduce fourgram LM probabilities in the resulting final word graph. This latter one defines a reduced search space over which word hypotheses are rescored using the acoustic models of the second pass adapted with the supervision of the second pass itself.

2.1. Acoustic models

Acoustic models used in all decoding steps are state-tied, cross-word, gender-independent triphone HMMs. Output probability densities are defined by mixtures of Gaussian functions having diagonal covariance matrices. A phonetic decision tree was used for tying states and for defining the context-dependent allophones.

The acoustic front-end employed in the first decoding step is different from the one utilized in the second and third passes. The details of acoustic modeling adopted in the various decoding passes are given below.

First decoding step. The observation vectors for HMMs consist of 13 Mel Frequency Cepstral Coefficients (MFCCs) extracted using a 20ms Hamming window and a frame step of 10ms. Cluster-based Cepstral Mean and Variance Normalization (CMVN) is performed to ensure that each segment cluster contains acoustic observations exhibiting zero mean and unit variance. Successively, first and second order time derivatives are computed to form a 39-dimensional feature vector.

To train HMMs a variant of the Constrained MLLR based Speaker Normalization (CMLSN) procedure (Stemmer et al., 2005) is adopted, as briefly summarized below.

- A simple target model, that is a GMM with 1024 components, is trained.
- For each speech segment cluster in the training data, a constrained MLLR (CMLLR) transform (Gales, 1998) is estimated w.r.t. the target GMM.
- The CMLLR transforms are applied to the feature vectors. The resulting transformed/normalized feature vectors are supposed to contain less speaker, channel, and environment variability.
- A conventional Maximum Likelihood (ML) training procedure is used to initialize and train the recognition models on the normalized data, including state tying and the definition of the context-dependent allophones.

Second and third decoding steps. Segment-based Cepstral Mean Normalization (CMN) is applied to the 13 MFCCs, obtaining zero mean acoustic observations for each segment (no variance normalization is performed in this case). Then, first, second and third order time derivatives are computed to form a 52-dimensional feature vector. HLDA projection is then performed to obtain observation vectors with 39 components.

The HMM adaptive training procedure is described in (Giuliani et al., 2004; Giuliani et al., 2006; Stemmer and Brugnara, 2006) and is summarized below.

- The HLDA transformation is estimated w.r.t. a set of reference models. Reference models are triphone HMMs with a single Gaussian density for each state; they are estimated on the 52-dimensional acoustic observation space.
- The HLDA transform is applied to training data to obtain 39-dimensional acoustic observation vectors.
- A set of target models is generated for the acoustic space obtained through HLDA projection. The target models are tied-states triphone HMMs with a single Gaussian density for each state.
- For each cluster of speech segments in the training data, a CMLLR transform is estimated w.r.t. the target models.
- The CMLLR transforms are applied to the feature vectors.
- ML training is performed on the normalized features.

By exploiting the output ot the first decoding step, data are normalized through cluster based CMLSN normalization, and acoustic model adaptation is carried out. Just Gaussian means are adapted through the application of a number of simple 'shift' transformations estimated in the MLLR framework. A regression class tree is employed, in conjunction with a low occupancy count threshold (i.e. 100), for dynamic allocation of regression classes. For each regression class only a bias vector is estimated.

3. English transcription system

This section describes specific features of the system used to cope with EPPS English task in the '06 TC-STAR evaluation campaign.

3.1. Acoustic models

For training acoustic models the EPPS English training corpus, released for the '06 TC-STAR evaluation campaign, was exploited. This data set consists of about 176 hours of audio recordings partitioned into: about 101h of transcribed audio data and about 75h of untranscribed audio data. Untranscribed speech data were automatically transcribed using an early version of the transcription system.

The acoustic models used in the first decoding pass have about 7600 tied states and about 243000 output Gaussian densities.

Acoustic models used in the second and third decoding steps have about 7700 tied states and about 244000 Gaussian densities.

3.2. Language models

The trigram Language Model, used in the first and second decoding steps, was trained on:

- English EPPS final text edition corpus, about 36M words (from parallel texts).
- An out-of-domain text corpus of about 200M words, released by LDC, containing broadcast news transcriptions.

The resulting LM was then adapted to the manual transcriptions of the EPPS audio data released for acoustic model training of '06 TC-STAR evaluation. These texts consisted of about 0.7M words. The LM adaptation algorithm is the modified shift-beta one described in (Bertoldi and Federico, 2004). In a similar way, a fourgram LM, to be used in the third decoding step, was trained.
The trigram LM and the lexicon were used to build the static decoding graph with about 25M states and 23M labeled arcs. The network has a tree based topology, and exploits the tail sharing technique to reduce redundancy (Brugnara and Cettolo, 1995);
The use of additional training text data (i.e. those stemming from the LDC corpus mentioned above), corresponds to the "public" training condition defined in the TC-STAR evaluation. For comparison purposes, trigram and fourgram LMs were also trained exploiting only the English EPPS text data. This corresponds to the "restricted" training condition defined in the TC-STAR evaluation.
In all cases, "true case" word transcription was ensured by adopting a "true case" recognition vocabulary. This vocabulary was shared by all LMs.

3.3. Pronunciation lexicon

The pronunciations in the lexicon are based on a set of 45 phones. The lexicon contains 49k words, and was generated by merging different source lexica for American English (LIMSI '93, Cmudict, Pronlex). Furthermore, phonetic transcriptions for few hundreds of words were manually generated. Note that not all of the words in the manual transcriptions of '06 EPPS English acoustic training set are present in the lexicon.
Finally, additional HMMs were used for modeling the following acoustic events: one HMM for "silence", five HMMs for filler words and one HMM for out of vocabulary words (used only during training/adaptation).

3.4. Punctuation module

Punctuation is added to the recognized word sequence in a final post-processing step. Punctuation marks (including an empty symbol) are assigned to each recognized word according to the score provided by an artificial neural network (ANN) previously trained. Training is performed exploiting both the speech signal of the word sequence and the related temporal word boundaries provided by the speech recognizer: prosodic features are extracted on a word basis. Almost all of these features are related to word duration, word energy, and pause following the word itself (if present). For the future we plan to add also the pitch. Other non-acoustic features include Part of Speech (PoS) tags, which are added using the SVMTool[1] developed at the Universitat Politecnica de Catalunya (UPC). Finally, the set of features related to each word is augmented with the features of the ± 2 adjacent words. We have used 90 prosodic features for English and 152 for Spanish (102 prosodic plus 50 PoS).
Punctuation symbols are grouped into 4 classes: *full stop* (which includes "." and "!"), *comma* ("," ":" and ";"), *question mark* ("?") and *no mark*. The training of the ANNs is performed on 40 hours of transcribed speech data, both for English and Spanish. During training, a single neural network learns to associate a punctuation mark with each word. During classification, the neural network adds a punctuation mark, with an associated probability, to each recognized word.

4. Spanish transcription system

For Spanish, EPPS TC-STAR'06 acoustic training corpus consists of about 173 hours: about 100h of transcribed audio data and about 73h of untranscribed audio data.
Similarly to English, an automatic system was used to transcribe untranscribed data.
HMMs used in the first decoding pass have about 4100 tied states and about 65000 Gaussian densities, while HMMs used in both second and third decoding steps have about 5300 tied states and about 168000 Gaussian densities.
For Spanish, only the restricted training condition has been exploited. A trigram language model was trained on: the Spanish EPPS final text edition, the Spanish Parliament Texts and EPPS parallel corpora, about 79M words in total.
Also in this case, the resulting LM was adapted using about 880K words coming from manually transcribed EPPS audio data.
The LM and the lexicon allowed to generate the static Finite State Network (FSN) with about 8.6M states and 8.1M labeled arcs, used in both the first and second decoding steps.
Similarly to English, a "true case" recognition vocabulary and LM was used to ensure "true case" word transcriptions.

4.1. Pronunciation Lexicon

The phone units used in the pronunciations lexicon are 31; in addition, there are models representing: hesitations, background noises and breaths. The lexicon contains 57K words, whose phonetic transcriptions were automatically generated using a set of grapheme-to-phoneme rules. This tool can handle acronyms and multiple pronunciations; some rules were added to handle some common foreign patterns. Finally, some hundreds of foreign names were manually corrected.

4.2. Punctuation module

For Spanish punctuation is added to word transcriptions in a manner similar to English. The punctuation module, based on a neural network, exploits 152 features (102 prosodic + 50 PoS). The neural network is trained on about 40 hours of transcribed speech data.

[1] http://www.lsi.upc.es/~nlp/SVMTool/

	dev05	*eval05*	*dev06*	*eval06*
Res.3-gram'05 OOV	0.6%	0.7%	0.9%	0.8%
Res.3-gram'06 OOV	0.4%	0.4%	0.4%	0.4%
Pub.3-gram'06 OOV	0.4%	0.4%	0.4%	0.4%
Pub.4-gram'06 OOV	0.4%	0.4%	0.4%	0.4%
Res.3-gram'05 PP	90	106	122	141
Res.3-gram'06 PP	87	103	119	136
Pub.3-gram'06 PP	92	106	129	138
Pub.4-gram'06 PP	85	97	118	127

Table 1: *OOV rates and perplexities (PP) on the EPPS English development and evaluation data sets (dev05, eval05, dev06 and eval06) with several language models.*

5. Experimental results

5.1. EPPS English

Table 1 gives both OOV rates and perplexities computed with three trigram and one fourgram LMs on the EPPS English test sets of both '05 and '06 TC-STAR evaluation campaigns. The Restricted'05 trigram LM was used by the primary system developed for the official '05 TC-STAR evaluation. Major differences between the '05 and '06 restricted LMs are listed below:

- the different number of words in the recognition vocabulary;
- the use of a "true case" recognition vocabulary in the Restricted'06 LM and of a "case insensitive" vocabulary in the Restricted'05 LM;
- the different amounts of manual transcriptions available for LM adaptation, about 0,37M words for the Restricted'05 LM and about 0,7M words for the Restricted'06 LM.

The Public'06 LMs (both trigram and fourgram) have been used in the official '06 TC-STAR evaluation. They were trained, as described in Section 3.2., using additional publically available, out-of-domain text data.
In computing perplexities and OOV rates transcriptions of truncated words were omitted. It can be noted that the Restricted'06 trigram LM gives lower perplexities than both the Restricted'05 trigram LM, and the Public'06 trigram LM.

	dev05	*eval05*	*dev06*	*eval06*
1st decoding pass	15.8	16.3	23.6	20.3
2nd decoding pass	13.6	13.4	17.6	14.9

Table 2: *Recognition results (% WER) on the EPPS English development and evaluation data sets (dev05, eval05, dev06 and eval06) achieved with the primary English system developed for the '05 TC-STAR evaluation campaign. Intermediate recognition results, after the first decoding step, are also reported.*

Tables 2 and 3 reports recognition results achieved, on the EPPS English development and evaluation data sets, by the English primary systems developed for the '05 and '06 TC-STAR Evaluation campaigns. Intermediate recognition results are also given in the Tables.

	eval05	*dev06*	*eval06*
1st decoding pass	12.5	16.7	14.9
2nd decoding pass	10.3	13.6	11.7
3rd decoding pass	9.7	13.0	11.0

Table 3: *Recognition results (% WER) on the EPPS English development and evaluation data sets (eval05, dev06 and eval06) achieved with the primary EPPS English system developed for the '06 TC-STAR evaluation campaign.*

In addition to the different amount of EPPS training data available in the two evaluation campaigns and exploited for system training, the '05 and '06 primary systems differ in the following aspects:

- Audio segmentation. For the development and evaluation data sets released for the '05 TC-STAR evaluation boundaries of speech segments to be transcribed were provided. All recognition experiments reported in this paper exploited this information. Instead, for the development and evaluation data sets released for the '06 TC-STAR evaluation speech segments boundaries were not provided and needed to be detected automatically.
- Acoustic feature extraction. As seen above, in the '06 system, an HLDA acoustic feature projection is embedded into the acoustic front-end for acoustic models used in the second and third decoding steps. Effectiveness of HLDA was proved by training and testing transcription systems under the restricted conditions defined for the '05 TC-STAR evaluation campaign. Comparative results, using or not using HLDA, are given in Table 4 for both English and Spanish *dev05* an *eval05* evaluation sets.

	English		Spanish	
	dev05	*eval05*	*dev05*	*eval05*
Reference'05	13.6	13.4	11.8	12.4
HLDA'05	13.2	12.6	10.8	12.0

Table 4: *Recognition results (% WER) on the EPPS English and Spanish '05 development and evaluation data sets (dev05, eval05) achieved with the reference English and Spanish systems, either exploiting or not HLDA.*

- Decoding strategy. The '06 system adopts a three pass decoding strategy, with a fourgram LM rescoring in the third step, while a two pass decoding strategy exploiting a trigram LM was adopted by the '05 system. A relative WER reduction of about 5%, as can be derived from Table 3, has been reached on the English evaluation set with the addition of the third decoding step.
- Language model. The '06 system features a "true case" recognition vocabulary and the LM was trained exploiting additional out-of-domain language resources (public training condition). Instead, the '05 primary system was trained only on the EPPS text data (restricted training conditions) and had a case insensitive vocabulary.

	Public'06	Restricted'06
1st decoding pass	14.9	15.7
2nd decoding pass	11.7	12.0
3rd decoding pass	11.0	11.7

Table 5: *Recognition results (% WER) on the '06 EPPS English evaluation data set achieved with the primary '06 system (Public'06) and with a contrastive '06 system exploiting LMs estimated only on EPPS text data (Restricted'06).*

Case ins.	*Case sens.*	*Case ins. + punct.*	*Case sens. + punct.*
11.0	11.9	18.7	19.5

Table 6: *Recognition results (% WER) on the '06 EPPS English evaluation data set achieved with the '06 primary system. Recognition results are computed in several manners: case insensitive, case sensitive, case insensitive with punctuation and case sensitive with punctuation.*

- Punctuation. The automatic punctuation module described in Section 3.4 has been utilized. Note that no punctuation was provided by the '05 system.

To measure the benefit of adding out-of-domain text data to train the LMs, Table 5 reports recognition results achieved in both public and restricted conditions. As previously seen, the public training condition exploits the 200M words broadcast news LDC corpus, in addition to the EPPS final text edition corpus, used in the restricted training condition.

Table 6 reports recognition results computed in four different manners: case insensitive, case sensitive, case insensitive with punctuation and case sensitive with punctuation.

It can be noted that in case sensitive scoring recognition results are worse than those obtained with case insensitive scoring. However, the decrease in performance is not dramatic. Taking into account punctuation during case insensitive scoring affect substantially recognition results for two main reasons. Firstly, the punctuation module used is still in an initial version and further work is needed. Secondly, punctuation in reference transcriptions is ambiguous and many punctuation marks should be considered as optional deletable by the scoring tool. Automatic punctuation is a critical issue that will merit further attention in the TC-STAR future activity.

Table 7 reports the execution time for the different processing steps performed by the EPPS English primary system in transcribing the EPPS English evaluation data set. The source signal duration for this data set was 11566.1s. The total execution time was 266736s, as measured on an Intel(R) Xeon(TM) 3.00GHz processor with 1024 KB cache and 4GB memory, roughly corresponding to 23.0 times the source signal duration.

Execution time for the first decoding pass includes audio stream partitioning and acoustic data normalization, while the execution time of the second decoding pass includes the execution time for acoustic data normalization and acoustic model adaptation.

1st decoding pass	128700 (11.1xRT)
2nd decoding pass	96883 (8.40xRT)
word graph generation	9240 (0.80xRT)
3rd decoding pass	29375 (2.50xRT)
punctuation	2538 (0.21xRT)
Total	266736 (23.0xRT)

Table 7: *Execution times (in seconds) of the different processing steps for the English transcription system on the '06 EPPS English evaluation data set.*

	dev05	*eval05*	*dev06*	*eval06*
Res.3-gram'06 OOV	0.6%	0.7%	0.6%	0.6%
Res.4-gram'06 OOV	0.6%	0.7%	0.6%	0.6%
Res.3-gram'06 PP	84	97	107	102
Res.4-gram'06 PP	77	89	97	93

Table 8: *OOV rates and perplexities (PP) of the Spanish development and evaluation data sets (dev05, eval05, dev06 and eval06) with the restricted '06 LMs.*

5.2. EPPS Spanish

Table 8 gives OOV rates and perplexities computed with the restricted condition LMs (Res.[3-4]gram'06) on the Spanish dev and test sets of both '05 and '06 TC-STAR evaluation campaigns. Similarly to English, note the increased perplexities between dev-eval'05 and dev-eval'06 tasks.

Tables 9 and 10 report recognition results achieved, on the Spanish development and evaluation data sets, with the Spanish primary systems developed for the '05 and '06 TC-STAR Evaluation campaigns. As previously mentioned, the Spanish '06 dev and eval sets include speech data stemming from both EPPS and Spanish Parliament. Intermediate recognition results are also reported in the tables. Note that differences between the '05 and '06 primary Spanish systems are comparable to the corresponding ones given for the English case.

Table 11 reports recognition results computed in four different manners: case insensitive, case sensitive, case insensitive with punctuation and case sensitive with punctuation. Comments similar to the English case can be given for this table.

Table 12 reports the execution time for the different processing steps performed by the Spanish primary system in transcribing the Spanish evaluation data set. The source signal duration for this data set was 26003s. The total execution time was 222983s, as measured on an Intel(R) Xeon(TM) 3.00GHz processor with 1024 KB cache and 4GB memory, roughly corresponding to 8.58 times the source signal duration.

	dev05	*eval05*	*dev06*	*eval06*
1st decoding pass	14.0	14.7	23.7	25.8
2nd decoding pass	12.7	13.7	18.9	18.7

Table 9: *Recognition results (% WER) on the Spanish development and evaluation data sets achieved with the primary Spanish system developed for the '05 TC-STAR evaluation campaign. Intermediate recognition results, after the first decoding step, are also reported.*

	dev05	*eval05*	*dev06*	*eval06*
1st decoding pass	13.3	14.4	21.9	24.4
2nd decoding pass	10.0	11.1	13.7	13.9
3rd decoding pass	9.8	10.7	13.3	13.3

Table 10: *Recognition results (% WER) on the Spanish development and evaluation data sets achieved with the primary Spanish system developed for the '06 TC-STAR evaluation campaign.*

Case ins.	*Case sens.*	*Case ins. + punct.*	*Case sens. +punct.*
13.5	14.7	22.6	23.6

Table 11: *Recognition results (% WER) on the Spanish '06 evaluation data set achieved with the '06 primary system. Recognition results are computed in several manners: case insensitive, case sensitive, case insensitive with punctuation and case sensitive with punctuation.*

6. Conclusions

In this paper we have described the transcription systems used in the TC-STAR '06 evaluation campaign and we have given the results obtained on the related evaluation data sets. Differences with the systems developed for the '05 evaluation campaign have also been presented and their impact on recognition performance discussed.

Results show that systems developed for the '06 evaluation campaign outperform systems developed for the '05 evaluation campaign, both for English and Spanish languages.

Table 13 reports the relative WER reduction, between the '06 and 05' primary transcription systems, obtained on both the English and Spanish '05 and '06 evaluation sets.

Gain in performance is partially due to the availability of additional training data and partially to system improvements. In particular, the use of HLDA-derived acoustic features and the adoption of a three decoding pass strategy ensure a tangible performance improvement over systems developed for the '05 evaluation campaign.

Furthermore, using additional out of domain text data for LM training showed to be effective for the EPPS English system. Future work will be devoted to better exploit out of domain data to train fourgram LMs to be used in the third decoding step. Future activities will also encompass effort to improve unsupervised cluster-based acoustic model adaptation by exploiting, for example, the word lattice instead of only the best word hypothesis. Another future activity will address the usage of a fourgram LM already in the first and second decoding steps, avoiding the generation of word graphs for LM rescoring.

Not much attention has been devoted so far to the computational costs of the different processing steps of the transcription process. Future work will be also devoted to speed up the transcriptions process having in mind real application requirements.

7. Acknowledge

This work was in part funded by the European Union under the integrated project TC-STAR Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738).

1st decoding pass	95398 (3.67xRT)
2nd decoding pass	89164 (3.43xRT)
word graph generation	20231 (0.78xRT)
3rd decoding pass	13490 (0.52xRT)
punctuation	4700 (0.18xRT)
Total	222983 (8.58xRT)

Table 12: *Execution times (in seconds) of the different processing steps for the Spanish transcription system on the Spanish '06 evaluation data set.*

	eval05	*eval06*
English	27.6	26.2
Spanish	21.9	28.9

Table 13: *% WER relative reduction, between the '06 and '05 English and Spanish transcription systems, obtained on the '05 and '06 evaluation sets.*

8. References

N. Bertoldi and M. Federico. 2004. Broadcast News LM Adaptation Over Time. *Computer Speech and Language*, 18(1):417 – 435.

F. Brugnara and M. Cettolo. 1995. Improvements in tree-based language model representation. In *Proceedings of the 4th European Conference on Speech Communication and +Technology*, pages 2075–2078, Madrid, Spain.

F. Brugnara, M. Cettolo, M. Federico, and D. Giuliani. 2002. Issues in Automatic Transcription of Historical Audio Data. In *Proc. of ICSLP*, pages 1441–1444, Denver, CO, September.

M. Cettolo. 2002. Porting an Audio Partitioner Across Domains. In *Proc. of ICASSP*, pages I–301–304, Orlando, Florida, May.

M. J. F. Gales. 1998. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, 12:75–98.

D. Giuliani, M. Gerosa, and F. Brugnara. 2004. Speaker Normalization through Constrained MLLR Based Transforms. In *Proc. of INTERSPEECH/ICSLP*, pages 2893–2897, Jeju Island, Korea, Oct.

D. Giuliani, M. Gerosa, and F. Brugnara. 2006. Improved automatic speech recognition through speaker normalization. *Computer Speech and Language*, 20:107–123.

G. Stemmer and F. Brugnara. 2006. Integration of Heteroscedastic Linear Discriminant Analysis (HLDA) into Adaptive Training. In *Proc. of ICASSP*, Toulouse, France, May.

G. Stemmer, F. Brugnara, and D. Giuliani. 2005. Adaptive Training Using Simple Target Models. In *Proc. of ICASSP*, pages I–997–1000, Philadelphia, PA, March.

F. Weng, A. Stolcke, and A. Sankar. 1998. Efficient Lattice Representation and Generation. In *Proceedings of ICSLP98*, volume 6, pages 2531 – 2534, Sydney, Australia.