THE IBM 2007 SPEECH TRANSCRIPTION SYSTEM FOR EUROPEAN PARLIAMENTARY SPEECHES

Bhuvana Ramabhadran[*], *Olivier Siohan*[*†]

[*] IBM T. J. Watson Research Center
Yorktown Heights, NY

Abhinav Sethy[‡]

[‡]Speech Analysis and Interpretation Lab
Department of Electrical Engineering-Systems
Viterbi School of Engineering
University of Southern California, CA

ABSTRACT

TC-STAR is an European Union funded speech to speech translation project to transcribe, translate and synthesize European Parliamentary Plenary Speeches (EPPS). This paper describes IBM's English speech recognition system submitted to the TC-STAR 2007 Evaluation. Language Model adaptation based on clustering and data selection using relative entropy minimization provided significant gains in the 2007 Evaluation. The additional advances over the 2006 system that we present in this paper include unsupervised training of acoustic and language models; a system architecture that is based on cross-adaptation across complementary systems and system combination through generation of an ensemble of systems using randomized decision tree state-tying. These advances reduced the error rate by 30% relative over the best-performing system in the TC-STAR 2006 Evaluation on the 2006 English development and evaluation test sets, and produced one of the best performing systems on the 2007 evaluation in English with a word error rate of 7.1%.

Index Terms: LVCSR, Language Modeling, TC-STAR, unsupervised learning.

1. INTRODUCTION

The TC-STAR (Technology and Corpora for Speech to Speech Translation) project financed by the European Commission within the Sixth framework Program is a long-term effort to advance research in speech to speech translation technologies[1]. The primary goal of the TC-STAR project is to produce an end-to-end system in English and Spanish that accepts parliamentary speeches in one language, transcribes, translates and synthesizes them into another language, while significantly reducing the gap between the performance of a human (interpreter) and a machine. To support this goal, the performance of each component technology, namely, speech recognition (ASR), machine translation (MT) and text-to-speech (TTS) is optimized to produce the best output at their respective stages [1]. The EPPS corpus comprises of over 800 politicians discussing current affairs during several public sessions of the European Parliament in multiple languages and the minutes of these sessions edited by the European parliament also known as the Final Text Editions (FTE). In the 2007 evaluation, the training, development and evaluation data comprised of recordings made between April 1996 and May 2006. Within the TC-STAR project, the evaluation is done under three different conditions:

- *public*, which allows the use of any data that is publicly available, such as Broadcast news, data mined from the web released by University of Washington, and data from the British Parliament sessions in addition to the EPPS acoustic training data and Final Text Editions;
- *restricted*, which allows the use of EPPS data only;
- *open*, which allows the use of publicly available and any in-house material in addition to the EPPS data.

This paper describes the IBM systems submitted under the *open*, *public* and *restricted* conditions. The key design characteristics for all evaluation conditions include:

- Language model adaptation using a relative entropy minimization scheme for data selection and Latent Dirichlet Allocation (LDA) [2] for clustering;
- Use of Unsupervised data in acoustic and language model building;
- System combination via ROVER of multiple ASR systems built using a randomized decision-tree growing procedure [3];
- A basic set of models that use VTLN and SAT training followed by fMPE+MPE training [4] and speaker adaptation using MLLR;
- A two-pass decoding strategy that uses an in-domain language model for the first pass using an SI system followed by a second-pass with the above models that uses an in-domain LM for the restricted condition and out-of-domain language models for the *open* and *public* conditions. This is the only step that uses non-EPPS training material;
- Dynamic decoding with quinphone context; and
- Training of acoustic models using EPPS material only.

2. ALGORITHMS

The 2007 IBM TC-STAR speech recognition system is organized around an architecture that combines multiple systems via ROVER. This section explains the new algorithms we introduced in the 2007 Evaluation.

2.1. Language Model Adaptation

In addition to the existing sources of text data for language modeling, namely the EPPS text, Broadcast news material released from LDC, Web data released from University of Washington and the

[†]Currently in Google Inc.
[‡]Currently in IBM T.J Watson Research Center
[1]Project No. FP6-506738

HANSARD (British Parliament session) corpus, we explored techniques to exploit the enormous resources available on the web that can enhance the performance under the *open* condition. The fundamental idea is to adapt the language model using data selected from any resource via a relative entropy based data selection scheme [5].

2.1.1. Data Selection

We used a relative entropy (R.E) minimization criterion[5] to select adaptation text relevant to the domain. Data selection methods which rank sentences based on perplexity, BLEU score or similar criteria and select the top sentences induce a bias towards the high density regions of the in-domain data distribution. R.E based selection addresses the bias problem by using a distributional similarity criterion and optimizing selection of subsets of sentences in place of individual sentences.

R.E based data selection[5] is an incremental and greedy algorithm which selects a sentence if adding it to the already selected set of sentences reduces the relative entropy with respect to the in-domain data distribution. To understand the core idea behind the algorithm, let us denote the probability of word i in the language model built from in-domain data by $P(i)$. To describe the algorithm we will employ the paradigm of unigram probabilities though the method generalizes to higher n-grams also.

We denote the count of word i in the current sentence that the algorithm is considering for inclusion in the set of relevant sentences by $m(i)$. Let $W(i)$ be the count for word i in the set that has already been selected. Let $n = \sum_i m(i)$ be the number of words in the sentence and $N = \sum_i W(i)$ be the total number of words already selected. The relative entropy of the maximum likelihood estimate of the language model of the selected sentences to the initial model P is given by

$$D = \sum_i P(i) \ln \frac{P(i)}{W(i)/N}$$

If we select the current sentence, the updated R.E is given by

$$D^+ = \sum_i P(i) \ln \frac{P(i)}{(W(i)+m(i))/(N+n)}$$

Direct computation of R.E using the above expressions for every sentence in the web-data will have a very high computational cost since $O(V)$ (where V is the size of the vocabulary) computations per sentence in the web-data are required. The number of sentences in the web-data can be very large and can easily be on the order 10^8 to 10^9. The computation cost for moderate vocabularies (around 10^5) would be very large. In addition, if we include bigrams and trigrams the computation would be infeasible.

However given the fact that $m(i)$ is sparse, we can split the summation D^+ into

$$\begin{aligned} D^+ &= \sum_i P(i) \ln P(i) + \\ &\quad - \sum_i P(i) \ln \frac{W(i)+m(i)}{N+n} \\ &= D + \underbrace{\ln \frac{N+n}{N}}_{T1} \\ &\quad \underbrace{- \sum_{i, m(i)\neq 0} P(i) \ln \frac{(W(i)+m(i))}{W(i)}}_{T2} \end{aligned}$$

Intuitively, the term $T1$ measures the decrease in probability mass because of adding n words more to the corpus and the term $T2$ measures the in-domain distribution P weighted improvement in probability for words with non-zero $m(i)$.

With the use of the relative entropy based data selection method, we were able to achieve with just one-third of the selected data the same performance achieved with the entire data [6].A detailed description of the theory and implementation of the R.E criterion can be found in [5],[6].

2.1.2. Clustering

In order to ensure that we had adequate coverage of adaptation material for the range of topics inherent in the TC-STAR domain, we first use unsupervised methods to build document clusters that represent topics covered in the training data. Traditional text clustering methods based on Latent Semantic Analysis (LSA)[7] represent a document as a vector comprising of weighted word counts. SVD (Singular Vector Decomposition) of the document/word-count matrix is then used to identify the topic centers. The central assumption is that a Euclidean distance metric will cluster similar documents. However, a more natural model for documents is a multinomial distribution. Recently proposed probabilistic methods such as pLSI[8] and Latent Dirichlet Allocation(LDA)[2] can represent documents in terms of generative multinomial factors and have been shown to have a superior performance compared to Euclidean distance based measures.

We represented each speaker turn as a document and used LDA for clustering the in-domain data. The sessions in the European Parliament consist of various politicians addressing a specific issue before moving on to the next topic of interest. This implicit structure naturally allowed for the use of speaker turns as document boundaries. Stop words and disfluencies were removed for clustering. We then acquired additional data separately for each cluster. This additionally acquired data was reclustered using LDA. From every document cluster, we selected domain relevant data using a relative entropy minimization algorithm (Section 2.1.1). The clustered language models were merged with weights optimized on Dev06 test set. The number of clusters was varied from three to eight and perplexity on the dev-set was found to be minimal for five clusters. Table 1 shows the top ten words in three of the five topics. As can be seen from the table, the top words for each topic seem to be semantically consistent. It is important to note that the clusters were obtained from an unsupervised clustering method, which makes the clean separation between keywords more interesting.

Topic1	**Topic2**	**Topic3**
EUROPEAN	PEOPLE	DEVELOPMENT
STATES	WAR	HEALTH
RIGHTS	GOVERNMENT	INFORMATION
EU	IRAQ	ECONOMIC
INTERNATIONAL	PRESIDENT	PUBLIC
UNION	WORLD	SYSTEM
COUNTRIES	BUSH	RESEARCH
COMMISSION	POSTED	MARKET
COUNCIL	COMMENTS	SERVICES
EUROPE	MILITARY	POLICY

Table 1. Top 10 words from each cluster in LDA assignment for three of the topics generated by automatic clustering

2.2. Unsupervised Training

The untranscribed EPPS material from 2005 and 2006 was used for unsupervised acoustic model training. This untranscribed material was first decoded using the Eval06 system submitted under the *public* condition. No form of lightly supervised training was done using the FTE corresponding to the untranscribed material. However, a histogram based thresholding of poorly scoring utterances was used and this resulted in a rejection of 5.5% of the data. The same decoded material was also used for data selection in language modeling and its use as a separate component in the interpolated language model was explored. The use of this data resulted in the doubling of the acoustic training material. Table 2 illustrates the gains obtained at each stage of acoustic modeling with the additional data. The unsupervised data doubled the overall acoustic training material and yielded gains of 5% relative at WERs of 10%. The language model used at this point was built using the EPPS data only.

	Dev 06		Evl 06	
System	Sup.	Unsup.	Sup.	Unsup.
SI	17.9	16.1	16.1	14.8
VTLN	16.6	15.6	14.7	14.0
FSA	14.4	13.6	12.5	11.6
fMPE+MPE	12.7	12.2	11.1	10.5

Table 2. Comparison of WERs on the Dev06 and Evl06 EPPS test sets when using supervised and unsupervised training

Increased number of parameters to account for the additional data did not provide any significant gains in performance.

2.3. Automated Speaker Segmentation and Clustering

The EPPS recordings contain speeches from politicians and interpreters in different languages, both, from native and non-native speakers of English. Similar to the 2006 Evaluation, in this year's evaluation, the number of speakers and their speech boundaries were not provided. Therefore, the first step in the recognition system is a segmentation of each session's audio file into speech and non-speech segments. We use an HMM-based segmentation system that models speech and non-speech segments with five-state, left-to-right HMMs with no skip states. The speech and non-speech models are obtained by applying a likelihood-based, bottom-up clustering procedure to the speaker independent acoustic model used in the first pass of the decoding step. This is followed by a clustering procedure to cluster the segments into clusters using a likelihood thresholding scheme to decide the number of clusters that can then be used for speaker adaptation. All homogeneous speech segments are modeled using a single Gaussian density and clustered into a pre-specified number of clusters using K-means and a Mahalanobis distance measure.

2.4. Cross system adaptation

Cross-system adaptation involves computing speaker-specific transforms from ASR transcripts generated from different systems. Unlike the 2006 system [9] where each branch in the overall decoding scheme was cross-adapted across two different segmentation schemes, in this year's scheme, cross-system adaptation was achieved differently. The transcripts generated from a single system using one speaker clustering scheme was aligned to the acoustic models used in the second system which were then adapted with the newly aligned transcripts. The transcripts generated by the adapted second system are subsequently used to adapt a third system and so on. This procedure is continued breadth-wise across the different systems. This simpler scheme involves fewer decoding steps (6) than the one used in the 2006 evaluation (24) and provides similar gains ranging from 0.3% to 0.5%.

2.5. Ensemble of ASR systems using randomized decision trees

A characteristic of our system architecture is the use of an ensemble of ASR systems whose decisions are combined using ROVER [10] to obtain a single recognition hypothesis (see Fig. 3). The ROVER voting approach is most effective when the individual ASR systems of the ensemble make uncorrelated errors. A typical procedure to build such systems is to use different acoustic front-ends (e.g. PLP vs MFCC) or different phone sets across systems. In this work however, we adopt a more systematic approach to build multiple systems by randomizing the training procedure identical to the system used in the 2006 evaluation. Randomness is introduced by replacing the classical decision-tree state-tying procedure used to tie context-dependent acoustic units, by a randomized decision tree growing procedure [3]. Randomized decision trees are grown by randomly selecting the split at each node, from the top N-best split candidates. In contrast, standard decision trees are grown by selecting the best split candidate. ASR systems built on different sets of randomized decision trees will model different clusters of context-dependent units. Multiple systems can then be systematically built simply by changing the random number generator seed. We have experimentally observed that such systems are good candidates to be used with the ROVER voting procedure [3]. In our experiments, we found that combining four or five systems yielded the best performance and merging additional systems beyond this provided little or no gain. It should be noted that the word error rates on these test sets are rather low and the gains seen from combining systems is increased when the word error rates are higher. Table 3 demonstrates a 0.4% reduction in WER on the Dev06 test set where the top 5 candidates were considered for the split. Four different systems were built with 6000 states and 150K Gaussians and combined with the baseline system of the same size.

	Baseline	R1	R2	R3	R4	ROVER
Dev06 WER	8.0	7.9	8.0	8.0	7.9	7.6
Evl06 WER	5.8	5.7	5.7	5.7	5.8	5.5

Table 3. Comparison of WER: Effect of Randomized Decision Tree tying based system combination

3. TRAINING DATA

The manually transcribed English training data comprises of 101 hours of the English portion of the EU plenary sessions with approximately 75 hours of speech from over 1900 speakers (politicians and interpreters) . This data covers sessions from May 2004 through May 2005. The Dev06 development test set on which the acoustic and language models were optimized consists of approximately 3 hours of data from 42 speakers (mostly non-native speakers). The 2006 English Evaluation corpus (Evl06) comprises of 3 hours of data from 41 speakers. Given the nature of the task, it is only natural that there is an overlap between the training and test corpora. Table 4 illustrates the overlap between the training and test corpora. Care was

taken to ensure that the time intervals of training and test material did not overlap.

	Dev06	Evl06
Matching speakers	15 (34.1%)	18 (43.9%)
Matching Data	30%	36.8%
Average duration per speaker	2.22 min	2.52 min

Table 4. Overlap between training and test speakers

The language model (LM) separates the systems submitted under the three evaluation conditions. For the *restricted* LM, the training material consisted of only EPPS acoustic transcripts (755K words) and FTE texts (34M words). The *public* LM was built from the *restricted* LM data, the British Parliament text also known as the HANSARD corpus [2] (40M words), Broadcast news (204M words) and Web-based data (525M words) released by the University of Washington. The *open* LM was built using data selected from the web (12G words) using a relative entropy based scheme [6].

4. BASIC SYSTEM DESCRIPTION

4.1. Acoustic Modeling

The acoustic models used for all three evaluation conditions are explained below. The acoustic front-end employs 40-dimensional, LDA-ed, perceptual linear prediction (PLP) features that are mean- and variance-normalized on a per-utterance basis. The speaker-independent (SI) acoustic models used in the system consist of multiple sets of Hidden Markov Models (HMMs) all of which have been trained on all-transcribed acoustic material from the European Parliamentary Plenary Sessions (EPPS) domain, and available for training as released by RWTH for this project. The evaluation system employs Vocal Tract Length Normalization (VTLN) [11, 12], with a piecewise linear frequency warping and a breakpoint of 6500Hz. The speaker-adaptive training (SAT) model [13, 14] is trained on features in a linearly transformed feature space resulting from applying feature-based Maximum Likelihood Linear Regression (fMLLR) transforms computed on a per speaker basis to the VTLN-normalized features. Five different sets of SAT HMMs were built for the English system using randomized decision tree clustering of quinphone statistics, each with 6000 tied states and 150K Gaussians. The Minimum Phone Error (MPE) model is trained on features obtained from a feature-space minimum phone error (fMPE) transformation [4]. The acoustic model training is described in detail in [9].

4.2. Language Modeling

4.2.1. Selection based on Topics and R.E

All LMs were 4-gram modified, Kneser-Ney models built using the SRI LM toolkit [15]. A perplexity minimizing mixing factor was computed using the Dev 06 reference text. The data for the *open* LM was selected from a May 2006 snapshot of the web using the architecture described in [16]. All domains under europa.eu were blocked in our web-crawling setup. To ensure that development data was not accidentally included we removed all web pages that had more than two 8-grams or more then six 6-grams common with the Dev 06 test set (This ended discarding about 0.25% of web pages crawled). To increase the coverage of web-content, we split the training data (and decodes from the unsupervised training data) into 5 topics using Latent Dirichlet Allocation [2] and gathered data for each topic (approx. 4G words). The downloaded documents were re-clustered into 5 topics and LMs were built on the data selected from each topic cluster. In all, a total of 12G filtered words were used in the LM build. Perplexity on the Dev06 test set reduced from 120 (with the restricted condition language models) to 63 with inclusion of web data. This data was merged with the 204M words of Broadcast News. This resulted in an LM containing 150M n-grams after pruning. The next few subsections provide a brief description of the various steps in building the language model for the *open* condition.

The 57k lexicon was obtained by taking all words occurring at least twice in the text corpus and once in the the acoustic training transcripts. The OOV rate on the Dev 06 test set was slightly under 0.4%. Pronunciations are based on a 45 phone set (42 speech, 1 silence phone and 2 noise phones). Pronunciations were obtained from the Pronlex lexicon and verified manually. In addition politician names were also included in the lexicon.

4.2.2. Query generation and web-crawl

A key step in acquiring data from the web is the generation of domain relevant queries. We identified n-grams with high occurrence probabilities in the training data to serve as query terms. We found that using language models to identify query terms gave better performance than using counts directly. We believe that this can be attributed to the smoothing and back-off algorithms involved in building language models. The query term selection criteria that worked best for us was KL divergence, given by the expression $p \ln \frac{p}{q}$ where p is probability of the term in the in-domain model and q the probability in a conversational (switchboard) language model.

The top URLs returned by the search engine for the queries are then downloaded and filtered[16]. Analysis of the top URLs returned by the search engines yielded an interesting observation. As can be seen from Table 5 more than half of the domain addresses for the top URLs were blog sites. This could be an indication that blog sites are especially beneficial for language modeling.

www.blogger.com
shaan.typepad.com
encycl.opentopia.com
eureferendum.blogspot.com
www.parliament.uk
www.samizdata.net
atangledweb.typepad.com
www.haloscan.com
www.eureferendum.com
news.bbc.co.uk

Table 5. Top 10 referred URLs in the web-crawl with 6 of them pertaining to blog sites

4.2.3. Decontamination

It is important to make sure that the out-of-domain corpus does not include the development test set as this can lead to overestimation of the out-of-domain LM's weight. To detect accidental overlap between out-of-domain corpus and development set, we calculated the number of 6-grams overlapping with the development set set for each document. Based on the histogram of overlap (see Figure 1) we

[2] Released by ELDA

fixed the threshold for maximum allowed 6-gram overlap as six and removed all documents with more than six 6-grams common with development set. Additionally we removed all documents with two or more matching 8-grams. Overall we removed 0.25% of the data. Since this data was well matched to the development set the corresponding increase in perplexity was 1.5%.

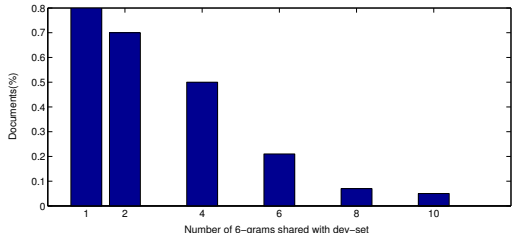

Fig. 1. Percentage of documents and the number of 6-grams shared with development set

4.2.4. Adaptation data aging

European parliamentary speeches are based on recent European issues and world events. Their textual content shows significant temporal variation. In a task of this nature, it is expected that the utility of out-of-domain data will decay with time. We collected transcripts (from the web) of European parliamentary speeches from June 2006 to January 2007 and for each month we compared the perplexity obtained from the language model built for the 2007 Evaluation under the *open* condition with the language model built for the 2007 Evaluation under the *public* condition. As described earlier, all the data for the *open* condition system was acquired before June 2006. Figure 2 shows that the perplexity improvement from inclusion of the out-of-domain data in the *open* LM decays with time. The reduction in perplexity gains from out-of-domain data with time, reinforces the need to keep the adaptation corpora updated for dynamic tasks like TC-STAR and broadcast news.

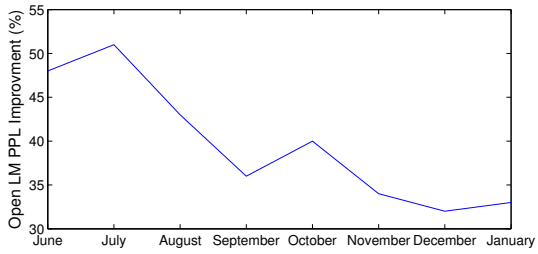

Fig. 2. Decay in perplexity improvement with time over the period June 2006 to January 2007

5. OVERALL SYSTEM ARCHITECTURE

This section describes the overall system architecture (see Figure 3) detailing the decoding steps and acoustic models (described in Section 4.1) that include the algorithms described in Section 2. The first step in the recognition system is the segmentation of each audio file into speech and non-speech segments followed by a clustering procedure to combine segments into clusters that can then be used for speaker adaptation. After the speaker clusters are determined, the final system output is obtained in 4 steps:

a) The SI pass uses the SI model and the LDA projected PLP features.

b) Using the transcript from a) as supervision, warp factors are estimated for each cluster using the voicing model.

c) Using the transcript from a) as supervision, fMLLR transforms are estimated for each cluster using the SAT model. A new transcript (*A-fsa-ctm*) is obtained by decoding using the SAT model and the language model built for the restricted condition is used for decoding.

d) Subsequently, the features are subjected to the fMPE transform. MLLR transforms are computed using *A-fsa-ctm* as supervision and a new transcript (A-ctm) is obtained by decoding using the MPE+MLLR model. The language model built for the *open* or *public* condition is used at this step for decoding.

The *A-fsa-ctm* transcript is used as the reference script by the ensemble of ASR systems (denoted by R1 through R4) used in the combination scheme to determine the warp factors and the transformations. Cross system adaptation across the baseline and complimentary systems is achieved in the following manner:

The *A-ctm* transcript is used in the last decoding step by system R1 to produce *R1-ctm*. *R1-ctm* is used for MPE+MLLR decoding using model R2 to produce *R2-ctm*. *R2-ctm* is used for MPE+MLLR decoding using Model R3 to produce *R3-ctm*. *R3-ctm* is used for MPE+MLLR decoding using Model R4 to produce *R4-ctm*.

6. RESULTS

The final numbers for the *open*, *public* and *restricted* conditions on the Eval 07 test set are given in Table 6. The English system uses the combination of multiple, cross-adapted ASR systems according to the procedure described in Section 5. The IBM ASR system was one of the top performing systems in the evaluation with 28% relative gain over the 2006 ASR system with an overall WER of 7.1%. A simple ROVER combination of the IBM 2007 ASR system with that of the partner sites served as input for the subsequent translation stage [1].

System	Open	Public	Restricted
Dev 06	7.6	9.6	10.0
Eval 06	5.5	7.7	8.3
Eval 07	7.1	8.9	9.8

Table 6. WERs on the Dev06, Evl06 and Evl07 EPPS test sets using the 2007 Evaluation System.

7. CONCLUSIONS

We have presented the IBM 2007 TC-STAR English ASR system. Overall, a relative 30% improvement in performance over the 2006 ASR system was obtained. Enhancements in language modeling such as clustering, acquiring and carefully selecting adaptation text account for a significant portion of the gain (18% relative) while

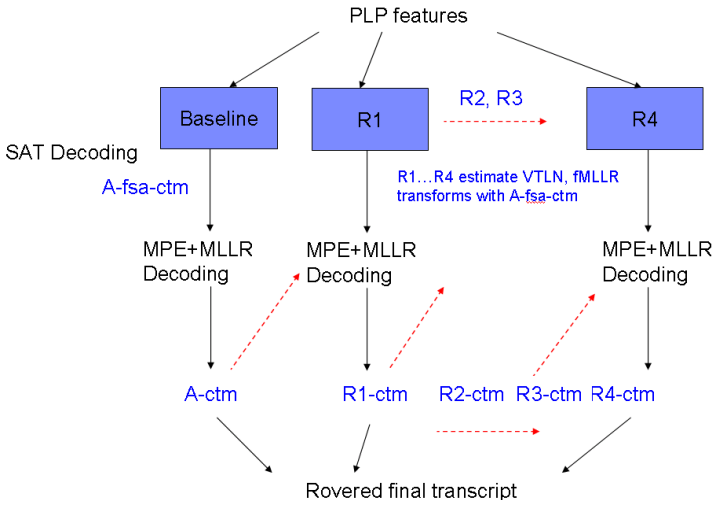

Fig. 3. Overall System Architecture

other enhancements such as unsupervised training, cross system adaptation and ASR system combination contributed a modest, additive gain. Significant gains can be achieved if we simultaneously boost the coverage and the relevance of the adaptation corpus when building language models.

8. REFERENCES

[1] ELDA, "TC-STAR: Technology and corpora for speech to speech translation," http://www.tc-star.org, 2004.

[2] Michael Jordan David M Blei, A. Y. Ng, "Latent dirichlet allocation," in *Journal of Machine Learning Research*, 2003.

[3] O. Siohan, B. Ramabhadran, and B. Kingsbury, "Constructing ensembles of asr systems using randomized decision trees," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Philadelphia, USA, 2005.

[4] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fMPE: Discriminatively trained features for speech recognition," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Philadelphia, USA, 2005.

[5] Abhinav Sethy, Panayiotis G. Georgiou, and Shrikanth Narayanan, "Text data acquisition for domain-specific language models," in *Proceedings of EMNLP*, 2006.

[6] Abhinav Sethy, Bhuvana Ramabhadran, and Shrikanth Narayanan, "Data driven approach for language model adaptation using stepwise relative entropy minimization," .

[7] J.R Bellegarda, "Large vocabulary speech recognition with multispan statistical language models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, 2000.

[8] Thomas Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning Journal*, vol. 42, 2001.

[9] B. Ramabhadran, O. Siohan, L. Mangu, G. Zweig, M. Westphal, H. Schulz, and A. Soneiro, "The IBM 2006 speech transcription system for European parliamentary speeches," in *Proceedings of ICSLP*, 2006.

[10] J. G. Fiscus, "A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (ROVER)," in *Proc. IEEE ASRU Workshop*, Santa Barbara, 1997, pp. 347–352.

[11] S. Wegman, D. McAllaster, J. Orloff, and B. Peskin, "Speaker normalization on conversational telephone speech," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, Georgia, USA, 1996.

[12] G. Saon, M. Padmanabhan, and R. Gopinath, "Eliminating inter-speaker variability prior to discriminant transforms," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Trento, Italy, 2001.

[13] M. F. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer Speech and Language*, , no. 12, pp. 75–98, 1998.

[14] G. Saon, G. Zweig, and M. Padmanabhan, "Linear feature space transformations for speaker adaptation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Salt Lake City, Utah, USA, 2001.

[15] A. Stolcke, "SRILM - an extensible language modeling toolkit," in *Proc. Int. Conf. on Spoken Language Processing*, Denver, Colorado, USA, 2002.

[16] Abhinav Sethy, Panayiotis Georgiou, and Shrikanth Narayanan, "Building topic specific language models from web-data using competitive models," in *Proceedings of Eurospeech*, 2005.