EXPERIMENTS ON CROSS-SYSTEM ACOUSTIC MODEL ADAPTATION

Diego Giuliani and Fabio Brugnara

FBK-irst Fondazione Bruno Kessler - Centro per Ricerca Scientifica e Tecnologica*
38050 Pantè di Povo, Trento, Italy
{giuliani,brugnara}@itc.it

ABSTRACT

Most state-of-the-art automatic transcription systems generate word transcriptions of the incoming audio data through two or more decoding passes interleaved by adaptation of acoustic models. It was proved that better results are obtained when the adaptation procedure exploits a supervision generated by a system different than the one under adaptation. In this paper, cross-system adaptation is investigated by using supervisions generated by several systems built varying the phoneme set and the acoustic front-end. Furthermore, an adaptation procedure is presented that makes use of multiple supervisions of the audio data for adapting the acoustic models within the MLLR framework. The gain achieved with cross-system adaptation and by adapting the acoustic models exploiting multiple, intra-site and cross-site, supervisions is demonstrated on the English European parliamentary speeches task.

Index Terms— cross-system acoustic model adaptation, ASR system combination, automatic speech recognition

1. INTRODUCTION

It has often been observed that different Automatic Speech Recognition (ASR) systems can make errors of different nature, while demonstrating similar Word Error Rates (WER). This characteristic is often exploited to improve recognition performance through cross-system Acoustic Model (AM) adaptation and system combination via ROVER (Recognizer Output Voting Error Reduction) or confusion network combination [1, 2, 3, 4].

When performing unsupervised AM adaptation of a system, better results are obtained if the supervision is generated by a different system, provided the latter ensures an adequate level of recognition performance. Several approaches have been proposed for building complementary ASR systems able to produce sufficiently different word transcriptions with uncorrelated errors. For example, the use of different acoustic front-ends and/or pronunciation lexica [5], or the randomization of the training procedure by randomizing the phonetic decision tree growing procedure [3]. When possible, word transcriptions generated by systems developed by different sites are exploited [2, 6].

In this paper[1], cross-system adaptation is first investigated by exploiting the output of several intra-site ASR systems each one using a different phoneme set and/or acoustic front-end.

To further improve recognition performance we propose an adaptation procedure which exploits multiple supervisions, that is word hypotheses generated by different ASR systems, to adapt AMs before the final decoding pass. The method is conceptually straightforward: it consists of performing the adaptation step on as many replicas of the audio data as there are supervisions, assigning to each replica a different supervision. In other words, it cumulates the counters that result from adapting the AMs on each individual supervision. AM adaptation is carried out in the Maximum Likelihood Linear Regression (MLLR) framework [7]. This adaptation procedure is compared to ROVER combination of the same system outputs. ROVER is the most widely used system combination approach. It is a post-recognition process which combines word hypotheses, generated by different ASR systems and ideally annotated with word confidence information, to generate a word hypothesis with reduced error rate [8].

The gain achieved with cross-system adaptation and by adapting the AMs exploiting multiple supervisions is demonstrated on the English European parliamentary speeches task by using the FBK-irst transcription system developed for the 2007 TC-STAR ASR evaluation campaign.

To further validate results of intra-site adaptation experiments, cross-site adaptation is investigated by exploiting as supervision the outputs of several systems developed by another participant in the 2007 TC-STAR evaluation campaign. Experiments reported confirm that better recognition results are achieved when supervisions for adaptation are generated by systems developed independently.

The paper is organized as follows. In Section 2 the FBK-irst transcription system is described while the procedure for AM adaptation with multiple supervisions is presented in Section 3. Intra-site system adaptation experiments are presented in Section 4 while cross-site system adaptation experiments are reported and discussed in Section 5. Finally, we summarize our conclusions in Section 6.

2. TRANSCRIPTION SYSTEM DESCRIPTION

In this section we summarize the main features of the FBK-irst systems developed for the 2007 TC-STAR ASR evaluation. One of the task of the evaluation was transcription of speeches delivered in English by politicians at the European Parliamentary Plenary Sessions (EPPS).

Word transcription is generated in two decoding passes after partitioning of the input audio stream. The input audio signal is first divided into homogeneous non overlapping segments using an acoustic classifier, based on Gaussian Mixture Models (GMMs), followed by a segment clustering method based on the Bayesian information criterion. The resulting segmentation and clustering is then exploited by the recognition system to perform cluster-wise feature normalization and AM adaptation.

Data for AM training were released to the participants in the

*Fondazione Bruno Kessler formerly known as ITC-irst.

[1]This work was partially funded by the European Union under the TC-STAR project (FP6-506738).

Table 1. *Recognition results (% WER) achieved by using several one-pass decoding systems.*

Data Set	One-Pass Decoding System			
	MFCC-USLex(1)	MFCC-BEEPLex(1)	PLP-USLex(1)	PLP-BEEPLex(1)
dev06en	13.6	13.5	14.1	13.6
eval07en	12.8	12.7	13.1	12.9

evaluation: about 101h of transcribed audio recordings data and 200h of untranscribed recordings. The untranscribed recordings were automatically transcribed using a preliminary version of the system. In total, about 250h of speech data were used for training.

AMs for the first and second decoding pass were both trained exploiting a variant of the speaker adaptive training scheme proposed by Gales [7], called Constrained MLLR-based Speaker Normalization (CMLSN) [9]. In the CMLSN method there are two sets of AMs, the target models and the recognition models. The method makes use of an affine transformation to normalize acoustic features on a cluster by cluster basis with respect to the target models. For each cluster of speech segments an affine transformation is estimated through CMLLR [7] with the aim of minimizing the mismatch between the cluster data and the target models. Once estimated, the affine transformation is applied to cluster data. Recognition models are then trained on normalized data. Leveraging on the possibility that the structure of the target and recognition models can be determined independently, a GMM can be adopted as target model for training AMs used in the first decoding pass [10]. This has the advantage that word transcriptions of test utterances are not required for estimating feature transformations. Instead, target models for training recognition models used in the second pass are usually triphones with a single Gaussian per state [9]. The same target models are used for estimating cluster-specific transformations during training and recognition.

In the current version of the system, a projection of acoustic feature space, based on Heteroscedastic Linear Discriminant Analysis (HLDA), is embedded in the feature extraction process as follows. A GMM with 1024 Gaussian components is first trained on an extended acoustic feature set consisting of static acoustic features plus their first, second and third order time derivatives. Acoustic observations in each, automatically determined, cluster of speech segments, are then normalized by applying a CMLLR transformation estimated w.r.t. the GMM. After normalization of training data, an HLDA transformation is estimated w.r.t. a set of state-tied, cross-word, gender-independent triphone Hidden Markov Models (HMMs) with a single Gaussian per state, trained on the extended set of normalized features. The HLDA transformation is then applied to project the extended set of normalized features in a lower dimensional feature space. Recognition models used in the first and second decoding pass are trained on normalized, HLDA projected, features. HMMs for the first decoding pass are trained through a conventional maximum likelihood procedure. Models used in the second decoding pass are trained through the CMLSN method exploiting as target-models triphone HMMs with a single Gaussian density per state.

At recognition stage, the output of the first decoding pass is exploited as supervision for CMLSN-based feature normalization and MLLR-based acoustic model adaptation. In order to investigate cross-system adaptation, we trained several sets of AMs considering two acoustic front-ends and two pronunciation lexica. The two acoustic front-ends were:

- MFCC: 13 Mel-frequency Cepstral Coefficients, including the zero order coefficient.
- PLP: 13 Perceptual Linear Prediction acoustic features.

In both cases, acoustic features were computed every 10ms using a Hamming window of 20ms length. First, second and third order time derivatives were computed, after cluster-based mean and variance normalization, to form 52-dimensional feature vectors. Acoustic features were normalized and HLDA projected to obtain 39-dimensional feature vectors as described above.

The two different lexica used to provide phonetic transcriptions of words were as follows:

- *USLex*: Pronunciations in the lexicon are based on a set of 45 phones. The lexicon was generated by merging different source lexica for American English (LIMSI '93, CMU dictionary, Pronlex).
- *BEEPLex*: This lexicon was generated by exploiting the British English Example Pronunciations (BEEP) lexicon. Pronunciation models in this lexicon are based on a set of 44 phones. Transcription for a number of missing words were obtained by exploiting the pronunciation models in the *USLex* lexicon and mapping phonetic symbols into the BEEP phone set.

By considering the different possible combinations of the two acoustic front-ends with the two lexica, four sets of state-tied, cross-word, gender-independent triphone HMMs were trained for each decoding pass. Around 300,000 Gaussian densities, with diagonal covariance matrices, were allocated for each model set.

Two different fourgram Language Models (LMs) were used in the first and second decoding pass. For the first pass, a background LM was trained on texts from news agencies, about 164M words, released by the Linguistic Data Consortium (LDC) in addition to texts from the EPPS Final Text Edition corpus, about 36M words, released for the TC-STAR evaluation campaign. The EPPS Final Text Edition texts are the official transcriptions of the parliamentary debates. The LM included 49k unigrams, 11M bigrams, 17M trigrams and 23M fourgrams. For the second decoding pass, the background LM was trained on an extended selection of public texts from several sources. In total, the training corpus consisted of 674M words. This LM included 65k unigrams, 27M bigrams, 29M trigrams and 27M fourgrams.

Both background LMs were adapted to the EPPS domain by exploiting a text corpus consisting of the manual transcriptions of the EPPS audio data released for training of the AMs (consisting of about 0.8M words) plus texts, about 4M words, corresponding to the EPPS Final Text Edition texts covering the same period of the acoustic training data.

3. MULTIPLE SUPERVISION ADAPTATION

The unsupervised AM adaptation procedure proposed here aims at mitigating the effect of errors in the recognition hypotheses relying on the fact that sufficiently different systems should produce different recognition errors, thus providing supervisions with complementary information. It assumes that several different system outputs are available for the test data and consists in performing adaptation on as many replicas of the audio data as there are supervisions, assigning to each replica a different supervision. In other words, it cumulates

Table 2. *Recognition results (% WER) of cross-system adaptation experiments achieved on the* dev06en *development set.*

First Pass Decoding System	Second Pass Decoding System			
	MFCC-USLex(2)	MFCC-BEEPLex(2)	PLP-USLex(2)	PLP-BEEPLex(2)
MFCC-USLex(1)	11.5	10.6	11.4	10.6
MFCC-BEEPLex(1)	11.0	11.1	10.9	11.0
PLP-USLex(1)	11.4	10.7	11.6	10.8
PLP-BEEPLex(1)	10.8	11.1	11.0	11.1

Table 3. *Recognition results (% WER) of cross-system adaptation experiments achieved on the* eval07en *evaluation set.*

First Pass Decoding System	Second Pass Decoding System			
	MFCC-USLex(2)	MFCC-BEEPLex(2)	PLP-USLex(2)	PLP-BEEPLex(2)
MFCC-USLex(1)	10.2	9.7	10.3	9.8
MFCC-BEEPLex(1)	9.7	10.1	9.8	10.1
PLP-USLex(1)	10.1	9.6	10.4	9.8
PLP-BEEPLex(1)	9.7	10.0	9.8	10.4

the counters that result from adapting the AMs on each individual supervision.

Another approach that directly exploits multiple supervisions in the adaptation procedure is the lattice-based unsupervised MLLR speaker adaptation [11]. In this case, however, multiple supervisions are represented as a word lattice generated by a single system.

With respect to the ASR system described in the previous section, the proposed adaptation procedure, slightly complicated by the fact that segmentations and lexica may be not aligned among the different systems generating the supervisions, is as follows [6]:

- All the word hypotheses generated by different systems are time aligned with a reference segmentation, that in our case is provided by the FBK-irst audio partitioner. The segmentation includes a cluster label for each segment, that will be used in the following steps for performing cluster-based acoustic feature normalization and model adaptation.
- To identify pronunciation variants, a forced alignment of the audio data with the word-level transcriptions is performed by applying the pronunciation model. Words in the transcriptions that are outside the FBK-irst lexicon are mapped to an out-of-vocabulary acoustic model.
- Clusters of speech segments are built according to the reference segmentation. Each cluster includes as many copies of its speech segments as there are supervisions, each copy having assigned a different supervision.
- Cluster-wise CMLLR normalization of audio data is performed with respect to target HMMs. Target HMMs are triphone HMMs with a single Gaussian per state and trained on normalized, HLDA projected, acoustic features.
- Cluster-wise adaptation of acoustic models used in the final decoding pass is performed on normalized acoustic data resulting from the previous step.

With this approach, the same portion of the audio data can contribute to counters of different model states, and its influence is weighted both by agreement among word hypotheses and by the acoustic match with the reference models.

The latter step in the adaptation procedure, that is cluster-wise AM adaptation, is performed in the MLLR framework [7] to adapt Gaussian means of triphone HMMs. Instead of a few affine transforms, a variant which is based on many simple "shift" transforms is adopted. Based on past experience [6], we consider transformations that consist in a shift vector added to the Gaussian means, that is $\mu' = \mu + c$. For this kind of transforms, a reliable estimate can be achieved on a small amount of data.

Regression classes are determined dynamically based on adaptation data. For this purpose a full regression class tree is top-down explored and a regression class is defined at the lower level for which the class occupancy counter is still over a fixed minimal threshold. The minimal occupancy threshold adopted can be much smaller than the one commonly used for full matrix estimation, e.g. 50 or 100 frames instead of 1000. This adaptation set up is common to all experiments reported in this paper, including adaptation experiments with a single supervision.

4. INTRA-SITE ADAPTATION EXPERIMENTS

Table 1 reports on recognition results achieved on the TC-STAR '06 development (*dev06en*) and '07 evaluation (*eval07en*) sets by performing a single decoding pass with systems using different acoustic front-ends and lexica. Each test set consists of about 3 hours of speech data.

In the table, and in the following ones, systems are identified with a label that specifies the acoustic front-end (MFCC or PLP), the lexicon (*BEEPLex* or *USLex*), and the presence of adaptation, denoting with (1) the decoding with unadapted AM and with (2) the decoding after AM adaptation.

Results show that, for single pass decoding systems, MFCC derived features are better than the PLP derived features and that the *BEEPLex* lexicon provides slightly better pronunciation models than the *USLex* lexicon on this task.

Tables 2 and 3 report on results achieved by performing two decoding passes with several systems. Results reported on the main diagonals correspond to the case in which systems used in the first and second decoding pass make use of the same acoustic front-end and lexicon. We have to point out however that, in all experiments reported, the language models as well the acoustic models used in the first and second decoding pass were different. Preliminary experiments showed that this may also induce some cross-adaptation effects leading to improved recognition performance.

In the tables cross-adaptation effects can be observed when the system used for the second decoding pass makes use of different lexicon or/and acoustic front-end with respect to the system generating the supervision exploited for acoustic feature normalization and AM adaptation. Results show that cross-adaptation effect is more visible when systems used in the two decoding passes exploit different lexica, resulting in WER relative reductions around 4-5%.

Table 4. *Recognition results (% WER) of cross-system adaptation experiments achieved on the* dev06en *development set with three decoding passes.*

Pass	System	WER	System	WER	System	WER	System	WER
1	MFCC-BEEPLex(1)	13.5	MFCC-USLex(1)	13.6	MFCC-BEEPLex(1)	13.5	MFCC-USLex(1)	13.6
2	MFCC-BEEPLex(2)	11.1	MFCC-USLex(2)	11.5	PLP-USLex(2)	10.9	PLP-BEEPLex(2)	10.6
3	MFCC-BEEPLex(2)	11.0	MFCC-BEEPLex(2)	10.4	MFCC-BEEPLex(2)	10.5	MFCC-BEEPLex(2)	10.3

Table 5. *Recognition results (% WER) of cross-system adaptation experiments achieved on the* eval07en *evaluation set with three decoding passes.*

Pass	System	WER	System	WER	System	WER	System	WER
1	MFCC-BEEPLex(1)	12.7	MFCC-USLex(1)	12.8	MFCC-BEEPLex(1)	12.7	MFCC-USLex(1)	12.8
2	MFCC-BEEPLex(2)	10.1	MFCC-USLex(2)	10.2	PLP-USLex(2)	9.8	PLP-BEEPLex(2)	9.8
3	MFCC-BEEPLex(2)	10.0	MFCC-BEEPLex(2)	9.4	MFCC-BEEPLex(2)	9.3	MFCC-BEEPLex(2)	9.5

Best results achieved on the development and evaluation sets are 10.6% and 9.6% WER, respectively. In Table 3, a 9.7% WER is reported for the configuration using in the first decoding pass the MFCC-USLex(1) system and in the second decoding pass the MFCC-BEEPLex(2) system, which corresponds to the configuration chosen for the 2007 TC-STAR evaluation submission. Due to a post evaluation refinement, the recognition score reported here is 0.1% absolute better than the score reported in the official results of the evaluation.

Additional recognition experiments were then carried out to ascertain whether better recognition results could be obtained through an additional recognition pass. Tables 4 and 5 report recognition results achieved on the development and evaluation sets, respectively, by performing three decoding passes interleaved by acoustic feature normalization and acoustic model adaptation. Intermediate recognition results, achieved with one and two decoding passes, are also reported. In all recognition experiments the final decoding pass was performed with the MFCC-BEEPLex(2) system. In both tables, the first column corresponds to the case in which all the three decoding passes exploit the same acoustic features and the same lexicon. It can be seen that little improvement is achieved after the second decoding pass. Instead, in all the other cases, in which acoustic features and/or the lexicon change across decoding passes a more tangible performance gain is achieved. Best recognition results achieved with three decoding passes are 10.3% and 9.3% WER for the development and the evaluation sets, respectively.

By exploiting the adaptation procedure described in Section 3, we carried out AM adaptation experiments by exploiting multiple supervisions of the same acoustic data. We used the outputs of two or more systems corresponding to system outputs scored in Tables 2 and 3 for the development and evaluation sets, respectively:

- *2Sups*: 2 supervisions corresponding to system outputs generated by performing an initial decoding pass with the MFCC-USLex(1) system and two parallel second decoding passes with the PLP-USLex(2) and PLP-BEEPLex(2) systems.
- *3Sups*: 3 supervisions corresponding to system outputs generated by performing an initial decoding pass with the MFCC-USLex(1) system and three parallel second decoding passes with the MFCC-USLex(2), PLP-USLex(2) and PLP-BEEPLex(2) systems.
- *4Sups*: 4 supervisions generated by adopting in the first and second decoding pass systems using the same acoustic front-end and lexicon, for example the combination MFCC-USLex(1) and MFCC-USLex(2). Scores of the system outputs included in the supervision set are reported on the main diagonals of Tables 2 and 3.
- *8Sups*: 8 supervisions generated by adopting in the first and second decoding pass systems using different lexica, that is the combinations MFCC-USLex(1) with MFCC-BEEPLex(2), MFCC-USLex(1) with PLP-BEEPLex(2), etc.

For comparison purposes, supervisions in each set were first combined through ROVER leading to the recognition performance reported in Table 6. In this work, ROVER combinations were always performed taking into account confidence scores associated to word hypotheses. We note that performing ROVER on just two or three system outputs (*2Sups* and *3Sups* columns) does not always improve performance w.r.t. the best system output entering in the combination. When four or eight system outputs enter in the ROVER combination some advantage is ensured.

Table 6. *Recognition results (% WER) achieved by combining the outputs of several systems through ROVER.*

Data Set	ROVER Input			
	2Sups	3Sups	4Sups	8Sups
dev06en	10.7	10.8	10.6	10.2
eval07en	9.7	9.7	9.4	9.1

After having performed acoustic feature normalization and AM adaptation by exploiting multiple supervisions, a final decoding pass was carried out with the MFCC-BEEPLex(2) system. Recognition results are reported in Table 7, upper part. Results show that the use of multiple supervisions for AM adaptation is effective and gives better performance than performing ROVER of the same system outputs (see Table 6). With respect to the three decoding passes experiments whose results are reported in Tables 4 and 5, multiple supervision adaptation always provides some benefit, and tangibly better results are achieved when exploiting eight supervisions, leading to 9.9% and 9.0% WER on the development and evaluation sets, respectively. The 9.2% WER achieved on the *eval07en* evaluation set exploiting the *3Sups* and *4Sups* supervision sets were obtained with the same system configurations chosen for the contrastive systems participating in the 2007 TC-STAR evaluation.

The ROVER output was used in its turn as a supervision for AM adaptation. Results achieved are reported in the lower part of Table 7. They show that adapting the system by using only the ROVER output is less effective than adapting with multiple supervisions.

5. CROSS-SITE ADAPTATION EXPERIMENTS

Leveraging on the fact that cross-system adaptation should be more effective when the system generating the word transcription is independently developed with respect to the system to be adapted, in [2]

Table 7. *Recognition results (% WER) achieved adapting the MFCC-BEEPLex(2) system by exploiting multiple supervisions, upper part, and by exploiting the output of the ROVER combination as a single supervision, lower part.*

Data Set	Supervision Set			
	2Sups	3Sups	4Sups	8Sups
dev06en	10.2	10.1	10.1	9.9
eval07en	9.3	9.2	9.2	9.0
Data Set	Supervision			
	ROVER of 2Sups	ROVER of 3Sups	ROVER of 4Sups	ROVER of 8Sups
dev06en	10.3	10.3	10.3	10.3
eval07en	9.4	9.4	9.2	9.2

it was proposed to use a cascaded cross-site adaptation scheme in which the output of a system developed at a given site was used to adapt a system developed by another site.

In this work, for cross-site adaptation experiments we exploited the outputs of several transcription systems developed by RWTH, Aachen University. Table 8 reports recognition performance achieved by the RWTH systems. Four systems, denoted as s1, s2, s3 and s4, were designed in order to be sufficiently different one from each other in view of their use in system combination schemes. The 1-best outputs of these systems were combined through ROVER. RWTH was able to further improve recognition results by combining the word lattices generated by the four systems s1-s4 and performing frame-based, minimum fWER decoding for generating the final hypotheses [12]. In the table recognition results achieved with lattice and 1-best combination techniques are denoted with *fWER* and *ROVER*, respectively. The result achieved with the fWER combination was the best reported in the 2007 TC-STAR ASR evaluation under condition "public" for which any publicly available training data could be used for LM and AM training.

Table 8. *Recognition results (% WER) achieved by each individual system developed by RWTH (s1-s4) and combining the four systems through ROVER and minimum fWER combinations.*

Data Set	RWTH System					
	s1	s2	s3	s4	ROVER	fWER
dev06en	11.5	12.0	11.8	10.7	10.1	9.7
eval07en	10.1	10.9	11.8	9.8	9.3	9.0

Table 9 reports recognition results achieved adapting the MFCC-BEEPLex(2) system by exploiting as supervision word transcriptions generated by the RWTH systems. It can be seen that cross-site adaptation leads to better recognition results with respect to results achieved by performing three decoding passes with intra-site systems (see Tables 4 and 5). These results confirm that cross-system adaptation is more effective when systems generating the supervisions are independently developed with respect to the system to be adapted. Noticeable are the results achieved using as supervisions the outputs of the s4 system which are only little worse than results achieved using as supervisions the outputs of the ROVER and fWER combinations.

Results of cross-site adaptation experiments with multiple supervisions are reported in Table 10. We first carried out adaptation experiments using the supervisions generated by the four systems, s1-s4, developed by RWTH and scored in Table 8. Finally, we performed adaptation experiments exploiting supervisions from two sites: the outputs of the four s1-s4 RWTH systems and the outputs of

Table 9. *Recognition results (% WER) achieved adapting the MFCC-BEEPLex(2) system by exploiting as supervision word transcriptions generated by the RWTH systems.*

Data Set	RWTH Supervision					
	s1	s2	s3	s4	ROVER	fWER
dev06en	9.6	9.9	9.6	9.5	9.3	9.2
eval07en	8.6	8.8	9.0	8.4	8.3	8.3

the four FBK-irst systems (*4Sups* set) scored on the main diagonals of Tables 2 and 3 for the two test sets. It can be noted that when using the s1-s4 RWTH supervisions, multiple supervision adaptation provides better performance than ROVER and fWER combinations (see Table 8) even when the outputs of these combinations are used as single supervisions (see Table 9).

Results reported in Table 10 also show that when both internal and external system outputs are used as supervisions no advantage with respect to using only external supervisions is achieved. Furthermore, in this case, comparison between multiple supervision adaptation and ROVER is slightly in favor of the latter: 8.9% and 8.1% WER for the development and evaluation sets, respectively. These results were not improved by exploiting the ROVER output as a single supervision for system adaptation. Effectiveness of ROVER in this case can be explained by considering that while ROVER is neutral with respect to the different systems generating the word hypotheses, multiple supervision adaptation is biased towards the intra-site systems and therefore is less effective in exploiting the additional information provided by their outputs.

Table 10. *Recognition results (% WER) achieved adapting the MFCC-BEEPLex(2) system by exploiting the four s1-s4 RWTH supervisions (s1+s2+s3+s4) and by exploiting eight supervisions generated by systems developed in two sites: the four s1-s4 RWTH supervisions and the four FBK-irst supervisions corresponding to the 4Sups supervision set (s1+s2+s3+s4+4Sups).*

Data Set	RWTH Supervision	
	s1+s2+s3+s4	s1+s2+s3+s4 + 4Sups
dev06en	9.0	9.0
eval07en	8.2	8.2

6. CONCLUSIONS

In this paper we have presented a series of experiments concerning cross-system adaptation and we have proposed a method for exploiting multiple supervisions for AM adaptation. Cross-system adaptation experiments confirmed the importance of designing systems able to produce sufficiently different word transcriptions, with uncorrelated errors, to be used as a supervision for acoustic model adaptation in a multi-pass transcription process.

The proposed method for acoustic model adaptation with multiple supervisions is effective and can provide an alternative to the well known ROVER combination.

7. REFERENCES

[1] M.J.F. Gales, D.Y. Kim, P.C. Woodland, D. Mrva, R. Sinha, and S.E. Tranter, "Progress in the CU-HTK Broadcast News

Transcription System," *IEEE Trans. on Audio Speech and Language Processing*, vol. 14, no. 5, pp. 1513–1525, 2006.

[2] R. Prasad, S. Matsoukas, C.-L. Kao, J. Ma, D.-X. Xu, T. Colthurst, G. Thattai, O. Kimball, R. Schwartz, J.-L. Gauvain, L. Lamel, H. Schwenk, G. Adda, and F. Lefevre, "The 2004 BBN/LIMSI 20xRT English Conversational Telephone Speech System," in *Proc. DARPA RT04*, Palisades NY, Nov. 2004.

[3] 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 *Proc. of INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 1225–1228.

[4] B. Hoffmeister, D. Hillard, S. Hahn, R. Schlülter, M. Ostendorf, and H. Ney, "Cross-site and Intra-Site ASR System Combination: Comparisons on Lattice and 1-Best Methods," in *Proc. of ICASSP*, Honolulu, Hawai'i, April 2007, pp. IV–1145–1148.

[5] S. Stüker, C. Fügen, S. Burger, and M. Wölfel, "Cross-system adaptation and combination for continuous speech recognition: the influence of phoneme set and acoustic front-end," in *Proc. of INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 521–524.

[6] D. Giuliani and F. Brugnara, "Acoustic Model Adaptation with Multiple Supervisions," in *TC-STAR Workshop on Speech-to-Speech Translation*, Barcelona, Spain, June 2006, pp. 151–154.

[7] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Computer Speech and Language*, vol. 12, pp. 75–98, 1998.

[8] J. Fiscus, "A Post-Processing System to Yield Reduced Word Error Rates: Recogniser Output Voting Error Reduction (ROVER)," in *Proc. of ASRU*, Santa Barbara, CA, 1997, pp. 347–352.

[9] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Computer Speech and Language*, vol. 20, pp. 107–123, 2006.

[10] G. Stemmer, F. Brugnara, and D. Giuliani, "Adaptive Training Using Simple Target Models," in *Proc. of ICASSP*, Philadelphia, PA, March 2005, pp. I–997–1000.

[11] M. Padmanabhan, G. Saon, and G. Zweig, "Lattice-Based Unsupervised MLLR for Speaker Adaptation," in *ISCA ITRW ASR2000*, Paris, 2000, pp. 128–131.

[12] B. Hoffmeister, T. Klein, R. Schlülter, and H. Ney, "Frame based system combination and a comparison with weighted ROVER and CNC," in *Proc. of INTERSPEECH*, Pittsburgh, PA, Sept. 2006, pp. 537–540.