Acoustic Model Adaptation with Multiple Supervisions

D. Giuliani, F. Brugnara

ITC-irst - Centro per la Ricerca Scientifica e Tecnologica, 38050 Pantè di Povo, Trento, Italy
{giuliani,brugnara}@itc.it

Abstract

This paper reports on several experiments performed during the development of the ITC-irst transcription system for the TC-STAR '06 evaluation campaign. The aim is to find methods of exploting a set of alternative hypotheses produced by different systems to derive a transcription that is more acccurate than any of these. The most used technique for combining alternative hypotheses relies on the ROVER technique. In this work we found that another approach, based on adaptation of a reference system, may provide some advantage over that technique.

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). The most notable example of a method for exploiting this feature is the ROVER system (Fiscus, 1997), a postprocessing module that first combines all the different hypotheses in a single graph, and then rescores the paths along it by a "voting" procedure that takes into account mutual agreement between word hypotheses, as well as, potentially, confidence scores attributed by the systems.
ROVER combination of the output of different systems has been shown to result in a significant reduction in WER with regard to any of the system entering in the combination, provided they are different enough to produce complementary errors. This characteristic is also exploited systematically in some systems, including the IBM and UKA system described elsewhere in these proceedings. While this is a convenient and effective way of exploiting multiple recognition hypotheses, it has some drawbacks. For example, it can only produce an improved single-best hypothesis, and no additional information such as word-graphs, hence it makes sense to consider different ways of combining system outputs.

2. Multiple-supervision adaptation

Most of the state-of-the-art automatic transcription systems, including all the systems used in the TC-STAR Evaluation, are based on multi-stage processing. After the first decoding, one or more additional steps are performed, each one involving adaptation of an acoustic model and using the output of a previous step as supervision, therefore using it as if it was the correct transcription of the input. The quality of the supervision can greatly influence final performance. It is obvious that a lower WER in the supervision can make the adaptation step more effective, but the behavior is not only related to this performance index. The benefit is significantly larger if the distribution of the errors in the supervision is well diversified with respect to the errors made by the model under adaptation. For example, during the development of the Eval06 system, we reduced the WER of the first step from 18.9 to 17.1 observing only a 0.2 improvement in the output of the second step. In this particular case, the improvement was obtained by introducing text-independent Constrained Maximum Likelihood Speaker Normalization (CMLSN) (Giuliani et al., 2006; Stemmer et al., 2005) in the first step, while the second step already exploits text-dependent CMLSN. Probably, this resulted only in moving to the first stage some discrimination capability already present in the second stage, without adding anything really new.

In order to be able to deliver a complete system output, while still exploiting information provided by several complementary systems, one can consider to put a reference system in the best conditions to perform the last step, providing it with an improved supervision. The latter can be obtained by ROVER combination of the system outputs. This is what was proposed within the TC-STAR project, and it has been shown to be a valid technique, able to improve the ROVER combination, as will be shown in some of the following experiments. With this approach, the relative importance of the systems is no longer as symmetrical as it is with the usual ROVER combination. The final performance depends, of course, on the accuracy of the system used in the last stage, and the weighting of the plausibility of the alternative hypotheses is still left to ROVER.

The method presented in this work, is still based on a last step of processing via a "reference" system. However, it also leaves to the latter system the task of weighting the plausibility of the alternative hypotheses.

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 merges the counters that result from adapting the acoustic model on each individual supervision. Adaptation can be made according to any technique, and in this work we compare results obtained by standard 4-class MLLR, Shift-MLLR, and their combination. It turns out that the adaptation setup can significantly change the final performance. It is noteworthy that, except in the case of experiments reported in the following section and labeled as "Baseline", adaptation in our system always includes preliminary CMLSN data normalization, performed with the same supervision. This can affect the comparison of effectiveness between the methods. The procedure for each input file, slightly complicated by the fact that segmentations and lexica are not aligned among the different systems, is as follows:

- All the word hypotheses generated by different systems are aligned with a reference segmentation, that in our case is provided by the ITC-irst audio partitioner. The segmentation includes a cluster label for each segment, that will be used in the following steps for performing cluster-based normalization and adaptation.
- A forced-alignment of the audio data with the word-

	Supervision	MLLR	S-MLLR	Cascade	Alternate	MLLR6	S-MLLR6
Baseline	16.7	15.0	15.5	15.0	15.0	15.0	15.6
Normalized		13.8	13.6	13.4	13.5	13.8	13.7

Table 1: *Performance (WER %) of different adaptation procedures on EPPS-DEV06EN, using the ITC-irst system.*

level transcriptions is performed, to identify pronunciation variants. Word in the transcriptions that are outside the ITC-irst lexicon are mapped to an OOV model.

- Cluster-wise CMLSN normalization is then performed, where each cluster is made of copies of the segments with identical labels, possibly with different supervisions.
- Cluster-wise acoustic model adaptation is performed with the same partitioning of the previous step.

The procedure is certainly more time consuming than a ROVER combination of the hypotheses, given the need to carry out adaptation on an expanded audio file, but the time required for this step is still only a small fraction of the time needed to generate the input hypotheses.
With this approach, the same portion of the audio data can contribute to counters of different states, and its influence is weighted both by agreement between hypotheses and by the match with the reference model. If the reference system is accurate enough, this can be a more detailed balancing among hypotheses with respect to what can be done by using hypotheses agreement and confidence measures.
As will be shown by the experiments outcomes, the accuracy of the final transcription also depends on the adaptation technique. In this context, we observed significant variations of performance by using, instead of the classical MLLR based on a few affine transforms, a variant which is based on many simple transforms, as described in the following.

3. Shift-MLLR

A widely used, effective technique for acoustic model adaptation is Maximum Likelihood Linear Regression (Leggetter and Woodland, 1995). This technique assumes that the probability density b_s associated to an HMM state s is a mixture of Gaussian densities, that is: $b_s(x) \equiv \Pr[x|s] = \sum_{k=1}^{N_s} w_{s,k}\mathcal{N}(x;\mu_{s,k},\Sigma_{s,k})$, where x is a feature vector.
In its original form, MLLR tranforms each mean μ of a Gaussian density by means of an affine tranformation, $\mu' = A\mu + c$, so that the likelihoods of observation vectors are given by $\sum_{k=1}^{N_s} w_{s,k}\mathcal{N}(x; A\mu_{s,k} + c, \Sigma_{s,k})$.
The parameters (A, c) of the transform are estimated, following a Maximum Likelihood criterion, so as to maximize the likelihoods of the adaptation data according to the transformed model. The technique assumes that the transform is shared among several Gaussians, otherwise it would simply return the ML estimate of the means on the adaptation data, only for the Gaussians that are actually observed. The appropriate degree of tying depends on the size of the adaptation set. When used for refining the models at recognition time, it often happens that the adaptation data are not enough to reliably estimate a large number of parameters, and the total number of degrees of freedom is controlled by limiting the number of different transforms. Though this helps in avoiding the risk of overtraining, it may also compromise the detail of the adaptation process, because Gaussian means that are scattered in the acoustic space are forced to be transformed by a common transform.
Another way of reducing the degrees of freedom, while preserving a higher level of acoustic resolution, is that of exploiting a larger number of transform, but impose a smaller number of parameters to each of them. In this work, we consider simple tranformations that consist in a shift vector added to the means, that is $\mu' = \mu + c$. For this kind of transforms, a reliable estimate can be achieved on a small amount af data, say a few tens of frames, intead of the hundreds required for estimating a full matrix. The criterion used for estimation is still ML but, given the strong constraints on the form of the transforms, the reestimation formula is particularly simple. If $g \equiv \mathcal{N}(.;\mu_g,\Sigma_g)$ is any Gaussian that shares the shift parameter c, γ_g is the overall posterior of g on the adaptation data, and $\bar{\mu}_g$ is the ML reestimate of the mean of g, we have:

$$c = \left(\sum_g \gamma_g {\Sigma_g}^{-1}\right)^{-1} \sum_g \gamma_g {\Sigma_g}^{-1}(\bar{\mu}_g - \mu_g)$$

If the covariance matrices of Gaussians are diagonal, as is often the case, the above equation translates straightforwardly in a component-wise expression.
We have found that the idea of using such simple transformation is not new, having already been proposed in (Digalakis et al., 1999). In that work, however, the shift (or *bias*) transform is used together with dependency modeling to assign different degrees of tying to matrices and offsets.
In the present work, the basis for tying the transform is the usual regression tree, built by agglomerative clustering of the Gaussians. What changes between full tranforms and shift transforms is the depth at which the tree is exploited. When using full transforms, at most the first two levels are used, giving rise to two or four regression classes, while in the case of the shift transform the full tree is searched until the node occupancy falls below a threshold. The threshold itself is much smaller than the one used for matrix estimation, e.g. 50 or 100 frames instead of 1000. The number of transforms can therefore vary considerably from cluster to cluster. For example, by looking at the effective usage of nodes in a typical experiment, we observed a range from 2 to 750, with an average of 155. As already suggested in (Digalakis et al., 1999), the matrix and bias tranforms could be estimated together, with different tying schemes. However, this complicates the expression of the objective function making optimization harder to achieve. As an easily realizable approximation, they propose to use a *cascade* combination. That is, first estimate the full tranforms with a stronger tying, and then estimate the shift tranforms with a looser tying. In this work we present results both using this procedure, and also a slightly different one, which is aimed at better approximating the joint estimation, that is *alternate* estimation of full and shift transforms. Performance difference is very limited between this two variants.
As will be shown by the experiments, one can not expect large benefits from this technique alone in a typical system. On the other hand, it appears to be advantageous in the context of cross-system adaptation or multiple supervisions.

Table 1 summarizes results obtained on the development set of the EPPS'06 English task (EPPS-DEV06en), by using the adaptations variants described above. All the experiments exploit the same supervision, that was provided by the first step of the baseline system after text-independent CMLSN normalization. The features of the system correspond to those reported in the description of the ITC-irst evaluation system elsewhere in these proceedings, except that no 4-gram LM is used in these experiments. The table presents results relative to the use of two different AMs in the adaptation step. The first row ("Baseline") applies adaptation to the same AM used for generating the supervision. This AM only exploits text-independent CMLSN normalization, while the second row ("Normalized") reports the performance of the AM that includes HLDA and supervised CMLSN.

The first column refers to the use of a standard setup with three iterations of MLLR adaptation of means and variances using a 4-class regression tree. It can be seen that this delivers an improvement of about 10% relative to the supervision, a fairly common result observed in many tasks. The second column shows the performance obtained by three iterations of Shift-MLLR, using the full regression tree, but imposing a threshold for occupancy of 100 frames. It turns out that this procedure provides worse performance than the previous one, confirming that it is not intrinsically superior to the usual method. We observed this in similar experiments as well, even though the difference was usually smaller. The following two columns show results obtained by combining full transforms and shift transforms in two different ways. For "Cascade", three iterations of 4-class full-transform MLLR are followed by three iterations of Shift-MLLR, while for "Alternate" the two estimations are interleaved for three times. Results show that any of the combinations performs similarly, recovering the performance of the standard method.

The second row exhibits different relative performance. In this case the Shift-MLLR method provides a small improvement, which becomes more visible when the technique is used jointly with full transform MLLR. The different trend can be explained by considering that in this case the data already undergo a HLDA+CMLSN normalization, providing most of the benefit with respect to the baseline. The additional full-transform MLLR step increases performance only marginally, being too similar to this processing. In contrast, Shift-MLLR influences model parameters in a more complementary way.

The last two columns presents contrastive results, where only one estimation method is applied with six iterations, since the latter is the total number of iterations performed when methods are combined. They show that the increased number of iterations does not affect the behavior of either technique, and in the case of Shift-MLLR a slight degradation occurs, probably due to overtraining.

From this and similar experiments, we conclude that the Shift-MLLR technique, while not being a substitute for the standard full-transform MLLR, is able to add some adaptation capability in the framework of a system that already exploits variations of the MLLR technique. In any case, performance is never lost with respect to the best case when Shift-MLLR and MLLR are used in combination.

4. System Combination Experiments

During the preparation for the 2006 TC-STAR Evaluation, it was decided to study the effect of techniques for combining different system outputs, so we run several cross-system adaptation experiments to systematically measure the benefits of using cross-site supervisions in an adaptive recognition system.

IBM05	15.3
IRST-p1	19.2
LIMSI05	14.0
RWTH	18.3
UKA	17.3

(a)

IRST-p1+IRST-p2	14.4
IBM05+IRST-p2	12.4
LIMSI05+IRST-p2	12.5
RWTH+IRST-p2	13.1
UKA+IRST-p2	13.6

(b)

All + IRST-p2	11.6
IRST-p1,RWTH,UKA + IRST-p2	12.8
IBM05,LIMSI05 + IRST-p2	11.5

(c)

Table 2: *Performance achieved with the first set of cross-system adaptation and system combination experiments on EPPS-DEV06en.*

Tables 2(a), 2(b) report on the first set of such experiments, performed on EPPS-DEV06en. Table2(a) shows the performance of each TC-STAR partner's intermediate system as available on the project web site during the initial stages of development. The first row of Table 2(b) shows the performance of the reference ITC-irst system used for the experiments, exploiting as supervision the output of IRST-p1, an unadapted system. The remaining rows show the performance obtained by using each of the alternative supervisions. All experiments were based on three iterations of Shift-MLLR, with an occupancy threshold of 50 for the nodes of the regression class tree. In this kind of experiments, Shift-MLLR always outperformed 4-class full-transform MLLR. This aspect will be better exemplified in the last set of experiments.

In Table 2(b), a significant advantage of cross-site systems can be observed over the ITC-irst system, and also over the best single-site system. As already noted in Section 2., this is not strictly related to WER. For example, comparing the effect of supervisions generated by the RWTH and UKA systems, one sees that, in spite of the former having a higher WER, it appears to be more effective for the ITC-irst system. A similar consideration applies to the comparison between the effects of IBM and LIMSI supervisions.

The cross-system adaptation experiments in Table 2(b) confirm and quantify a phenomenon which is already well known. They are useful for setting up a reference for what was the real goal of the work, namely to develop a technique to effectively combine multiple systems. To this end, we exploited a few grouping of the outputs to perform multiple-supervision adaptation with the procedure described in Section 2.. Results are reported in Table 2(c). The first row shows that, by using the supervisions of all

IBM_p	11.7
IRST_r	13.8
LIMSI_p	10.9
RWTH_r	14.2
UKA_p	14.6

(a)

ROVER	10.2
ROVER+LIMSI_p	9.7
ROVER+IRST-i	10.1
MulSup+IRST-i	9.9

(b)

Table 3: *Results of system combination experiments on EPPS-DEV05en.*

IBM_v4	10.6	IRST_v4	13.0
LIMSI_v4	10.1	RWTH_v4	12.9
UKA_v4	12.7	ROVER	8.7

(a)

	MLLR	S-MLLR	Cascade	Alternate	MLLR6	S-MLLR6
ROVER Sup.	12.8	10.6	10.4	10.4	12.7	10.4
Multiple Sup.	13.1	9.9	9.9	9.8	13.1	9.8

(b)

Table 4: *Performance on EPPS-DEV06EN of different adaptation procedures in system combination experiments. The reference system is the same used for producing the results in the second row of Table 1.*

systems, the combination achieves a 17% improvement relative to the best single-site system, and a 19% improvement relative to the reference system. The following rows report on results obtained by dividing the systems in two classes, those with a WER above 16% and those below that threshold. It can be seen that, even using the output of the less accurate systems, multiple supervision adaptation approaches the performance obtained with the best supervisions, and still outperforms any single-site system. However, the best performance overall is achieved by using the two most accurate systems.

The multiple supervisions technique presented in this paper attributes most relevance to the system used in the final step, therefore, to exploit the method at its best, this should be among the most accurate. To verify the validity of the method while approximating this condition, we performed some experiments on the EPPS-DEV05en data, using last year's system outputs, and an improved version of our system (IRST-i), with a WER of 11.2%, not far from the best single-site system. In this case, we are also able to compare similar experiments performed at different sites.

Table 3 summarizes the outcomes of this second set of experiments. On the left are the performance of the systems used by different sites in the EPPS'05 evaluation, on the right are results obtained by different combination techniques. The first row of Table 3(b) shows the performance of the ROVER combination of the single-site outputs, achieving a 6.4% relative WER reduction over the best single-site system. The second and third rows report on similar experiments performed at LIMSI and ITC-irst, in which the ROVER output was used as supervision for an adaptation step of the acoustic model.

While the performance of ROVER+IRST-i is not as good as ROVER+LIMSI_p, there is still a small gain when compared with the ROVER result. The last row shows that an improvement was obtained by using the multiple supervision method. Even if the absolute performance is still worse than the ROVER+LIMSI_p combination, due to a less accurate reference system, this result represents a 9.2% relative WER reduction over the best single-site system, and suggests that, for a given reference system, the multiple supervision method has the potential of better exploiting the information contained in multiple transcriptions.

The last set of experiments was run on EPPS-DEV06en after improved systems for the 2006 Evaluation became available. Table 4(a) shows the performance of each single-site system, together with the performance of their ROVER combination. The reference system for the combination experiments is the same used for the experiments in the second row of Table 1. It differs from IRST_v4 in that it does not include fourgram rescoring of the word-graphs. In this case, we compare the different variants of the adaptation technique in a system combination framework, both by using the supervision obtained by ROVER, and the multiple supervision method. The first thing to notice is that in this case the output with the multiple supervisions method can still be better than any single-site system, even if by a narrow margin, but is definetely worse than ROVER combination. In fact, in this case the assumption that the reference system is among the best is clearly violated, so we dicuss relative performance of the adaptation methods.

Two main observations can be made observing Table 4: the first is that, in this context, Shift-MLLR, either alone or together with full-transform MLLR, exhibits a clear advantage; the second is that the multiple supervision approach is actually superior to the ROVER supervision only in combination with Shift-MLLR. These can be explained by considering the difference in the characteristics of the tasks and of the adaptation methods.

There is a trade-off between acoustic resolution and robustness, and the relative influence of these two opposites is different in a single-site setup and a system combination setup. In a single-site system, the risk of overtraining is higher, since the model that generated the supervision and the one that is adapted according to its supervision are more likely to be similar, if not the same. Hence, robustness has to be preserved by means of a stronger tying. When using a cross-site supervision, or multiple supervisions, overtraining is much less likely to occur, therefore a looser tying is beneficial, assuming that the ratio between data sample and parameters is adequate for a reliable statistic estimation.

5. Conclusion

In this paper we presented a series of experiments concerning cross-adaptation between different ASR systems, and discussed a method for combining the output of multiple systems. It turns out that the latter can provide some advantage over the well known ROVER combination, if some conditions are met. Results also show that a simple variant of the MLLR technique is effective in this context.

6. References

V. Digalakis, S. Berkowitz, E. Bocchieri, C. Boulis, W. Byrne, H. Collier, A. Corduneanu, A. Kannan, S. Khudanpur, and A. Sankar. 1999. Rapid speech recognizer adaptation to new speakers. In *Proc. of ICASSP*, pages 765–768, Phoenix, AZ, USA.

J. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recogniser output voting error reduction (rover). In *Proc. of ASRU*, pages 347–352, Santa Barbara, CA.

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

C. J. Leggetter and P. C. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9:171–185.

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.