SPEAKER ADAPTATION BY VARIABLE REFERENCE MODEL SUBSPACE AND APPLICATION TO LARGE VOCABULARY SPEECH RECOGNITION

Wen Xuan Teng [1], *Guillaume Gravier* [2], *Frédéric Bimbot* [2], *Frédéric Soufflet* [1]

[1] TELISMA, FRANCE
{wteng,fsoufflet}@telisma.com

[2] IRISA (CNRS & INRIA) / METISS
{ggravier,bimbot}@irisa.fr

ABSTRACT

Recently, we presented a rapid speaker adaptation technique, reference model interpolation (RMI), which is based on the linear interpolation of speaker-dependent models and the *a posteriori* selection of reference models. The approach uses the *a priori* knowledge provided by a set of representative speakers to guide the estimation of a new speaker model in the speaker space. RMI achieved rapid supervised adaptation in phoneme decoding tasks. In this paper, we present two new results of RMI: firstly, we apply the RMI technique in a practical large vocabulary continuous speech recognition (LVCSR) system with unsupervised instantaneous adaptation. Secondly, we propose an evolutional subspace scenario which integrates the slow update of reference models with RMI rapid adaptation to achieve incremental adaptation. The unsupervised adaptation experiments carried out on broadcast news transcription task show encouraging results for both instantaneous and incremental adaptation.

Index Terms— speaker adaptation, reference models, LVCSR

1. INTRODUCTION

In a recent paper [1], we showed that most of rapid acoustic model adaptation techniques can be seen as a linear interpolation of some reference models. These reference models can be eigen-vectors issued from PCA [2], or *centroid* vectors of Gaussian mixtures [3], or directly a set of speaker dependent (SD) models. We also showed experimentally that for a particular speaker, there is a great variance in performance with randomly selected reference models. This implies that when reference models are fixed, these interpolation-based adaptation techniques are limited such that they cannot provide robust improvements for a particular adaptation target.

We present in this paper the notion of variable reference model subspace to address this limitation. Considering an acoustic model space in which each acoustic model is represented as a point, the reference models can be viewed as some "anchor models" in this space, and a reference model subspace can be formed by these anchors. The interpolation-based adaptation techniques define the adapted model within this subspace by a linear combination. When the dimension of the subspace is limited, the target model may not be found within the subspace. This can be improved by increasing the number of reference models used for the linear combination. However, it is often the case that we do not have enough data to build large numbers of SD models. Another solution is to make the reference model subspace variable at runtime during the adaptation. It is motivated by our experimental finding: the target model of a new acoustic condition is more likely to be found in the space formed by those reference models which have characteristics more similar to the target. In [1], we proposed *a posteriori* selection of reference models, which makes the subspace variable by dynamically selecting the pertinent reference models with respect to the test/adaptation data. In this paper, we present the evolutional reference model subspace to improve RMI for different speakers and acoustic conditions. The idea of evolutional subspace is to move each reference model within the acoustic model space to try to form more pertinent subspaces for different adaptation targets. This can be easily implemented by combining RMI rapid adaptation with the slow update of reference models in an incremental adaptation scenario.

In this paper, we first review the RMI rapid adaptation technique, applied to a large vocabulary ASR system for automatic broadcast news transcriptions. We also analyze in details the RMI combination coefficients and propose a dominant model selection method. Then, we present the notion of evolutional subspace which extends RMI to carry out incremental adaptation. Finally, experimental results are reported.

2. REFERENCE MODEL INTERPOLATION

2.1. Introduction

The basic idea behind the RMI adaptation technique is that an adapted model for a new speaker or acoustic condition is defined as a linear interpolation of a set of SD models with an optional translation by a bias vector. Assuming a combination of K SD reference models, a Gaussian mean vector of the adapted model has the form

$$\mu^{RMI} = \sum_{k=1}^{K} w_k \mu_k + b$$

where μ_k is the mean vector of the k^{th} reference model, w_k *is* the k^{th} combination coefficient and $b = [b_1 \cdots b_D]$ is the global bias vector (where D is the dimension of acoustic feature vectors). The matrix of the mean vectors of the K SD models $[\mu_1^T \cdots \mu_K^T]$ forms a reference model subspace (RMS) in the acoustic model space, and the combination coefficients $[w_1 \cdots w_K]$ are the coordinates of the adapted model in this subspace. The global bias vector, which acts as a compacted bias model, performs a translation of the adapted model. We use the EM algorithm to estimate the combination coefficients and the bias vector by maximizing the likelihood of adaptation data o_t.

2.2. *A posteriori* selection of reference models

RMI enables the *a posteriori* selection of reference models due to the simple form of the adaptation scheme. We use the coordinates (estimates of the combination coefficients) as selection indices to dynamically select a pertinent subset of reference models for each test speaker or utterance, because these coordinates directly represent the importance of each reference model in the adapted model. One iteration is firstly performed to estimate all the K combination coefficients $[w_1 \cdots w_K]$ using the current test data. Then the N ($N < K$) reference models with the largest combination coefficients (in absolute values) are selected to form the reference model subspace for RMI adaptation.

2.3. Relation to model selection adaptation

Model selection is commonly used in today's LVCSR systems (*e.g.* system with gender-dependent models). The idea is to select, at runtime, one of the pre-prepared acoustic models according to some criteria (*e.g.* maximum likelihood), and to decode with the selected model. However, the major drawback is that if none of the pre-prepared acoustic models fits well the test data, performance cannot be improved, or can even be degraded.

We can see that model selection is a special case of the *a posteriori* selection of reference models where only one model is selected. We have observed that when a test utterance fits well one of the reference models, the combination coefficient associated with this reference model becomes dominant among all the combination coefficients, as shown in Figure 1. It is thus very easy to identify the dominant reference model using a threshold. By selecting a dominant reference model, the selection of reference models falls into the framework of model selection. However, contrary to model selection, performance can still be improved if no dominant speaker is found.

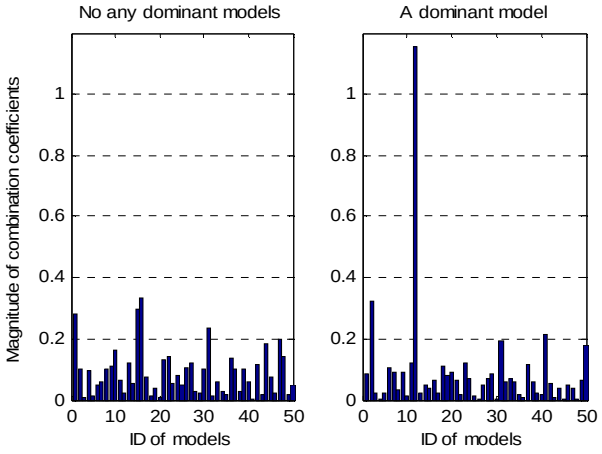

Figure 1 Magnitudes of the combination coefficients for all the reference models: (left) none of reference models fits well the test data (no dominant speaker); (right) one model fits well the data

This automatic switch between dominant model selection and model combination is in particular useful for automatic broadcast news transcription applications in which a significant amount of speech is uttered by journalists whose voices can be known to the system. We report the performance of the dominant selection on ESTER in Section 4. In the next section, we present another method of variable subspaces by updating rather than selecting reference models.

3. EVOLUTIONAL REFERENCE MODEL SUBSPACE

The idea behind the evolutional reference model subspace is to slowly update the reference models using the increasing amount of adaptation data, and then to regularly apply RMI adaptation using the updated subspaces. We posit that the target model is more likely to be found in the subspace which is "close" to the adaptation data. As illustrated in Figure 2, the reference models which are used to form the subspace can evolve into the so-called evolutional subspace. The resulting subspaces are updated towards the region of the model space which is close to the adaptation data; RMI adaptation is then applied to locate the target model in the updated subspaces.

In this way, the evolutional subspace combines rapid adaptation using RMI with slow adaptation. This allows us to achieve incremental adaptation to improve adaptation performance with the increase in the amount of data. The implementation of the evolutional subspace scenario is described in Figure 3. The input speech utterance is firstly decoded using speaker independent models. The speech utterance, together with the recognized transcription, is then used to update the reference models by MAP or MLLR [4]. Once the reference models are updated, RMI uses the updated mean vectors of the reference models and the speech data together with the recognized transcriptions to perform adaptation on the current test utterance.

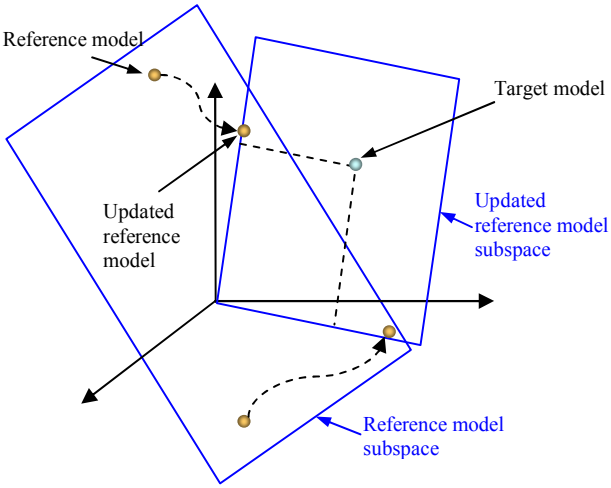

Figure 2 Illustration of the evolutional reference model subspace

Note that RMI relies on the correspondence among the Gaussian components from the different reference models. In order to keep this correspondence with MLLR update of reference models, we should use a one-class MLLR adaptation or apply the RMI adaptation separately for each MLLR class. We use one-class MLLR in this work to simplify the implementation of the algorithm.

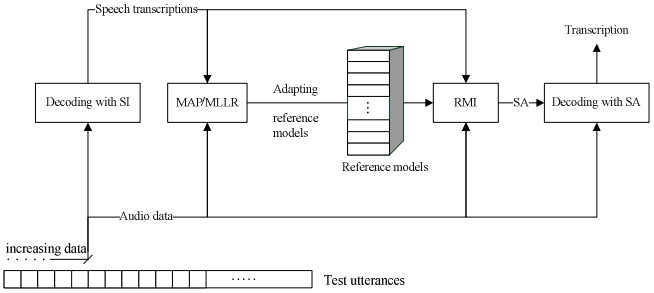

Figure 3 Incremental adaptation system using RMI with the evolutional reference model subspace

4. EXPERIMENTS

4.1. Experimental setup and reference models creation

Experiments are carried out using the IRENE automatic transcription system initially developed by IRISA and ENST for the ESTER automatic transcription evaluations [5]. IRENE is a multi-pass system in which an input broadcast news show is firstly divided into segments of a few seconds of speech. Speaker clustering is then applied to group speech segments into speaker clusters, which can be used for speaker adaptation purposes. The speech segments are then decoded using context-independent acoustic models and a trigram language model. This first pass generates a word graph on which rescoring is performed using a quadrigram language model and context-dependent tied-state triphones.

The experiments were conducted on the ESTER I phase 2 corpus, which includes material broadcasted by different French radio stations in 2003. A large part of the speech was uttered by a few journalists who were reporting headlines, brief news or short stories. Nevertheless, there are also some interviewed people with more spontaneous speaking styles. The 50 speakers with the largest amount of training speech data are selected to create reference models. Each speaker has between ten minutes and one hour and a half of speech. The 50 reference models are generated using MAP adaptation of SI models, adapting only the mean vectors. According to the adaptation *scenarios* (instantaneous or incremental), the use of the test data can be different. We thus describe the profiles of the test and adaptation data in each subsection.

4.2. Instantaneous adaptation

In the instantaneous adaptation scenario, we perform RMI adaptation utterance by utterance on about 5 hours of materials. Unsupervised adaptation is performed respectively on the first and second pass to adapt context-independent phoneme (monophone) and context-dependent phoneme (triphone) models.

As presented above, the ESTER corpus contains a large part of speech uttered by a number of principal speakers (journalists). Thus, the test speech data also contains some principal speakers whose data have been used to create the reference models. Hence, results are reported for two groups of speakers. Table 1 shows the percentage of these principal speakers who are known to the system.

MLLR adaptation is carried out for comparison. We use MLLR with three block-diagonal transformations and a bias vector in the experiments because it gave the best performance level compared to other MLLR variations in our experiments.

	Num. of speakers	Num. of words	Word percentage
Principal speakers	20	21,062	46.23%
Other speakers	141	24,502	53.77%

Table 1 Percentage of principal speakers in the evaluation data

4.2.1. Performance on the first pass

Table 2 lists the performance of the utterance-level adaptation on monophone models. RMI with dominant selection is also performed by applying a high threshold (0.9) on the estimated combination coefficients to select dominant reference models for current test utterance.

	Principal speakers	Other speakers	Overall
SI	23.6%	37.1%	30.9%
MLLR (3bd)	23.5%	36.7%	30.6%
RMI	20.0%	35.2%	28.2%
RMI (dominant selection)	19.8%	35.2%	28.1%

Table 2 Adaptation performance in word error rate on monophone model at the 1st pass

These results show that significant improvements can be achieved by the RMI methods at the end of the first pass. The RMI utterance adaptation gives improvements for both groups. For principal speakers, it has a 3.6% absolute improvement. More importantly, for the others speakers who are unknown to the system, RMI can still achieve an absolute improvement of 1.9%.

Usually, the performance on the first pass is measured by lattice word error rate which stands for the minimum word error count of any path through a word graph. Table 3 lists the lattice word error rates along with the sizes of word graphs.

	Lattice word error rate	Size of word graph	
		avg. # nodes / frame	avg. # arcs / frame
SI	11.0%	8.41	56.25
MLLR (3bd)	12.0%	6.45	38.77
RMI	10.8%	5.74	34.34

Table 3 Lattice word error rates at the end of the 1st pass against sizes of word graphs

We can see that the quality of the word graphs is improved after RMI adaptation: lattice word errors are reduced even though the graphs are much smaller.

4.2.2. Performance on the second pass

Table 4 shows the performance of RMI adaptation applied on both monophone and triphone models. The final performance has a 1.3% absolute improvement in word error rate compared to the SI system. This is better than the improvement of 0.6% obtained by performing adaptation solely at the second pass. MLLR adaptation on both passes degrades the performance. It may be due to the fact that the lattice word error rates at the end of the first pass are

degraded. It is not very surprising because the classic MLLR technique is not designed for instantaneous utterance adaptation.

	Principal speakers	Other speakers	Overall
SI	16.8%	29.4%	23.6%
MLLR (3bd)	17.5%	29.6%	24.0%
RMI	15.5%	28.2%	22.3%

Table 4 Adaptation performance on both monophone and triphone model (word error rates at the end of the 2nd pass)

4.3. Incremental adaptation

Incremental adaptation is applied on the first pass. In order to gather enough test utterances for each test speaker, we have selected the speakers who uttered the most utterances. The 7 speakers, other than the 50 speakers used for creating reference models, with more than 60 utterances are selected for the test. Each test speaker has an average of 68 utterances.

During the adaptation for each test speaker, RMI adaptation is performed on each test utterance using incremental speech data. In the meantime, the MLLR adaptation is triggered each ten utterances to update the reference models.

MLLR speaker adaptation and RMI instantaneous adaptation are performed for comparison. For MLLR adaptation, one global transformation with bias is used. For RMI instantaneous adaptation, utterance level rather than speaker level adaptation is performed on each of the test utterances.

4.3.1. Adaptation performance

Figure 4 reports the incremental adaptation performance. Note that the SI performance in Figure 4 is much better than the one in Table 2. This is because only 7 selected speakers are used in this test. The performance of RMI performed with accumulated adaptation data but without the regular update of reference models (RMI incr.) is also shown for comparison.

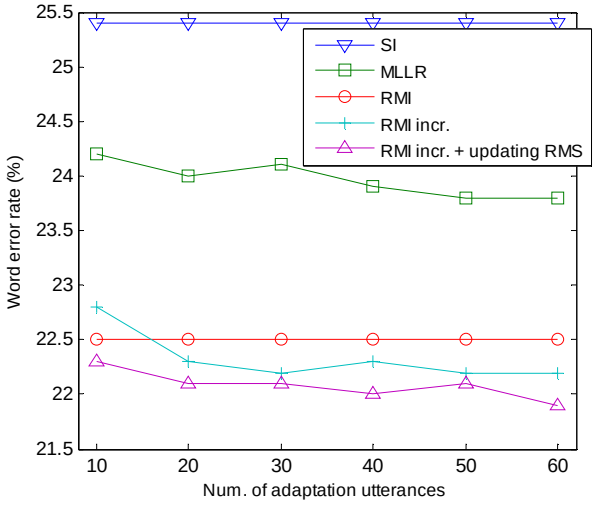

Figure 4 Incremental adaptation performance

At first glance, all the adaptation strategies outperform the SI performance. Instantaneous RMI adaptation already achieves an absolute improvement of 3% in word error rate. Incremental RMI without updating the reference models exhibits a slightly worse performance level than instantaneous RMI adaptation at the beginning, when only 10 adaptation utterances are used. With the increase in the number of test utterances, incremental RMI outperforms instantaneous RMI adaptation. However, performance is not much improved after 20 utterances. There is only a 0.1% improvement between 20 utterances (22.3%) and 60 utterances (22.2%). Updating the reference models further improves incremental RMI adaptation. By updating the reference models, the performance of the incremental RMI is 22.3% when 10 utterances are used. From 20 utterances to 60 utterances, *RMI incr.+updating RMS* has a 0.2% improvement, which is actually the same amount of improvement achieved by MLLR adaptation.

5. CONCLUSION

We have applied RMI adaptation in a practical LVCSR system, and have illustrated instantaneous and incremental adaptation implemented by RMI. The experimental results show that RMI can achieve utterance by utterance instantaneous adaptation in a large vocabulary task. This is useful when the grouping of speech segments is difficult, *e.g.* in streaming mode, or when speakers are unknown to the system such as online telephone services. In addition, the dominant model selection method based on the *a posteriori* selection in RMI is shown to be useful for the applications in which there are some principal speakers who can be known to systems, *e.g.* journalists in broadcast news. We have also presented the idea of evolving subspaces by updating at runtime the reference models. The incremental adaptation system implemented using RMI with the update of reference models is presented and evaluated. Experiments show that incremental RMI with the update of reference models outperforms instantaneous RMI adaptation and can further improve the performance with the increase of adaptation data. Besides, we have observed recently that the bias model of RMI has an important effect on adaptation performance. The RMI incremental adaptation may also be achieved by using an evolutional bias model. This point will be studied in future works.

6. REFERENCES

[1] W. X. Teng, G. Gravier, F. Bimbot, F. Soufflet, “Rapid Speaker Adaptation by Reference Model Interpolation”, INTERSPEECH, 2007

[2] R. Kuhn, P. Nguyen, J.-C. Junqua, L. Goldwasser, N. Niedzielski, S. Fincke, K. Field, M. Contolini, “Eigenvoices for speaker adaptation”, ICSLP, 1998

[3] T. J. Hazen, J. R. Glass, “A comparison of novel techniques for instantaneous speaker adaptation”, EUROSPEECH, 1997

[4] C. J. Leggetter, P. C. Woodland, “Flexible speaker adaptation for large vocabulary speech recognition”, EUROSPEECH, 1995

[5] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, K. Choukri. “Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news”, LREC, 2006