CROSSLINGUAL ADAPTATION OF SEMI-CONTINUOUS HMMS USING ACOUSTIC SUB-SIMPLEX PROJECTION

Frank Diehl, Asunción Moreno, Enric Monte

TALP Research Center
Universitat Politècnica de Catalunya (UPC)
Jordi Girona 1-3, 08034 Barcelona, Spain
{frank,asuncion,enric}@gps.tsc.upc.edu

ABSTRACT

With the demand on providing automatic speech recognition (ASR) systems for many markets the question of porting an ASR system to a new language is of practical interest. Transferring already existing hidden Markov models (HMM) from a source to the target language is seen as a key step to cope with this task. Typically, such a crosslingual model adaptation task consists of a three step procedure. It starts by polyphone decision tree specialisation (PDTS), specialising the phonetic-acoustic decision tree of the source models to the target language. In a second step initial target language models are predicted out of the adjusted decision tree. Finally, the predicted acoustic models are adapted to the target language using a limited amount of target data.
In this work we focus on the final model adaptation step in the case of a system architecture employing semi-continuous HMMs (SCHMM). In contrast to continuous density HMMs (CDHMM), adaptation techniques for SCHMMs are not as well developed. In particular, no powerful transformation based adaptation method for adjusting the information bearing mixture weights of the common prototype densities is on-hand. To overcome this problem we introduce a novel adaptation scheme for SCHMM. The method relies on the projection of retrained model parameters to a solution sub-simplex which is obtained through acoustic regression classes derived from the decision tree of the source models. The performance of the procedure is demonstrated by the transfer of multilingual Spanish-English-German models to Slovenian and to French. In the full paper, reference results for a standard maximum likelihood linear regression (MLLR) approach are given too.

1. INTRODUCTION

With the demand on providing automatic speech recognition (ASR) systems for many markets the question of porting an ASR system to a new language is of practical interest. However, a common trouble, developers working in this field are faced with, is the availability of adequate speech material to train the acoustic models. To alleviate this problem crosslingual speech recognition became an active research area. Instead of relying on a complete speech database in the target language one tries to manage with less target material by transforming existing acoustic models of a source language to the target language [1], [2], [3].
In the work on hand we follow the approach which was pointed out in [3]. It consists in a three step procedure using a limited amount of target data to convert acoustic models from the source language to the target language. The method starts by polyphone decision tree specialisation (PDTS) [4], specialising the phonetic-acoustic decision tree of the source models to the target language. In a second step initial target language models are predicted out of the adjusted decision tree. Finally, the predicted acoustic models are adapted to the target language using a limited amount of target data.
Until now almost all results reported on language adaptation of hidden Markov models (HMM) refer to the use of continuous density HMMs (CDHMM). However, semi-continuous HMMs (SCHMM) are still widely-used. Their lower complexity paired with high performance make them an attractive alternative to CDHMMs especially for small and medium scale systems. Unfortunately, adaptation techniques for SCHMMs are not as well-developed as in the case of CDHMMs. Besides MAP adaptation proposed in [5], a powerful transformation based method is still not established. Of course, MLLR [6] may be applied. However, its use is of limited effect due to SCHMMs rely on one common codebook. Applying regression class specific MLLR transformations for different model groups is impossible. Only common transformations are feasible weakening the strength of MLLR significantly.
The basic problem when adapting SCHMMs lies in the way they code the acoustic information. The probability density functions of the individual states are modelled as superpositions of prototype densities taken from a common codebook. Each state contains a probability vector of mixture weights referring to the prototype densities. Hence, adapting SCHMMs means adapting these vectors under standard probabilistic constraints hampering the design of an appropriate method significantly.
Recently two methods were introduced to overcome this problem. In [7] it was proposed to decompose the B-matrix – i.e. the weights matrix composed of the weights vectors of the states of all HMMs – by probabilistic latent semantic analysis (PLSA) [8]. The adaptation procedure is then reduced to a subset of the resulting component matrices. For solving the final optimisation problem expectation maximisation is used, implicitly satisfying the probabilistic constraints on the weights.
A second solution to the problem was presented in [9]. The method tries to identify reasonable solution sub-simplexes according to a regression scheme using the underlying phonetic-acoustic decision tree of the source models. Having defined these solution spaces, so called measurements, actually the source models retrained by the adaptation data, are projected onto the solution spaces resulting in the final adapted models.

This work was granted by the CICYT under contract TIC2002-04447-C02.

In this work we pick up the second method. First we recall its derivation and emphasise its maximum likelihood character justifying the name, maximum likelihood convex regression (MLCR), of the method. In a second step the method is extended by introducing prior information to take better advantage of the adaptation data. The resulting method is called maximum a posteriori convex regression (MAPCR). Finally, the performance of the two methods is demonstrated for a crosslingual model adaptation task with and without the use of PDTS.

2. THE DATA MODEL

In a SCHMM speech recognition system the output densities $p(x|s)$ associated to the states $s \in \{1, ..., S\}$ are expressed as superpositions of prototype densities $G_k(x)$ taken from a common codebook

$$p(x|s) = \sum_{k=1}^{K} c_{sk} G_k(x) \qquad s \in \{1, ..., S\} \tag{1}$$

with $k \in \{1, ..., K\}$ naming the prototypes. For calculating an adapted version $\underline{\overline{c}}_s = [\overline{c}_{s1}, ..., \overline{c}_{sK}]^T$ of the mixture weights $\underline{c}_s = [c_{s1}, ..., c_{sK}]^T$ of state s we model them as a convex combination

$$\underline{\overline{c}}_s = \underline{U}_s \underline{\alpha}_s \tag{2}$$

of a set of L prototype weights vectors $\underline{u}_{sl}$ forming matrix $\underline{U}_s$. I.e., matrix $\underline{U}_s$ models our belief of the acoustic neighbourhood of $\underline{c}_s$. The L-dimensional vector $\underline{\alpha}_s$ represents the combination weights which need to be estimated under standard probabilistic constraints. As indicated by the subscript s the $\underline{u}_{sl}$ and the $\underline{\alpha}_s$ depend on the current state s. We set $L << K$ to get the desired reduction in the number of free parameters.

3. MAXIMUM LIKELIHOOD CONVEX REGRESSION

For estimating the mixture weights $\overline{c}_{sk}$ under the Baum-Welsh reestimation framework one finds [6] the well known auxiliary function

$$Q(\lambda, \overline{c}_{sk}) = \sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{sk}(t) \log(\overline{c}_{sk}) \tag{3}$$

which has to be maximised with respect to the $\overline{c}_{sk}$ subject to

$$\sum_{k=1}^{K} \overline{c}_{sk} = 1 \qquad \forall s \in \{1, .., S\},$$
$$\overline{c}_{sk} \geq 0 \qquad \forall s \in \{1, ..., S\} \wedge \forall k \in \{1, .., K\}.$$

In (3) the term $\gamma_{sk}(t)$ defines the occupation probability of mixture component k of state s at time t given the observation sequence $O = o_1 ... o_T$ and the current parametrisation λ. $\overline{c}_{sk}$ is the objective to estimate, namely the mixture weight for the k^{th} prototype density of state s. In a standard Baum-Welsh training the solution for the mixture weights is given by

$$c_{sk} = \frac{\sum_{t=1}^{T} \gamma_{sk}(t)}{\sum_{t=1}^{T} \sum_{k=1}^{K} \gamma_{sk}(t)} \tag{4}$$

[10]. Dividing (3) by the denominator of (4) we get the modified auxiliary function

$$\tilde{Q}(c_{sk}, \overline{c}_{sk}) = \sum_{k=1}^{K} c_{sk} \log(\overline{c}_{sk}). \tag{5}$$

In (5) we have replaced the general placeholder λ by c_{sk} too. Expressing (5) in vector notation and plugging in (2) the auxiliary function changes to

$$\tilde{Q}(\underline{c}_s, \underline{\overline{c}}_s) = \underline{c}_s^T \log(\underline{U}_s \underline{\alpha}_s). \tag{6}$$

The interpretation of (6) is as follows. The $\underline{c}_s$ correspond to measurements taken from the adaptation data. The $\underline{\alpha}_s$ are the parameters to estimate, and $\underline{U}_s$ represents the model constraint in form of a solution sub-simplex which is based on prior knowledge taken from the underlying source models.
In light of (6) the final optimisation problem is stated as

$$\arg\min_{\underline{\alpha}_s} -\underline{c}_s^T \log(\underline{U}_s \underline{\alpha}_s) \tag{7}$$

$$\begin{aligned} &subject\ to && \sum_{l=1}^{L} \alpha_{sl} = 1 \\ &and && \alpha_{sl} \geq 0 \qquad \forall l \in \{1, .., L\} \end{aligned}$$

which need to be solved for each $s \in \{1, ..., S\}$. Though we found no closed form solution for the problem, it is identified as convex and can be solved by convex optimisation [11].
As depicted in Fig. 1, solving problem (7) consists in projecting the measurement $\underline{c}_s$ to the probabilistic sub-simplex spanned by the convex combination of $\underline{U}_s$, minimising the distance, i.e. the cross-entropy, between the measurement $\underline{c}_s$ and the solution $\underline{\overline{c}}_s$.

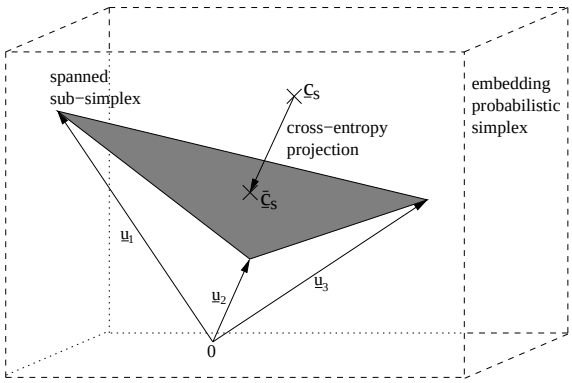

Fig. 1. *Graphical interpretation of the optimisation problem.*

It is illuminating to remember that the c_{sk} can be interpreted as normalised counts [10]

$$c_{sk} = \frac{n_{sk}}{n_s}, \tag{8}$$

with n_{sk} denoting the expected number of times in state s and prototype density k is active, and n_s the expected number of times in state s. Plugging (8) into (5) and removing the state count n_s, not affect the maximisation of (5), we get the log-likelihood expression

$$\mathcal{L}(\underline{\alpha}_s) = \sum_{k=1}^{K} n_{sk} \log(\overline{c}_{sk}) \tag{9}$$
$$= \log \prod_{k=1}^{K} \overline{c}_{sk}^{\,n_{sk}} \tag{10}$$
$$= \log \prod_{k=1}^{K} \overline{c}_{sk}^{\,c_{sk} n_s}. \tag{11}$$

Defining an observation as "a feature is caused by prototype density k given state s", and remembering that c_{sk} is actually defined as the probability of such an observation, we see that minimising expression (7) is equivalent to maximising the likelihood of the sequence of these observations as given by (10). The event counts n_{sk} or c_{sk}, respectively, provide the necessary statistics and are estimated by retraining the source models with the adaptation data of the target language justifying the notation *"measurement"* for the c_{sk}.

4. TARGET MODEL PREDICTION AND ACOUSTIC REGRESSION CLASSES

Crosslingual model adaptation starts with the prediction of suitable target models out of the decision tree of the source language. In our system we use a phonetic-acoustic decision tree for state tying. It is constructed during the training of the source models constituting a function from a generic phonetic feature space to a state space. The input domain consists of phonetic feature vectors assigned to the central phone and the phonetic contexts of a state. The features are generic, i.e. to a large degree independent of the used language. An example might be $(plosive, bilabial, voiced)$. The output domain holds the weights vectors of the states.

With the input domain being of generic nature the tree can also be used to predict the tied states of the models of a new language. After setting up the feature vectors for the new language one calls the tree applying the features. Afterwards the predicted weights vectors are assigned to the models of the target language. This procedure effectively defines the target models $\underline{c}_s$ and initialises their training. In a further step the decision tree is also used to define the solution sub-simplexes $\underline{U}_s$ by exploiting the acoustic neighbourhood knowledge given by neighbouring leaves of the tree. Acoustic regression classes are defined by cutting the tree above its leaves constructing a set of subtrees. In Fig. 2 we depict this situation. It shows a

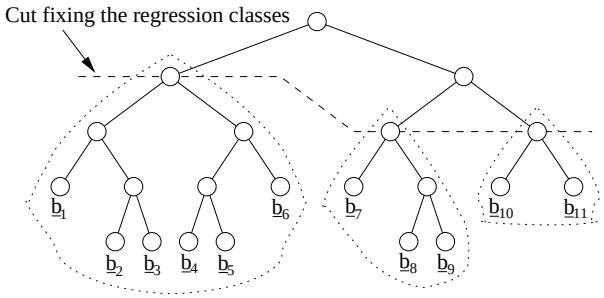

Fig. 2. *Sub-simplex construction.*

fictitious decision tree which is cut by the dashed line. It results in three subtrees with 6, 3, and 2 leaves and the corresponding mixture weights of the base states $\{\underline{b}_1, \underline{b}_2, \underline{b}_3, \underline{b}_4, \underline{b}_5, \underline{b}_6\}$, $\{\underline{b}_7, \underline{b}_8, \underline{b}_9\}$ and $\{\underline{b}_{10}, \underline{b}_{11}\}$. The cut is accomplished searching the complete tree for the subtree giving minimal accumulated model entropy. The search, starting from the tree's root, stops after having reached the number of desired regression classes, i.e. nodes.

All leaves of a subtree share one or more common phonetic features or, in other words, are to some degree acoustically similar. In consequence, the mixture weights vectors $\underline{b}_i$ associated to the leaves of a subtree form a more or less consistent acoustic sub-simplex. Furthermore, for the adaptation task we assume that the base models we use for initialisation are already close to the target models. From these considerations we claim that each target model should lie within the space spanned by its initialisation state and its neighbouring states occupying the same subtree.

Although each set of base states $\underline{b}_i$ already constitutes a possible solution space for a $\underline{\bar{c}}_s$ we do not stack them directly to $\underline{U}_s$. We actually prefer to have the same number of base vectors for each $\underline{U}_s$, and we would like to reduce stochastic dependencies within $\underline{U}_s$. To cope with these demands we use PLSA to retrieve the $\underline{U}_s$ out of the $\underline{b}_i$ assigned to a regression class.

5. PROBABILISTIC LATENT SEMANTIC ANALYSIS

The starting point for PLSA is the so called aspect model [8]. For a pair of random variables $(x, y) \in (X, Y) = \{(x_k, y_s)|1 \leq k \leq K \wedge 1 \leq s \leq S\}$ an underlying production mechanism involving a hidden variable $z \in Z = \{z_1, .., z_L\}$ is assumed. The hidden variable is called factor or latent variable, and the production mechanism assumes conditional independence between X and Y given the latent variable Z. With this model the joint probability $P(x, y)$ is expressed as

$$P(x, y) = \sum_{z \in Z} P(x|z)P(y|z)P(z). \tag{12}$$

The parameters of (12) are estimated using the EM-algorithm as described in [8]. With the terms $\underline{P} = [P(x_k, y_s)]_{k,s}$, $\underline{\tilde{U}} = [P(x_k|z_l)]_{k,l}$, $\underline{\tilde{\Sigma}} = diag[P(z_l)]_l$, and $\underline{\tilde{V}} = [P(y_s|z_l)]_{s,l}$, (12) can be arranged in matrix form as

$$\underline{P} = \underline{\tilde{U}}\underline{\tilde{\Sigma}}\underline{\tilde{V}}^T. \tag{13}$$

Equation (13) shows some formal similarities to singular value decomposition (SVD). The matrices $\underline{\tilde{U}}$ and $\underline{\tilde{V}}$ constitute bases for the X- and Y-spaces. The L elements of matrix $\underline{\tilde{\Sigma}}$ act similar to singular values, controlling the relative strength of the base vectors when mixing them together to matrix $\underline{P}$.

But there are also some significant differences with respect to a SVD. Besides being solely a model provided with a model error the mixture approximation gives well defined probabilistic distributions. I.e. the column vectors of $\underline{\tilde{U}}$ and $\underline{\tilde{V}}$ as well as the diagonal of $\underline{\tilde{\Sigma}}$ accomplish all probabilistic constrains to be interpreted as discrete distribution. As a consequence, the base vectors given by $\underline{\tilde{U}}$ form a probabilistic sub-simplex instead of a subspace as in the case of a SVD. Furthermore, we are free to choose the model complexity. That means especially that we can choose the number of latent variables L for controlling the size of $\underline{\tilde{U}}$.

In light of our search for suitable basis vectors $\underline{u}_{sl}$, we use PLSA as follows. After having identified the neighbouring base vectors for a specific state s they are stacked together forming matrix $\underline{B}_s$. E.g. for the first subtree shown in Fig. 2 we get

$$\underline{B}_s = [\underline{b}_1, \underline{b}_2, \underline{b}_3, \underline{b}_4, \underline{b}_5, \underline{b}_6]. \tag{14}$$

The components of matrix $\underline{B}_s$ stand for the conditional probabilities $P(k|s, T_s)$, i.e. the probability of mixture density k given state s and subtree T_s. To get the joint probabilities $P(k, s|T_s)$ we multiply the columns of $\underline{B}_s$ by the $P(s|T_s)$, the occupation probabilities of the states conditioned on the current subtree T_s. Usually, the $P(s|T_s)$ can easily be derived during the training of the base models. Finally, after having fixed the model complexity by the number of latent variables L, the resulting matrix $\underline{B}_s$ is decomposed by PLSA providing $\underline{\tilde{U}}$ being actually the desired base $\underline{U}_s$ in (6).

6. MAXIMUM A POSTERIORI CONVEX REGRESSION

A critical point of the ML solution is the definition of the solution sub-simplex $\underline{U}_s$. It embodies our expectation that a good solution is more likely to lie close to $\underline{U}_s$ than to $\underline{c}_s$. Though this might be a reasonable assumption in case of little adaptation data, it loses its justification as more data becomes available. If we had plenty of data we would anticipate that the solution for the adaptation problem is $\underline{c}_s$ itself. A natural way to take this relationship into consideration is the introduction of prior information about our confidence in the solution sub-simplex $\underline{U}_s$, i.e. extending the ML approach to a MAP approach.
For a MAP formulation of the problem we start by including $\underline{c}_s$ into the solution sub-simplex by extending $\underline{U}_s$ by $\underline{c}_s$, demanding the extension of $\underline{\alpha}_s$ by α_{sL+1}, too. I.e. the data model changes to

$$\overline{\underline{c}}_s = [\underline{U}_s, \underline{c}_s]\,[\alpha_{s1}, ..., \alpha_{sL}, \alpha_{sL+1}]^T. \tag{15}$$

In case of the prior distribution $p(\underline{\alpha}_s)$ we are actually only interested in weighting the solutions between $\underline{U}_s$ and $\underline{c}_s$. In practise this means that we can set $p(\underline{\alpha}_s) = p(\alpha_{sL+1})$, being equivalent to a non-informative, uniform prior for the original αs.
Hence, starting from (11), the objective function $\mathcal{M}(\alpha)$ to maximise can be stated as

$$\mathcal{M}(\alpha) = p(\alpha_{sL+1}) \prod_{k=1}^{K} \overline{c}_{sk}{}^{c_{sk} n_s} \tag{16}$$

where $\overline{c}_{sk}$ refers to the extended data model of (15). The prior is chosen in an ad hoc manner as a gamma distribution

$$p(\alpha_{sL+1}) = C \alpha_{sL+1}^{\mu} exp(-\eta \alpha_{sL+1}) \tag{17}$$

with $\alpha_{sL+1} \in [0, 1]$, $\mu, \eta \geq 0$ and C a suitable normalisation constant. Parameter μ and η serve to control the shape of the prior. For η close to zero the prior gets uniform expressing our uncertainty respective $\underline{U}_s$. But, as η gets bigger the prior changes its shape to a peak concentrating its probability mass near zero. In this case α_{sL+1} is forced to be small, reflecting our suspicion regarding the measurement $\underline{c}_s$. The value of μ is of minor importance especially if η gets big. Using vector notation and definition (17), the final optimisation problem according to (16) becomes

$$\begin{aligned} \arg\min_{\underline{\alpha}_s} \quad & -n_s \underline{c}_s^T \log \; [\underline{U}_s, \underline{c}_s]\,[\alpha_{s1}, ..., \alpha_{sL+1}]^T \\ & -\mu \log \alpha_{sL+1} + \eta \alpha_{sL+1} \end{aligned} \tag{18}$$

$$\begin{aligned} subject\ to \quad & \sum_{l=1}^{L+1} \alpha_{sl} = 1 \\ and \quad & \alpha_{sl} \geq 0 \qquad \forall l \in \{1, .., L+1\} \end{aligned}$$

which is convex, i.e. can also be solved by convex optimisation.
A key role in the interpretation of (18) inheres in the state count n_s. It is the factor balancing the information reflected by the measurement and the prior. As a reliable measurement comes with a high state count, a high n_s gives emphasise to the likelihood part of (18) shifting the solution to $\underline{c}_s$. On the other hand, if n_s is small the prior terms will dominate. This forces α_{sL+1} to be close to zero leading to a solution close to the original not extended sub-simplex.

7. POLYPHONE DECISION TREE SPECIALISATION

In [4] it was pointed out that the different phonotactic of languages significantly hamper the use of context dependent acoustic models in crosslingual speech recognition. A phonetic context which appears frequently in one language may be rare in another language or might even not exist. When predicting target models out of the decisions tree of the source language one is therefore confronted by the problem that some models which are needed to model adequately the target language can not be predicted. Instead, one assigns acoustically significant different target models to the same tree leaves. This results in immoderate broad models unable to model adequately the acoustic properties of the target language.
To cope with this problem the authors of [4] proposed PDTS. PDTS consists of the crosslingual adaptation of a phonetic-acoustic decision tree to a target language. One restarts the tree growing process of the source tree, using some adaptation data of the target language. By this, one introduces phonetic context information into the decision tree which is not present in the source language but is important for the target language.
In the present work and in the light of MAPCR, PDTS is applied as follows. For a given source tree the tree growing process is restarted applying the adaptation data of the target language. Afterwards, the models associated to the new leaves are trained by one iteration Baum-Welsh training on the adaptation data. The resulting models are the measurements $\underline{c}_s$. Finally, MAPCR is applied for each measurement $\underline{c}_s$. The only difference to the case of MAPCR without applying PDTS is the presence of more leaves i.e. measurement. In other words, we need to adapt more models.

8. SYSTEM OVERVIEW

We use a SCHMM system calculating every 10ms twelve mel-cepstrum coefficients (MFCC) (and the energy) using cepstral mean subtraction. First and second order differential MFCCs plus the differential energy are employed. For each stream a codebook is constructed consisting of 256 and 32 (delta energy) Gaussian mixtures, respectively. Hence, each stream is adapted, meaning that the adaptation procedure is run four times for each state.
The model topology consists of 3-state state-tied left-to-right demiphones. Demiphones [12] can be thought of as triphones which are cut in the middle giving a left and a right demiphone. For state tying we apply a binary decision tree to each state position but over all source phonemes resulting in six trees. Thus, beside context questions also questions respective the central phoneme of a model are asked. The phoneme sets for Slovenian and French consist of 47 and 43 phonemes, respectively. The questions for the decision tree are of phonetic character and are derived from the IPA-chart.

9. THE ADAPTATION TASK

Starting point for the crosslingual model adaptation is a set of multilingual speaker independent source models trained on Spanish, English and German data. Slovenian and French serve as target languages. For training and testing we apply SpeechDat-II fixed telephone databases. The multilingual base system is trained on phonetically rich sentences of 3000 speaker, 1000 from each language. For the crosslingual adaptation we use a model set comprising 1500 tied source states.
For both target languages two different sized adaptation set are used. One comprises 20 and the other 50 speakers. The adaptation sets are

balanced respective sex. They consist of phonetically rich sentences containing 170 and 426 sentences in case of Slovenian, and 169 and 422 sentences in case of French. The two independent test sets, 50 women and 50 men for each language, consist of phonetically rich words mixed with application words. They comprise 614 sentences for Slovenian and 670 sentences for French. The resulting grammar, just a word list, exhibit a word based perplexity of 372 and 445 for Slovenian and French, respectively.

10. EXPERIMENTS

The experiments carried out to test the proposed adaptation procedure group into the ones with and the ones without applying PDTS. Beside the reference results stemming from pure monolingual systems and systems with predicted but not adapted models, all other results divide in two groups. One for the small, 20 speaker, and one for the big, 50 speaker, adaptation set. Except for the monolingual reference results all other outcomes are based on the use of the same multilingual source tree comprising 1500 leaves, i.e. a model set of 1500 tied states. In case of MLCR and MAPCR the number of regression classes, i.e. sub-simplexes, was chosen to 100. PLSA was carried out on the $\underline{B}_s$, i.e. assuming equal probable source states. After some initial testing the PLSA-order was fixed to 25 for MLCR and to 10 for MAPCR. During these tests we also fixed the hyper parameters. Finally they were set to $\mu = 0.5$ and $\eta = 7$. All simulation results are reported by word error rates (WER). The corresponding confidence intervals range from by ca. $\pm 2\%$ WER up to ca. $\pm 3.5\%$ WER.

10.1. Experiments without applying PDTS

Tab. 1 summarise the results obtained for the tests without using PDTS. The MONO results refer to pure monolingual systems trained with 900 speakers. PRED and PRED-I1 state the results when directly using the predicted source models and after retraining them by one iteration Baum-Welsh training on the adaptation data. MLLR, MLCR and MAPCR give the WERs after applying the corresponding adaptation technique. In case of MLLR the PRED-I1 models serve to initialise the MLLR training. In case of MLCR and MAPCR the PRED-I1 models are the measurements as explained in section 3 and 6.

Table 1. *Tests not applying PDTS, WERs in* [%].

	Slovenian		French	
#Speaker	20	50	20	50
MONO	9.61		6.12	
PRED	50.49		45.37	
PRED-I1	26.71	20.68	27.91	22.84
MLLR	26.38	21.50	27.01	21.64
MLCR	32.41	32.08	31.19	31.79
MAPCR	20.03	18.89	22.84	19.40

When inspecting Tab. 1 one finds WERs of 50.49% and 45.37% for the not adapted, solely predicted models (PRED). Though these numbers are too bad for any reasonable application, they are in line with analogue experiments reported in [2], and [3]. Comparing the PRED with the PRED-I1 results it turns out that simple retraining of the predicted models by one iteration Baum-Welsh training on the adaptation data reduces the WER by ca. $40\% - 60\%$.

As expected and lined out in Section 1, MLLR does hardly help. Though, up to 1.20% improvement is obtained for French, in case of Slovenian yet degradation of 0.82% is observed. Hence, merely adapting the common codebook, as done by MLLR, is barely a good adaptation policy for SCHMMs.

Proceeding by inspecting the MLCR and MAPCR results one finds that MLCR significantly worsens the situation. On the other hand, MAPCR gives the best results, providing, respective the PRED-I1 case, improvements in WER of 5.07% and 6.68% for the small, and 1.79% and 1.69% for the big adaptation set. The results for MLCR and MAPCR show up, that directly assuming a good solution in the sub-simplexes defined by the predicted source models, as in the case of MLCR, is too simple. But, searching the solution between the MLCR solution space and the PRED-I1 models, as done by MAPCR, turns out to be very effective, resulting in the best adapted models. Comparing the best adapted models with the monolingual ones, one still finds a performance gap of ca. $9\% - 18\%$. We attribute this behaviour, at least partly, to the crosslingual phonetic context mismatch as lined out in Section 7.

10.2. Experiments applying PDTS

To investigate the influence of PDTS on the acoustic modelling a second test series was carried out. Tab. 2 summarises the corresponding test results. Beside the MAPCR results obtained without PDTS, results when using PDTS without adaptation, PDTS-5/10/15, and with adaptation, MAPCR-PDTS-5/10/15, are given. In fact, the PDTS-5/10/15 models are the measurements needed for MAPCR-PDTS-5/10/15, i.e. obtained by one iteration Baum-Welsh training on the adaptation data. In both cases the numbers 5/10/15 refer to the stopping criterion for extending the tree by PDTS. They give the minimum number of models which need to fall into a leaf. As smaller this number as more additional leaves and thus contexts are generated. But, as higher this number as more reliable are the new, retrained PDTS-5/10/15 models.

Table 2. *Tests applying PDTS, WERs in* [%].

	Slovenian		French	
#Speaker	20	50	20	50
MAPCR	20.03	18.89	22.84	19.40
PDTS-5	32.57	26.06	21.19	14.03
PDTS-10	26.71	20.36	19.40	12.39
PDTS-15	25.57	19.22	19.25	11.94
MAPCR-PDTS-5	28.50	23.94	18.21	14.33
MAPCR-PDTS-10	23.13	19.71	16.12	11.79
MAPCR-PDTS-15	21.01	18.40	16.27	11.79

We start our discussion of Tab. 2 with the French results. We find that PDTS always outperforms MAPCR, and, expect of one case, MAPCR-PDTS always outperforms PDTS. In case of MAPCR-PDTS the final WERs are given by 16.27% and 11.79%, corresponding to an absolute reduction of the WER of 6.55% and 7.61% respective simple MAPCR without PDTS. Interestingly, in case of the small adaptation set the performance gain stems to more or less equal parts form PDTS and MAPCR. In contrast, in case of the big adaptation set the gain stems nearly completely from PDTS. This observation underlines that adaptation is most efficient if little adaptation data is on hand. On the other hand, as more data is available as more effective becomes simple retraining of the models.

The results also indicate that robust measurements are favoured over an improved context modelling. Comparing the recognition performances of Tab. 2 with the corresponding tree sizes, i.e. number of leaves, see Tab. 3, it is obvious that smaller, more robust trees are

Table 3. *Number of leaves after applying PDTS.*

	Slovenian		French	
#States	1500/1017		1500/696	
#Speaker	20	50	20	50
PDTS-5	1884	2672	1890	2516
PDTS-10	1468	2118	1516	2112
PDTS-15	1260	1828	1315	1834

preferred over the big, highly specialised PDTS-5 trees.
Next we focus on Slovenian. When analysing Tab. 2 we conclude that PDTS fails completely in that case. Running PDTS and retraining the models always results in models performing significantly worse than the MAPCR models. Though, MAPCR is able to remedy this outcome to some extend, a significant improvement over the MAPCR case is never obtained.
It is not the first time that such behaviour is reported for PDTS. In [13] the authors describe their attempt to port English and Spanish models to Indonesian. Though they do not report degradation, PDTS did not lead to any improvement. As possible explanation the authors doubt on the way PDTS adapted a decision tree. It is argued that the most important splits of a decision tree happen near to the tree's root, whereas PDTS merely results in a refinement of the leaves.
Also we believe that the bad PDTS performance for Slovenian is related to the fact that PDTS develops on a decision tree which was initially constructed for a different language. Considering that Slovenian, as a Slavic language, belongs to a different language group than any of the source languages whereas French as well as Spanish are Romanic, one would expect that the source tree matches better the French than the Slovenian acoustics. Though this is confirmed by the PRED results of Tab. 1, the #States-line of Tab. 3 indicates the opposite. With Slovenian using 1017 of the original 1500 leaves but French just 696 it looks like Slovenian takes better advantage of the source tree than French. This antagonism may be interpreted in such a manner that predicting initial Slovenian models out of the decision tree yet consumes useless-proven questions without improving the system performance. On the other hand, the wasted questions are missing during PDTS resulting in a badly adapted tree. We therefore conclude that PDTS has to be applied with caution. It is not guaranteed that it improves crosslingual model performances. Further research is necessary to understand its interaction with the decision tree it is based on.

11. SUMMARY

In this work we studied crosslingual acoustic model adaptation in the context of a speech recognition system applying SCHMMs. We lined out that a traditional technique as MLLR might be inadequate for this task and introduced MLCR and MAPCR as new adaptation techniques for SCHMMs. During evaluation both techniques were tested by converting multilingual Spanish-English-German models to Slovenian and to French. Analysing the test results MAPCR turned out to be the most efficient method making more effective use of limited adaptation data than the comparison methods.

The use of PDTS resulted ambivalent. Though greatly improving the French outcomes, significant degradations were observed for Slovenian. From where the bad PDST behaviour in the Slovenian case stems from remains an open question. Though we suspect an inadequate underlying decision tree, no clear final conclusion can be drawn.

ACKNOWLEDGMENTS

The authors would like to acknowledge Dr. Albino Nogueiras Rodríguez for his collaboration in the development and set up of the tools used in this work.

12. REFERENCES

[1] C. Nieuwoudt and E. C. Botha, "Cross-language use of acoustic information for automatic speech recognition," *Speech Communication*, vol. 38, no. 1, pp. 101–113, 2002.

[2] A. Zgank, Z. Kacic, and B. Horvat, "Comparison of acoustic adaptation methods in multilingual speech recognition environment," *International Conference On Text, Speech and Dialogue*, vol. 2807, no. 6, pp. 245–250, 11 2003.

[3] T. Schultz and A. Waibel, "Language portability in acoustic modeling," *Workshop On Multilingual Speech Communication*, pp. 59–64, 10 2000.

[4] T. Schultz and A. Waibel, "Language adaptive lvcsr through polyphone decision tree specialization," *Multi-Lingual Interoperability in Speech Technlogy*, pp. 97–102, 9 1999.

[5] Qiang Huo, Chorkin Chan, and Chin Hui Lee, "Bayesian adaptive learning of the parameters of hidden Markov model for speech recognition," *Transactions On Speech and Audio Processing*, vol. 3, no. 5, pp. 334–345, 9 1995.

[6] C. J. Leggetter and P. C. Woodland, "Speaker adaptation of HMMs using linear regression," Tech. Rep., Cambridge CB2 1PZ, 6 1994.

[7] A. Raux and R. Singh, "Maximum-likelihod adaptation of semi-continuous HMMs by latent variable decomposition of state distributions," *International Conference On Spoken Language Processing*, , no. 8, 10 2004.

[8] T. Hofmann, "Probabilistic latent semantic analysis," *Proceedings of Uncertainty in Artificial Intelligence*, 1999.

[9] F. Diehl, A. Moreno, and E. Monte, "Crosslingual adaptation of semi-continuous HMMs using acoustic regression classes and sub-simplex projection," *COST278 and ISCA Tutorial and Research Workshop (ITRW) On Applied Spoken Language Interaction in Distributed Environments*, 11 2005.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol. 77, no. 2, pp. 257–286, 2 1989.

[11] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge CB2 2RU, UK, 2004.

[12] J. B. Mariño, P. Pachès-Leal, and A. Nogueiras, "The demiphone versus the triphone in a decision-tree state-tying framework," *International Conference On Spoken Language Processing*, , no. 5, 11 1998.

[13] T. Martin and S. Sridharan, "Cross-language acoustic model refinement for the indonesian language," *International Conference On Acoustics, Speech, and Signal Processing*, vol. 1, pp. 865–868, 3 2005.