

CROSSLINGUAL ACOUSTIC MODEL DEVELOPMENT FOR AUTOMATIC SPEECH RECOGNITION

Frank Diehl, Asunción Moreno, and 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

In this work we discuss the development of two crosslingual acoustic model sets for automatic speech recognition (ASR). The starting point is a set of multilingual Spanish-English-German hidden Markov models (HMMs). The target languages are Slovenian and French. During the discussion the problem of defining a multilingual phoneme set and the associated dictionary mapping is considered. A method is described to circumvent related problems. The impact of the acoustic source models on the performance of the target systems is analyzed in detail. Several crosslingual defined target systems are built and compared to their monolingual counterparts. It is shown that crosslingual build acoustic models clearly outperform pure monolingual models if only a limited amount of target data is available.

Index Terms— crosslingual, acoustic modelling

1. INTRODUCTION

Enterprises engaged in ASR are usually faced with the question of globalizing their products. This does not only concern big international companies but also smaller business. Companies which are operating telephone assistant systems or automobile manufacturers demand from their suppliers system components that can be used worldwide. This may mean monolingual operability for multiple languages but also multilingual usability for multilingual markets or applications.

As state-of-the-art ASR technology greatly relies on the availability of adequate language resources, big efforts were undertaken to construct and distribute publically available speech and text databases. Although these efforts were highly successful in terms of covered languages and environmental conditions, companies are still faced with the problem of unavailable training data and the inflexible handling of new languages. A typical scenario is the demand to extend an ASR system to a minority language which is not yet covered by available databases, or, a speech database in the target language is available but does not match the environmental or dialectal conditions of the target application.

In this work we address the issue of porting an ASR system from one language to an other. We examine two target languages, Slovenian, and French, and assume that a limited amount of speech material in these target languages is available. The acoustic models of a multilingual Spanish-English-German system serve as a starting point. The chosen application scenario consists of a typical medium scale task, trying to recognize a list of so-called phonetically rich words, and application words. For the experiments, tied-mixture HMMs

are used, also reflecting the idea of a medium scale, or even embedded application.

2. BASIC CONCEPTS

With few exceptions, [1], recent work on crosslingual acoustic modelling assumes the availability of a certain, though limited, amount of speech material in the target language. Under the additional presumption that speech material and some well formed acoustic models of one or more source languages are available, three main research lines for crosslingual modelling can be identified. They are:

- Feature compensation
- Model combination
- Model adaptation

In feature compensation the focus lies directly on the acoustic data. The main idea is to transform speech material from a source language to the feature space of the target language, [2], [3]. As a result the sparse target language speech material is augmented, broadening the database for the subsequent HMM training. As feature compensation acts on the feature stream prior to acoustic model definition and training we name it a pre-processing technique.

The approach of model combination is quite contrary to feature compensation. Instead of building dedicated acoustic models for the target language, acoustic models of several source languages are chosen. That is, multiple source language ASR systems are run in parallel, each configured to recognize the target language. In a post-processing step the hypotheses of all systems are then combined, and the task is to extract the best from each outcome. For the post-processing ROVER [3] or discriminative model combination (DMC) [4] was explored.

Model adaptation may be seen as an intermediate technique, located between feature compensation and model combination. Differences in the acoustics between languages are seen as an acoustic mismatch problem similar to the one of speaker adaptation. Thus, instead of directly acting on the acoustic data (as in case of feature compensation), classical model adaptation techniques are applied to port the acoustic models of the source language to the target language [5], [6]. In contrast to model combination, only one source model set is used. This model set might be the one of a dedicated source language, or, preferably, a multilingual model set based on several source languages.

In addition to the acoustic mismatch, crosslingual problems exhibit also a structural mismatch. Caused by the different phoneme sets and the different phonotactics of the involved languages, a language specific definition of the acoustic model set is needed. To overcome this problem an adaptation of the model set by so-called poly-

This work was granted by the CICYT under contract TIC2006-13694-C03-01/TCM and contract TIN2005-08852.

phone decision tree adaptation was proposed, [6].

In this work we follow the idea of model adaptation. The starting point is a set of multilingual Spanish-English-German hidden Markov models with their associated decision tree.

3. SOURCE MODEL DEFINITION

In crosslingual acoustic modelling, the question arises which source language one should choose for a specific target language. In previous work it was found that a language which is *close* to the target language tends to be a good choice. In [5], for example, Spanish turned out to be the best choice for building an Italian system. However, in practical situations a *close* language is often not available. In the current case, Slovenian, as one of the target languages, belongs to the Slavic language group, and there are no other Slavic languages in the source language portfolio. In such a case it has been shown that a set of multilingual source models tends to outperform monolingual acoustic source models [6]. For this reason, we decided to use a set of trilingual Spanish-English-German HMMs as source models for the Slovenian, and also for the French mapping task.

A common practice for ASR systems using context dependent acoustic models, is to define the model sets by a phonetic decision tree [7]. Usually, instead of building one big decision tree over the complete acoustic space, one sub-tree per central phoneme and state position is built. This action is justified by the fact that acoustic correlations between different state positions and different phonemes are expected to be small.

In the multilingual case, the a-priori assignment between central phonemes and decision trees can not be overtaken directly. When applying several source languages one is confronted by the problem that the phoneme sets associated with the different source languages are, in general, quite different. For all phonemes which do not have exact counterparts (the same SAMPA symbols) in one of the other languages, pure monolingual trees would be built. To cope with this problem the phonemes of the individual languages are usually clustered to a *multilingual phoneme set*, [8], and [6]. Next, the original phonemes are mapped to the corresponding *multilingual phonemes* which are defined by each cluster. Finally the tree growing process is carried out by using the *multilingual phonemes* as the trees' roots.

However, the use of *multilingual phonemes* has its own disadvantages. There is the problem of the fixed assignment of models with the same central phoneme to one specific tree. Phoneme clustering in general, and in particular in the multilingual case, is far from being unequivocal. Depending on the context the quality of a central phoneme may change in such a way that some of its polyphones may better be assigned to other trees. However, the a-priori assignment of models having the same central phoneme to one predefined set of trees does not permit considering such peculiarities.

Furthermore, *multilingual phonemes* require a dictionary mapping. In the multilingual case this is easily accomplished. Simply, the mapping from the monolingual to the multilingual phonemes defined by the clustering needs to be applied. In the crosslingual case, however, the situation is different. The concept of a *multilingual phoneme set* needs to be extended to the phonemes of the target language which may yet introduce further uncertainties into the model definition process.

To remedy these problems, in this work we apply for each state position one decision tree which covers all central phonemes. Beside applying context questions the tree also uses questions with respect to the central phonemes. For the question set itself, generic features as defined by the International Phonetic Alphabet (IPA), are used, e.g. plosive, bilabial, et cetera. Compared to commonly used broad phonetic classes, such features have the advantage that, in general, they

can simply be picked out of a textbook. Furthermore, crosslingual model definition results much simpler. Instead of having to map the target phonemes to the broad phonetic classes used to construct the source model tree, the tree can directly be entered by the IPA features associated to the target language phonemes. In addition, the fact that common tree roots rather than a *multilingual phoneme set* were used to construct the decision tree pays off twice. First, there is no need to map the target language phonemes to a *multilingual phonemes set*, and, as a corollary, no mapping of the target language dictionary is necessary. The use of common root nodes naturally results in a direct assignment of source to target models.

4. ACOUSTIC SOURCE MODELS

The starting point for the crosslingual model transfer is a tri-lingual Spanish-English-German set of HMMs. All source language data as well as the adaptation and test data of the target languages stem from SpeechDat fixed telephone databases. In the case of the source language data, from each database a 1000 speaker training part was extracted. Only so-called *phonetically rich sentences* were used for the training. Table 1 gives and overview of the data.

Lang.	#Phrases	Hours	Speakers	
			Female	Male
SP	7994	6.9	500	500
EN	8089	7.2	500	500
GE	7540	9.2	500	500

Table 1. Training data used for building the multilingual source models. All data stem from the Spanish (SP), English (EN), and German (GE) SpeechDat fixed telephone databases. In each case a 1000 speaker subset was extracted. The data chunks are balanced respective sex, giving 500 female and 500 male speakers.

The data is parametrized by calculating every 10ms twelve mel-cepstrum coefficients (MFCC) (and the energy). Cepstral mean subtraction is applied. First and second order differential MFCCs plus the differential energy are employed too. For each of the four data streams a codebook is constructed consisting of 256 and 32 (delta energy) Gaussian mixtures, respectively.

For acoustic modelling, a 3-state left-to-right demiphone topology is used, see figure 1. Demiphones [9] can be thought of as tri-phones which are cut in the middle giving a left and a right demiphone. In contrast to triphones, they neglect the influence the left context of a phone might have on the right and vice versa. This drawback in its modelling capabilities is, at least partly, compensated by its improved trainability due to the reduced number of models. Assuming N phonemes, we get N^3 possible triphones, but only $2N^2$ demiphones. In light of the amount of the available training data this might be seen as advantageous.

As lined out in section 3, a phonetic decision tree is used for

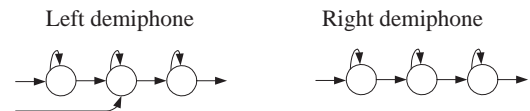


Fig. 1. Demiphone topology

state tying. According to the model topology and the fact that common tree roots for all phonemes are used the overall tree consists

of six sub-trees, see figure 2. The question space of the tree is con-

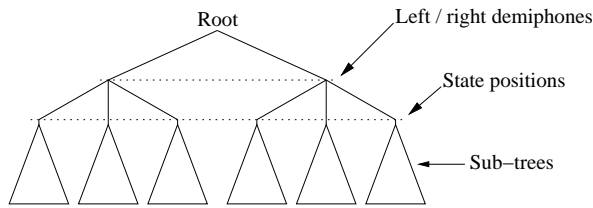


Fig. 2. Structural layout of the decision tree adopting one binary sub-tree for each state position but over all central phonemes. The use of the demiphone topology results in an additional differentiation for left and right demiphones.

structed by so-called IPA-features. As the SpeechDat databases come along with the definition of phoneme sets in SAMPA, the SAMPA were mapped back to the IPA and the associated characteristics, as e.g. plosive, bilabial, unvoiced for a p , were assigned to the SAMPA. This was done for all phonemes, 31 for Spanish, 44 for English, and 47 for German, resulting in 122 individual feature vectors. During tree induction, up to two of the individual attributes were combined to form questions. Bearing in mind that questions with respect to the central phoneme were also asked, valid questions were of the form: Is the central phoneme a plosive or unvoiced?, or: Is the right context phoneme bilabial?

In [6] it was found that the use of multilingual source models is advantageous for crosslingual acoustic modelling. However, in [10] the same authors also report that, when provided with dedicated language information in the form of language tags, a decision tree tends to cluster the language information out when applying corresponding language questions. To develop an idea of this effect, two trees are grown. The first tree applies only linguistically motivated questions. In the case of the second tree, additional questions which ask for the language are used. Figure 3 shows the impact of these so-called language questions on the two resulting decision trees. For growing tree sizes the number of pure monolingual tree leaves are plotted over the total number of leaves.

The lower plot in figure 3 corresponds to a tree grown by the ex-

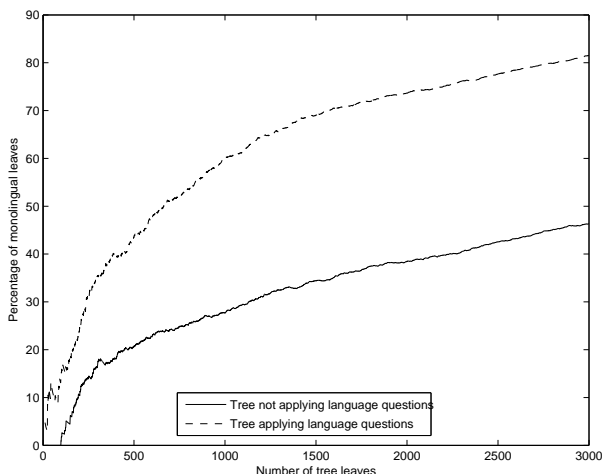


Fig. 3. Development of the *monolinguality* of the tree's leaves for growing tree size. The two plots correspond to two trees which were constructed with and without applying language questions.

clusively use of linguistically motivated questions as described above. In case of the upper plot, beside the linguistically motivated questions also language questions were used. Figure 3 basically confirms the findings of [10]. The use of the language information leads rapidly to a high amount of monolingual tree leaves. For just 1500 leaves, i.e. tied HMM states, already 70% of the leaves are pure monolingual. Without asking language questions the situation changes significantly. For 1500 leaves, the amount of monolingual leaves drops down to 35%.

The behavior of the tree which does not apply language questions is, at least partly, caused by construction. As explained above, linguistic attributes are assigned to individual phonemes according to the assigned SAMPA symbol. Phonemes of the three languages which are assigned the same SAMPA symbol own therefore identical IPA-features. This makes it yet impossible for the tree growing algorithm to distinguish corresponding models of these languages, even if these models were different from an acoustic point of view.

In this work we follow the results of [6]. That is, we base all crosslingual modelling experiments on multilingual models which are defined by a phonetic decision tree not applying language questions. In contrast to [6] we use common decision trees for the central phoneme of the context dependent models. To investigate the influence of different sized source model sets we built two HMMs, one with 1000 and the other with 3000 tied states.

5. BASIC CROSSLINGUAL CONSIDERATIONS

In section 1 we mentioned, that in this work, tied-mixture HMMs are used. In fact, a semi-continuous systems is used, which is mainly driven by the circumstance that such systems are still widely used in real world applications where restrictions with respect to CPU and memory consumption exist. In the case of crosslingual applications, the use of semi-continuous HMMs results in some additional problems. As a semi-continuous systems codes the discriminative information between acoustic units by mixture weights associated to the codebook entries, it is questionable if the commonly used codebook adaptation (updating of the means and covariances) is a useful strategy for porting the source models to the target language. In the following we thus investigate this topic. In addition, some baseline systems for Slovenian and French are developed. These systems will provide benchmarks for the judgment of the final crosslingual systems.

First a pure monolingual Slovenian and a pure monolingual French system is built. As in the case of the source languages, the training and the test data is taken from the corresponding SpeechDat fixed telephone databases. However, because of the Slovenian SpeechDat database consists of merely 1000 speakers, only 900 speakers (450 female, 450 male) are used for the system build. The remaining 100 are used for testing. The training data consists of so called *phonetically rich sentences*. In case of the test data, a word list consisting of *phonetically rich words* and *application words* is used. Also the test data is balanced with respect to sex. To keep the French system comparable to the Slovenian one, also for the French system a 900 speakers training set and a 100 speaker test set was defined. The design of the French training and test sets follow the considerations for the Slovenian ones.

The column called CB-mono of table 2 presents the performance of these systems. We attribute the worse Slovenian performance mainly to the smaller amount of Slovenian training data (see also table 2). One needs also to take into account the fact that Slovenian is modeled by 47 SAMPA, whereas for French 43 SAMPA are used, making the Slovenian model space potentially bigger. Table 2 also provides the results of two additional systems named CD-multi. The

Lang.	Hours	CB-mono	CB-multi
SL	4.7	9.60	9.61
FR	7.6	6.12	6.57

Table 2. Training data and system performance for 900 speaker monolingual Slovenian (SL) and French (FR) target systems, WER in [%]. CD-mono indicates the use of dedicated monolingual codebooks, and CD-multi indicates the use of the multilingual source language codebooks.

CD-multi systems serve to investigate the impact of the multilingual source language codebooks when used instead of dedicated target language codebooks. According to table 2 this impact is in fact negligible. Switching from a pure monolingual setup, applying Slovenian and French codebooks, respectively, to a mixed set-up, that is, applying the multilingual codebooks of the source languages which have never seen any data of the target languages, hardly affects the results. Hence, as we do not expect any significant harm from the use of the multilingual instead of dedicated monolingual codebooks, all subsequent tests on building a crosslingual Slovenian or French target system are based on the use of the multilingual codebooks composed from the three source languages.

Next, the basic crosslingual model mapping step is done. After assigning corresponding linguistic features to all Slovenian and French phonemes, the two multilingual trees described in section 4 are entered and two sets of Slovenian and French models are defined. Table 3 shows that, independent of the language and model set size

	Slovenian		French	
	1000	3000	1000	3000
PRED	48.85	45.88	47.14	47.65
MONO _{pred}	13.68	10.75	21.04	15.52

Table 3. Predicted models before and after a complete re-training, WERs (MONO_{pred}) and mWERs (PRED) in [%].

after pure model prediction (PRED) mean word error rates (mWER) of approximately 50% are achieved¹. These PRED model constitute the base for all subsequent model refinements.

In table 3 we also present so called MONO_{pred} results. These results are obtained by retraining the PRED models using the complete 900 speaker Slovenian, and French training data. Compared to the CD-multi results of table 2 the MONO_{pred} models distinguish only in the underlying decision tree. They serve therefore to judge the crosslingual modelling capability of the underlying source languages decision trees. Comparing the CB-multi with the MONO_{pred} results we observe for Slovenian solely 1-4%, but for French 9-15% loss in performance. Bearing in mind that significantly more French training material is available, it is clear that the underlying Spanish-English-German decision trees match the Slovenian phonotactics better than the French.

To further investigate this issue, we calculate the demiphone overlap [6] between the source and the target languages. Table 4 presents the percentage of demiphone types found in the target language databases which are covered by source language demiphones (upper numbers), and the corresponding coverage of demiphone tokens (lower numbers). The coverage of demiphone tokens is calculated from the percentage of demiphone types by weighting the individual demiphones by their normalized occurrence counts. Focusing on the actual demiphone coverage by the multilingual configuration, it is striking

¹For a description of the term *mean word error rate* see section 6.

	Coverage Measure	SP	EN	GE	MU
SL	Types	21.36	28.70	60.77	75.65
	Tokens	36.95	34.68	82.49	92.47
FR	Types	17.76	17.02	25.80	39.75
	Tokens	28.54	27.86	32.56	43.10

Table 4. Demiphone coverage, in [%], for Slovenian (SL) and French (FR) by a Spanish (SP), a English (EN), a German (GE), and a combined Multilingual (MU) model set.

that for Slovenian the coverage numbers are approximately a factor of two higher than in the French case. This confirms the conclusions drawn from table 3. Examining also the individual language-specific results, it is clear that the good Slovenian behavior basically stems from the German models. However, the highest coverage numbers obtained for French stem also from the German source models.

The findings described by table 3, and 4 are further confirmed when analyzing the phonetic decision trees after target model prediction. In case of accessing a source language tree with linguistic characteristics of a target language typically not all parts of the tree are used. This is caused by linguistic features and feature combinations which are present in the source language but not in the target language. This results in sub-trees of the source language tree which can not contribute to the target model definition. The number of target models (leaves) predicted from the source language decision tree are therefore expected to be smaller than the total number of tree leaves, and, as poorer the match between source and target language gets, as less target models should result. To probe this assumption, we compare the number of tree leaves of the two multilingual source trees to the corresponding number of leaves which are actually used by the target language. Table 5 presents the results of this analysis. Also ta-

	Slovenian		French	
	1000	3000	1000	3000
#Leaves _{source}	1000	3000	1000	3000
#Leaves _{target}	718	1832	508	1130

Table 5. Number of tree leaves, i.e. tied model states, of the multilingual source and target trees. The target trees are obtained by target model prediction from the source trees, and are thus sub-trees of a sources trees.

ble 5 confirms the previous findings. For Slovenian we find that about 60-70% of the original leaves are used by the predicted Slovenian models. Though these numbers appear already low, in case of French they even drop down to 35-50% confirming one more time that the multilingual source models combined out of Spanish, English, and German fit much better to Slovenian than to French.

6. CROSSLINGUAL MODEL REFINEMENT

From table 3, it appears to be quite clear that pure crosslingual model prediction does not lead to reasonable system performances. A common strategy to overcome this problem consists in acoustic model adaptation by a limited amount of target data [2], [5].

Also in this work we investigate therefore acoustic model adaptation. In a first step classical acoustic model adaptation is applied. Such techniques are yet not able to overcome the structural modeling problems introduced by predicting the target models from a source tree which has never seen any target language. To overcome this problem, also so-called polyphone decision tree specialisation (PDTS) [6] is investigated.

For all subsequent tests two adaptation sets per target language

are used. A small one comprising data from 10 speakers and a big one comprising data from 50 speakers. Table 6 gives a detailed overview of the adaptation sets. Note that, though the same number of speakers is used for both languages, in terms of recording time, the amount of Slovenian adaptation data is actually significantly smaller than the French one.

When testing the systems we were faced by the problem that

#Speaker		10	50
SL	#Phrases	85	426
	Time	3.1	15.4
FR	#Phrases	84	422
	Time	5.4	27.6

Table 6. Amount of Slovenian (SL) and French (FR) adaptation data. The times are given in minutes and exclude silence. All data sets are balanced respective sex.

quite high error rates were observed (see the PRED results of table 3). As a consequence, the confidence margins for the word error rates (WER) were quite big, and we could therefore not conclude that a single system was better than another. To overcome this problem, we based the system evaluation on a two-way analysis of variance test (ANOVA) which tests for the hypothesis that the mean WERs (mWER) (calculated over several test sets) of two systems are equal [11]. Thus, instead of a single test, 8 similar but independent tests were run for each system configuration. Afterwards the mean WERs of two system configurations were compared by ANOVA. In the following sections we therefore present mean WERs. Statistically significant (95% confidence interval) different results, with respect to some reference results, are marked boldface. The 8 test sets themselves consist of single phonetically rich words and application words and comprise between 662 and 678 sentences for French, and between 619 and 646 sentences for Slovenian. The resulting grammars, just word lists, consist of between 438 and 452 words per French test set, and between 360 and 383 words per Slovenian test set.

6.1. Crosslingual Acoustic Model Adaptation

In section 5, the use of dedicated target language codebooks did not result in any significant performance improvement. Hence, we actually do not expect to see any significant performance gains by adapting the codebooks to the target languages. Instead adapting mean and covariance parameters, we therefore adapt the mixture weights of the Gaussian mixture densities. Adaptation itself is performed by maximum a-posteriori convex regression (MAPCR) [12]. Contrasting the

#Speakers	Slovenian		French	
	10	50	10	50
PRED	48.85	45.88	47.14	47.65
MAPCR ₁₀₀₀	25.12	18.89	27.55	22.90
MAPCR ₃₀₀₀	26.66	18.73	26.24	19.64

Table 7. Contrast of MAPCR adapted models with the predicted ones (PRED), mWER in [%]. The subscripts denote the model set size of the underlying multilingual model set.

previously obtained PRED results with the MAPCR results, see table 7, reductions in mWER of nearly a factor of two for the small 10 speaker adaptation sets, and more than a factor of two for the 50 speaker adaptation sets are obtained.

It is interesting to note that all Slovenian systems perform better than their French counterparts, though significantly less adaptation

material is available for them. We attribute this behavior one more time to the structural shortcomings of the predicted French model sets. In fact, only approximately 30-50% of the source model states are used by the French target models. This results in a considerable limitation of the French modelling capabilities which can not made up for by more adaptation data.

6.2. Polyphone Decision Tree Specialisation

To cope with the problem of the phonetic context mismatches, PDTS was proposed [6]. PDTS consists of the crosslingual adaptation of a phonetic-acoustic decision tree to a target language. The tree growing process of the source tree restarts using some adaptation data of the target language. Consequently, PDTS permits introducing 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 of Baum-Welsh training on the adaptation data. The resulting models may directly be used as a final model set. In our system we also use them as a starting point for a MAPCR stage on top of the newly generated states.

When running PDTS one is confronted with the problem of deciding when to stop the tree growing process. We decided to test for two configurations, stopping PDTS when the minimum occupation count fell below 5, and 15 model tokens per leaf. In table 8 the resulting PDTS adapted and retrained model sets are compared with the MAPCR adapted model sets already presented in table 7. Note in particular the subscripts given in table 8. They specify the tree sizes before and after applying PDTS.

Drawing our attention first on the Slovenian models we see that

#Speakers	Slovenian		French	
	10	50	10	50
MAPCR ₁₀₀₀	25.12 ₇₁₈	18.89 ₇₁₈	27.55 ₅₀₈	22.90 ₅₀₈
MAPCR ₃₀₀₀	26.66 ₁₈₃₂	18.73 ₁₈₃₂	26.24 ₁₁₃₀	19.64 ₁₁₃₀
PDTS ₁₀₀₀ ⁵	34.21 ₁₂₀₁	23.53 ₂₅₅₄	27.20 ₁₃₈₆	16.34 ₁₄₅₉
PDTS ₃₀₀₀ ⁵	47.65 ₂₀₂₆	26.39 ₃₀₂₉	37.66 ₁₇₅₀	18.04 ₂₆₇₈
PDTS ₁₀₀₀ ¹⁵	28.81 ₈₃₂	18.03 ₁₆₈₆	26.49 ₈₈₅	14.84 ₁₇₅₈
PDTS ₃₀₀₀ ¹⁵	45.94 ₁₈₅₈	21.13 ₂₃₄₁	34.21 ₁₃₆₁	15.82 ₂₀₆₉

Table 8. Contrast of retrained PDTS adapted models versus MAPCR adapted models, mWER in [%]. The subscripts denote the number of tied states of the source and of the target model sets. The superscripts denote the minimum occupation counts.

all PDTS adapted models sets which are significantly different to MAPCR adapted models perform much worse than their MAPCR counterparts. This is definitely caused by the circumstance that, after PDTS, the available adaptation data spreads over more states (see the subscripts), and the model parameters can no longer be estimated robustly. In case of French, the picture changes a lot. It is striking that for the 50 speaker adaptation set the PDTS systems always perform better than the MAPCR systems. Relative reductions in mWER of up to 25% are achieved. The reason for this behavior is twofold. At first, PDTS improves the French model definition significantly. Second, the amount of adaptation data is large enough to give reasonable model estimates of the, by PDTS, increased model set. This is in line with the circumstance that the French adaptation data is about the double of the Slovenian one.

In the Slovenian but also in the French 10 speaker adaptation case, the disappointing performances obtained by PDTS are expected to be caused by the inappropriate acoustic adaptation. Running just one iteration of Baum-Welch training with the available amount of adaptation data and an increased model space leads to poorly estimated models. To remedy this problem, the PDTS defined model sets were refined by MAPCR. Table 9 contrasts the final PDTS and

#Speakers	Slovenian		French	
	10	50	10	50
MAPCR ₁₀₀₀	25.12	18.89	27.55	22.90
MAPCR ₃₀₀₀	26.66	18.73	26.24	19.64
APDTS ₁₀₀₀ ⁵	28.57	23.28	22.93	15.17
APDTS ₃₀₀₀ ⁹	26.70	23.52	24.47	15.58
APDTS ₁₀₀₀ ¹⁵	23.72	17.44	21.80	13.70
APDTS ₃₀₀₀ ¹⁵	25.67	18.04	23.17	14.47

Table 9. Contrast of PDTS and MAPCR adapted models (APDTS) versus MAPCR adapted models, mWER in [%]. The subscripts denote the number of tied states of the source model sets. The superscripts denote the minimum occupation counts.

MAPCR adapted models to the MAPCR only adapted models.

Now, also in the case of Slovenian, significant improvements over the MAPCR-only adapted models are achieved. The best results with 23.72% and 17.44% mWER are obtained for the APDTS₁₀₀₀¹⁵ models. Also in case of French the best results are achieved for the APDTS₁₀₀₀¹⁵ system. They are yet with 21.80% and 13.70% mWER significantly better than the Slovenian counterparts. Here the larger amount of adaptation data combined with PDTS pays off. It is notable that crosslingual acoustic modelling clearly favors broad robust source models. The best results are always obtained for the systems which are based on the smallest source tree, 1000 leaves, and the smallest target language tree (highest minimum occupation count).

It is worthwhile to compare the best crosslingual models to monolingual models which are built exclusively on the adaptation data. We thus build corresponding monolingual Slovenian and French model sets exclusively using the adaptation data. Table 10 contrasts these systems with the best crosslingual, i.e. the APDTS₁₀₀₀¹⁵ ones. From

#Speakers	Slovenian		French	
	10	50	10	50
Monolingual	54.06	21.33	33.28	14.61
Crosslingual	25.67	18.04	23.17	14.47

Table 10. Contrast of monolingual and crosslingual trained target systems, mWER in [%].

table 10 we can draw two conclusions. In case of very limited target data (10 speakers) crosslingual acoustic modelling provides a powerful method to build reasonable target systems. Compared to pure monolingual built models, relative reductions in mWER of 30-50% are achieved. When more adaptation data becomes available this advantage may decrease rapidly. In case of French, 50 speaker adaptation data (27.6 minutes) are yet enough to train a system which performs as good as a crosslingual defined one.

Finally, comparing the best crosslingual results with the monolingual references from table 2, a performance gap of 7.8% for French and 8.4% for Slovenian are observed. We attribute this to the small amounts of adaptation data, but also to the fact that, as PDTS builds upon the stem of a given decision tree, the model refinement by PDTS gives only suboptimal results.

7. CONCLUSIONS

This paper has described the crosslingual acoustic model development for a Slovenian and a French ASR system. The paper has concentrated on the definition of a suitable set of acoustic source models facilitating an easy transfer of the source to the target models. The interaction of the phonetic source language decision tree with the target languages was investigated in detail. After target model prediction the target models were refined in two steps by polyphone decision tree specialization (PDTS) and maximum a-posteriori convex regression (MAPCR). In the case of the small adaptation sets, the crosslingual model sets outperformed their monolingual counterparts always significantly. For the big adaptation sets the corresponding monolingual systems were outperformed too. However, significant differences could only be detected for the Slovenian systems. Finally, the performance of well trained (900 speakers) pure monolingual models could not be reached. We attribute his behaviour to inherent limitations of PDTS, but also to the very small amount of adaptation data.

8. REFERENCES

- [1] C. Liu and L. Melnar, "An automated linguistic knowledge-based cross-language transfer method for building acoustic models for a language without native training data," *International Conference on Speech and Language Processing*, pp. 1365–1368, Sep. 2005.
- [2] 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.
- [3] W. Byrne, P. Beyerlein, J.M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri, and T. Wang, "Towards language independent acoustic modeling," Tech. Rep., 1999.
- [4] D. Vergyri, S. Tsakalidis, and W. Byrne, "Minimum risk acoustic clustering for multilingual acoustic model combination," *International Conference on Speech and Language Processing*, vol. 3, pp. 873–876, Oct. 2000.
- [5] A. Žgank, Z. Kačič, 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, Nov. 2003.
- [6] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1-2, pp. 31–51, Aug. 2001.
- [7] S.J. Young, J.J. Odell, and P.C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," *ARPA Workshop on Human Language Technology*, pp. 307–312, 1994.
- [8] P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward, "Towards a universal speech recognizer for multiple languages," *Automatic Speech Recognition and Understanding*, pp. 591–598, Dec. 1997.
- [9] J. B. Mariño, A. R. Nogueiras, P. Pachès-Leal, and A. Bonafonte, "The demiphone: An efficient contextual subword unit for continuous speech recognition," *Speech Communication*, vol. 32, no. 3, pp. 187–197, Oct. 2000.
- [10] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," *International Conference on Speech and Language Processing*, pp. 1819–1822, 1998.
- [11] G. E. P. Box, W. G. Hunter, and J. S. Hunter, *Statistics for Experimenters*, Wiley, New York, 1978.
- [12] F. Diehl, A. Moreno, and E. Monte, "Crosslingual adaptation of semi-continuous HMMs using maximum likelihood and maximum a posteriori convex regression," *Proceedings of the 14th European Signal Processing Conference*, Sep. 2006.