POLYPHONE DECISION TREE SPECIALIZATION FOR LANGUAGE ADAPTATION

T. Schultz and A. Waibel

Interactive Systems Laboratories
University of Karlsruhe (Germany), Carnegie Mellon University (USA)
tanja@ira.uka.de

ABSTRACT

With the distribution of speech technology products all over the world, the fast and efficient portability to new target languages becomes a practical concern. In this paper we explore the relative effectiveness of adapting multilingual LVCSR systems to a new target language with limited adaptation data. For this purpose we introduce a polyphone decision tree specialization method. Several recognition results are presented based on mono- and multilingual recognizers. These recognizers are developed in the framework of the project GlobalPhone. In this project we investigate speech recognition in the 15 languages Arabic, Mandarin and Shanghai Chinese, Croatian, English, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish.

1. INTRODUCTION

With the distribution of speech technology products all over the world, the fast and efficient portability to new target languages becomes a practical concern. One of the major time and costs factor for developing LVCSR systems in new languages is the need of large amounts of training data. In this paper we describe a multilingual acoustic model combination for the purposes of porting these models to a new target language. We address three aspects of research due to the amount of available data in the target language:

Cross-language transfer refers to the technique where a system developed in one language is applied to recognize another language *without using any training data of the new language*. Experimental results indicate a relation between language similarity and cross-language performance [1], [2]. Furthermore it is shown that multilingual acoustic transfer models perform better than monolingual ones [2], [3]. The key idea in the **bootstrapping** approach is to initialize a recognizer in the target language by using acoustic models from other languages as seed models. After initialization the system is completely rebuilt *using large training data of the target language*. This idea was first proposed by Zue and evaluated by [4] and [5] showing that cross-language seed models outperform flat starts or random models. Recently the usefulness of multilingual phoneme models as seed models have been demonstrated by [6], [7]. The **language adaptation** technique lies between the two extremes in terms of training data. With this technique an existing recognizer is adapted to the target language *with only very limited data*. [5], [6], [8] proved that the language adaption performance is strongly related to the amount of adaptation data. [6] and [8] investigated the effectiveness of multilingual acoustic models showing that monolingual models were outperformed.

Language	Word based			Phoneme based		
	ER[1]	Vocab	PP	ER	Vocab	PP
Ch-Mandarin	14.5	45K	207	45.2	141	12.5
Croatian	20.0	15K	280	36.7	32	9.6
English	14.0	64K	150	46.4	46	9.2
French	18.0	30K	240	36.1	38	12.1
German	11.8	61K	200	44.5	43	9.0
Japanese	10.0	22K	230	33.8	33	7.9
Korean	14.5	64K	137	36.1	43	9.9
Spanish	20.0	15K	245	43.5	42	8.2
Turkish	16.9	15K	280	44.1	31	8.5

Table 1: LVCSR systems in nine languages

Previous systems which combine multilingual acoustic models have been limited to small tasks and context independent modeling. The extension to wider context modeling across languages was first proposed by [9] and [3]. However, when porting those wide context models to new languages, the problem of phonetic context mismatches is an open issue. We present a new approach to overcome this problem and will report the results below.

2. LANGUAGE ADAPTATION

2.1. Multiple language recognition

As a starting point for adaptation to new languages, we developed language dependent LVCSR systems in nine languages using our Janus Recognition Toolkit (JRTk). These monolingual recognizers were trained and tested on the GlobalPhone database, which was modelled on the WSJ task. GlobalPhone currently consists of the languages Arabic, Chinese (Mandarin and Shanghai dialects), Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. Along with the English WSJ and French Bref task the database covers 9 of the 12 most widespread languages of the world. In each of the GlobalPhone languages about 15 hours of high quality speech was collected, spoken by 100 native speakers per language. For further details refer to [8].

For each language the baseline recognizer consists of a fully continuous 3-state HMM system with 3000 triphone models. Each HMM-state is modeled by a codebook containing a mixture of 32 Gaussians. The preprocessing is based on 13 Mel-scale cepstral coefficients with first and second order derivatives, power and zero crossing rate. After cepstral mean subtraction a linear discriminant analysis reduces the input to 32 dimensions.

[1] Mandarin and Korean is given in character based error rate , Japanese in hiragana based error rate

Throughout the experiments 80% of the database speakers were used for training the acoustic models, 10% were defined as a test set, and the remaining 10% were kept as further cross-validation set. In table 1 we arranged the word based error rates (ER) vocabulary size (Vocab) and trigram perplexities (PP) for the monolingual recognizers. Since the core engines are the same across the languages, differences in the recognition performance are due to either language specific inherent difficulties or to differences in quality and quantity of the used knowledge sources and data. In our opinion it is misleading to infer from the given word error rates to language difficulties. On the one hand the concept of a word does not hold for each language (Chinese, Japanese, and Korean), on the other hand the word error rates are strongly affected by available corpus data and by the human language expertise, which in our case is not comparable in all languages.

In order to give a more reliable measure of the acoustic difficulties of the nine languages table 1 presents the phoneme based recognition rates using a phoneme recognizer without any phonotactic constraints. The results indicate differences in acoustic confusability between languages. We found two groups, one lead by Japanese which seems to be the easiest task, followed by French, Korean, and Croatian. The second group is significant harder to recognize, with English to be bottom of the group.

2.2. Multilingual acoustic model combination

In previous experiments we found that multilingual acoustic models outperform monolingual ones for the purposes of language adaptation [3]. Therefore, we briefly summarize our multilingual acoustic model combination. We intend to share acoustic models of similar sounds across languages. Those similarities can be either derived from international phonemic inventories like IPA, by data-driven methods or by a combination of both. We defined a *global phoneme set* based on the phonemic inventory of the monolingual systems. Sounds which are represented by the same IPA symbol share one common phoneme category. Starting with five languages we get 171 language specific phonemes and pooled them together into 85 phoneme categories. For nine languages we pooled 339 language dependent phonemes into 140 categories. Thus the phone-set compression rate of 49% in the five-lingual case increases to 41% in the nine-lingual case.

Multilingual context dependent models are built by assigning one model to each phoneme category and training this model by sharing the data of all languages belonging to that phoneme category. For context-dependent modeling we use a divisive clustering algorithm that builds context querying decision trees. As selection measure for dividing a cluster into two subclusters we use the maximum entropy gain on the mixture weight distributions. This clustering approach gave significant improvements across different tasks [10] and languages [11]. In previous experiments we investigated two methods of data sharing for model combination: Training the models either by blind sharing the data of all languages which belong to the phoneme category or by preserving the information about to which language the data belong. The latter means that the set of context questions for the decision tree clustering is enhanced by adding questions about the language or language group to which a phoneme belongs. The decision whether language information is more important than phonetic context information becomes data-driven. When the recognition of known languages is the target, our experience showed -coincident to other studies [9], [6]- that the acoustic model combination achieves better results if the language information is preserved. However, blind data shared models perform better if the recognition of new languages is the target, which can be explained by an augmented language robustness by sharing all information across languages (see [3] and [8] for details). In the following experiments we apply the blind shared multilingual acoustic models to the new target language Portuguese.

2.3. Phonetic Context Mismatch

Using larger phonetic context windows (polyphonic modeling) for multilingual acoustic models, a mismatch arise due to the phonotactic differences between languages. In order to examine the polyphone mismatch between languages, we define the non symmetric polyphone coverage measure as the number of polyphone occurrences in one language covered by polyphones in another language. Figure 1 shows the coverage for Portuguese phonemes resulting from different context width of a nine- and five-language pool. The calculation of the plotted coverage proceeds as follows: first we select the language among all pool languages which achieves the highest coverage for Portuguese and plot that value tagged with the language name. We then remove this language from the pool and calculate the coverage between Portuguese and each language pair resulting from the combination of removed language plus remaining pool language. The procedure is repeated for triples and so forth. Thus in each step we find the language which maximally complements the polyphone set.

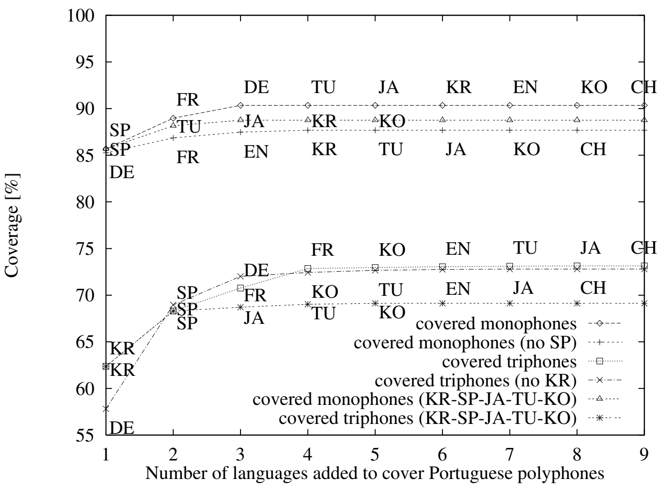

Figure 1: Portuguese polyphone coverage by nine languages

From figure 1 we observed that as expected the coverage dramatically decrease for larger context (for quintphones which are not ploted in figure 1 a maximal coverage of 46% could be attained). After incorporating three languages the coverage of Portuguese monophones can not be increased any further, limited to 91% with the nine language pool and dropping to 85% when Spanish (SP) is removed. The contribution of the Spanish phoneme set to the monophone coverage can not be compensates by other languages remaining in the pool. Second we found a saturation in coverage for four languages after increasing the context width to 1. Further increasing the context width to 2 we observed that at least five languages contribute to the quintphone coverage rate. For triphones the main contribution comes from the Croatian language. Removing this language from the pool is nearly completely compensate by German and Spanish triphones. This indicate that Croatian, German, and Spanish polyphones covers a similar portion of the Portuguese triphones set. Whereas the curve (KR-SP-JA-TU-KO) indicates that the French language contribute unique polyphones which can not be recruited from other languages. In this case the lacking phonemes belong to the categories of nasal vowels.

We conclude from these observations that for the design of a language pool for adaptation purposes it is more critical to find a complement set of languages than to cover a large number of languages. It can be easily seen from the coverage rates that using a polyphone tree even based on several languages can not be applied successfully to a new target language without adapting it to the new contexts.

3. POLYPHONE DECISION TREE SPECIALIZATION

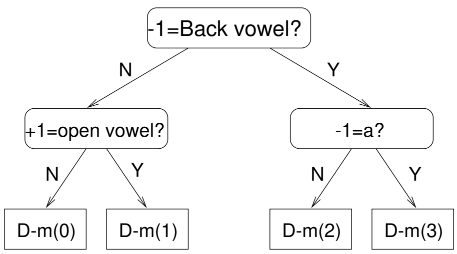

Figure 2: Polyphone Cluster Tree for middle state of monophone D **before** Polyphone Decision Tree Specialisation (PDTS)

In order to overcome the problem of the observed mismatch between represented context in the multilingual polyphone decision tree and the observed polyphones in the new target language, we introduce the Polyphone Decision Tree Specialisation (PDTS) approach. In this approach the clustered multilingual polyphone decision tree is adapted to the target language by restarting the decision tree growing process according to the limited adaptation data available in the target language. Figure 2 illustrates the polyphone cluster tree for the middle state of the monophone D before adaptation. During the clustering procedure only three splits resulting in four leaf nodes were decided to present properly the phonetic context of D in the multilingual data. However, in the Portuguese language this phoneme is very frequent and occurs in very different contexts. Traversing this non-adapted tree during decoding Portuguese speech would lead to very poorly estimated residual class models, since the context questions do not reflect the Portuguese contexts.

Figure 3 shows the decision tree for the middle state of the same phoneme D after applying PDTS. The former tree was spitted further according to matching questions, resulting in 18 leaf nodes. The regrowing process was completed after reaching a predefined number of new leaf nodes depending on the amount of training data. The adapted decision tree now represents valid contexts of the Portuguese D and is expected to improve the recognition results for Portuguese input.

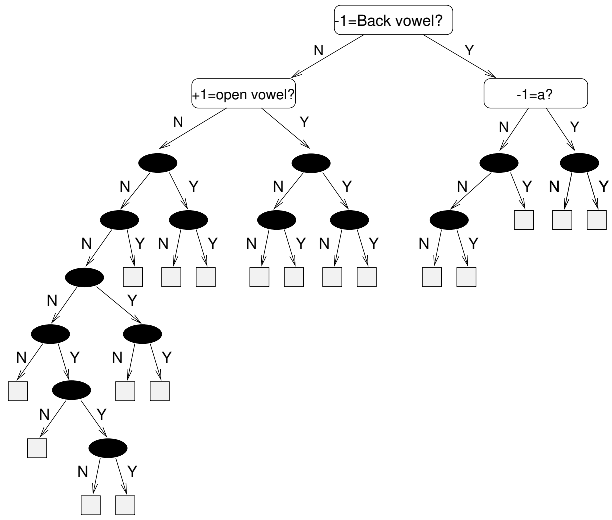

Figure 3: Polyphone Cluster Tree for middle state of monophone D **after** Polyphone Decision Tree Specialisation (PDTS)

4. EXPERIMENTS

We adapt a five-lingual recognizer containing Croatian, Japanese, Korean, Spanish and Turkish data to the Portuguese language and investigate the benefit of the described methods. For adaptation we assume that only 200 Portuguese spoken utterances (25 minutes) and their transcription (3370 words) are given. Although [5] found that the number of speakers is more critical than the number of utterances we decide to use only 7 different Portuguese speakers for adaptation since in our dictation task it is more expensive to get single utterances of many different speakers than to get many utterances spoken by one speaker. A subset of 96 randomly selected utterances from 3 test speakers was used to carry out our experiments. The test dictionary has 7300 entries, the OOV-rate is set to 0.5% by including the most common test words. A trigram language model was calculated on 10 million word text corpus from Agence France Press (LDC95T11) interpolated with the GlobalPhone data leading to a trigram perplexity of 297.

System	Data	Labels	Technique	Ptree
	Cross-language transfer			
S1	0	-	-	ML
S2	0	-	-	CI
	Language adaptation			
S4	100	initial	MLAdapt	CI
S5	100	initial	Viterbi	ML
S6	100	initial	MLAdapt	ML
S7	100	good	MLAdapt	ML
S8	200	good	MLAdapt	ML
S9	200	good	PDTS	ML-PO
	Bootstrap			
S3	100	initial	Rebuild	PO
S10	6600	good	Rebuild	PO

Table 2: Description of systems ported to Portuguese

Table 2 describes the systems used for our porting experiments, their performance on Portuguese is compared in figure 4.

The column **Data** in table 2 refers to the number of recordings used as training data. Transfer without any training data results in the cross-language approach as performed in the systems S1 and S2. Whereas the training based on 6600 utterances (S10) represents the bootstrap technique. For the systems S3 to S9 we used very limited data of 100 and 200 utterances. **Labels** explains whether the phonetic transcriptions of the recordings are created based on the multilingual recognition engine (Labels = initial) or based on good phonetic alignments which we taking to be already given (Labels = good). The latter was used to accelerate our adaptation process.

The term **Technique** is related to the training approach applied to the systems. Viterbi refers to one iteration of viterbi training along the given labels. MLAdapt means Maximum Likelihood Adaptation technique, Rebuild refers to the iterative procedure of writing labels, viterbi training, model clustering, training, and writing improved labels. PDTS is the described Polyphone Decision Tree Specialization. The **Ptree** item describes the origin of the polyphone decision trees. CI refers to context independent modeling, meaning that no polyphone tree is used, ML is the multilingual decision tree with 3000 polyphones and PO is a polyphone tree build exclusively on Portuguese data. ML-PO refers to the regrown polyphone tree applying PDTS.

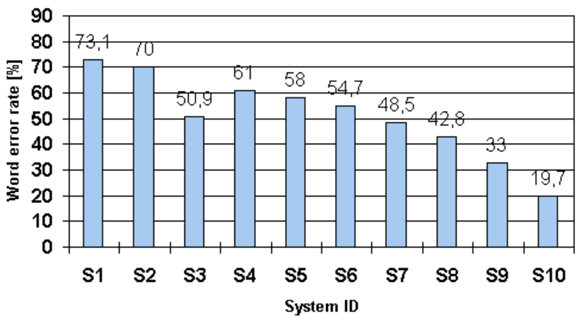

Figure 4: Language adaptation to Portuguese

As expected the cross-language transfer using the five-lingual recognizer without any training on Portuguese data results in extremely high word error rates of 73.1% for the context dependent system (S1) and slightly better error rates of 70% for the context independent system (S2). Therefore, the initial labels are written with system S2. Using 100 of these initial labels for adapting the context independent multilingual system (S4) and the context dependent system by MLA (S6) or viterbi training (S5) shows a significant gain. In S3 the initial labels are used to completely rebuild a Portuguese system after bootstrapping from multilingual seed models. The comparison of S6 and S3 indicates that the adaptation without polyphone decision tree specialization is outperformed by the bootstrap technique (S3) even if data are very limited. Nevertheless the word error rate of the winning system S3 achieving 50.9% is still unsatisfying.

We obtain the next performance boost from using improved labels (S7) and double amount of adaptation data (S8). Finally we applied our PDTS approach (S9) which leads to significant improvements achieving 33% word error rate. This performance compares to 19.7% word error rate (S10) resulting from bootstrapping and rebuilding a Portuguese LVCSR system using 16 hours of speech spoken by 78 speakers. To summarize we get the highest performance gain in language adaptation from the PDTS technique, enlarging adaptation data, and improved labels, in this order.

5. CONCLUSION

In our language adaptive approach we explore the relative effectiveness of multilingual context dependent acoustic models in combination with a polyphone decision tree specialization (PDTS). We examine the benefit when porting a multilingual engine to new target languages with very limited training data. The results are very promising achieving 33% word error rate for an Portuguese LVSCR system when using only 200 spoken utterances for adaptation.

6. ACKNOWLEDGMENT

The authors gratefully acknowledge all members of the GlobalPhone team for their great enthusiasm. We also wish to thank the members of the Interactive Systems Laboratories, especially Roald Wolff for his active support, great encouragement and contribution to this research.

7. REFERENCES

[1] A. Constantinescu et al.: *On Cross-Language Experiments and Data-Driven Units for ALISP*, Proc. ASRU, pp. 606-613, St. Barbara, CA 1997.

[2] U. Bub et al.: *In-Service Adaptation of Multilingual Hidden-Markov-Models*, Proc. ICASSP, pp. 1451-1454, Munich 1997.

[3] T. Schultz et al.: *Multilingual and Crosslingual Speech Recognition*, Proc. DARPA Workshop on Broadcast News Transcription and Understanding, pp. 259-262, Lansdowne, VA 1998.

[4] J. Glass et al.: *Multi-lingual Spoken Language Understanding in the MIT Voyager System*, Speech Communication (17), pp. 1-18, 1995.

[5] B. Wheatley et al.: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language*, Proc. ICASSP, pp. 237-240, Adelaide 1994.

[6] J. Köhler: *Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks*, Proc. ICASSP, pp. 417-420, Seattle, 1998.

[7] T. Schultz et al.: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets*, Proc. Eurospeech, pp. 371-374, Rhodes 1997.

[8] T. Schultz et al.: *Language independent and language adaptive LVCSR*, Proc. ICSLP, pp. 1819-1822, Sydney 1998.

[9] P. Cohen et al.: *Towards a Universal Speech Recognizer for Multiple Languages*, Proc. ASRU, pp. 591-598, St. Barbara CA, 1997.

[10] K.-F. Lee: *Large-vocabulary Speaker-independent Continuous Speech Recognition: The SPHINX System*, PhD Thesis, Carnegie Mellon University, 1988.

[11] M. Finke et al.: *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech*, Proc. ICASSP, pp. 1743-1746, Munich 1997.