Language Portability in Acoustic Modeling

Tanja Schultz and Alex Waibel
{*tanja@cs.cmu.edu*}

Interactive Systems Laboratories
Carnegie Mellon University (USA), University of Karlsruhe (Germany)

ABSTRACT

With the distribution of speech technology products all over the world, the portability to new target languages becomes a practical concern. As a consequence our research focuses on the question of how to port LVCSR systems in a fast and efficient way. More specifically we want to estimate acoustic models for a new target language using speech data from varied source languages, but only limited data from the target language. For this purpose we introduce different methods for multilingual acoustic model combination and a polyphone decision tree specialization procedure. Recognition results using language dependent, independent and language adaptive acoustic models are presented and discussed in the framework of our GlobalPhone project which investigates LVCSR systems in 15 languages.

1. Introduction

The state of the art in large vocabulary continuous speech recognition (LVCSR) has advanced substantially for quite a number of languages. Recognition systems developed originally for one language have been successfully ported to several languages, including systems developed by IBM [6], Dragon [2], BBN [5], Cambridge [21], Philips [8], MIT [10], and LIMSI [13]. The transformation of English systems to such diverse languages like German, Japanese, French, and Mandarin Chinese illustrates that speech technology generalizes across languages and that similar modeling assumptions hold for various languages.

To date, however, extensions have only been performed with well known languages for which large amounts of data are available. To build a recognizer, this data usually includes dozens of hours of recorded and transcribed speech. Unfortunately the assumption that large speech databases can be provided on demand does not hold for many reasons. As a consequence, our research has focused on the question of how to build a LVCSR system for a new target language using speech data from varied source languages, but only limited data from the target language.

In our present research we focus mainly on *language independent acoustic modeling* problems and assume that text resources and pronunciation dictionaries are given in the target language. This is a reasonable assumption since acquiring the training data for acoustic models is usually the most expensive part of a data collection. However, we are aware of the fact that appropriate large text material are, to date, only available in hundreds of languages and pronunciation dictionaries in some tens of the most spread and studied languages.

To achieve the goal of adaptation to new target languages we investigate multilingual LVCSR systems, i.e systems capable of *simultaneously recognizing languages* which have been presented during the training procedure. Particularly we define a global unit set which is suitable to cover 12 languages. Based on this global unit set we evolve and evaluate different techniques to combine the acoustic models of varied languages and call the resulting multilingual acoustic models *language independent*. These language independent acoustic models allow the data and model sharing of various languages to reduce the complexity and number of parameters of a multilingual LVCSR system. Furthermore, these models will be used as seed models for the initialization of acoustic models in a new target language.

In *language adaptation* experiments these preexisting models are adapted towards an optimal recognition of a new target language, using only limited adaptation data from this target language. Given the data limitation we face a new problems: the large phonetic mismatch between varied source languages and the target language when extending the phonetic context window for building context dependent acoustic models. Phoneme model of arbitrary context width are called *polyphones*. The use of large phonetic context windows has proven to increase the recognition performance significantly in the monolingual setting. Therefore, it seems natural to extend this idea to the multilingual setting as well. In order to solve this problem we introduce a procedure of adapting multilingual polyphone decision trees to a target language with very limited adaptation data. In summary we present techniques which enable us to set up a LVCSR recognition engine in a new target language by borrowing speech data from varied source languages but only limited data from the target language itself.

2. The GlobalPhone project

GlobalPhone is a project undertaken at the Interactive Systems Labs which investigates multilingual LVCSR in various languages. We collected a database consisting of the languages Arabic, Chinese (Mandarin and Shanghai dialects), Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. Along with the English Wall Street Journal (WSJ0, distributed by LDC and French BREF (BREF-Polyglot sub-corpus, distributed by ELRA) databases, this covers 9 of the 12 most widespread languages of the world (see for example [20]). In each of the languages about 15-20 hours of high quality speech was collected, spoken by 100 native speakers per language. Each speaker read several articles about political and economical topics chosen from national newspapers. All the newspapers are accessible via Internet, so that large text corpora for language modeling can be easily downloaded. Further details about the GlobalPhone project are given in [18].

2.1. Global Unit Set

Our research in language independent and adaptive LVCSR is based on the assumption that the articulatory representations of phonemes are so similar across languages that phonemes can be considered as units which are independent from the underlying language. Based on this assumption the language specific phoneme inventories of N languages can be unified into one global set $\Upsilon = \Upsilon_{L_1} \cup \Upsilon_{L_2} \cup \ldots \cup \Upsilon_{L_N}$. This idea was first proposed by the International Phonetic Association [11] then transfered to automatic speech recognition by Andersen and Dalsgaard [1]. According to this idea we differentiate between the group of language independent *polyphonemes*[1] Υ_{LI}, containing phonemes occurring in more than one language, and N remaining groups of language dependent *monophonemes* $\Upsilon_{LD_{L_1}}, \ldots \Upsilon_{LD_{L_N}}$, which contain phonemes only occurring in one language.

Shared by	#	Modeled Phonemes (IPA symbols)	
	83	Polyphonemes shared across ≥ 2 languages	
		Consonants	Vowels
All	4	m,n,s,l	-
11	7	p,b,t,d,k,g,f	-
10	3	-	i,u,e
9	6	ŋ,v,z,j	a,o
8	1	ʃ	-
7	3	r,h,tʃ	-
6	1	-	ɛ
5	9	ɲ,ʒ,x,ts,dʒ	iː,y,ɘ,ɔ
4	4	-	ɨ,ø,ɑ,ei
3	11	ʎ,w,ç	ɪ,uː,eː,œ,oː,æ,ai,aʊ
2	34	pʰ,tʰ,dʲ,kʰ,gʲ,ʁ,ɾ$_r$,θ,ð,sʲ,zʲ,ʂ,ɻ,tsʰ,tʃʲ	ˈi,yː,ɯ,ʊ,ˈe,ɛː,øː,aː,ˈa,ɑː,ˈu,ˈo,aɪ,au,ia,io,eu,oi,oʊ
	79	Monophonemes belonging to *one* language	
		Consonants	Vowels
CH	15	tʂ,tʰʂ,cç,cçʰ	iʊ,iɛ,ua,uɛ,uɔ,ya,yɛ,iao,uɛi,uai,ioʊ
EN	5	ɾ$_d$	ʌ,ɜ˞,ɔi,ə˞
FR	5	ɥ	ɛ̃,œ̃,ɑ̃,ɔ̃
GE	3	-	ɐ,ʏ,ɔʏ
JA	2	ʔ	ɯː
KO	14	p˺,p’,t˺,t’,k˺,k’,s’,c’ʰ	ie,iɘ,iu,ɨi,oa,uɘ
KR	1	dʒʲ	-
PO	8	-	ĩ,ũ,ẽ,õ,ɐ̃,ew,ow,aw
RU	15	pʲ,bʲ,tʲ,mʲ,rʲ,vʲ,ʃʲ,ʒʲ,lʲ,ʃtʃ,ʃtʃʲ	ja,jɛ,jɔ,ju
SP	2	β,ɣ	-
SW	9	ʈ,ɖ,ɳ,ɭ,ks	œː,æː,ʉː,ɵ
TU	0	-	-
$\sum$	162	Silence and noises shared across languages	

Table 1: Global Unit Set for 12 languages

Similarities of sounds are documented in international phonetic inventories like IPA [11], which classify sounds based on phonetic knowledge. In our research we define a global unit set for 12 languages based on the IPA scheme. Sounds of different languages, which are represented by the same IPA symbol, share one common *IPA-unit*. Regarding Chinese sounds we abstain from handling tones separately, i.e. the 5 tonal variations of a Mandarin vowel are treated as one vowel. Table 1 summarizes the polyphonemes and monophonemes for all 12 languages. For each polyphoneme the upper half of Table 1 reports the number of languages which share one phoneme. The lower half of Table 1 contains the number and type of monophonemes for each language.

[1] polyphonemes should not be confused with polyphones

3. Language independent acoustic modeling

Based on the described global unit set we investigate different methods to combine the acoustic models of varied languages to one multilingual acoustic model. The main goals of the model combination are the reduction of the overall amount of acoustic model parameters and the improvement of the model robustness for language adaptation purposes.

3.1. Acoustic model combination

We introduce three different methods for acoustic model combination, the language separate *ML-sep*, the language mixed *ML-mix*, and the language tagged *ML-tag* combination method as illustrated in Figure 1. The evaluated systems applied the same preprocessing and acoustic modeling, i.e. they consist of fully continuous HMMs with 3000 sub-polyphones. The term *sub-polyphone* here refers to a polyphone which is divided into a begin, middle and end state. The probability $p(x|s_i)$ to emit x in state s_i is described by a mixture of $K_i = 16$ Gaussian components: $p(x|s_i) = \sum_{k=1}^{K_i} c_{s_i k} N(x|\mu_{s_i k}, \Sigma_{s_i k})$. The Gaussians are on 13 Mel-scale cepstral coefficients and power with first and second order derivatives. After ceptstral mean subtraction a linear discriminant analysis reduces the input vector to 32 dimensions. Figure 1 illustrates the three different acoustic model combination methods. The mixture weights c are symbolized as distributions and the Gaussian components $N(x|\mu, \Sigma)$ are symbolized as rounded boxes.

In the *ML-sep* combination method each language-specific phoneme is trained solely with data from its own language, i.e. no data are shared across languages to train the acoustic models. The multilingual component of *ML-sep* is the feature extraction, since one global LDA-matrix is calculated taking all language-specific phoneme models as LDA classes. Context dependent models are created by applying a decision tree clustering procedure which uses an entropy-based distance measure, defined over the mixture weights of the Gaussians, and a question set which consists of linguistically motivated questions about the phonetic context of a phoneme model [9]. In each step of clustering the question giving the highest entropy gain is selected when splitting the tree node. The splitting procedure is stopped after reaching the predefined number of 3000 sub-polyphone models. A schematic of the separate acoustic modeling method is shown in the left part of Figure 1 for the beginning state of phoneme "M".

In the *ML-mix* combination method shown in the middle part of Figure 1 we share data across different languages to train the acoustic models of polyphonemes, i.e. phonemes of different languages which belong to the same IPA-unit defined in our global unit set (see Subsection 2.1). During training we do not

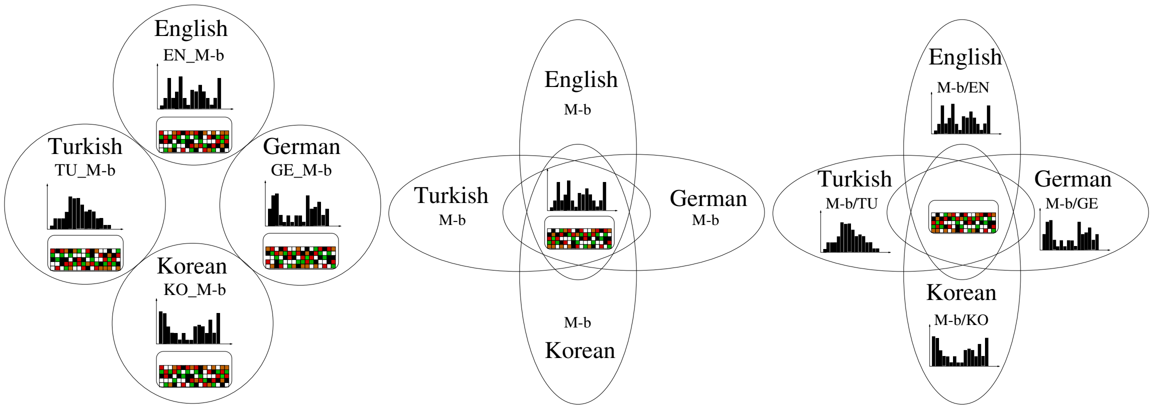

Figure 1: Separate **ML-sep** (left), mixed **ML-mix** (middle), and tagged **ML-tag** (right) acoustic modeling

preserve any information about the language. In other words, for each IPA-unit of the global unit set we initialize one mixture of 16 Gaussian components per state and train the model of this IPA-unit by sharing the data of all languages belonging to the IPA-unit. The context dependent models are created by applying the aforementioned clustering procedure. The splitting procedure is stopped after reaching a predefined number of 3000 language independent sub-quinphone models, which results in system *ML-mix3000*.

For the combination method *ML-tag* each phoneme receives a language tag attached in order to preserve the information about the language the phoneme belongs to. *ML-tag* is similar to *ML-mix* in the sense that they both share all the training data and use the same clustering procedure. But for *ML-mix* the training data are only labelled by phoneme identity, whereas for *ML-tag* the training data is labelled by both phoneme and language identity. The clustering procedure is extended by introducing questions about the language and language groups to which a phoneme belongs. The Gaussian components are shared across languages as in the *ML-mix* method but the mixture weights are kept separately. Therefore, the relative importance of phonetic context and language membership is resolved during the clustering procedure by a data-driven method.

3.2. Simultaneous recognition

We start with 650,000 different sub-quinphones defined over the five languages and create two fully continuous systems, *ML-tag3000* with 3000 models, and *ML-tag7500* with 7500 models, the latter one being of the same size as five monolingual systems each having 1500 models. We explore the usefulness of our modeling approach by comparing the recognition performance of the monolingual case with the performance which is achieved by the resulting systems from the *ML-sep*, *ML-tag*, and *ML-mix* combination method. The experiments are done for the five languages Croatian, Japanese, Korean, Spanish and Turkish. The comparison focus on the purpose of simultaneously recognizing these languages which are involved for training the multilingual acoustic models. First we compare the monolingual system to the system *ML-sep* which only differs in the multilingual LDA. Compared to the monolingual case the multilingual LDA slightly increase the word error rate but not significantly. When we compare the combination methods to each other we found that the system *ML-tag3000* outperforms the mixed system *ML-mix3000* in all languages by an average of 5.3% (3.1% - 8.7%) error rate. Since the collection of the GlobalPhone speech data is uniform in terms of recording and channel conditions we draw the conclusion that preserving the language information achieves better results with respect to simultaneous recognition. The *ML-tag3000* system reduces the model size to 40% compared to the monolingual case (3000 vs 5x1500 models), resulting in a 3.1% performance degradation on average (1.2% - 5.0%). However, not all of the degradation can be explained by the reduction of parameters. This can be derived from the comparison between the monolingual systems and *ML-tag7500*. We still observe an average performance gap of 1.1% (0.3% - 2.4%) when comparing the acoustic modeling with respect to simultaneous recognition of the relevant source languages. The finding coincides with other studies [3, 6, 12]. A detailed description of these experiments can be found in [16].

4. Language adaptive acoustic modeling

Previous approaches for language adaptation have been limited to context independent acoustic models [19, 7, 4]. Since for the language dependent case wider contexts increase recognition performance significantly, we investigate whether such improvements extend to the multilingual setting. The use of wider context windows raises the problem of phonetic context mismatch between source and target languages. To measure this mismatch we define the coverage coefficient. In order to approach the mismatch problem we introduce a method for polyphone decision tree adaptation.

4.1. Phonetic context mismatch

We define the coverage coefficient $\mathrm{cc_N(L_T)}$ of the target language L_T to be:

$$\mathrm{cc_N(L_T)} = \frac{|\Upsilon_{\mathrm{L_T}} \cap \Upsilon|}{|\Upsilon_{\mathrm{L_T}}|} = 1 - \frac{|\Upsilon_{\mathrm{LD_{L_T}}}|}{|\Upsilon_{\mathrm{L_T}}|} \qquad (1)$$

The coverage coefficient cc gives us the portion of phonemes in the target language L_T which are covered by phonemes of the global unit set. The coverage coefficient is zero, if no phoneme of the target language L_T has a counterpart in the global unit set, and one if each phoneme is covered, i.e. $0 \leq \mathrm{cc(L_T)} \leq 1$.

The idea of phoneme coverage can be extended naturally to models of various context width. Based on the above definition we now introduce monophone coverage, triphone coverage and in general polyphone coverage. We further distinguish between the coverage of polyphone types and polyphone occurrences. For the latter the frequency of a polyphone is taken into account to reflect that coverage of frequent polyphones is more important than coverage of less frequent ones with respect to recognition performance.

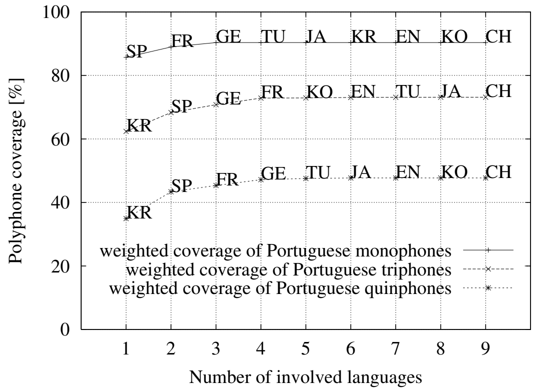

Figure 2: Portuguese polyphone coverage by nine languages

In the following we will apply the polyphone decision tree specialization procedure to adapt the multilingual recognition engine to the target language Portuguese. To examine how well the 46 Portuguese phonemes and resulting polyphones are covered by a given language pool, we calculated the coverage with respect to the global unit set (without Portuguese). The coverage indicates how well a generic polyphone decision tree fits to the target language Portuguese. The percentage coverage $\mathrm{cc}(\mathrm{Po}) \times 100$ is plotted in Figure 2 for context width zero (monophones), one (triphones) and two (quinphones). The calculation of plotted coverage proceeds as follows: We select the language among all pool languages which achieves the highest coverage for Portuguese. Then we 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 determine the language which maximally complements the polyphone set.

As expected, the coverage decreases dramatically for wider contexts. With a nine language pool, the coverage of Portuguese monophones achieves 91%, drops to 73% for triphones and to 47% for quinphones. After incorporating the three main contribution languages the coverage for monophones cannot be increased any further. When enlarging the context width to one, coverage saturates after four languages. For a context width of two we observed that at least five languages contribute to the quinphone coverage rate. Therefore, we expect that increasing the context width requires more languages.

4.2. Polyphone decision tree specialization

From analyzing the coverage in Figure 2 we draw the conclusion that a polyphone decision tree, even build on several languages, can not be applied successfully to a new language without adaptation. 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 propose the Polyphone Decision Tree Specialization (PDTS) procedure as described in [15]. In PDTS 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 3 illustrates the poly-

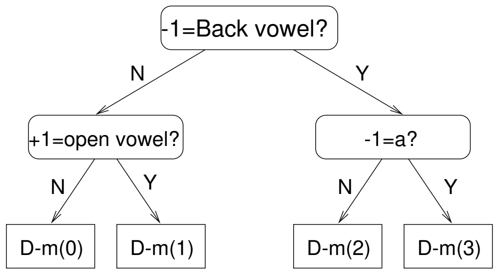

Figure 3: Tree **before** Polyphone Decision Tree Specialization

phone cluster tree for the middle state of the phoneme d^j before adaptation. During the clustering procedure only three splits resulting in four leaf nodes were used to capture the phonetic context of d^j 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.

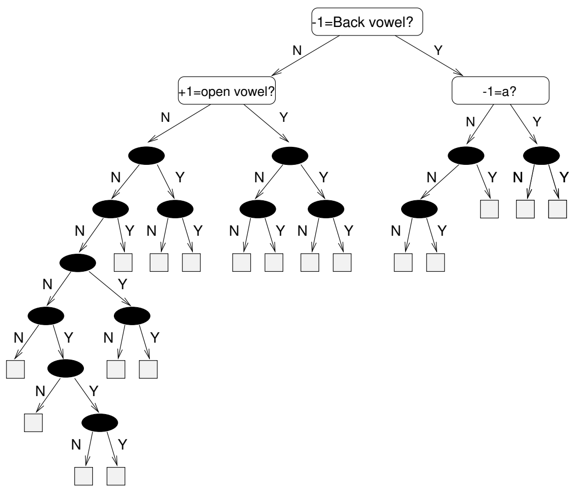

Figure 4: Tree **after** Polyphone Decision Tree Specialization

Figure 4 shows the decision tree for the middle state of the same phoneme d^j after applying PDTS. The former tree was further clustered according to 14 additional questions, resulting in 18 leaf nodes. The re-growing process is 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^j and is expected to improve the recognition results for Portuguese input.

5. Experiments

In the following experiments we investigate the benefit of the acoustic model combination and the polyphone decision tree specialization (PDTS) for the purpose of adaptation to the Portuguese language. The above-described five-lingual recognition systems are ported to Portuguese using different amounts of data. We assume that a Portuguese dictionary as well as the recordings and transcriptions of some spoken utterances are given. The dictionary mapping is done according to an heuristic IPA-based mapping approach [17]. A subset of 300 utterances from 10 test speakers is used to carry out the experiments. The test dictionary has about 7300 entries, the OOV-rate is set to 0.5% by including the most common words of the test set into the dictionary. A trigram language model with Kneser/Ney back-off scheme is calculated on a 10 million word corpus from Agency France Press (LDC95T11) interpolated with the GlobalPhone training data leading to a trigram perplexity of 297.

When using the entire portion of 16.5 hours of spoken Portuguese speech from the GlobalPhone database, we achieve a word error rate of 19.0% (SystemId S14) on the the aforementioned test set, dictionary, and language model.

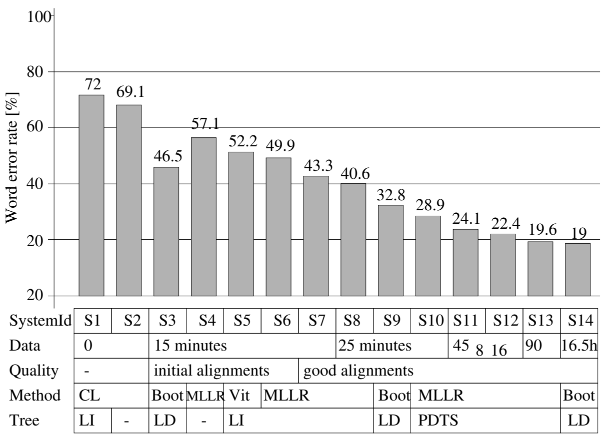

Figure 5: Language adaptation to Portuguese

Figure 5 summarizes the experiments which have been performed to improve the Portuguese LVCSR system. The **SystemId** identifies the developed systems. The row **Data** refers to the amount of adaptation data (0-90 minutes of spoken speech). **Quality** explains whether the phonetic alignments are *initially* created based on the multilingual recognition engine or assumed to be available in *good* quality. The term **Method** is related to the transfer approach which is applied: Cross-language (CL), adaptation (Viterbi or MLLR), and bootstrapping technique (Boot). *Viterbi* refers to one iteration of Viterbi training along the given alignments. *MLLR* is the Maximum Likelihood Linear Regression [14], and *Boot* refers to the iterative procedure: creating alignments, Viterbi training, model clustering, training, and writing improved alignments. The item **Tree** describes the origin of the polyphone decision tree: '–' refers to context independent modeling, *LI* is the generic language independent polyphone decision tree of system *ML-mix3000*, *LD* is the language dependent tree which is built exclusively on Portuguese data, and *PDTS* refers to the adapted LI polyphone tree after applying PDTS.

5.1. Transfer procedure

According to our finding that language independent models outperform language dependent ones and the fact that the *ML-mix* combination method performs better than *ML-tag* when using them as seed models for a new target language (see subsection 3.2), we use *ML-mix3000* as the basis system for the adaptation to Portuguese. For the systems S1 and S2 only the data of the five source languages has been applied for training the acoustic models, no adaptation is performed before decoding the Portuguese speech. The context independent system (S2) slightly outperforms the context dependent system (S1) as shown in Table 2. Therefore, the initial alignments are written with system S2. These initial alignments of 15 minutes Portuguese speech are used for adaptation, which leads to 17.4% word error rate reduction in the context independent (S2 → S4), and to 30.7% word error rate reduction in the context dependent case (S1 → S6). The improvement through context dependent modeling (S4 → S6) indicate that the language independent polyphone tree covers some parts of Portuguese phonotactics. However, system S3 which results from completely rebuilding a Portuguese system, outperforms system S6, i.e. a system with a polyphone decision tree build solely on Portuguese data achieves better results on 15 minutes adaptation data than a system with a non-adapted generic polyphone decision tree trained from various languages.

SystemId	Method	Word Error [%]		Improvement	
		CI	CD		
S2 / S1	Cross-language	69.1	72.0	17.4%	30.7%
S4 / S6	Adaptation	57.1	49.9	-	6.8%
S3	Bootstrapping	-	46.5		

Table 2: Transfer procedure

Provided that 15 minutes of Portuguese speech are given for adaptation, MLLR outperforms the Viterbi training by 4.4%, as the comparison of system S5 to S6 in Figure 5 indicates. Although, MLLR was originally designed for speaker adaptation, it can be successfully applied to language adaptation.

5.2. PDTS

Next we investigate the effect of specializing the polyphone decision tree according to the proposed PDTS procedure. In Table 3 we compare the results from PDTS specialized polyphone tree (S10) to non-adapted language independent trees (S6, S8) and to language dependent trees which are trained solely on Portuguese adaptation material (S3, S9). The language independent polyphone trees are outperformed by the language dependent ones if no tree specialization is applied. The performance difference increases from 6.8% to 19.2% after the amount of adaptation data is extended to 25 minutes. However, the PDTS adapted tree (S10) significantly outperforms even the language dependent tree in system S9 by 11.9%

SystemId	Tree	Alignments		Improvement	
		15 min initial	25 min good		
S6/S8	LI	49.9	40.6	6.8%	19.2%
S3/S9	LD	46.5	32.8	-	11.9%
S10	PDTS	-	28.9		

Table 3: The PDTS method [WE in %]

which means that the knowledge and phonotactics of several languages stored in the polyphone decision tree can be transfered successfully to a new target language.

5.3. Adaptation data

The phonetic alignments of the Portuguese adaptation utterances are initially created by the multilingual recognition system S2 (initial alignments). In order to accelerate our adaptation process we create improved phonetic alignments which we assume to be available (good alignments). This decreases the word error rate by 13.2% (S6 $\rightarrow$ S7) as can be seen from the upper part of Table 4. Furthermore, we evaluate the effect of extending the adaptation data, from 15 to 25, then to 45, and finally to 90 minutes of spoken speech. From this we achieve 16.6% (S10 $\rightarrow$ S11) and 12.5% (S12 $\rightarrow$ S13) improvement, whereas we achieved 7.1% by doubling the number of adaptation speakers (S11 $\rightarrow$ S12), reported in the lower part of Table 4. Further extension of the number of speakers to 32 and all 78 did not lead to any improvements.

SystemId	Data	Quality	WE [%]	Improvement
S6	15 min	initial	49.9	13.2%
S7	15 min	good	43.3	
SystemId	**Data**	**#Speakers**	**WE [%]**	**Improvement**
S10	25 min	8	28.9	16.6%
S11	45 min	8	24.1	7.1%
S12	45 min	16	22.4	12.5%
S13	90 min	16	19.6	

Table 4: Quality and amount of adaptation data

In combination we finally reach 19.6% word error rate applying the PDTS method based on 90 minutes adaptation data (S13). This result compares to 19.0% word error rate of our golden line (S14) given a large Portuguese database of 16.5 hours training data. The complete adaptation procedures runs on a 300MHz SUN Ultra and takes only 3-5 hours real-time.

6. Summary and Conclusion

In this paper we addressed language independent and language adaptive acoustic modeling for read speech recognition using a high number of different languages. Provided that speech databases are limited in general, we approached the problem of porting acoustic models to a new target language by borrowing models and data from various languages but using only a limited amount of adaptation data from the target language. We explored the relative effectiveness of language independent acoustic models with a wider context in combination with a polyphone decision tree specialization (PDTS) method.

The PDTS method gave 12% relative improvement compared to a recalculation of a language specific polyphone tree and 28% compared to a non specialized multilingual polyphone tree. In summary, we achieved 19.6% word error rate when adapting language independent acoustic models to Portuguese using only 90 minutes of spoken speech. This compares to 19.0% of a full trained system on 16.5 hours Portuguese speech. As a consequence the introduced techniques allow to set up LVCSR systems in a new target language without the need of large speech databases in that language.

References

1. O. Andersen et al.: *Data-Driven identification of Poly- and Mono-phonemes for four European Languages*, Proc. Eurospeech, Berlin 1993.
2. J. Barnett et al.: *Multilingual Speech Recognition at Dragon Systems*, Proc. ICSLP, Philadelphia 1996.
3. P. Bonaventura et al.: *Multilingual Speech Recognition for Flexible Vocabularies*, Proc. Eurospeech, Rhodes 1997.
4. U. Bub et al.: *In-Service Adaptation of Multilingual Hidden-Markov-Models*, Proc. ICASSP, Munich 1997.
5. Billa, J. et al.: *Multilingual Speech Recognition: The 1996 Byblos Callhome System*, Proc. Eurospeech, Rhodes 1997.
6. P. Cohen et al.: *Towards a Universal Speech Recognizer for Multiple Languages*, Proc. Workshop on Automatic Speech Recognition and Understanding, St. Barbara 1997.
7. A. Constantinescu et al.: *On Cross-Language Experiments and Data-Driven Units for Automatic Language Independent Speech Processing*, Proc. Automatic Speech Recognition and Understanding, St. Barbara 1997.
8. Dugast, C. et al.: *The Philips Large-Vocabulary Recognition System for American English, French, and German*, Proc. Eurospeech, Madrid 1995.
9. Finke, M. et al.: *Wide Context Acoustic Modeling in Read vs. Spontaneous Speech*, Proc. ICASSP, Munich 1997.
10. Glass, J. et al.: *Multi-lingual Spoken Language Understanding in the MIT Voyager System*, Speech Communication 17, 1995.
11. IPA: *The International Phonetic Association (revised to 1993) - IPA Chart*, Journal of the International Phonetic Association 23, 1993.
12. J. Köhler: *Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks*, Proc. ICASSP, Seattle 1998.
13. Lamel, L. et al.: *Issues in Large Vocabulary Multilingual Speech Recognition*, Proc. Eurospeech, Madrid 1995.
14. Leggetter, C. et al.: *Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models.*, Computer Speech and Language 9, 1995.
15. Schultz, T.: *Multilinguale Spracherkennung: Kombination akustischer Modelle zur Portierung auf neue Sprachen.*, Dissertation, Universität Karlsruhe, Institut für Logik, Komplexität und Deduktionssysteme, 2000.
16. T. Schultz et al.: *Multilingual and Crosslingual Speech Recognition*, Proc. DARPA Workshop on Broadcast News Transcription and Understanding, Lansdowne, VA 1998.
17. T. Schultz et al.: *Language Independent and Language Adaptive LVCSR*, Proc. ICSLP98, Sydney 1998.
18. T. Schultz et al.: *The GlobalPhone Project: Multilingual LVCSR with Janus-3*, Proc. SQEL, 2nd Workshop on Multi-lingual Information Retrieval Dialogs, Plzeň 1997.
19. B. Wheatley et al.: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new Language*, Proc. ICASSP, Adelaide 1994.
20. Webster's, New Encyclopedic Dictionary. Black, Dog & Leventhal, 1992.
21. S.J. Young et al.: *Multilingual large vocabulary speech recognition: the European SQALE project*, Computer Speech and Language, 1997, vol 11.