Experiments towards a Multi-language LVCSR Interface

Tanja Schultz and Alex Waibel
Interactive Systems Labs
University of Karlsruhe and Carnegie Mellon University
{tanja,ahw}@ira.uka.de

Abstract

This paper describes experiments towards a multi-language human-computer speech interface. Our interface is designed for large vocabulary continuous speech input. For this purpose a multilingual dictation database has been collected under GlobalPhone, *which is a project at the Interactive Systems Labs. This project investigates LVCSR systems in 15 languages of the world, namely Arabic, Chinese, Croatian, English, French, German, Italian, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. Based on a global phoneme set we build different multilingual speech recognizer and present several performance results in language independent and language adaptive setups.*

1 Introduction

With the distribution of speech technology products all over the world, multi-language human-computer interfaces are of increasing importance. Beyond it, new methods for the fast adaptation to new target languages with only limited training data becomes a practical concern. Monolingual phoneme sets are applied for cross-language adaptation [9] and also multilingual phonemic inventories has been demonstrated to give satisfactory results [7] within the same language family [3], [6], and limited tasks [4]. The focus of our work is the design of a multi-language interface for large vocabulary continuous speech which covers the most widespread and important languages of the world. Since one major limitation in developing recognition systems is the need of large training data, this work also explore the relative effectiveness of multilingual context dependent acoustic model combination for cross-language adaptation with limited training data. Furthermore we compare different approaches to adapt pronunciation dictionaries for the purpose of cross-language adaptation.

For all experiments we use our multilingual database GlobalPhone which is briefly introduced in the first section of this paper. In the second part, we describe the monolingual baseline systems trained and tested with this database. After that the global phoneme set and the multilingual acoustic model combination is introduced. In the last part of this paper, we address the problem of cross-language adaptation. Several recognition results will be presented in language independent and language adaptive setups.

Language	Speakers	Spoken units	Hours
Arabic	100	180,000	25
Ch-Mandarin	132	250,000	25.5
Ch-Shanghai	50	80,000	8
Croatian	98	130,000	18
Japanese	128	430,000	28
Korean	100	370,000	20
Portuguese	120	180,000	22.7
Russian	140	250,000	27.3
Spanish	100	200,000	22
Swedish	100	200,000	20
Tamil	50	-	10
Turkish	100	140,000	16.9

Table 1: The GlobalPhone database

2 The GlobalPhone project

The aim of the project GlobalPhone is the development of a multi-language human-computer speech interface for large vocabulary. For this purpose we recently started the collection of a large multilingual speech database which currently consists of the languages Arabic, Chinese (Mandarin and Shanghai dialect), Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil and Turkish. Considering the fact that German, English, and French are available in similar frameworks, we are able to cover 9 of the 12 most widespread languages of the world. In each language about 100 native speakers were asked to read 20 minutes of political and economic articles from a national newspaper. Their speech was recorded in office quality, with a close-talking microphone. The GlobalPhone corpus is fully transcribed including spontaneous effects like false

Language	Performance [ER]
Chinese	18.4%
Croatian	20.0%
Japanese	10.0%
Korean	47.3%
Spanish	20.0%
Turkish	16.9%

Table 2: Error Rates [ER] of monolingual systems

starts and hesitations. Up to now we collected 233 hours of spoken speech from about 1300 speakers in total. Further details of the GlobalPhone project are given in [8].

Table 1 summarizes the number of speakers, spoken units and the hours of recorded speech for the GlobalPhone database. Based on these data we train and test context-dependent models in six languages and context-independent in eight languages. The test sets consist of 100 utterances per language, the language adaptive experiments are evaluated on 200 German utterances. Because of the limited corpus size, we are not able to estimate reliable LVCSR n-gram models and vocabularies, which results in high out-of-vocabulary rates. Since we focus here on the multilingual acoustic modeling and compare error rates across languages, we reduced the OOV-rate to 0.0% by including all test words into the language model as monograms with small probabilities. We defined a 10K test dictionary by supplementing the test words with the most frequently seen training units.

3 Monolingual Baseline Systems

In the first step towards a multi-language interface we developed monolingual baseline systems in eight languages applying our fast crosslingual bootstrap technique [7]. For six languages Chinese, Croatian, Japanese, Korean, Spanish, and Turkish the resulting LVCSR recognizers consist of fully continuous 3-state HMMs with 1500 polyphone models. Each HMM-state is modeled by one codebook with a mixture of 16 Gaussian distributions. The preprocessing is based on 13 Mel cepstral coefficients with first and second order derivatives, power and zero crossing rate. After cepstral mean subtraction, a linear discriminant analysis is used to reduce the input to 24 dimensions. Table 2 shows the error rates for all languages. The systems performance ranges from 10% kana error rate for Japanese to 16.9% word error rate for Turkish, 18.4% pinyin error rate for Chinese, and 20% word error rate for Spanish and Croatian. The Korean performance is given in hangul syllables and achieves 47% error rate. For Portuguese and Russian so far only preliminary context independent systems have been developed. Their recognizers consist of 3-state HMMs with 53 and 34 monophone models. Each HMM-state is modeled by 32 Gaussian. The preprocessing is the same as in the context dependent counterparts.

4 Multi-language Systems

For the development of multi-language systems it is of great concern to combine the phonetic inventory of all languages to be recognized into one global acoustic model pool. Such a multilingual acoustic model combination leads to the following benefits:

1. Data sharing across languages reduces the total number of parameters in the system

2. One language independent acoustic model reduces the complexity of the multi-language interface compared to several language specific acoustic models

3. Data sharing results in a more reliable model estimation, especially for less frequent phonemes

4. Global phoneme pools allow a more accurate pronunciation modeling of -out of language- words, i.e. foreign proper names or brand names

5. Robust acoustic models enables a fast and efficient cross-language bootstrapping of systems in new languages even if only limited or no training data is available

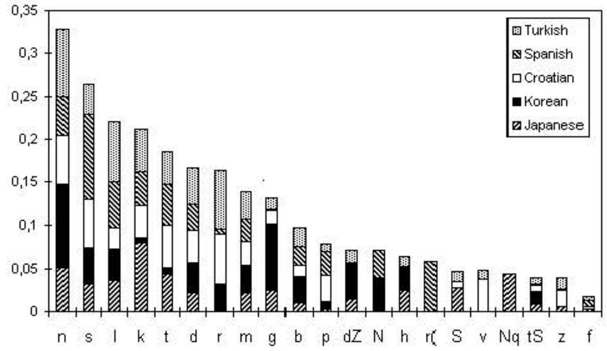

Figure 1: Relative frequencies of consonants

Figure 1 illustrates the relative frequencies of consonants for the five languages Croatian, Japanese, Korean, Spanish, and Turkish in our GlobalPhone training database (phoneme names are given in Worldbet notation). As can be seen the occurrences of sounds are highly dependent on the language. Some sounds are comparable frequent in all five languages like the

Phonemes [Worldbet]	KO	SP	CR	TU	JA	$\sum$
n,m,s,l,tS,p,b,t,d,g,k	X	X	X	X	X	
i,e,o	X	X	X	X	X	14
f,j,z		X	X	X	X	
r,u	X	X	X	X		
dZ	X		X	X	X	6
a	X	X	X			
S			X	X	X	
h	X			X	X	
4	X	X			X	4
$\tilde{n}$,x,L		X	X			
A				X	X	
N	X	X				
V,Z			X	X		
y,7	X			X		
ts			X		X	10
p',t',k',dZ',s',oE,oa,4i,	X					
E,uE,∧,i∧,u∧,	X					
iu,ie,io,ia	X					17
D,G,T,V,r(,ai,au,ei,		X				
eu,oi,a+,e+,i+,o+,u+		X				15
palatal c, palatal d			X			2
ix, soft				X		2
?,Nq,V[,A:,e:,i:,o:,4:					X	8
Monolingual $\sum$ 170	40	40	30	29	31	
Multilingual						78

Table 3: Global Phoneme Set [Worldbet notation]

sounds [n], [d], [m], and [b], whereas others are not,as for example the phoneme [g], which is frequent in Korean but extremel rare in Spanish. In the first case sharing the data results in language independent robust models of [n], [d], [m], and [b] and reduces the number of parameter of the final s stem. In the latter case the less frequent phonemes like the Spanish [g] would benefit from sharing the training data across the languages, since more reliable estimation for this model can be achieved. Sounds belonging to onl one language like [Nq] for Japanese or the flapped [r(] for Spanish help solving the language identification problem because these sounds are reliable predictors for one language.

4.1 Global Phoneme Set

Combining the phonetic inventor across languages into one *global phoneme set* requires the definition of similarities between sounds. Those similarities are documented in international phonemic inventories like Sampa, Worldbet, or IPA [5], which classif sounds based on phonetic knowledge. On the other hand data-driven methods are proposed for example in [1] and [3]. Previous s stems have been limited to context independent modeling. For the monolingual case context dependent modeling is proven to increase recognition performance significantl . Such improvements from context dependence extend naturall to the multilingual setting, but the use of context dependent models raises the question of how to construct a robust, compact, and efficient multilingual model set. In this paper we introduce a data-driven procedure for multilingual context dependent models. Based on the phonetic inventor of eight monolingual s stems we defined a global phoneme set. Sounds which are represented b the same IPA s mbol share one common phoneme categor . For eight languages this global set consists of 145 phoneme categories. Table 3 shows the global phoneme set for five languages in Worldbet notation. About half of the set consists of monophonemes belonging to onl one language, the other half is shared across at least two languages. Silence and the noise models are shared across all languages.

4.2 Acoustic model combination

Based on the above described categories we designed different multilingual s stems b combining language dependent acoustic models in different wa s. In the s stem *ML-mix* we share all models across languages without preserving an information about the language. For each categor model we initialize one mixture of 16 Gaussian distributions and train the models b sharing the data of five languages (*ML5-mix*), seven language (*ML7-mix*) and eight languages (*ML8-mix*) respectivel . For the *ML5-mix* s stem we create context dependent phoneme models b appl - ing a decision tree clustering procedure which uses an entrop based distance measure, defined over the mixture weights of the codebooks, and a question set which consists of linguisticall motivated questions about the phonetic context of a phoneme model. During clustering, the question with the highest entrop gain is selected when splitting the tree node according to this question. After reaching the predefined number of 3000 pol phones the splitting procedure ends.

Another wa to share phoneme models across languages is performed in the multilingual s stem *ML-tag*. Here each of the phoneme categories gets a language tag attached in order to preserve the information about the language. The above described clustering procedure is extended b introducing questions about the language and language groups to which a phoneme belongs. Therefore the decision if phonetic context information is more important than language information becomes data-driven. We started with 250,000 different quintphones over the five languages and created two full continuous s stems, one s stem *ML5-tag3* with 3000 models, and the second s stem *ML5-tag75* with 7500 models, which is of same size as five monolingual s stems with 1500 models each.

4.3 Recognition Results

We explore the usefulness of our modeling approach by comparing the performance of the multilingual systems for the five languages Croatian, Japanese, Korean, Spanish, and Turkish as given in figure 2.

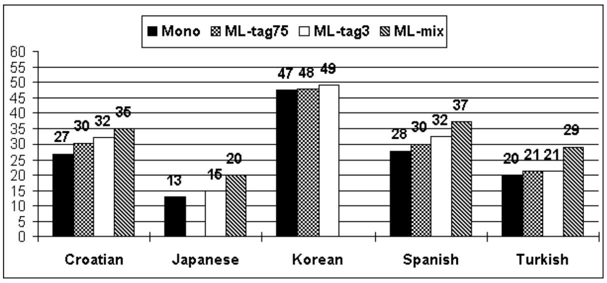

Figure 2: Results for multilingual setup [Word Error]

The experiments are twofold: first we explore which sharing method performs best, and second we examine the profit of sharing the acoustic parameters. The system architecture, the preprocessing and the training procedure are identical throughout this tests. To answer the first question we compare the performance of the multilingual system *ML5-tag3* to *ML5-mix* for all languages. Figure 2 shows that the tagged system outperforms the mixed system significantly in all languages by 5.3% error rate (3.1% - 8.7%). This indicates that preserving the language information and introducing questions about languages achieves significant improvements with respect to monolingual recognition.

To answer the second question we varied the number of polyphones modeled in the best multilingual system *ML5-tag*. In *ML5-tag3* the model number is reduced to 40% of the monolingual systems (3000 vs 5x1500), which leads to 3.14% (1.2% - 5.0%) performance degradation in average. But not all of the degradation can be explained by the reduced model number as the comparison with *ML5-tag75* shows. This system is of same model size like the 5 monolingual systems, but we still observe an average performance gap of 1.07% (0.3% - 2.4%). This finding is coincident to other studies [3], [6], and [2]. We therefore draw the conclusion that so far sharing data across languages decreases the performance with respect to monolingual speech recognition.

5 Language Information

In this section we intend to investigate the pertinence of language information coded in the acoustic models. We analyse the ratio of language questions compared to phonetic questions as well as the language information rate of polyphone models.

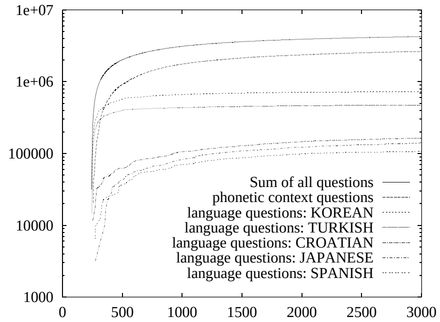

Figure 3: Importance of Language Questions

For the purpose of pertinence we computed the sum of entropy gain from the clustering procedure and plotted it over the number of splitted polyphones in figure 3. The curve *"sum of all questions"* gives the overall entropy gain of all questions asked during the clustering procedure, whereas the curve *"phonetic context questions"* shows the entropy gain belonging to non-language questions. The gap between both curves indicates that major parts of the entropy gain results from language questions. The remaining five curves give the contribution of questions belonging to only one language. It is shown that questions about Korean and Turkish are more important than about other languages, especially in the beginning of clustering. This indicates that sounds in those two languages seems to be different from the rest. Both results demonstrate that language questions are frequently asked and are especially in the beginning more important than questions about the phonetic context of a phoneme. It is also evident that the data-driven decision does not reflect the IPA-based classification across languages.

In table 4 we compile the detailed list of asked questions ranked by frequency, after clustering 500, 1500, and 3000 polyphone models. The highly frequent occurrence of the question about the language group Korean+Turkish sustains the above findings. Also the decreasing importance of language questions towards the end of splitting process can be seen from comparing column "500 models" to "3000 models".

Second, we want to analyze the language information rate of the resulting polyphone models. For this purpose we computed the language distribution for each split node as pictured in figure 4. We replaced the Gaussians distributions in the existing polyphone cluster tree by these language distributions and recalculated the entropy based distance. The cumulated distance is plotted over the number of nodes in figure

#	500 model	#	1500 model	#	3000 model
76	KO+TU	100	KO+TU	146	wordbound
38	KOREAN	73	KOREAN	131	back-vow
30	front-vow	73	back-vow	130	front-vow
27	back-vow	65	front-vow	128	consonant
23	vowel	61	wordbound	113	KO+TU
22	unvoiced	53	consonant	98	KOREAN
20	silence	48	unvoiced	97	voiced
19	fric-sibil	48	alveodental	90	vowel
16	wordbound	46	vowel	88	unvoiced
14	nasal	42	voiced	85	nasal
10	voiced	42	nasal	84	alveodental
10	round	36	silence	79	JAPAN.
10	JAPAN.	36	plos-unvoic	63	plos-unvoic
10	consonant	35	frik-sibil	59	frik-sibil
9	plos-unvoic	32	JAPAN.	59	close-vow
9	open-vow	29	round	56	silence
9	CR+JA+SP	28	plosive	55	round

Table 4: Prominence of asked questions

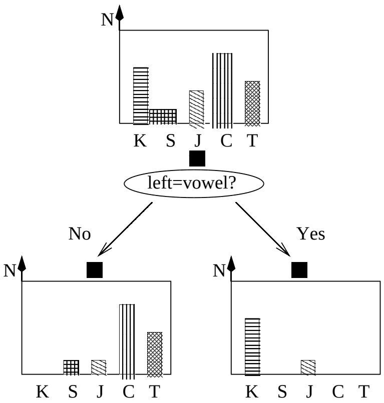

Figure 4: Language distribution of tree node

5. The most important finding is that main parts of the language information are clustered out after about 3000 splits, which means that in our case the multilingual system above 3000 polyphone models consists of mostly monolingual acoustic models.

6 Cross-language Adaptation

In the previous sections we examined the usefulness of multilingual acoustic modeling with respect to monolingual speech recognition. In this section we investigate the feasibility of the multi-language interface when applied to cross-language transfer, i.e. the adaptation to a new unseen input languages, in this case to the German language. For adaptation we used up to 14000 words (1000 utterances) spoken by 13 native German speaker, for testing 2500 words (200 utterances) spoken by 3 speakers. We performed two iteration of Viterbi to adapt to the target language and do not re-cluster the polyphone trees, but simply training the Gaussian and mixture weights of the language dependent models. The German baseline system achieves 15.8% word error rate tested on a 60k-dictionary.

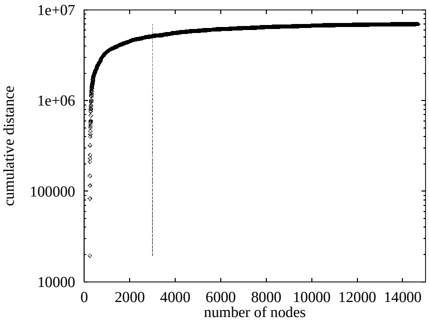

Figure 5: Language information rate

For recognizing a new target language we need a pronunciation dictionary suitable for this language in terms of the phoneme model set of the bootstrap engine. We investigate two different approaches for adaptation of target pronunciation dictionaries: the data-driven and the IPA-based approach, and compare them by running recognition tests using the resulting dictionaries. For our dictionary adaptation approaches we presuppose that either phonetic labels of a limited amount of data or a pronunciation dictionary in an arbitrary phoneme set is available. If none of them is given [4] introduced an algorithm which achieves promising results using the MMI-based criterion to initialize a phonemic representation and improve this representation iteratively applying a genetic algorithm. However this approach requires an isolated word task and thus is applicable to connected speech only if at least word labels are available.

In the data-driven approach (data-driven) we are running a phoneme recognizer of the bootstrap language to decode utterances spoken in the target language. The resulting hypotheses are than compared frame-wise to the reference phoneme string. A phoneme similarity matrix is calculated and every target phoneme is replaced by the counterpart given the highest frame confusion frequency. In the heuristic IPA-based approach (IPA-ML), the target language phoneme is related to that phoneme of the bootstrap set which is assigned to the same symbol in the IPA reference scheme. If no counterpart can be found that phoneme is chosen, which is as close as possible to the target phoneme in terms of the IPA classification. If we are using our five-lingual systems for bootstrapping a new language each sound can have up to five counterparts, one in each language. We explore dictionaries with different numbers of counterparts. In the

IPA-5L dictionary the decision for the best matching phoneme is left to the decoder by including 5 language dependent pronunciation variants, one variant for each language involved in the model combination.

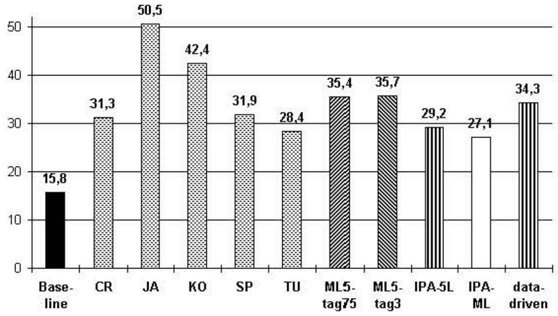

Figure 6: Cross-Language Performance

Throughout our recognition experiments we explore four questions: first we analyze the effect of cross-language transfer from monolingual versus multilingual models by comparing the *ML5-mix* systems to the monolingual systems. Second we investigate the usefulness of the different model combination (*ML5-tag3* vs *ML5-mix*), third we examine the effect of different parameter size (*ML5-tag75* vs *ML5-tag3*) and last but not least we compare the data-driven versus the IPA-based dictionary approach for the *ML5-mix* system. Figure 6 summarizes the results of the recognition tests (the three rightmost bars belong to the ML5-mix system).

One of the major outcome is that the multilingual system outperforms all monolingual systems. The average performance of the monolingual systems is 36.4% word error rate (47.4% - 28.4%), versus 27.1% for *ML5-mix*. From this result we can conclude that language transfer from multilingual acoustic models achieves better results, especially if few or nothing is known about the new language. Second, with regard to cross-language transfer the *ML5-mix* system outperforms the *ML5-tag* system. This indicates, that dedicated multilingual systems should be developed depending on whether cross- or multilingual speech recognition is projected. In the first case the *ML5-mix* system should be favored, in the latter the *ML5-tag* system. Third, increasing the number of model parameter do not improve the performance (*ML5-tag75* vs *ML5-tag3*) significantly. Finally we found from the experiments that the data-driven dictionary approach is clearly outperformed by the heuristic IPA-based approach.

7 Conclusion

In this paper a multi-language speech interface for up to eight languages, namely Chinese, Croatian, Japanese, Korean, Portuguese, Russian, Spanish, and Turkish is presented. To create multilingual context dependent acoustic models we evaluated different methods of parameter sharing. The resulting systems have been applied in language independent and language adaptive setups.

References

[1] O. Andersen, P. Dalsgaard, and W. Barry: *Data-Driven identification of Poly- and Mono-phonemes for four European Languages* in: Proc. Eurospeech, pp. 759-762, Berlin 1993.

[2] P. Bonaventura, F. Gallocchio, and G. Micca: *Multilingual Speech Recognition for Flexible Vocabularies* in: Proc. Eurospeech, pp. 355-358, Rhodes 1997.

[3] P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, T. Ward: *Towards a Universal Speech Recognizer for Multiple Languages* in: Proc. Workshop on Automatic Speech Recognition and Understanding, pp 591-597, St. Barbara 1997.

[4] A. Constantinescu, and G. Chollet: *On Cross-Language Experiments and Data-Driven Units for Automatic Language Independent Speech Processing* in: Proc. Automatic Speech Recognition and Understanding, pp. 606-613, St. Barbara 1997.

[5] The IPA 1989 Kiel Convention. in: Journal of the International Phonetic Association 1989(19) pp. 67-82

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

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

[8] T. Schultz, and A. Waibel: *Language Independent and Language Adaptive LVCSR* in: Proc. ICSLP98, Sydney 1998.

[9] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language* in: Proc. ICASSP, pp. 237-240, Adelaide 1994.