Language Independent and Language Adaptive Large Vocabulary Speech Recognition

T. Schultz and A. Waibel

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

ABSTRACT

This paper describes the design of a multilingual speech recognizer using an LVCSR dictation database which has been collected under the project GlobalPhone. This project at the University of Karlsruhe 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 built different multilingual speech recognition systems for five of the 15 languages. Context dependent phoneme models are created data-driven by introducing questions about language and language groups to our polyphone clustering procedure. We apply the resulting multilingual models to unseen languages and present several recognition results in language independent and language adaptive setups.

1. Introduction

As the demand for speech recognition systems in multiple languages grows, the development of multilingual systems which combine the phonetic inventory of many languages into one single acoustic model set is of increasing importance. The benefits of such an approach are:

1. Reduced complexity of systems by sharing models and parameters, adressed for example in [1]
2. Integrated language identification as for example described in [2] and [3]
3. Bootstrapping systems for *unseen* languages with limited adaptation data [4], [5], [6].

Combining acoustic models requires the definition of multilingual phonetic inventories. Previous systems have been limited to context independent modeling. For the monolingual case context dependent modeling is proven to increase recognition performance significantly. Such improvements from context dependence extend naturally 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. By applying a decision tree based clustering procedure we trained three context dependent systems which share their parameters in different ways. For one system we add language questions and afterwards analyze the resulting decision tree.

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 systems trained with this database. The multilingual acoustic modeling is introduced in the next section. In the last two sections we present results in monolingual, multilingual, and crosslingual setups based on the systems created.

Language	Utterances	Speakers	Spoken units
Chinese	5124	77	150,000
Croatian	2826	62	80,000
Japanese	5641	62	200,000
Korean	1587	22	140,000
Turkish	5371	82	112,000
Spanish	5455	79	160,000
German	1000	3	14,000

Table 1: GlobalPhone data used for training

2. The GlobalPhone Database

For the development of multilingual recognition systems, we have been collecting the GlobalPhone database which currently consists of the languages Arabic, Chinese (Mandarin and Wu), Croatian, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil and Turkish. 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 corpus is fully transcribed including spontaneous speech effects. Up to now we collected 233 hours of spoken speech from about 1300 speakers in total. Further details about the GlobalPhone project are given in [7].

Table 1 shows the part of the GlobalPhone database used for training. The monolingual and multilingual test sets consist of 100 utterances per language, the crosslingual 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 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 Systems

We developed monolingual LVCSR systems applying our fast crosslingual bootstrap technique [6] to initialize the not yet modeled languages. In each language the resulting baseline engine consists of a fully continuous 3-state HMM system with 1500 polyphone models. Each HMM-state is modeled by one codebook which contains 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 ceptral mean substraction, a linear discriminant analysis is used to reduce the input to 24 dimensions.

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 currently best monolingual systems

Table 2 shows the performance in error rates achieved by our currently best monolingual systems. The results for Chinese are given in terms of pinyin units, for Japanese in terms of hiragana words, and for the Korean language in morpheme based syllables.

4. Language Independent Speech Recognition

For multilingual speech recognition we intend to share acoustic models of similar sounds across languages. Similarities of sounds are documented in international phonemic inventories like Sampa, Worldbet, or IPA [8], which classify sounds based on phonetic knowledge. On the other hand data-driven methods are proposed for example in [9]. In this paper we introduce a data-driven procedure for multilingual context dependent acoustic modeling.

4.1. Global Phoneme Set

Based on the phonetic inventory of five monolingual systems we defined a *global phoneme set* for the languages Croatian, Japanese, Korean, Spanish and Turkish. Sounds which are represented by the same IPA symbol share one common phoneme categorie. The resulting set is shown in table 3 in Worldbet notation. Altogether it consists of 78 phonemes plus a silence and two noise models for spontaneous speech effects. 14 phonemes are shared across all five languages, but half of the set consists of mono-phonemes belonging to only one of the five languages.

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					
uE,E,∧,i∧,u∧,iu,ie,io,ia	X					17
D,G,T,V,r(,ai,au,ei,eu,oi		X				
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]

4.2. Multilingual Acoustic Modeling

Based on these 78 phoneme categories, we build three different multilingual systems: *ML5-mix*, *ML5-sep*, and *ML5-tag*. In the first one we share all models across languages without preserving any information about the language. For each of the 78 phonemes we initialize one mixture of 16 Gaussian distributions and train the models by sharing the data of all five languages. The resulting recognizer *ML5-mix* is a fully continuous system with 3000 models mixed over all languages. In the second multilingual system *ML5-sep* each element is modeled separately for each language. No data are shared, all models except silence and noise are language dependent. For each of the 170 phonemes we initialize one mixture of 16 Gaussian distributions, after training this results in a fully continuous system with 3000 language dependent models. In the third multilingual system *ML5-tag* we attached a language tag to each of the 78 phoneme categories in order to preserve the language information.

To achieve context dependent phoneme models we apply a decision tree clustering procedure which uses an entropy based distance measure, defined over the mixture weights of Gaussians, and a question set which consists of linguistically motivated questions about the phonetic context of a phoneme model. During clustering, the question with the highest entropy gain is selected when splitting the tree node according to this question. After reaching the predefined number of polyphones the splitting procedure ends. We extended this clustering routine to the multilingual case by introducing questions about the language and language groups to which a phoneme belongs. Therefore the decision whether phonetic context information is more important than language information becomes data driven. We started with 250,000 quintphones over the five different languages and created two fully continuous systems, *ML5-tag3* with 3000 models and *ML5-tag75* with 7500 models which is exactly the same size as the five monolingual systems (5x1500).

4.3. Analysis of Language Questions

Before reporting recognition results using the multilingual systems we intend to describe the pertinence of language questions compared to phonetic questions as well as the language information rate of polyphone models.

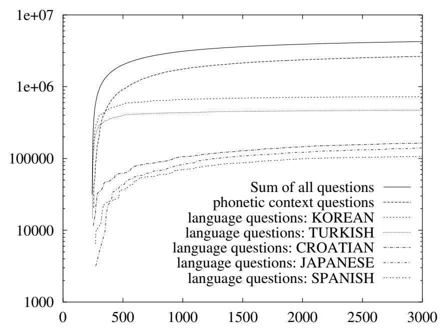

Figure 1: Importance of Language Questions

#	500 models	#	1500 models	#	3000 models
76	KO+TU	100	KO+TU	146	word bound
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	word bound	113	KO+TU
22	unvoiced	53	consonant	98	KOREAN
20	silence	48	unvoiced	97	voiced
19	fric-sibil	48	alveodental	90	vowel
16	word bound	46	vowel	88	unvoiced
14	nasal	42	voiced	85	nasal
10	voiced	42	nasal	84	alveodental
10	round	36	silence	79	JAPANESE
10	JAPANESE	36	plos-unvoic	63	plos-unvoic
10	consonant	35	frik-sibil	59	frik-sibil
9	plos-unvoic	32	JAPANESE	59	close-vow
9	open-vow	29	round	56	silence
9	CR+JA+SP	28	plosive	55	round

Table 4: Prominence of asked questions

For the purpose of pertinence we computed the sum of entropy gain and plotted it over the number of splitted polyphones in figure 1. 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 are definetely 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 occurence 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".

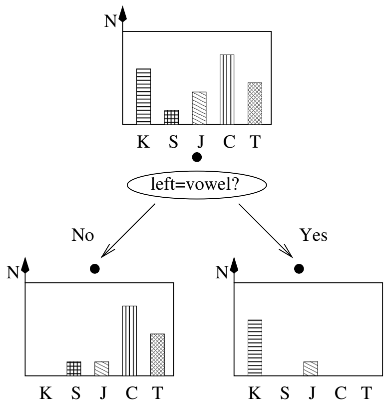

Figure 2: Language distribution of tree node

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 2. 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 3. The most important finding is that most parts of language information are clustered out after about 3000 splits, which means that a multilingual system with 3000 polyphone models and more consists of mostly monolingual acoustic models.

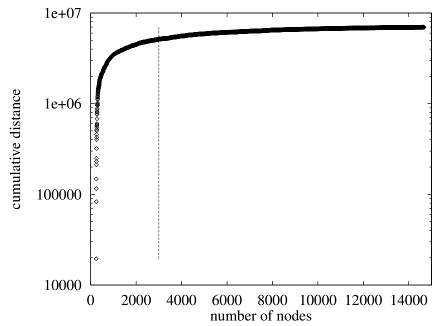

Figure 3: Language information rate of clustered polyphones

4.4. Multilingual Experiments

The following multilingual 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 4 shows that the tagged system outperformes the mixed system significantly by 5.3% error rate (3.1% - 8.7%). This indicates that preserving the language information and introducing questions about languages and language groups leads to significant improvements in the multilingual setup.

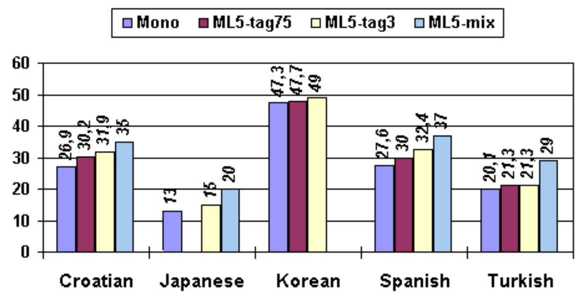

Figure 4: Results for multilingual setup [Error Rate]

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 in average to 3.14% (1.2% - 5.0%) performance degradation. 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%). We therefore conclude that language independent modeling decreases the model precision for recognition of seen languages which coincident to other studies e.g. [1].

5. Language Adaptive Speech Recognition

In this section we investigate the multilingual systems' performance when applied to recognize *unseen* languages. Our goal is to recognize German spoken speech. Experiments with and without language adaptation are performed. For adaptation we used up to 1000 utterances, for testing 200 utterances of 3 speakers from our GlobalPhone database. The German baseline system achieves 15.8% word error rate tested on a 60k-dictionary. For our experiments we presume that the German pronunciation dictionary is given.

5.1. Dictionary Adaptation

For recognizing unseen languages we need to define an appropriate mapping from our global phoneme set to the target phonemes. In our experiments we replaced the German phonemes by the corresponding IPA-based phoneme category. Since the global phoneme set contains models from five languages, a German sound can have up to five counterparts. In the first experiment we therefore explore different pronunciation dictionaries: Three dictionaries where the German phonemes are mapped to language dependent phonemes (Japanese, Spanish, and Turkish) and a 5-lingual dictionary containing the pronunciation variants of all five languages. In the 5-lingual case the decision for the best matching pronunciation variant is left to the decoder. We choose the system *ML5-tag3* to compare the four dictionaries because it performs best in the multilingual setup. It achieves 50% word error rate on the 5-lingual dictionary which clearly outperformes the Japanese (65.0%) and the Spanish (59.5%) but not the Turkish (49.5%) one. The results can be explained by the fact that Japanese phonotactic does not cover the German one because of its mora structure; Turkish tends to have long words and fits better into the German phonotactic whereas the 5-lingual dictionary has 5 times more entries which leads to higher confusion. We looked into the pronunciation variants used by the decoder, and found that Spanish models are preferred for short function words, which might result from the fact that 20% of the Spanish corpus consists of 2 phoneme long words.

Second we compared the multilingual systems based on the 5-lingual dictionary to each other. Surprisingly *ML5-mix* performed best with 41.5% word error rate. It outperforms *ML5-sep* which gave lowest performance (53%), also *ML5-tag3* (50.0% error rate) and even *ML5-tag75* which achieved 47.5% word error rate. *ML5-tag75* has more than twice as many models as *ML5-mix* and gave best results on the multilingual setup. This indicates that it is useful to develop dedicated multilingual systems depending on whether multilingual or crosslingual speech recognition tests are projected.

5.2. Crosslingual Training

We intend to adapt the recognizers to the German data. For this purpose we took the system *ML5-mix* and train it on 1500, 7500, and 14000 German spoken words. Two iterations Viterbi training are performed. The same procedure was applied to the monolingual recognizers. Figure 5 compares the performance of the monolingual systems to *ML5-mix* for different amount of adaptation data. Obviously the word error rate decreases during training the acoustic models on German speech but in case of the monolingual systems no further significant improvement seems to be achieved by adding more data. Since we do not recompute the polyphone tree, the remaining gap between the best crosslingual result and the German baseline turned out to be reasonable. The major outcome is that the multilingual system outperformes 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*. We therefore can conclude that bootstrapping from a multilingual recognition system achieves better results, especially if nothing is known about the new unseen language.

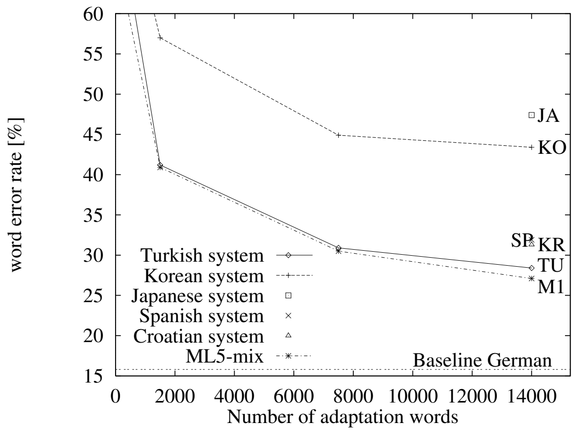

Figure 5: Results for language adaptation [Word Error]

6. Conclusion

In this paper multilingual LVCSR systems for five languages, namely Croatian, Japanese, Korean, Spanish, and Turkish are presented. To create multilingual context dependent acoustic models we evaluated different methods of parameter sharing, among other things questions about languages and language groups are introduced. We applied the trained systems to monolingual, multilingual and crosslingual setups. The results indicate that the method of parameter sharing shoud be decided depending on whether multilingual or crosslingual speech recognition is projected.

References

1. P. Cohen et al: *Towards a Universal Speech Recognizer for Multiple Languages* in: Proc. Automatic Speech Recognition and Understanding, pp. 591-598, St. Barbara 1997.
2. C. Corredor-Ardoy et al: *Language Identification with Language-independent Acoustic Models* in: Proc. Eurospeech, pp. 355-358, Rhodes 1997.
3. O. Andersen et al: *Language Identification based on Cross-language Acoustic Models and Optimized Information Combination* in: Proc. Eurospeech, pp. 67-70, Rhodes 1997.
4. B. Wheatley et al: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language* in: Proc. ICASSP, pp. 237-240, Adelaide 1994.
5. J. Köhler: *Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks* in: Proc. ICASSP, pp 417-420, Seattle 1998.
6. T. Schultz et al: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets* in: Proc. Eurospeech, pp. 371-374, Rhodes 1997.
7. T. Schultz et al: *The GlobalPhone Project: Multilingual LVCSR with Janus-3* in: Proc. SQEL, pp. 20-27, Plzeň 1997.
8. The IPA 1989 Kiel Convention. in: Journal of the International Phonetic Association 1989(19) pp. 67-82
9. O. Andersen et al: *Data-Driven identification of Poly- and Mono-phonemes for four European Languages* in: Proc. Eurospeech, pp. 759-762, Berlin 1993.