Development of Multilingual Acoustic Models in the GlobalPhone Project

Tanja Schultz and Alex Waibel

Interactive Systems Laboratories
Universität Karlsruhe, Carnegie Mellon University
{tanja,ahw}@ira.uka.de

Abstract. This paper describes our recent effort in developing the GlobalPhone recognizer for multilingual large vocabulary continuous speech. This project investigates LVCSR systems in 15 languages, namely Arabic, Chinese (Mandarin and Wu), Croatian, English, French, German, Japanese, Korean, Portuguese, Russian, Spanish, Swedish, Tamil, and Turkish. Based on five languages we developed a global phoneme set and built multilingual speech recognizer by variing the method of acoustic model combination. Context dependent phoneme models are created using questions about languages and language groups. Results of a multilingual system which can handle five languages are presented.

1 Introduction

As the demand for speech recognition systems in multiple languages grows, the development of multilingual systems which combine the phonetic inventory of all 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
2. Integrated language identification as for example described in [1] and [2]
3. Bootstrapping systems for new languages with limited adaptation data [3],[4].

Combining acoustic models requires the definition of multilingual phonetic inventories. Previous systems with combined acoustic phonetic models 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. We apply a polyphon tree based clustering procedure to make the decision of model sharing across languages data driven. Questions about language and language groups are introduced to the linguistic motivated question set and the resulting decision tree is analysed.

For our experiments we use the GlobalPhone database which is briefly described in the first section of this paper. In the second part we introduce the multilingual acoustic context modeling. The last section gives results for the monolingual and different multilingual speech engines.

2 The GlobalPhone Database

For the development of the multilingual recognition systems we use our recently collected GlobalPhone database. The current status of the database is described in table 1. In each language about 100 speakers were asked to read 20 minutes of political and economic articles from a national newspaper in their native language. Their speech was recorded in office quality, with a close-talking microphone. Up to now we collected 233 hours of spoken speech from about 1300 speakers in total. The (+spk)-sign in the table indicates that the data collection for this language is still ongoing. The corpus is fully transcribed including spontaneous effects like false starts and hesitations. Further details of the GlobalPhone database are given in [5].

Language	#speakers	#utterances	speech [hrs]	#words	vocabulary
Arabic	100	u.w.	20	u.w.	u.w.
Croatian	(+30) 83	4019	14	106K	20K
Japanese	121	9785	25	204K	23K
Korean	100	6868	18	80K	40K
Mandarin	132	9103	28	250K	12K
Portuguese	(+50) 74	6726	17	126K	6K
Russian	(+50)100	10229	20	155K	22K
Schwedish	100	u.w.	20	u.w.	u.w.
Spanish	100	6866	22	176K	21K
Tamil	49	u.w.	12	u.w.	u.w.
Turkish	100	6872	17	112K	16K
Wu	40	3000	10	80K	8K
German	(+80) 19	3300	10	47K	10K

Table 1. The GlobalPhone Database (u.w.=under way)

For training and testing we use Croatian, Japanese, Korean, Spanish, and Turkish data. The test set consists of 100 utterances per language. Because of the currently limited corpus size of about 80K to 250K spoken words, we are not able to estimate reliable LVCSR n-gram models. This results in high out-of-vocabulary rates. Since we focus on acoustic modeling here and want to make word error rates comparable across languages, we restricted the OOV-rate in the case of Croatian, Korean, Spanish, and Turkish to 0.0% by including all test words into the language model as monograms with small probabilities. In these languages we defined a 10K test dictionary by supplementing the test word set with the most frequently seen training words.

3 Multilingual Acoustic Modeling

For multilingual speech recognition we intend to share acoustic models of similar sounds across languages. Similarities of sounds are documented in international phonetic inventories like Sampa, Worldbet, or IPA [6], which classify sounds

based on phonetic knowledge. On the other hand data-driven methods are proposed for example in [7], [8]. In our work we introduce a data-driven procedure for multilingual context dependent modeling.

3.1 Monolingual Baseline Systems

For baseline we developed five monolingual LVCSR systems applying our fast crosslingual bootstrap technique. For each language the resulting 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 with a 24 dimensional feature space. 16 cepstra, power, and their first and second derivatives are calculated from the 16kHz sampled input speech. Mean subtraction is applied. The number of features is reduced to 24 coefficients by computing a linear discriminant analysis.

3.2 Global Phoneme Set

Phonemes [Worldbet]	KO	SP	CR	TU	JA	$\sum$
n,m,s,l,tS,p,b,t,d,g,k,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,uE,E,Λ,iΛ,uΛ,iu,ie,io,ia	X					17
D,G,T,V,r(,ai,au,ei,eu,oi,a+,e+,i+,o+,u+		X				15
palatal c, palatal d			X			2
ix, weichzeichen				X		2
?,Nq,V[,A:,e:,i:,o:,4:					X	8
Monolingual phonemes $\sum = 170$	40	40	30	29	31	
Multilingual phonemes						78

Table 2. IPA-based phoneme classes [Worldbet notation]

Based on the phonetic inventory of the monolingual systems we defined a *global phoneme set* for the five languages Croatian, Japanese, Korean, Spanish and Turkish. Sounds which are represented by the same IPA symbol share one

common phoneme class. The resulting phoneme set is shown in table 2 in Worldbet notation. Over all it consists of 78 phonemes plus a silence and two noise models for spontaneous speech effects. 14 models are shared across all five languages, but half of the set consists of mono-phonemes belonging to only one of the five languages.

3.3 Multilingual Context Modeling

Using the global phoneme set we built two different multilingual systems. In the first system *ML5-mix* we combined all acoustic models of the same phoneme class without preserving any information about to which language the model belongs. The context independent base system therefore consists of 78 language independent phoneme models plus 3 silence/noise models. In the second multilingual system *ML5-tag* each of the 78 phonemes get a language tag attached in order to preserve the language information, which results in a base system with 170 language dependent phoneme models plus 3 silence/noise models.

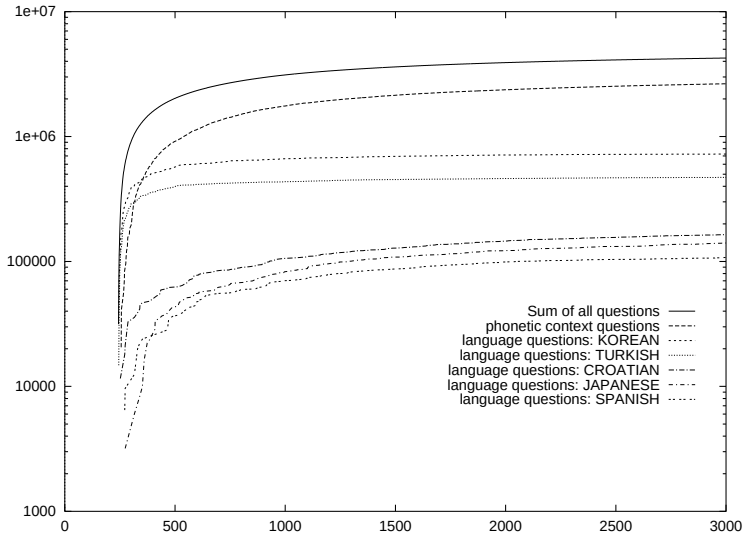

Fig. 1. Analysis of language questions

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 the codebooks, and a question set which consists of linguistic motivated questions about the phonetic context of a phoneme model. During clustering, the question which gives 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 enhanced this clustering routine to the multilingual case by 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. To analyse the importance of the language questions during the polyphone clustering procedure we calculated the sum of entropy gain over the number of splitted polyphones plotted in figure 1. The curve *"sum of all questions"* gives the over all entropy gain of all questions, whereas the curve *"phonetic context questions"* shows the entropy gain belonging to phonetic questions but not to language questions. Comparing both sums indicates that major parts of the entropy loss results from language questions therefore these questions are much more important than questions about the phonetic context. The other five curves give the relevance of questions belonging to only one language. It is shown that questions about Korean and Turkish are more important than those of the other languages, and also arise much earlier during the splitting process which indicates that in our database Korean and Turkish models are definitely different from Spanish, Croatian and Japanese models.

In summary one can say that the language information plays a significant part in creating context dependent models. From figure 1 it can be concluded that a multilingual system with more than 1500 models consists of mainly monolingual models.

4 Multilingual LVCSR System

Based on the above described methods we built three multilingual systems, two tagged and one mixed system. The *ML5-tag75* system consists of the same model size (7500 models) compared to the five monolingual system (5x1500 models). System *ML5-tag3* is a tagged system with only 3000 models whereas *ML5-mix* is a mixed system with 3000 models. Figure 2 compares the performance of the

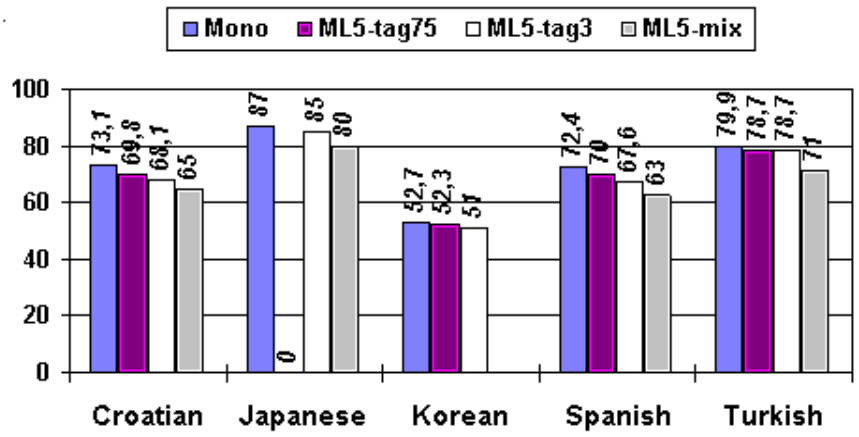

Fig. 2. Systems performance [Word accuracy]

three multilingual systems to the five monolingual systems. The system architecture, the preprocessing and the training procedure are identical in all systems. The number of parameters in the multilingual system *ML5-tag3* is reduced to 40% of the monolingual systems, it leads to 1.2% - 5% performance degradation. But not all of the degradation can be explained by the number of models as the performance of *ML5-tag75* shows. Comparing the *ML5-tag3* system to the *ML5-mix* system it can be seen that preserving the language information and introducing language questions leads to significant improvements in all five languages.

5 Conclusion

In this paper, multilingual LVCSR systems are presented which can handle five languages namely Croatian, Japanese, Korean, Spanish, and Turkish. In order to create context dependent multilingual acoustic models we introduced questions about language and language groups to our clustering procedure, which improve the recognition rate significantly.

6 Acknowledgement

The authors gratefully acknowledge all the members of the GlobalPhone team for their great enthusiasm during data collection and validation. We also wish to thank the members of the Interactive Systems Labs for useful discussion and active support.

References

1. C. Corredor-Ardoy, J.L. Gauvain, M. Adda-Decker, and L. Lamel: *Language Identification with Language-independent Acoustic Models* in: Proc. Eurospeech, pp. 355-358, Rhodes 1997.
2. O. Andersen, and P. Dalsgaard: *Language Identification based on Cross-language Acoustic Models and Optimised Information Combination* in: Proc. Eurospeech, pp. 67-70, Rhodes 1997.
3. 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.
4. T. Schultz and A. Waibel: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets* in: Proc. Eurospeech, pp. 371-374, Rhodes 1997.
5. T. Schultz, M. Westphal, and A. Waibel: *The GlobalPhone Project: Multilingual LVCSR with Janus-3* in: Proc. SQEL, pp. 20-27, Plzeň 1997.
6. The IPA 1989 Kiel Convention. In: Journal of the International Phonetic Association 1989(19), pp. 67-82.
7. 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.
8. J. Köhler: *Multi-lingual Phoneme Recognition exploiting Acoustic-phonetic Similarities of Sounds* in: Proc. ICSLP, pp. 2195-2198, Philadelphia 1996.

This article was processed using the LaTeX macro package with LLNCS style