# Multilingual and Crosslingual Speech Recognition

*T. Schultz and A. Waibel*

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

## 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. For our experiments we used six of these languages to train and test several recognition engines in monolingual, multilingual and crosslingual setups. Based on a global phoneme set we built a multilingual speech recognition system which can handle five different languages. The acoustic models of the five languages are combined into a monolithic system and context dependent phoneme models are created using language questions.

## 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 to be recognized into one single global acoustic model set is of increasing importance because of the following benefits:

1. Reducing the number of parameters and the complexity of the system by sharing codebooks and joining parameters across languages.
2. Language identification as described for example in [1] and [2].
3. Fast and efficient bootstrapping of recognition systems in new languages even if only a small amount of training data is available [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 decision tree based clustering procedure in order to achieve context dependent models and develop three systems which share their parameters in different ways. For clustering we add language questions to the linguistically motivated question set and 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 bootstrap and design of the monolingual systems. The experimental sections give results for the monolingual, multilingual, and crosslingual tests based on the systems created.

| Training Data | | | |
|---|---|---|---|
| Language | Utterances | Speakers | Word units |
| 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 |

Table 1: GlobalPhone database used for experiments

## 2. The GlobalPhone Database

For the development of the multilingual recognition systems and for the evaluation of the experiments, we used our recently collected database GlobalPhone which currently consists of the languages Arabic, Chinese (Mandarin), 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 GlobalPhone corpus is fully transcribed including spontaneous effects like false starts and hesitations. Further details of the GlobalPhone project are given in [5].

Table 1 shows the parts of the GlobalPhone database used for training. The test set consists of 100 utterances per language. Because of the currently limited corpus size of about 60K to 300K spoken words, we are not able to estimate reliable LVCSR n-gram models. This results in high out-of-vocabulary rates. Since we focus here on acoustic modeling and want to make word error rates comparable across languages, we restricted the OOV-rate in the case of Turkish, Croatian, Korean, and Spanish 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. Fast Systems Bootstrapping

For a baseline we developed five monolingual LVCSR systems for Croatian, Japanese, Korean, Spanish, and Turkish, applying our fast crosslingual bootstrap technique [4] to initialize the acoustic models for the not yet modeled languages. For each language the resulting context dependent monolingual system consists of a fully continuous 3-state HMM system with 1500 polyphone models. Each HMM-

| Language | Performance [WER] |
|---|---|
| Croatian | 26.9% |
| Japanese | 13.0% |
| Korean | 47.3% |
| Spanish | 27.6% |
| Turkish | 20.1% |

Table 2: Word Error [WER] of the monolingual systems

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.

Table 2 shows the performance in word error rates achieved by the monolingual systems. The results for Japanese are given in terms of hiragana words. The performance for the Korean system is given in morpheme based units, which explains the lower accuracy.

Figure 1 shows the number of phonemes per word in the training corpus and the dictionary which gives an impression of the difference in the modeled languages. The first plot shows the properties of the Spanish language. It can be seen that more than 20% of the words seen in the training set are only 2 phonemes long. This results in a high confusability of such words. The second graph shows the properties for Turkish. This concatenating language tends to have long words which might make it easier to distinguish Turkish words from each other but results in high out-of-vocabulary rates. The last plot reflects the mora structure of the Japanese language. Units consisting of 2, 4, or 6 phonemes are much more likely than others.

# 4. Multilingual Experiments

For multilingual speech recognition we wish to combine acoustic models of similar sounds across languages into one *multilingual phoneme set*. 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, it might be useful to find the similarities of sound in a data-driven way [7], [8]. In our work we defined a global phoneme set based on the IPA scheme. Sounds of different languages which are represented by the same IPA symbol share one common phoneme in this global phoneme set. The set consists of 82 different phonemes including silence and two noise models for spontaneous effects. About half of the global phoneme set is shared across the languages (polyphonemes), the other half consists of monophonemes belonging to only one of the five modeled languages. Table 3 shows the global phoneme set in Worldbet notation and gives the number of shared languages for each phoneme.

Based on these 82 phonemes in the global set, we build three different multilingual systems: *LangMix*, *LangSep*, and *LangTag*. In the first one we share all models across languages without preserving any information about the language. For each of the 82 phonemes we initialize one mixture of 16 Gaussian distributions and train the models by sharing the data of all five languages. The resulting multilingual recognizer is a fully continuous system with 3000 models mixed over all languages *(LangMix)*. In the second multilingual system

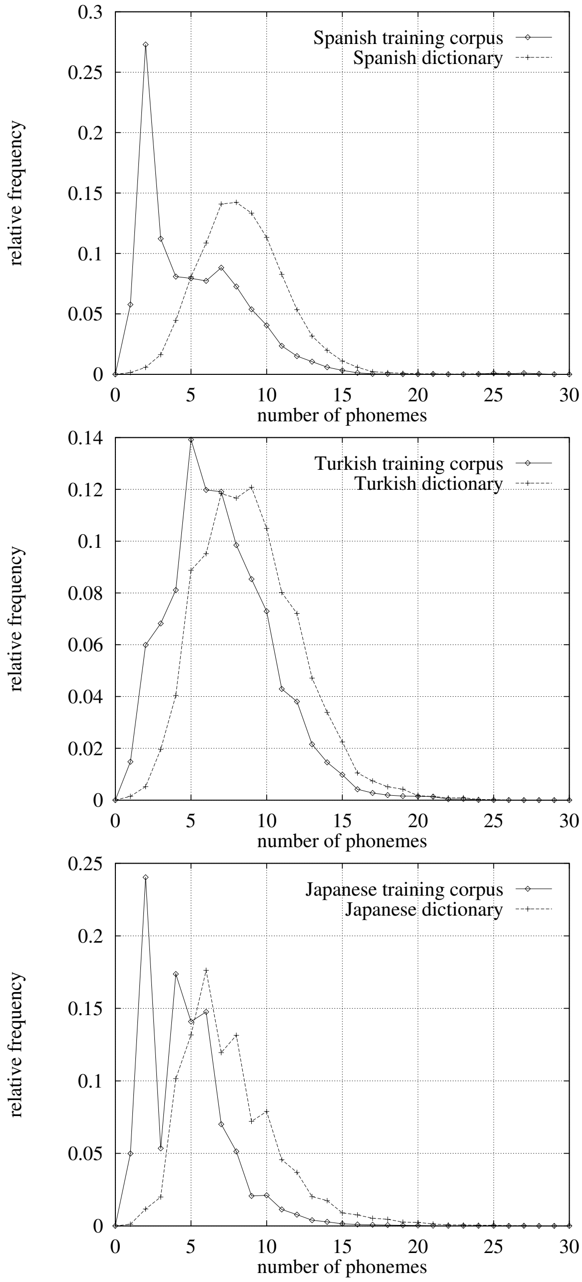


Figure 1: Phonemes per word in corpus and dictionary

*(LangSep)* each element of the global set which occurs in more than one language is modeled separately for each language. No data are shared, all models except silence and noise are language dependent. For each of the 173 ($173 = 30Cr + 33Ja + 41Ko + 40Sp + 29Tu$) phonemes we initialize one mixture of 16 Gaussian distributions. Finally, this results in a fully continuous system with 3000 language dependent models. In the third multilingual system *(LangTag)*, each of the 82 phonemes in the global set gets a language tag attached in order to preserve the information about the language. The decision whether to share models or not becomes data-driven by introducing questions about the language to which a phoneme belongs. We started with 250,000 different quintphones over the five differ-

| Shared across Languages | # | Phonemes (Worldbet symbols) Consonants | Vowels |
|---|---|---|---|
| All 5 | 14 | p,b,t,d,k,g, n,m,s,l,tS | i,e,o |
| 4 | 6 | r,f,z,j,dZ | u |
| 3 | 4 | S,h | a,4 |
| 2 | 10 | n∼,v,Z,x,L,ts,N | y,7,A |
| 1 Spanish | 15 | D,G,T,V,r( | ai,au,ei,eu,oi a+,e+,o+,i+,u+ |
| 1 Japanese | 8 | ?,Nq,V[ | A:,e:,i:,o:,4: |
| 1 Korean | 18 | p',t',k',dZ',s' | E,∧ ,i∧ ,iu,ie,io,ia iE,oE,oa,4i,u∧,uE |
| 1 Croatian | 2 | palatal c and d | |
| 1 Turkish | 2 | ix ,softer | |
| Sum | 79 | plus 3 models for silence and noise | |

Table 3: Global Phoneme Set

ent languages and created two fully continuous systems, one with 3000 models *LangTag3000* and the other one with the same number of models as the five monolingual systems together 5x1500 models *LangTag7500*.

## 4.1. Language Questions

To build context dependent phoneme models in our multilingual recognizer *LangTag* according to our polyphonic clustering procedure, we add questions about the language to the phonetic question set. During the splitting of the polyphonic tree we let the data decide whether the language question is better than a question about the phonetic context. The best question is selected based on the entropy distance between the mixture weights. The growth of the polyphone tree stops when the predefined number of leaves is reached. To analyze the importance of language questions during the decision process, we calculated the distribution over the languages during the splitting process for the parent node and the successor nodes as visualized in figure 2.

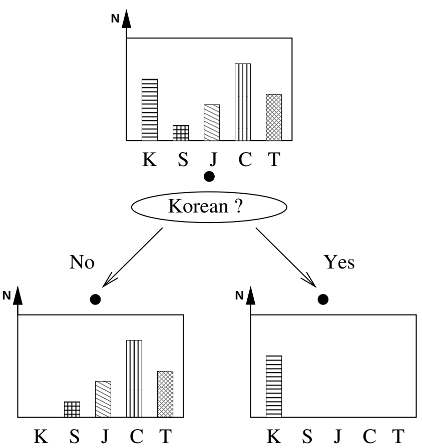


Figure 2: Language distribution in parent and successor nodes

The distance between the entropies of these language distributions are calculated and plotted over the number of leaves in the polyphone tree in figure 3. One important finding is that the language questions play a significant role in the decision of splitting the polyphone tree. Table 4 exemplifies the first ten node splits ranked by the entropy based distance. For instance the very first decision which was made separates Korean models of /a/ from the Croatian and Spanish ones. Conspicuously, most of the first split questions belong to Korean and Turkish, which indicates that there may be some sounds in those two languages which are different from the rest. The results might also indicate that the data-driven decision does not reflect the IPA-based classification across languages. Furthermore, monophonemes are the most useful for language identification, the sounds that are being split by language questions very early could also prove useful for language identification purposes.

| Dist | Phoneme | Question | Split apart |
|---|---|---|---|
| 0.2 | a | Korean | Ko from Cr,Sp |
| 0.8 | A | Turkish | Tu from Ja |
| 1.1 | s | Turkish | Tu from Cr,Ja,Ko,Sp |
| 1.4 | e | Turkish | Tu from Cr,Ja,Ko,Sp |
| 2.1 | n | Consonant | asymmetric split in Ja,Ko,Tu |
| 2.3 | k | Turkish | Tu from Cr,Ja,Ko,Sp |
| 2.5 | i | Korean | Ko from Cr,Ja,Sp,Tu |
| 3.2 | o | Korean | Ko from Cr,Ja,Sp,Tu |
| 4.0 | s 2nd | Korean | Ko from Cr,Ja,Sp |
| 4.2 | A 2nd | Unvoiced | split within Tu |

Table 4: Split prominence ranked by distance ($\times 10^6$)

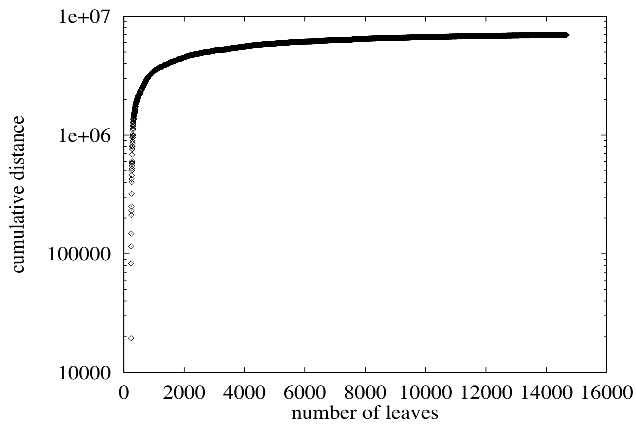


Figure 3: Entropy based Distance

Table 5 compares the performance of the described multilingual systems for the five languages. The system architecture, the preprocessing and the training procedure are identical to the monolingual engines. The number of parameters in the multilingual system *LangTag3000* is reduced to 40% of the monolingual systems (5x1500) and leads to only 1.2% - 5% performance degradation. But not all of the degradation can be explained by the number of models as the performance of *LangTag7500* shows.

Comparing the *LangTag3000* system to the *LangMix* system it can be seen that preserving the language information and introducing the language questions leads to significant improvements in all languages.

| Language | Monolingual | LangTag | LangTag | LangMix |
|---|---|---|---|---|
| Parameter | 5 x 1500 | 7500 | 3000 | 3000 |
| Croatian | 26.9 | 30.2 | 31.9 | 35.0 |
| Japanese | 13.0 | | 15.0 | 20.0 |
| Korean | 47.3 | 47.7 | 49.0 | |
| Spanish | 27.6 | 30.0 | 32.4 | 37.0 |
| Turkish | 20.1 | 21.3 | 21.3 | 29.0 |

Table 5: Results for multilingual setup [Word Error]

# 5. Crosslingual Experiments

We examined what performance can be achieved when the developed multilingual systems are applied to recognize *new unseen* languages without any additional training. For this purpose we have to define an appropriate mapping from the phoneme set of the unseen language to the phoneme model set used in the recognizer.

How can we find such a pronunciation dictionary for crosslingual tests? In the following experiments our goal is to recognize German spoken sentences. Therefore a German pronunciation dictionary has to be expressed in terms of our global phoneme set. The mapping was applied as follows: A German phoneme in the dictionary is replaced by a global phoneme corresponding to the same IPA-symbol. If no counterpart can be found, that global phoneme is chosen manually which is as close as possible to the German sound. Since our global phoneme set contains models from five different languages a German sound can have up to five counterparts - one in each language. We created four different pronunciation dictionaries for the crosslingual tests. For three language dependent dictionaries a mapping was produced from one specific language (Japanese, Spanish, and Turkish) to German. In the fourth dictionary we add the pronunciation variants from all five languages to the dictionary. For phonotactic reasons we do not use phonemes of different languages within one pronunciation. In this fourth dictionary we let the data decide which of the language specific variants are used for recognition.

Figure 4 shows the crosslingual recognition tests on the German part of the GlobalPhone database. Only a small set of German sentences with OOV-rate set to 0% was used for these experiments. The recognition rate of the German dictation system trained with German data is about 20% WER. The best system *LangMix* achieves 41.5% word error rate. It outperforms even the system *LangTag7500* which has more than twice as many models and gave best results in the closed multilingual test setup. This result indicates that different multilingual systems might be developed depending on the purpose.

Comparing the various pronunciation dictionaries for *LangTag3000*, we found that the 5-lingual dictionary outperforms the Japanese but not the Turkish dictionary. In the first case the reason might be that the Japanese phonotactic does not cover the German one because of its mora structure. The latter case might be due to the fact that the 5-lingual dictionary has 5 times more entries which leads to a higher confusability. The pronunciation variants used in the decoding indicate that for the German test sentences the Spanish models are preferred for short function words and Croatian models for longer words.

# 6. Conclusion

In this paper, multilingual LVCSR systems are presented which can handle five languages namely Croatian, Japanese, Korean, Spanish

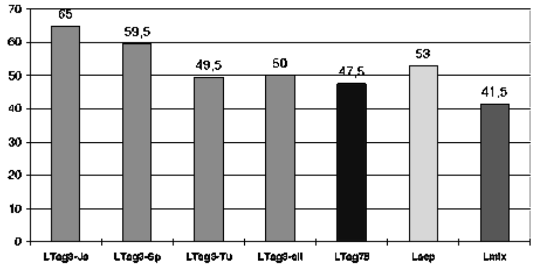


Figure 4: Crosslingual results on German test set [Word Error]

and Turkish. In order to create context dependent multilingual models we introduced language questions which improve the recognition rate significantly. Overall multilingual acoustic modeling leads to relatively small performance degradation even if the number of parameters is reduced to 40%. The multilingual systems are used to recognize a new unseen language without any training or adaptation and give promising results.

# 7. Acknowledgment

The authors gratefully acknowledge all 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 Laboratories, especially Detlef Koll, Michael Finke, Sondra Ahlén, Ivica Rogina, Martin Westphal and Thomas Kemp 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.