Adaptation of Pronunciation Dictionaries for Recognition of Unseen Languages

T. Schultz and A. Waibel

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

ABSTRACT

This paper studies the relative effectiveness of different methods for multilingual model combination and dictionary mapping for recognizing a new unseen target language if training data are limited. We examine the crosslanguage transfer from monolingual and multilingual models to German and Russian language for large vocabulary speech recognition using a 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 create recognizer which combine up to eight languages and perform recognition results in language independent and adaptive setups. We found that multilingual context dependent models outperform monolingual models for the purpose of crosslanguage transfer. Two dictionary mapping approaches are compared. Results show that the IPA-based mapping produces better results than a data-driven procedure.

1. Introduction

With the distribution of speech technology products all over the world, new methods for transfer of speech recognition systems across languages become a practical concern. [1] performed crosslanguage transfer from a language dependent system. Recently the usefulness of multilingual phonemic inventories have been demonstrated to give satisfactory results [2], [3], [4], but the use of context dependent multilingual models was not yet evaluated.

One major limitation in developing recognition systems is the need of large training data. This work explore the relative effectiveness of multilingual context dependent model combination for crosslanguage transfer with limited training data. Further a data-driven approach to adapt pronunciation dictionaries for the purpose of crosslanguage testing is compared to an heuristic mapping method. Multilingual phonemic sets so far are applied for crosslanguage transfer within the same language family [2], [5] and limited tasks [3]. The focus of our research is a multilingual recognizer for large vocabulary continuous speech which covers the most widespread and important languages of the world.

2. Multilingual LVCSR

For all experiments we use our multilingual database GlobalPhone which is briefly introduced in this section. The language dependent LVSCR systems and the multilingual context dependent acoustic modeling are also described in this section. In the second part of this paper, we address the problem of dictionary mapping. The last two sections give recognition results for crosslanguage transfer to German and Russian in monolingual and multilingual setups.

Language	Utterances		Speakers		Word units
Training Data					
	TrCD	TrCI	TrCD	TrCI	
Chinese	-	5149	-	79	150K
Croatian	2826	2616	62	64	80K
Japanese	5641	6419	62	82	200K
Korean	1587	4021	22	52	140K
Portuguese	-	6519	-	53	130K
Russian	-	7139	-	84	170K
Turkish	5371	5426	82	82	112K
Spanish	5455	5417	79	82	160K
Adaptation Data					
German	1000		13		14K
Test Data					
German	200	-	3	-	2.5K
Russian	-	100	-	12	1K

Table 1: GlobalPhone Database used for Experiments

2.1. The Multilingual Database GlobalPhone

For the multilingual speech recognition research, 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 articles from a national newspaper. Up to now we collected 233 hours of fully transcribed office quality speech. Further details of the GlobalPhone project are given in [6].

Table 1 shows the part of the database used for training the monolingual and multilingual systems. We trained context-dependent models in five languages, namely Croatian, Japanese, Korean, Spanish, and Turkish using the training set TrCD. For the experiments on up to eight-lingual context independent systems we add Chinese, Portuguese and Russian data. This training set is labeled TrCI. The tests are evaluated on 200 German and 100 Russian utterances. Since we focus here on dictionary adaptation and acoustic modeling we reduced the OOV-rate to 0.0% by including all test words into the language model as mono-grams with small probabilities. For German recognition tests we defined a 10K test dictionary by supplementing the test words with the most frequently seen training units. For Russian we run experiments on a very preliminary dictionary with 500 word entries.

2.2. Language Dependent LVCSR

In the first step we developed monolingual LVCSR systems in eight languages applying our fast bootstrap technique [4]. For six languages Chinese, Croatian, Japanese, Korean, Spanish, and Turkish the resulting baseline recognizer 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 cepstral mean subtraction, a linear discriminant analysis is used to reduce the input to 24 dimensions. The systems performance ranges from 10% kana error rate for Japanese to 16.9% for Turkish, 18.4% for Chinese, and 20% word error rate for Spanish and Croatian. The Korean performance is given in hangul syllables and achieves 47% error rate. A more detailed description of the systems can be found in [7].

For Portuguese and Russian so far only preliminary context independent systems have been developed. Their recognizer consists of 3-state HMMs with 53 and 34 monophone models. Each HMM-state is modeled by 32 Gaussian. The preprocessing is identical with the context dependent counterparts. Except for the Chinese system the context dependent systems are applied for multilingual context dependent acoustic modeling.

2.3. Multilingual Acoustic Modeling

For the purpose of crosslanguage transfer and dictionary mapping we intend to share acoustic models of similar sounds across languages. Those similarities can be either derived from international phonemic inventories documented in Sampa, Worldbet, or IPA [8] or by data-driven methods proposed for example in [9].

In this work we use a data-driven procedure for multilingual context dependent acoustic modeling. We defined a *global phoneme set* based on the phonemic inventory of the eight monolingual systems. Sounds which are represented by the same IPA symbol share one common phoneme category. Altogether this global set consists of 145 phoneme categories. Based on these categories we design different multilingual systems by combining language dependent acoustic models in different ways.

In system *ML-mix* we share all models across languages without preserving any information about the language. For each category model we initialize one mixture of 16 Gaussian distributions and train the models by sharing the data of five languages (*ML5-mix*), seven language (*ML7-mix*) and eight languages (*ML8-mix*) respectively. For the *ML5-mix* system we build a context dependent system by applying 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. We stop the splitting procedure after reaching 3000 polyphone models, which results in system *ML5-mix3000*.

Another way to share phoneme models across languages is performed in the multilingual system *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 enhanced 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. We started with 250,000 different quintphones over the five languages and created two fully continuous systems, one with 3000 models (*ML5-tag3000*) like *ML5-mix3000* and the other one with 7500 models (*ML5-tag7500*) which is comparable in size to five monolingual systems with 1500 models each.

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. The *ML5-tag3000* outperforms the mixed system *ML5-mix3000* in all languages by 5.3% (3.1% - 8.7%) error rate, which indicates that preserving the language information achieves better results with respect to monolingual recognition. The *ML5-tag3000* system reduces the model size to 40% compared to the monolingual case (3000 vs 5x1500 models), resulting in a 3.14% performance degradation in average (1.2% - 5.0%). But not all of the degradation can be explained by the reduction of parameter which can be derived from the comparison between the monolingual systems and *ML5-tag7500*. We still observe an average performance gap of 1.07% (0.3% - 2.4%). This finding is coincident to other studies [5], [2], and [10]. We therefore draw the conclusion that so far multilingual modeling decreases the performance with respect to monolingual recognition.

3. Crosslanguage Dictionary Mapping

In the previous section we examine the usefulness of multilingual acoustic modeling with respect to monolingual speech recognition. Now we investigate the benefit of multilingual models for crosslanguage transfer to new target languages. For this purpose we need a pronunciation dictionary suitable for the target language in terms of the phoneme model set of the bootstrap engine. How to create such kind of pronunciation dictionary for crosslanguage transfer?

We investigate two different approaches for adaptation of target pronunciation dictionaries, and compare them by running recognition tests using the resulting dictionaries. We perform tests with and without training on limited data of the target language. 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 [3] 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.

We perform a data-driven and an heuristic IPA-based mapping approach: In the data-driven approach 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, 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. In case of monolingual crosslanguage transfer the target phoneme set may collapse

System	Dictionary	Word Error
Baseline	German	15.8%
Mono-Croatian	Croatian	31.3%
Mono-Japanese	Japanese	50.5%
Mono-Korean	Korean	42.4%
Mono-Spanish	Spanish	31.9%
Mono-Turkish	Turkish	28.4%
Best of 5	sentence-based	21.8%

Table 2: Monolingual Transfer to German

System	Dictionary	Word Error	
		noTrain	Train
Baseline	German	15.8%	
ML5-tag3000	IPA-5L	69.4%	35.7%
ML5-tag7500	IPA-5L	69.1%	35.4%
ML5-mix3000	IPA-5L	63.0%	29.2%

Table 3: Multilingual Transfer to German

if the bootstrap set is only a small subset of the target set. On the other hand if the target set is a subset of the bootstrap set one target phoneme can have more than one counterpart. Especially 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. In the following two sections experiments are performed for crosslanguage transfer to German and to Russian language.

4. Crosslanguage Transfer to German

In this section we investigate the feasibility of mono- and multilingual recognition systems for crosslanguage transfer from five languages to German. Furthermore we explore the effect of adaptation to the German language using limited data. For training we used 14000 words (1000 utterances) spoken by 13 native German speaker, for testing 2500 words (200 utterances) spoken by 3 speakers. The German baseline system achieves 15.8% word error rate tested on a 60k-dictionary.

4.1. Monolingual Transfer

In previous work we demonstrated that bootstrapping a Japanese system from German system leads to very good results. For this application it was known that the German phonetic and phonological structure fits well to the Japanese one (but not vice versa). [3] found some evidence for the correlation between the similarity of two languages and the recognition rate when bootstrapping one from the other. They give results for 5 Indo-European languages. However the definition of similarity is not trivial if such completely different languages like Japanese, Turkish, Croatian, Korean etc. are involved. These languages belong to different language families, the have phonemic inventories which range from 30 to 58 phonemes. The average phonemic length of words in corpus varies from 2 to 7. The syntactical and morphological structure covers concatenating, inflecting and isolating style. Even the writing systems are completely different. Thus one interesting point is to explore, if and how these differences are reflected in recognition rate when using for crosslanguage transfer to German. For this experiment we apply the above described IPA-based dictionary mapping. Based on the 1000 German train utterances we perform two iteration of Viterbi training to adapt to the target language. We do not re-cluster the polyphone trees, but simply training the Gaussian and mixture weights of the language dependent models.

Table 2 gives the crosslanguage performance when using these five languages for crosslanguage transfer to German language. The performance ranges from 50.5% to 28.4% word error rate. The poor results for the monolingual Japanese system might be due to the fact that Japanese context dependent models do not cover the German phonology because of the Japanese mora structure. Since German is a language with high frequent consonant sequences this leads to an extremely mismatch. The Turkish language tends to have very long words and fits better into the German phonology. In our experiments we found that Spanish models are preferred for short function words, which might result from the fact that 20% of the Spanish corpus words consists of only two phonemes.

Additional we ran all five systems in parallel and calculate the "Best of 5" (assuming that the best is known). This results in 21.8% word error rate and outperforms the language dependent systems. However the decision for the best matching system remains utterance based.

4.2. Multilingual Transfer

Since one of the main drawback of monolingual transfer is the low coverage of phonemic and phonological structure as seen above, a more encouraging approach is the crosslanguage transfer based on multilingual models. Throughout the following experiments we like to explore 3 questions: first we analyse the effect of training with limited data (column noTrain vs Train), second we investigate the usefulness of the different model combination (*ML5-tag3000*) vs *ML5-mix3000* and third we examine the effect of different parameter size (*ML5-tag7500* vs *ML5-tag3000*). Table 3 summarizes the results of the recognition tests.

First, the effect of adaptation with limited training data is overwhelming which is of course not surprising. In a former study [7] we demonstrated that training with only 2000 spoken words nearly halves the error rate. Since we do not recalculate the polyphone trees, the gap between the best crosslanguage result and the German baseline system seems to be reasonable. Further research will explore the effect of recalculating the trees. Second, with regard to crosslanguage transfer sharing the phoneme models without preserving any information about language leads to best results. In view of monolingual recognition the tagged system outperforms the mixed system (see section 3) which indicates, that dedicated multilingual systems should be developed depending on whether cross- or monolingual speech recognition is projected. In the first case the *ML-mix* system should be favored, in the latter the *ML-tag* system. Third, increasing the number of model parameter improves the performance not significantly.

System	Dictionary	Word Error	
		noTrain	Train
Baseline	German	15.8%	
ML5-mix3000	IPA-ML	66.7%	27.1%
ML5-mix3000	IPA-5L	63.0%	29.2%
ML5-mix3000	data-driven	74.5%	34.3%

Table 4: IPA-based vs Data-driven Dictionary Mapping

4.3. Heuristic vs Data-driven Mapping

Finally we compare the quality of the heuristic IPA-based mapped dictionary to our data-driven mapping approach as described in section 3.

Table 4 shows the performance for the best system *ML5-mix* under three different dictionary conditions. The IPA-ML dictionary results from mapping the multilingual models to the best matching IPA counterpart. Thus language dependent phonological properties are ignored. Whereas in the IPA-5L dictionary the phonological properties are preserved, because for every language we keep one pronunciation variant in the dictionary. The results indicate that the data-driven approach is clearly outperformed by the heuristic one and second that ignoring the phonological properties leads to better results. One reason might be that the IPA-5L contains four times more dictionary entries, another reason could be that the robustness for crosslanguage transfer is increased.

5. Crosslanguage Transfer to Russian

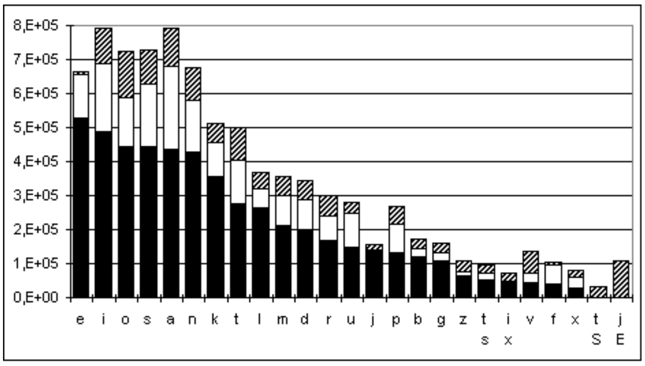

Figure 1: Data for the most frequent phonemes [10ms Frames]

We explore the crosslanguage transfer technique to a second language Russian. For the following experiments no training is applied. We focus here on the effect of increasing the number of languages combined in the multilingual global phoneme set. For this purpose we added Chinese, Portuguese and Russian data to train a seven- and eight-lingual context independent system. Figure 1 plots the accession of training data for the most frequent Russian phonemes for ML7 and ML8 compared to ML5. For baseline we set up a very preliminary Russian speech recognizer. The Russian dictionary was created from scratch by applying a simple grapheme-to-phoneme mapping. We build three Russian pronunciation dictionaries based on the global phoneme set of five (IPA-ML5), seven (IPA-ML7) and eight languages including Russian phonemes.

System	Dictionary	Word Error
Baseline-CI	Russian	37.6%
ML5-mix	IPA-ML5	61.7%
ML7-mix	IPA-ML7	59.3%
ML8-mix	Russian	42.3%

Table 5: Crosslanguage Transfer to Russian

The results in table 5 point out that including two new languages Chinese and Portuguese into the bootstrap system increases the crosslanguage performance (ML5-mix vs ML7-mix). But from this point it is not clear whether the performance gain results from the more training data or more language information. Further research will investigate these questions.

6. Conclusion

In this paper, multilingual LVCSR systems in eight languages are presented and applied to crosslanguage transfer to German and Russian language. The study indicates that multilingual acoustic models outperform monolingual models with respect to crosslanguage transfer. The method produces satisfactory results, requiring very little human effort.

References

1. 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.
2. J. Köhler: *Language Adaptation of Multilingual Phone Models for Vocabulary Independent Speech Recognition Tasks* in: Proc. ICASSP, pp. 417-420, Seattle 1998.
3. 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.
4. T. Schultz and A. Waibel: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets* in: Proc. Eurospeech, pp. 371-374, Rhodes 1997.
5. P. Cohen, S. Dharanipragada, J. Gros, M. Monkowski, C. Neti, S. Roukos, and T. Ward: *Towards a Universal Speech Recognizer for Multiple Languages* in: Proc. Automatic Speech Recognition and Understanding, pp. 591-598, St. Barbara 1997.
6. T. Schultz, M. Westphal, and A. Waibel: *The GlobalPhone Project: Multilingual LVCSR with Janus-3* in: Proc. SQEL, pp. 20-27, Plzeň 1997.
7. T. Schultz, and A. Waibel: *Language Independent and Language Adaptive Large Vocabulary Speech Recognition* in: Proc. ICSLP, to appear, Sydney 1998.
8. The IPA 1989 Kiel Convention. In: Journal of the International Phonetic Association 1989(19) pp. 67-82
9. 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.
10. P. Bonaventura, F. Gallocchio, and G. Micca: *Multilingual Speech Recognition for Flexible Vocabularies* in: Proc. Eurospeech, pp. 355-358, Rhodes 1997.