FAST BOOTSTRAPPING OF LVCSR SYSTEMS WITH MULTILINGUAL PHONEME SETS

T. Schultz and A. Waibel

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

ABSTRACT

In this paper we described an efficient method to bootstrap continuously spoken, large vocabulary speech recognition systems by multilingual phoneme sets. To evaluate this techniques we collected the multilingual database GlobalPhone which currently consists of 9 different languages. A multilingual recognizer (MULTI) based on the four languages German, English, Japanese and Spanish was developed to serve as a source system. Likewise this system is very useful for language identification and achieves 100% language identification rate. Based on the MULTI system we evaluated our bootstrap technique on such completely different languages as Chinese, Croatian, and Turkish.

1. INTRODUCTION

As the demand for speech recognition and translation systems in multiple languages grows, the development of multilingual systems is of increasing concern. On the one hand a multilingual system can be used as a language independent speech recognition and translation system with integrated automatic language identification. On the other hand it can serve as a phoneme pool for rapid bootstrapping of speech recognition systems into other languages.

The development of reliable multilingual phoneme sets and the evaluation of the rapid cross language bootstrapping technique requires uniform speech data of high quality in several languages. The high quality guarantees that the only difference of the acoustic data is the spoken language itself. As could be seen in [1] different quality conditions can influence the language identification significantly. The corpus domain and the collection scenario should ensure consistency: Vocabulary and task should be comparable across all languages. Existing databases like the OGI corpus or CALL HOME are collected in telephone quality. Other databases of high quality data like the Spontaneous Scheduling Task [2] cover too few languages for our approach. Therefore we started the collection of a multilingual database called GlobalPhone which is described in the first section of this paper. In the second part of the paper the experiments based on the GlobalPhone corpus are described. These experiments pursue two aims: first a multilingual phoneme set is produced and it is clarified if and how much performance is lost when combining four different language dependent systems into one unique multilingual system. Second we determine how well models trained on spontaneous spoken speech fit to bootstrap read speech from a new domain in a new language. For that latter part the created multilingual recognizer serves as the bootstrap engine.

2. THE GLOBALPHONE DATABASE

We have collected a multilingual high quality speech corpus called GlobalPhone, which is suitable for the development of multilingual large vocabulary continuous speech recognition systems. For the present this database consists of 9 different languages namely Arabic (Tunisia), Chinese (Mandarin), Croatian (Croatia and Bosnia), Japanese, Korean, Portuguese (Brazil), Russian (Belorussia), Spanish (Costa Rica), and Turkish. Considering the fact that English, French, and German are already available in similar frameworks, the database covers 9 out of the 12 most frequent languages of the world.

Language	Speaker	Rec. Hours	Spoken units	Vocab Size
Arabic	93	28	-	-
Chinese	132	40	125K	4K
Croatian	85	18	89K	17K
Japanese	121	41	182K	21K
Korean	70	32	-	-
Portuguese	75	33	182K	6K
Russian	99	26	186K	20K
Spanish	89	20	164K	21K
Turkish	100	18	110K	15K

Table 1: The GlobalPhone Database

Transcribing conversational speech is one of the most expensive and time consuming step of a database collection. Though we decided to collect speech data read from a text source already electronic available, which also allows to collect additional consistent text data for the training of n-gram language models. These constraints were fulfilled by collecting articles of national newspapers available via Internet with national and international political and economic topics.

The corpus consists of continuous spoken speech read by about 100 native speakers per language. Each speaker read about 20 minutes recorded in an office environment, with a Sennheiser close-speaking microphone and a portable DAT recorder. Table 1 shows the current status of the GlobalPhone database. Further details are given in [3].

3. MULTILINGUAL PHONEME SET

In this study the development of the multilingual speech recognition engine pursued two purposes: On the one hand such a system is needed for a multilingual speech recognition and translation system. This implies the need of solving the language identification problem. On the other hand we want to build an engine which offers the opportunity to serve as a bootstrap machine for new -not yet modeled- languages. For this approach we aspire a multilingual phoneme set that preserves the language specific characteristics of each model. Thus the term "multilingual phoneme sets" is used here to define a conglomerate set of language dependent phonemes into one global set. It does not mean a mixture of similar phonemes to combined language independent models as proposed for example in [4] and [5].

Language	Words	Vocab	Phonemes	WE
German	158K	5438	65	14%
English	280K	2601	53	23%
Japanese	92K	1879	39	9.3%
Spanish	91K	3939	47	17%
MULTI		20082	204+SIL	

Table 2: Language dependent systems; latest Word Error rates [WE]

We developed a multilingual recognition engine MULTI which covers German, English, Japanese and Spanish data based on the spontaneously spoken appointment scheduling task [2]. Table 2 gives some information about the used systems together with the word error rate of the currently best recognition engines.

To create MULTI we combined the acoustic and language models of the four existing recognizer. The phoneme set contains 205 language dependent phoneme models. Each phoneme is modeled by a context-independent 3-state HMM, where each HMM-state is modeled by one codebook. Each codebook contains 16 mixture Gaussian distributions of a 24 dimensional feature space. This feature space results from a Linear Discriminant Analysis calculated to reduce the dimension of the input feature vector consisting of mel frequency cepstral coefficients, power and their first and second derivatives. The dictionary of MULTI combines the four language specific dictionaries and has a size of 20K words.

To evaluate if and how much performance is lost in the MULTI system we run two experiments. In the language dependent experiment (LD) the performance of the context-dependent language specific systems which are used as source systems for creating the MULTI system is calculated. In the language independent experiment (LI) the MULTI system with context-independent phoneme models, language independent preprocessing, combined dictionary and language model is tested. Experiment LI results in the decrease of performance as can be seen in table 3 mostly due to the fact that only context-independent phoneme models are used.

	LD	LI	
Language	WE	WE	LID-rate
German	13.2%	35.5%	100%
English	31.4%	38.5%	100%
Japanese	13.0%	24.4%	100%
Spanish	37.8%	39.9%	100%
Total		35.0%	100%

Table 3: Word Error rates of LI vs LD

3.1. Language Identification

When using LVCSR systems for the identification of a spoken language we think of two approaches: A *parallel* architectures, in which for each language to be identified a language dependent system is trained. The language identification is performed by running all systems in parallel. Each system decodes the utterance to determine the best hypothesis. The language belonging to the system with the best score is hypothesized. In former studies we achieved 86.3% identification rate on the 4-language task with this approach.

In the *integrated* architecture a single language independent recognition system is applied. The language identification is implicitly done during the decoding by pruning the hypotheses of the wrong languages. We used our MULTI system for this approach which gives a language identification rate of 100% on the same 4-language task. Language identification with this approach seems to be very time consuming during the decoding process when the number of languages is large but [8] could show for his system that the hypotheses of the wrong languages are pruned away within the first 2 seconds.

4. BOOTSTRAPPING

In former experiments we bootstrapped a Japanese speech recognition system with the phoneme set of an existing German recognizer. This cross language transfer from one language to another produced very promising results [7]. We now generalize our approach to a bootstrap technique from a multilingual phoneme set based on German, English, Japanese, and Spanish:

- of spontaneously spoken speech to read speech
- in a new large vocabulary domain
- into various languages like Chinese, Croatian, and Turkish.

The selected languages are completely different from the languages of the MULTI recognizer and from each other. Only the Croatian language belongs to the same language family (Indo-European) as German, English, and Spanish. Like Croatian the Turkish language, which belongs to the Turk family is highly inflecting. Chinese is a tonal language and therefore totally different from the others. Table 4 summarizes the information about the used data which are part of the GlobalPhone corpus.

Language	Utts	Speech	Units	Vocabulary
Training data				
Chinese	2055	378 min	61832	5524
Croatian	1533	336 min	41955	10744
Turkish	4261	606 min	67106	14723
Test data				OOV-rate PP
Chinese	100	15min	2462	10.25% 486.4
Croatian	131	26min	3522	16.26% 241.1
Turkish	124	22min	2393	23.93% 316.9

Table 4: Data used for bootstrapping experiments

The definition of the unit "word" is not comparable for all languages. The Korean, Chinese and Japanese languages do not have a word concept in the general meaning. Since there is no white space between written words the phrases have to be segmented into units, requiring morphological analysis. For segmentation and romanization (*hanse* to *pinyin*) of the Chinese language we developed an algorithm which gives 98% correct pinyin mapping and 95% word segmentation accuracy. Our approach results in relatively small vocabulary growth rates for Chinese. In opposite the Croatian, Russian, and Turkish written language contains definite word boundaries, but these languages are highly inflecting. Since we count each flexion as one word we obtain very high growth rates for Croatian and Turkish. Figure 1 illustrates the growth rates and compares our read speech corpus GlobalPhone to the conversational speech data of the Spontaneous appointment Scheduling Task (SST). Unlike GlobalPhone, the SST database is a

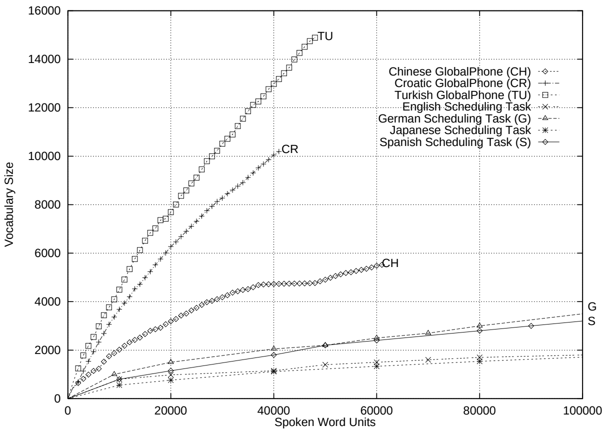

Figure 1: Vocabulary Growth of GlobalPhone vs SST corpus

very limited domain task which leads to a small vocabulary size with low growth rates and low Out-Of-Vocabulary (OOV) rates.

4.1. Experiments

The bootstrap mechanism (B-MULTI) is divided into 5 steps and works as follows:

- Step 0: Mapping of the language specific phones to those of the MULTI phoneme set motivated by a phonetic analysis done by native experts
- Step 1: Initialization of the acoustic models according to the mapping table
- Step 2: label boosting with MLLR, calculation of a language specific LDA, kmeans clustering to initialize the codebooks
- Step 3: Four training iterations
- Step 4: Repetition of step 2 and step 3

Figure 2 illustrates the results of B-MULTI. For each step the phoneme error rate of the according phoneme recognizer for every language is given. In the case of the Croatian language we compare B-MULTI to a second bootstrap techniques in which step 0 is replaced by a random phoneme initialization step. After 4 iteration training with Croatian data (step 3) the "random" system is still worse than the Croatian B-MULTI system, which has never seen any Croatian data (step 1). This indicates that the MULTI phoneme set covers the Croatian phoneme set sufficiently. Even if the MULTI phoneme set is a very coarse approximation as it is the case for the tonal language Chinese our bootstrap technique leads to good results. The low performance in step 1 indicates the mismatch between the Chinese and the MULTI phoneme set but in step 2 the accuracy of the Chinese system raises fast and in fact outperforms the Croatian system after step 3. .

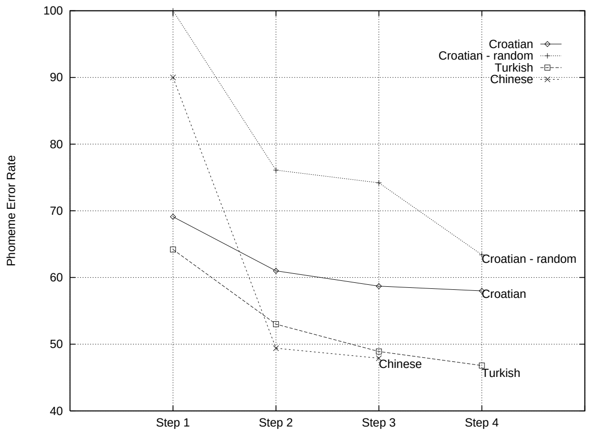

Figure 2: Bootstrap Performance

4.2. Preliminary LVCSR systems

Table 5 summarizes the performance of the resulting Chinese, Croatian, and Turkish LVCSR system after step 4 of B-MULTI. The results are promising considering the fact that currently not all speech data are processed and acoustic modeling is based on context independent phoneme models. Up to now we concentrate on the acoustic aspects of the fast bootstrapping mechanism. Thus some problems remain to be solved: The pronunciation dictionaries are build fully automatically and do not contain any pronunciation variants which might lead to suboptimal modeling. Some kind of language dependent tuning is to be done to take language specific characteristics into account like i.e. the tonal feature for the Chinese language and a morphological approach for the language modeling of such highly inflecting languages as Turkish and Croatian. More text data have to be processed to overcome problems with the high OOV rates and perplexities (see table 4). For the Chinese language we processed a 3.8 million word text and calculated a new language model (LM3.8). We compare the standard language model (LM) used so far to the LM3.8 model. Using LM3.8 leads to a word error reduction of 16% on the open vocabulary test, and to 23% on the closed vocabulary test.

Test	Chinese		Croatian	Turkish
	LM3.8	LM		
open vocab	52.6%	43.6%	40.0%	36.1%
closed vocab	60.6%	48.9%	48.8%	47.3%

Table 5: Word Accuracy of LVCSR systems

5. CONCLUSION

We described the development of a multilingual speech recognition system covering the languages German, English, Japanese, and Spanish, which achieves a language identification rate of 100% on the 4-language task. This multilingual system serves as a source engine for fast bootstrapping of LVCSR systems in a new domain into completely different languages like Chinese, Croatian, and Turkish. From that experiments we conclude that cross language bootstrapping is a very efficient technique even if the phonetic inventory mismatches. Further research will explore the extensibility of the multilingual approach to context dependent phoneme modeling.

6. ACKNOWLEDGMENTS

The authors wish to thank all members of the Interactive Systems Laboratories especially the GlobalPhone team: Olfa Karboul Zouari and Mohamed Zouari (Arabic), Tianshi Wei, Jing Wang, Jürgen Reichert and Jiaxing Weng (Chinese), Sanela Habibija and Stefan Raschke (Croatic), Laura J. Tomokyo, Hiroko Akatsu, and Sayoko Takeda (Japanese), Keal-Chun Cho and Sang-Hun Shin (Korean), Orest and Natalia Mikhailiouk (Russian), Raul Ivo Faller and Caleb Everett (Portuguese), Giovanni Najera Barquero (Spanish), Mutlu Yalcin and Kenan Carki (Turkish). This research would not have been possible without their great enthusiasm during collection and validation of the database.

7. REFERENCES

[1] T. Schultz, I. Rogina, and A. Waibel: *LVCSR-based Language Identification.* Proceedings of the ICASSP 96, Atlanta, USA, May 1996.

[2] B. Suhm et al.: *JANUS: Towards Multilingual Spoken Language Translation* in: DARPA Speech and Natural Language Workshop, 1995.

[3] T. Schultz, M. Westphal, and A. Waibel: *The GlobalPhone Project: Multilingual LVCSR with JANUS-3* in: Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzen April 1997.

[4] J. Köhler: *Multilingual Phoneme Recognition Exploiting Acoustic-Phonetic Similarities of Sounds* in: Proc. of ICSLP, pp. 2195–2198, Philadelphia 1996.

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

[6] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and Martin Westphal: *The Karlsruhe-Verbmobil Speech Recognition Engine* in: Proc. of ICASSP, Munich 1997.

[7] T. Schultz, D. Koll, and A. Waibel: *Japanese LVCSR on the Spontaneous Scheduling Task with JANUS-3* Eurospeech, Rhodes 1997.

[8] S. Harbeck: *Multilingual Speech Recognition* in: Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzen, April 1997.