EXPERIMENTS ON CROSS-LANGUAGE ACOUSTIC MODELING

T. Schultz and A. Waibel

Interactive Systems Laboratories
Carnegie Mellon University (USA), University of Karlsruhe (Germany)
{*tanja,ahw*}@*cs.cmu.edu*

ABSTRACT

With the distribution of speech products all over the world, the portability to new target languages becomes a practical concern. As a consequence our research focuses on rapid transfer of LVCSR systems to other languages. In former studies we evaluated the performance if limited adaptation data is available. Particularly for very time constrained tasks and minority languages, it is even reasonable that no training data is available at all. In this paper we examine what performance can be expected in this scenario. All experiments are run in the framework of the GlobalPhone project which investigates LVCSR systems in 15 languages.

1. INTRODUCTION

The extension of LVCSR systems to new target languages requires large amounts of transcribed audio data for training accurate acoustic models. For several reasons it is often not possible to provide much data and especially in time constrained tasks or for minority languages it might even become realistic that no training data is available at all. To accommodate potential variations in the amount of data available for the target language, we address three issues: 1) the *bootstrapping approach* where the key idea is to initialize the acoustic models of the target language recognizer with seed models developed for other languages [4, 8]. After this initialization, the resulting models are completely rebuilt using large amounts of target language training data. 2) The *language adaptation technique* where an existing recognizer is adapted to the target language using very limited data [8, 3], and 3) *cross-language transfer* which refers to the technique of using a recognition system to decode a new language without having ever seen any training data of the language in question [1, 2].

In former studies we applied acoustic models from four languages to bootstrap Chinese, Croatian, and Turkish [5]. We demonstrated the usefulness of a universal language independent model inventory for the adaptation to German [6], and recently introduced a method for multilingual polyphone tree specialization using Portuguese as an example [7]. In this paper we extend our work in two directions: firstly we explore a new language, Swedish; secondly we investigate what performance can be expected if no training data is available. We apply language dependent and independent models as seed models, and examine whether context dependent models are helpful. Furthermore, we compare different approaches to find an appropriate mapping from the universal to the Swedish phoneme set.

2. THE GLOBALPHONE FRAMEWORK

The following experiments are carried out using recognition engines developed in the GlobalPhone project [6, 7]. We applied monolingual systems from Chinese, Croatian, German, French, Japanese, Spanish and Turkish, as well as a multilingual phoneme recognizer. Swedish is treated as the target language, which does not imply that we consider Swedish to be a minority language. We chose Swedish since it is so far not studied heavily in the speech community and not studied in our group, which ensures that the acoustic models are not contaminated by Swedish data.

We focus on cross-language effects concerning the acoustic models and assume that a pronunciation dictionary and a language model are given. However, we are aware of the fact that the latter are critical issues for speech recognition in new languages. All Swedish resources used in the experiments are generated from scratch. The dictionary was created by a letter-to-sound approach, using about 250 pronunciation rules. Speech and text databases are collected in GlobalPhone style using the "Göteborgs-Posten" newspaper (http://www.gp.se). The corpora for generating the language model contains only 150k words, leading to a trigram perplexity of 1029 given a 24k vocabulary. For testing we used 200 utterances spoken by 10 native Swedish speakers.

3. LANGUAGE DEPENDENT SEED MODELS

In the following experiments we investigate the usefulness of *language dependent* acoustic models, i.e. models which are solely trained on a single language. We examine the correlation between language characteristics and cross-language transfer performance, and evaluate whether context modeling leads to improvements. To express the Swedish pronunciation dictionary in terms of the monolingual phoneme sets, we applied a heuristic mapping approach [6] based on the International Phonetic Association alphabet (IPA).

Language		context-independent	context-dependent	Δ
Chinese (CH)	45.2	75.2	76.0	-1.0%
Croatian (KR)	36.7	59.0	58.1	1.5%
French (FR)	36.1	69.6	70.3	-0.7%
German (DE)	44.5	64.9	63.2	2.6%
Japanese (JA)	33.8	76.0	74.1	2.5%
Spanish (SP)	43.5	69.6	67.1	3.6%
Turkish (TU)	44.1	59.9	59.9	0%
$Average_{L7}$		67.8	67.0	1.2%

Table 1. Cross-language transfer to Swedish [PER in %]

3.1. Language Differences

Table 1 shows the performance results of decoding Swedish utterances by seven monolingual recognizers without any prior training or adaptation. Since we are focusing basically on phonetic mismatches, and to counterbalance the effect of the weak language model we give all results in phoneme error rates. For the cross-language experiments the error rates are calculated within word boundaries, meaning that a word-based dictionary is used to guide the decoder. The third column in Table 1 gives a relative difference of up to 22% between the languages, ranging from 59% for Croatian to 76% error rate for Japanese. From this we conclude that the knowledge about the best-matched language is crucial for cross-language transfer.

However, we could not derive a reliable predictor for a suitable transfer language from our results. Column 2 of Table 1 gives the baseline error rates of the monolingual phoneme recognizers on the training language (recognizer without any language model constraints). These baseline error rates do not correlate to the cross-language performance. Furthermore, no relation between the language family and cross-language performance could be found. German as the closest family member achieves better results than the other Indo-European languages, however the best performance is achieved by Croatian and Turkish, which are not related to Swedish. The assumption that phoneme coverage [7] serves as an indicator does not hold either. We found that German, Japanese, and French contribute the most to cover Swedish monophones and triphones, but this is not reflected in the cross-language transfer performance. A relation between the compactness of the phoneme set and cross-language performance is counter-proved by the good results of German which has a large phoneme set and the poor results of Japanese which has the third-most compact phoneme set.

3.2. Modeling Context

For the monolingual case it is well known that context dependent modeling improves the performance significantly. We investigate whether this holds for cross-language transfer as well. Table 1 shows the performance for the seven speech recognizer applying context dependent models. Compared to context independent models on average only 1.2% relative improvement could be achieved. In our opinion, the main reason for the little improvements results from the poor context overlap across different languages. Therefore, the potential gain by the finer granularity and more accurate modeling is counteracted since the models do not fit to the new language. Indeed German and Japanese show slightly higher gains which correlates with the observed triphone coverage. From these results we conclude that significant gains can only be expected after the adaptation of the context dependent models to the new target language. Recently, we introduced the Polyphone Decision Tree Specialization method (PDTS) to overcome this problem [7] and achieved significant improvements.

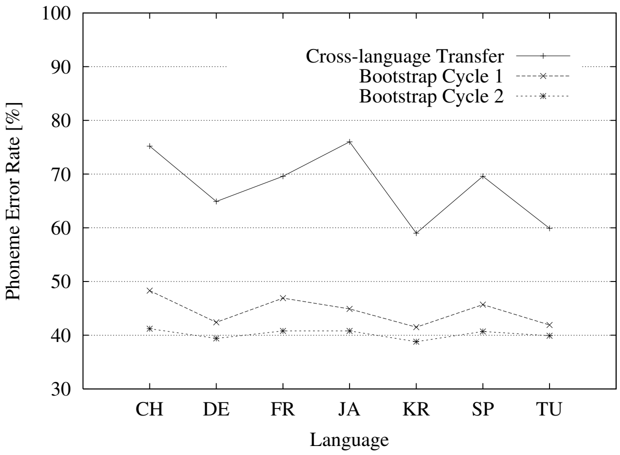

Fig. 1. Bootstrapping approach to Swedish [PER in %]

3.3. Bootstrapping

Figure 1 compares cross-language transfer to the bootstrapping approach assuming that a Swedish training database of about 17 hours spoken speech is available. After the cross-language initialization step the acoustic models are trained on these Swedish data by running two bootstrap cycles, each calculating a new Linear Discriminant Analysis, estimating new Gaussian mixtures through K-means clustering and performing 4 iteration of Viterbi training (see [5] for details). The results show an initial error range of 22% relative after the initialization step. The large difference between the languages reduces significantly to 5% and 6% respectively but does not level off completely. This result leads to two implication. First, as expected the choice for suitable seed models is more crucial in the cross-language than in the bootstrap scenario. Second, even in the bootstrap approach the right choice makes a difference of at least 5% relative, which is 2.4 percentage points in our experiments.

4. LANGUAGE INDEPENDENT SEED MODELS

In this section we describe experiments on language independent phonemes as seed models and compare the results to the language dependent case. Language independent models are obtained by sharing training data across languages whenever sounds of these languages are represented by the same IPA symbol (see [6] for details).

In the language independent case we have a variety of acoustic models to choose from as seed models. The previous results imply that choosing the appropriate seed models leads to significant performance improvements. Therefore, the definition of an appropriate mapping becomes an important issue. Once such a mapping is defined, it can be applied for bootstrapping: find suitable seed models, for adaptation: share data across languages, and for cross-language transfer: convert pronunciation dictionaries.

4.1. Phoneme Mapping

In the following experiments we compare a heuristic with two data-driven approaches to determine the mapping from language independent models to Swedish target models.

Knowledge-based Phoneme Mapping
Assuming that no training data in the target language is available, we applied a priori knowledge to find an appropriate mapping. A human expert defined this mapping according to the IPA scheme by picking the closest IPA counterpart of the Swedish target phoneme among all language independent phonemes. The first column of Table 2 gives the Swedish target phoneme in IPA convention. The second column "IPA-Map" presents the resulting heuristic phoneme mapping. Out of the seven-lingual phoneme set 39 of all 48 Swedish phonemes can be replaced with the exact matching IPA counterpart. The remaining nine phonemes which are marked by "(-)" are replaced by the closest possible match.

Data-driven Phoneme Mapping
When training data becomes available, a data-driven approach is an option. If phonetic transcription are provided by human experts or automatically derived from phoneme recognizer alignments even a supervised method is possible. In our experiments we assume that 500 spoken Swedish utterances (about 1 hour speech) and Viterbi-alignments from a Swedish phoneme recognizer are given. We decode the same utterances using a seven-lingual phoneme recognizer and calculated a confusion matrix between referenced (Swedish) and hypothesized (language independent) phoneme models. The confusions are computed by a frame-wise comparison of the alignments, normalized by the summed frequency of the hypothesized phoneme. The mapping for a Swedish target phoneme is derived by picking the hypothesized phoneme which leads to the highest normalized confusion score. The results are given in column 3 "Phone-Map" in Table 2.

Additionally, we extend the phone-based mapping to the sub-phone level to improve the potential of finding better seed models for the large Swedish phonetic inventory. To do so, the confusion matrix was calculated on the sub-phone instead of the phone sequence level. The results are shown in column 4 of Table 2 indicated by "Subphone-Map".

Target	IPA-Map	Phone-Map	Subphone-Map		
			-b	-m	-e
p	p	p	p-b	p-m	p-e
b	b	b	b-b	b-m	b-e
t	t	t	t-b	t-m	t-e
d	d	d	d-b	d-m	d-e
ʈ	t (-)	t	t-b	t-m	t-e
ɖ	d (-)	d	b-b	d-m	d-e
k	k	k	k-b	k-m	k-e
g	g	g	g-b	g-m	g-e
m	m	m	m-b	m-m	m-e
n	n	n	n-b	n-m	n-e
ɳ	n (-)	n	n-b	n-m	n-e
ŋ	ŋ	ŋ	ei-e	ɴ-m	ŋ-e
r	r	r	r-b	r-m	r-e
f	f	f	f-b	f-m	f-e
v	v	v	v-b	v-m	v-e
s	s	s	s-b	s-m	s-e
ʃ	ʃ	ʃ	ʃ-b	ʃ-m	k-e
ʂ	ʂ	ʃ	s-b	ʃ-m	ʃ-e
ç	ç	x	θ-b	ɯ-m	x-e
h	h	h	h-b	h-m	h-e
j	j	j	ʎ-b	j-m	j-e
l	l	l	l-b	l-m	l-e
ɭ	l (-)	l	l-b	l-m	l-e
ks	x (-)	s	ts-b	s-m	s-e
i	i	e	i:-b	i:-m	e-e
i:	i:	i	ʎ-b	i:-m	i-e
y	y	e:	e:-b	e:-m	i-e
y:	y:	e:	uei-m	e:-m	e:-e
ʉ:	u (-)	ø:	ø:-m	ɯ-m	ɯ-m
u	u	ʊ	ɔ-b	ʊ-b	ʊ-m
u:	u:	u	u-b	ʊ-m	'u-e
e	e	e:	e:-m	e:-m	e-e
e:	e:	e	e:-b	e:-m	e-e
ø	ø	œ	œ-e	œ-m	œ-e
ø:	ø:	œ	ɔ-m	œ-m	œ-e
ə	ə	e	i:-e	uei-m	e-e
ɵ	ə (-)	ɔ	y-m	ɔ-m	ɔ-e
o:	o:	o:	o:-b	o:-m	o:-e
ɛ	ɛ	e	e-b	e-m	e-e
ɛ:	ɛ:	e	e-b	ɛ-m	'e-e
œ	œ	ɐ	ɐ-b	ɐ-m	ɐ-e
œ:	œ (-)	eu	eu-b	eu-b	eu-m
ɔ	ɔ	o:	o:-b	o:-m	o:-e
æ	æ	e	e-b	e-m	ai-m
æ:	æ (-)	'a	œ-b	ɐ-m	'a-b
a	a	ɑ	a:-b	a:-m	'a-m
a:	a:	a:	œ-m	iao-m	au-m
ɑ:	ɑ	ɑ	a:-b	a:-m	'a-e

Table 2. Swedish phoneme mapping

Comparison
The comparison of the resulting mappings in Table 2 shows that Swedish consonants are mapped very consistently. However, in the group of vowels we found significant differences. The analysis of the confusion matrix confirms this

finding; the confusions between a reference consonant and its most frequent counterpart was always to an order of 10 higher than between reference and number two ranked counterpart, whereas for vowels the N-best confusion candidates are close together. This implies that consonants are less confusable across languages than vowels. One reason might be that vowels are more prone to coarticulatory variation.

Mapping approach		CL	Boot-1	Boot-2
Heuristic:	IPA-Map	65.8	43.9	40.2
Data-driven:	Phone-Map	60.9	43.8	39.7
	Subphone-Map	61.8	42.3	39.5

Table 3. Comparison of mapping approaches [PER in %]

Table 3 compares the error rates of the three mapping approaches for cross-language transfer (CL) and bootstrapping. It shows that the data-driven approach leads to better performance. In cross-language transfer the heuristic mapping is outperformed by the phone-based mapping by 7.4%, and by the sub-phone-based mapping by 6%. After the first bootstrapping cycle the difference drops down to 3.6% (Boot-1) and after the second cycle further down to 1.7% (Boot-2). While the phone-based mapping achieves better results in cross-language transfer, the sub-phone based one is better in the bootstrap approach.

4.2. Language Dependent vs Independent Models

Finally, we compare the multilingual recognizer consisting of language independent models trained across seven languages to the best-matched ($Best_{L7}$ =Croatian) and the average ($Average_{L7}$) of the seven monolingual engines. Since we share the data of seven languages to train the language independent models, more data is available to estimate the model parameters. Considering this fact we can apply a higher number of Gaussians for the acoustic models. In the previous experiments we used the same number of Gaussians for the language dependent and the independent models. In the following experiment we use seven times more Gaussians for the language independent than for the language dependent models. Table 4 compares the results of a multilingual phoneme recognizer using 128 Gaussians per model (Phone-map$_{128}$) with a system using 16 Gaussians (Phone-map). For cross-language transfer this leads to a 3.6% performance improvement. For the bootstrapping approach we gain 13.6% improvements, however this is obviously a result of the higher number of Gaussians allowing a finer granularity of the acoustic models concerning the Swedish training data.

5. CONCLUSION

In this paper we investigated what performance can be expected if no data is available to train acoustic models of a new target language. The results in the language dependent case imply that prior knowledge about the best suitable language makes a significant difference, since the performance variation is very high giving a 22% range. On the other hand if we use language independent models for cross-language transfer, prior knowledge becomes obsolete. We also examined different phoneme mapping approaches and showed that the language independent models outperform even the best-matched language dependent models when a data-driven phoneme mapping is applied.

Acoustic Models: Language		CL	Boot-1	Boot-2
Dependent:	$Best_{L7}$	59.0	41.5	38.8
	$Average_{L7}$	67.8	44.5	40.2
Independent:	Phone-Map	60.9	43.8	39.7
	Phone-Map$_{128}$	58.7	36.6	34.3

Table 4. Language dep. versus indep. models [PER in %]

6. REFERENCES

[1] A. Constantinescu and G. Chollet: *On Cross-Language Experiments and Data-Driven Units for ALISP*, Proc. ASRU, pp. 606–613, St. Barbara, CA 1997.

[2] S. Gokcen and J.M. Gokcen: *A Multilingual Phoneme and Model Set: Towards a universal base for Automatic Speech Recognition*, Proc. ASRU, pp. 599–603, St. Barbara, CA 1997.

[3] J. Köhler: *Language Adaptation of Multilingual Phone Models For Vocabulary Independent Speech Recognition Tasks*, Proc. ICASSP, pp. 417–420, Seattle, 1998.

[4] L. Osterholtz, C. Augustine, A. McNair, I. Rogina, H. Saito, T. Sloboda, J. Tebelskis, A. Waibel, and M. Woszczyna: *Testing Generality in JANUS: a Multilingual Speech Translation System*, Proc. ICASSP, San Francisco, CA 1992.

[5] T. Schultz and A. Waibel: *Fast Bootstrapping of LVCSR Systems with Multilingual Phoneme Sets*, Proc. Eurospeech, pp. 371–374, Rhodes 1997.

[6] T. Schultz and A. Waibel: *Language independent and language adaptive LVCSR*, Proc. ICSLP, pp. 1819–1822, Sydney 1998.

[7] T. Schultz and A. Waibel: *Polyphone Decision Tree Specialization for Language adaptation*, Proc. ICASSP, Istanbul 2000.

[8] B. Wheatley, K. Kondo, W. Anderson, Y. Muthusamy: *An Evaluation of Cross-language Adaptation For Rapid HMM Development in a new language*, Proc. ICASSP, pp. 237–240, Adelaide 1994.