CROSS-LANGUAGE BOOTSTRAPPING BASED ON COMPLETELY UNSUPERVISED TRAINING USING MULTILINGUAL A-STABIL

Ngoc Thang Vu, Franziska Kraus, Tanja Schultz

Cognitive Systems Lab (CSL), Institute for Anthropomatics, Karlsruhe Institute of Technology (KIT)

thang.vu@kit.edu, franziska.kraus@student.kit.edu, tanja.schultz@kit.edu

ABSTRACT

This paper presents our work on rapid language adaptation of acoustic models based on multilingual cross-language bootstrapping and unsupervised training. We used Automatic Speech Recognition (ASR) systems in English, French, German, and Spanish to build a Czech ASR system from scratch. System building was performed without using any transcribed audio data by applying three consecutive steps, i.e. cross-language transfer, unsupervised training based on the "multilingual A-stabil" confidence score [1], and bootstrapping. Based on the confidence score we selected 72% (16.6 hours) of the available audio data with a transcription WER of less than 14.5%. The cross-language bootstrap achieves a word error rate of 23.3% on the Czech development set and 22.4% on the evaluation set. These results are very promising as the performance compares favorably to the Czech ASR system which was trained on 23 hours of manually transcribed data (21.8% on the development set and 21.3% on the evaluation set).

Index Terms— rapid language adaptation of ASR, unsupervised training, multilingual A-Stabil

1. INTRODUCTION

With the distribution of speech technology products all over the world, fast and efficient portability to new languages becomes a practical concern. One of the major time and cost factors for developing large vocabulary continuous speech recognition (LVCSR) systems for new languages is the need for large amounts of transcribed training data. Detailed transcriptions require about 20-40 times real-time, and even after manual verification the final transcriptions are not free of errors. As described in [2] rapid development of an automatic speech recognition system (ASR) can greatly benefit from the use of unsupervised acoustic model training, i.e. the use of ASR hypotheses as transcriptions. Typically, unsupervised training is applied to improve an available ASR through the use of additional acoustic data. For best performance, confidence measures [3] [4] [5] [6] derived from the recognizer output are used to select or weight the contribution of the acoustic training data. If no suitable ASR system exists for a new language, the cross-language transfer technique [7] can be used, where a system developed for one language is applied to recognize another language without using any training data of the new language. Afterwards, an unsupervised training might be applied to improve the word error rate (WER) iteratively [8] [9]. Our results in [1] indicated that generating hypotheses by cross-language transfer based on acoustic models from several languages combined with the word-based confidence score "multilingual A-stabil" followed by unsupervised training is a very efficient ASR system building process. For our former experiments in [1] we had chosen source languages from the same language family as the target language. However, this choice may overestimate the results. Also, recognizers of the same language family may not always be at disposal. To study the generalization of our approach, we opted in this paper for a selection of resource-rich language recognizers in English, French, German and Spanish to build a Czech ASR system.

The remainder of this paper is organized as follows. In section 2 we describe the data resources and our baseline systems. Section 3 presents a comparison between the original and the modified cross-language transfer approach. In section 4 we introduce the confidence score "multilingual A-stabil" and the multilingual unsupervised training framework. Section 5 reports the experimental results on the Czech dataset. A summary in section 6 concludes the paper.

2. DATA RESOURCES AND BASELINE SYSTEMS

GlobalPhone is a multilingual text and speech corpus that covers speech data from 20 languages [10]. It contains more than 400 hours of speech spoken by more than 1900 adult native speakers. For this work we selected Czech, English, French, German, and Spanish from the GlobalPhone corpus. To retrieve large text corpora for language model building, we used our Rapid Language Adaptation Toolkit (RLAT) [11] for an up to twenty days crawling process [12]. For acoustic modeling, we applied the multilingual rapid bootstrapping approach which is based on a multilingual acoustic model inventory trained from seven GlobalPhone languages [13]. To bootstrap a system in a new language, an initial state alignment is produced by selecting the closest matching acoustic

models from the multilingual inventory as seeds. The closest match is derived from an IPA-based phone mapping. In this work, we did a phone mapping for each language and trained five different acoustic models, using the standard front-end by applying a Hamming window of 16ms length with a window overlap of 10ms. Each feature vector has 143 dimensions resulting from stacking 11 adjacent frames of 13 MFCC coefficient each. A Linear Discriminant Analysis transformation reduces the feature vector size to 42 dimensions. The model uses a fully-continuous 3-state left-to-right HMM. The emission probabilities are modeled by Gaussian Mixtures with diagonal covariances. Table 1 gives a breakdown of the trigram perplexities (PPL), Out-Of-Vocabulary (OOV) rate, vocabulary size, and WER for the selected languages.

Table 1. *PPL, OOV, vocabulary size, and WER for Czech, English, French, German, and Spanish*

Languages	PPL	OOV	Vocabulary	WER
Czech (CZ)	1,886	3.7%	276k	21.8%
English (EN)	284	0.5%	60k	15.4%
French (FR)	352	2.4%	65k	22.3%
German (GE)	148	0.4%	41k	13.2%
Spanish (SP)	224	0.1%	19k	23.3%

3. CROSS-LANGUAGE TRANSFER

Cross-language transfer refers to the technique where a system developed in one language is applied to recognize another language without using any training data of the new language [7]. In the original paper [7], the seed models were selected from a monolingual or multilingual acoustic model set to best match the target language phone set and dictionary. In [1] we introduced a modified cross-language transfer procedure in which we did not modify the acoustic model of the source language, but the pronunciation dictionary of the target language. I.e. we modeled Czech words with phones of the source languages by applying a manual phone mapping based on the IPA scheme. These mapped dictionaries allow for using the source language acoustic models in combination with the Czech pronunciation dictionary and language model in order to decode the Czech training data. Figure 1 depicts the modified cross-language transfer procedure with English as source and Czech as target language.

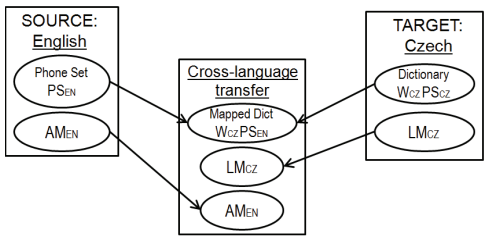

Fig. 1. Modified cross-language transfer (English to Czech)

Consequently, in contrast to [7], the modified approach will benefit from context similarities between languages by leveraging the context dependent acoustic models of the source language. The disadvantage of the modified method is that we adapt the phone models of the source language rather than those of the target language. This is compensated by fully retraining the target models as a final step of system building. In this work we apply the modified cross-language transfer from English, French, German, and Spanish as source languages to Czech as target language. Table 2 compares the performance between the original and the modified cross-language transfer approach based on the Czech development set. It also shows the percentage of polyphones from the target language covered by each source language, respectively. For comparison, we added our former results from [1] for Slavic languages as source language. The results indicate that modified cross-language transfer outperforms the original approach for those source language that belong to the same language family as the target language. This is most likely due to the fact that words (and contexts) are more similar among the Slavic languages and thus better leverage the context dependent acoustic models when mapping the dictionary. However, despite the significantly weaker performance for non-matching source languages, our results in section 5 will demonstrate that our approach generalizes well in combination with unsupervised training and cross-language bootstrapping.

Table 2. *Original vs modified cross-language transfer (WER)*

Languages	Original	Modified	abs. Δ	Polyphone Coverage
Bulgarian (BL)	67.0%	61.0%	6%	16.9%
Croatian (HR)	68.0%	57.2%	10.8%	15.6%
Polish (PL)	67.7%	55.8%	11.9%	13.2%
Russian (RU)	72.5%	64.3%	8.2%	10.0%
Spanish (SP)	85.4%	87.2%	-1.8%	6.8%
German (GE)	75.2%	75.2%	0%	6.4%
French (FR)	84.5%	95.2%	-10.7%	2.0%
English (EN)	87.4%	99.8%	-12.6%	0.4%

4. MULTILINGUAL UNSUPERVISED TRAINING FRAMEWORK

4.1. Multilingual A-Stabil

The basic idea of unsupervised training is to improve an acoustic model with transcriptions generated by an iterative decoding of audio training data. Automatically generated transcriptions are used to retrain the acoustic model using this data. However, to use available acoustic data effectively, it is crucial to utilize confidence measures for selecting or weighting the data contributions such that only almost correct training data is used. For this purpose we applied our method "multilingual A-stabil" [1] to compute confidence scores using n monolingual acoustic models. In our current experiments $n = 4$ as we used the acoustic models of English, French, German, and Spanish. Using a set of alternative

hypotheses derived from all four languages, we compute the frequency of each word of the reference output normalized by the number of alternative hypotheses. The best hypothesis of each language serves as reference output. In order to generate alternative hypotheses we build the word lattices first and use different weights of acoustic and language model of each language. As a result we get a multilingual arbiter which indicates the confidence for each word in the best hypothesis. Figure 2 illustrates the described method.

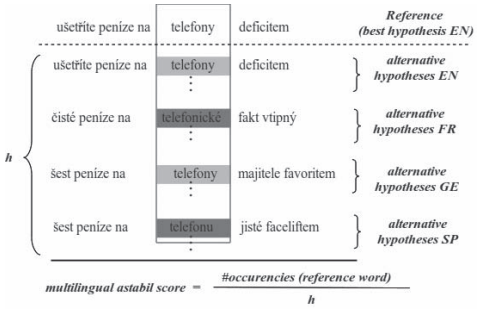

Fig. 2. "Multilingual A-stabil" confidence score computation

Figure 3 shows the plot of recognition error (WER) over this score which presents a very high correlation between multilingual A-Stabil and the recognition error for well-trained acoustic models as well as poorly estimated acoustic models.

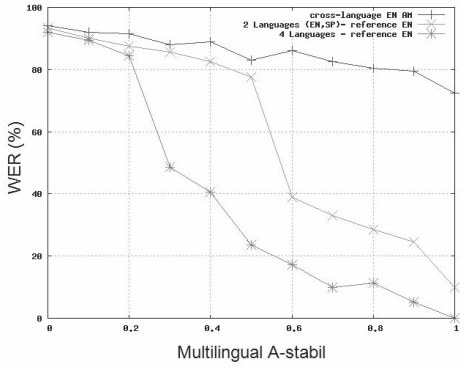

Fig. 3. WER over multilingual A-stabil using one (EN), two (EN, SP), and all four languages for cross-language transfer

4.2. Multilingual unsupervised training framework

In this section we present our multilingual unsupervised training framework, which mainly consists of two steps, in the following called initial and final step. The initial step is an iterative process, in which we use several acoustic models to generate automatic transcriptions. We applied cross-language transfer to decode the audio training and development data. Using the development set we evaluated "multilingual A-stabil" and estimated a suitable threshold. Afterwards all words that have a confidence score higher than this threshold were selected for acoustic model adaptation. In our work a MAP adaptation was applied iteratively to improve acoustic models and thus increase the amount of data. This process terminates if the gain in amount of adaptation data from one iteration to the next is smaller than 5% relative. By using this process we could enlarge the amount of automatic transcriptions with a high precision on one side and select data from many different contexts due to the multilingual effect on the other side. In the final step, we used the multilingual inventory which was trained earlier from seven GlobalPhone languages [13] to write the alignment for the selected data extracted in the intial step and train the acoustic model. The final acoustic model is the one with the best performance on the development set. Figure 4 illustrates the multilingual unsupervised training framework.

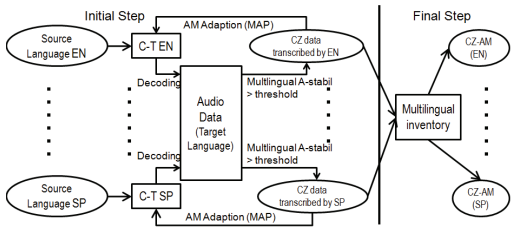

Fig. 4. Multilingual unsupervised training framework

5. EXPERIMENTAL RESULTS

5.1. Iterative Automatic Generation of Transcriptions

We started by applying the original cross-language transfer based on English (EN), French (FR), German (GE), and Spanish (SP) acoustic models without any retraining in order to recognize the Czech development set. WER is relatively high, with 87.35% for EN, 84.52% for FR, 75.30% for GE, and 85.42% for SP. With these initial models we decoded the Czech training data and selected appropriate adaptation data using the "multilingual A-stabil" confidence scores. As confidence threshold for data selection we heuristically picked 0.3 since for scores larger than this thresholds those words occuring in alternative hypotheses must originate from more than one language. Figure 3 compares the recognition error over the "multilingual A-stabil" score on the Czech development set for the first iteration using one language (EN), two languages (EN and SP), and four languages (EN, FR, GE, SP) for cross-language transfer, respectively. At a confidence score of 0.3, WER drops rapidly (for the first iteration from 82% to 50%). Furthermore, using four languages outperforms the one- and two-language transfer results. To our believe this indicates the benefit of our multilingual approach. We terminate the process after 4 iterations as gains seem to saturate. Table 3 summarizes the amount of selected data after each iteration given in percentage of all untranscribed words and shows the resulting transcription quality in terms of WER for the case of English and German.

Table 3. *Amount and quality of generated transcriptions*

Initial Step Iteration	Amount of data EN	Amount of data GE	% of all data EN	% of all data GE	Quality (WER in %) EN	Quality (WER in %) GE
1	1.6h	2.3h	7.0	10.1	33.5%	27.1%
2	7.6h	8.7h	33.1	37.8	23.5%	22.9%
3	9.4h	10.1h	41.1	43.6	22.8%	23.4%
4	9.7h	10.2h	42.2	44.2	23.2%	23.5%

5.2. Cross-language Bootstrapping

After Czech acoustic training data was generated and selected, an initial state alignment is produced by finding the closest matching acoustic models from the multilingual inventory as seeds. The closest match is derived from an IPA-based phone mapping. Then the Czech system is completely rebuilt using the seed acoustic models and the selected data for training (one data set per source language). We built a quintphone system with 2000 models by applying merge&split and Viterbi training. Table 4 summarizes the performance on the Czech development set for the systems trained with the four data sets. The WER ranges from 25.7% to 28.4% after the first iteration. To increase the amount of the acoustic training data, we again decode the training data. This time the acoustic models from the previous iterations were applied together with data selected from high multilingual A-stabil scores. We obtained automatic transcriptions of about 72% (16.6 h) training data with a quality of 14.5% WER. For the 2nd iteration we used the acoustic model from the 1st iteration to generate the state alignments and trained the system with the same parameters as in iteration 1 afterwards. The resulting best system achieves 23.3% WER on the Czech development set and 22.4% WER on the evaluation set. The results show that iterative unsupervised training with multilingual A-Stabil results in accurate automatic transcriptions that allow to further improve the acoustic model of the target language.

Table 4. *Cross-language Bootstrapping (Czech dev set)*

Final Step	English	French	German	Spanish
1st iteration	27.9	28.4	27.7	25.7
2nd iteration	23.9	23.3	23.5	23.7

6. SUMMARY

In this paper we investigated a multilingual unsupervised training procedure. We developed a Czech ASR without any transcribed training data using English, French, German, and Spanish acoustic models. A combination of cross-language transfer and unsupervised training was applied. We explored the relative effectiveness of using acoustic models from more than one language for cross-language transfer and "multilingual A-stabil" to select Czech audio data. The results are very promising achieving 16.6 hours (72%) of all available audio training data. The generated automatic transcriptions have a WER of about 14.5% WER. The best Czech ASR system has 23.3% WER on the development set and 22.4% on the evaluation set, which is very close to the performance of the Czech ASR trained with 23 hours audio data with manual transcriptions (21.8% on the development set 21.3% on the evaluation set). These results compare very favorably to our former results when using same-family languages and indicate that our approach generalizes well.

7. ACKNOWLEDGMENT

This work was partly realized as part of the Quaero Programme, funded by OSEO, French State agency for innovation.

8. REFERENCES

[1] N. T. Vu, F. Kraus and T. Schultz. Multilingual A-stabil: A new confidence score for multilingual unsupervised training. In IEEE Workshop on Spoken Language Technology, SLT 2010, Berkeley, California, USA, 2010.

[2] G. Zavaliagkos and T. Colthurst. Utilizing untranscribed training data to improve performance, in DARPA Broadcast News Transcription and Understanding Workshop, Landsdowne, VA, USA, Feb. 1998.

[3] T. Kemp and T. Schaaf. Estimating confidence using word lattices. In Proc. of Eurospeech, pp. 827-830, 1997.

[4] F. Wessel, K. Macherey and H. Ney. A comparison of wordgraph and N-best list based confidence measures. In Proc. of Eurospeech, Budapest, Hungary, 1999.

[5] G. Evermann and P. Woodland. Large vocabulary decoding and confidence estimation using word posterior probabilities. In Proc. ICASSP, Istanbul, Turkey, 2000.

[6] R. Zhang and A. I. Rudnicky. A New Data Selection Approach for Semi-Supervised Acoustic Modeling. In Proc. of ICASSP, Toulouse, France, 2006.

[7] T. Schultz and A. Waibel. Experiments on cross-language acoustic modeling. In Proc. Eurospeech, Aalborg, Denmark, 2001.

[8] J. Lööf, C. Gollan, and H. Ney. Cross-language Bootstrapping for Unsupervised Acoustic Model Training: Rapid Development of a Polish Speech Recognition System. In Interspeech, pages 88-91, Brighton, U.K., 2009.

[9] L. Lamel, J. Gauvain and G. Adda. Unsupervised acoustic modelling. In: Proc. ICASSP Orlando, USA, 2002.

[10] T. Schultz. GlobalPhone: A Multilingual Speech and Text Database developed at Karlsruhe University. In: Proc. ICSLP Denver, CO, 2002.

[11] T. Schultz and A. Black. Rapid Language Adaptation Tools and Technologies for Multilingual Speech Processing. In: Proc. ICASSP Las Vegas, USA 2008.

[12] N.T. Vu, T. Schlippe, F. Kraus, and T. Schultz. Rapid Bootstrapping of five Eastern European Languages using the Rapid Language Adaptation Toolkit. In Interspeech, Makuhari, Japan, 2010.

[13] T. Schultz and A. Waibel. Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition. In Speech Communication August 2001., Volume 35, Issue 1-2, pp 31-51.