Rapid building of an ASR system for Under-Resourced Languages based on Multilingual Unsupervised Training

Ngoc Thang Vu, Franziska Kraus, Tanja Schultz

Cognitive Systems Lab, 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 the six source languages English, French, German, Spanish, Bulgarian and Polish to build from scratch an ASR system for Vietnamese, an under-resourced language. 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. We investigated the correlation between performance of "multilingual A-stabil" and the number of source languages and improved the performance of "multilingual A-stabil" by applying it at the syllable level. Furthermore, we showed that increasing the amount of source language ASR systems for the multilingual framework results in better performance of the final ASR system in the target language Vietnamese. The final Vietnamese recognition system has a Syllable Error Rate (SyllER) of 16.8% on the development set and 16.1% on the evaluation set.

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

1. Introduction

In light of the world's globalization, one of the most important trends in present-day speech technology is the need to support multiple input and output languages, especially when applications are intended for international markets and linguistically diverse user communities. As a result, new algorithms and strategies are required, which support a rapid adaptation of speech processing systems to new languages. Currently, the time and costs associated with this task is one of the major bottlenecks in the development of multilingual speech technology. One of the major time and cost factor for developing LVCSR systems for new languages is the need for large amounts of transcribed training data. While the acquisition of large data resources is a challenging task in the case of under-resourced languages in general, the transcription of data typically involves language experts in particular, requires manual labor and is error prone. 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 [3] 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 [4] [5] [6] [7] 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 [8] 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 [9] [10]. Our results in [1] [2] 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. In this paper we applied our multilingual unsupervised training framework to Vietnamese, an under-resourced Asian language which is very different from all source languages. We assumed to not have manual transcriptions for the Vietnamese audio data and used ASR systems from six European languages, namely English, French, German, Spanish, Bulgarian, and Polish to build a Vietnamese ASR system from scratch and thereby prove the generalization of our approach. We investigated the correlation between the performance of "multilingual A-stabil" and the number of languages and improved the performance of "multilingual A-stabil" by applying it at syllable level. Furthermore, we explored the performance of the target language ASR system by increasing the amount of source language ASR systems for the multilingual framework.

The remainder of this paper is organized as follows. In section 2 we describe the data resources and our baseline systems. Section 3 describes our multilingual unsupervised training framework for rapid building an ASR system for a new language. In section 4 we introduce the confidence score "multilingual A-stabil", its performance on the syllable level, the correlation between the performance of "multilingual A-stabil" and the number of languages. Section 5 reports the experimental results on the Vietnamese 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 19 languages [11]. It contains more than 400 hours of speech spoken by more than 1900 adult native speakers. For this work we selected Vietnamese, English, French, German, Spanish, Bulgarian, and Polish from the GlobalPhone corpus. To retrieve large text corpora for language model building, we used our Rapid Language Adaptation Toolkit (RLAT) [12] for an up to twenty days crawling process [13]. For acoustic modeling, we applied the multilingual rapid bootstrapping approach which is based on a multilingual acoustic model inventory trained from seven GlobalPhone languages [14]. To bootstrap a system in a new language, an initial

Copyright © 2011 ISCA

state alignment is produced by selecting the closest matching acoustic models from the multilingual inventory as seeds. The closest match is derived from a manual IPA-based phone mapping. In this work, we did a phone mapping for each language and trained seven 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, Syllable Error Rate (SyllER) for Vietnamese and WER for the source languages. For the Vietnamese ASR system we improved the system by using tonal features for tone modeling and trained the acoustic model with trancriptions based on multisyllabic words to increase the context for cross-word context dependent modeling [15]. The performance of our best Vietnamese system is 11.8% SyllER on the development set. In this paper we aim at building a Vietnamese ASR system under the assumption that no transcriptions are given. This lack of transcribed audio data is compensated by decoding the available audio data and thereby generating transcripts in a fully unsupervised way.

For decoding we applied a trigram language model which includes the training data transcripts. However, different from [15] we set the cut-off parameter to 10 such that the effect of training transcripts is only moderate while the LM still covers the most frequent bigrams and trigrams.

Table 1: *PPL, OOV, vocabulary size, and ER for 7 languages (SyllER for Vietnamese and WER for other 6 languages)*

Languages	PPL	OOV	Vocabulary	ER
Vietnamese (VN)	323	0%	35k	14.3%
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%
Bulgarian (BL)	500	1.0%	274k	20.3%
Polish (PL)	1372	2.9%	243k	24.3%

3. 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 (C-T) [8] 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 give 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 iterative 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 trained the multilingual inventory with all data from the existing source languages via a phone merging based on IPA [16]. The resulting inventory is used to write forced alignments for the selected data extracted in the intial step and train the acoustic model. Finally we applied confusion network combination (CNC) [17] to merge the results from all acoustic models. Figure 1 illustrates the multilingual unsupervised training framework.

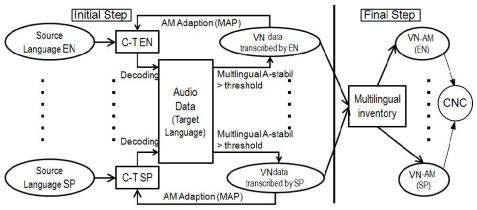

Figure 1: *Multilingual unsupervised training framework*

4. Multilingual A-stabil confidence score

4.1. Basic idea

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. However, to apply 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 [1] [2] it was proved that this confidence score is very stable even if the acoustic model is relatively weak. Using a set of alternative hypotheses derived from all six source 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. To evaluate this confidence

Figure 2: *"Multilingual A-stabil" confidence score*

score we computed the recognition error over score intervals. Figure 3 plots the recognition error (SyllER) over the confidence score intervals for one (EN), two (EN and SP), four (EN, SP, GE and FR) and six source languages (EN, SP, GE, FR, BG and PL). The results imply that with an increasing number of source languages the "multilingual A-Stabil" confidence score

becomes a better indicator for resulting transcription quality in terms of recognition error.

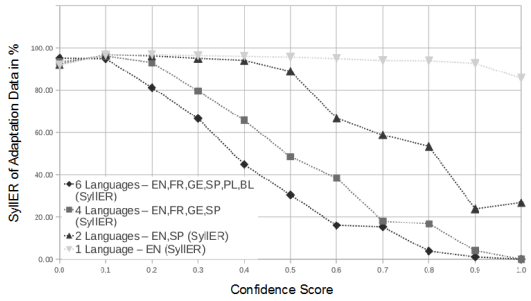

Figure 3: *SyllER over multilingual A-stabil using one (EN), two (EN, SP), four (EN, SP, GE, FR) and all six languages*

4.2. Syllable- vs. Word-based

In order to improve "Multilingual A-stabil" we compute this confidence score on the syllable level, i.e. we splitted Vietnamese words into syllables before computing the confidence score. We found the voting process to be more efficient at syllable level than at word level. Therefore, we can extract more data using the same confidence threshold. Further benefit of generating automatic transcriptions on syllable level is that coarticulation effects can be modeled by an adaptation or training process. Table 2 shows the amount of data and the quality of automatic transcriptions in terms of SyllER by application of "Multilingual A-stabil" at syllable and word level for four different languages (EN, SP, GE and FR) with threshold of 0.3 for the first iteration. It indicates that we gain 24% more training data by applying "Multilingual A-stabil" at syllable level while achieving almost the same transcription quality. We applied "Multilingual A-stabil" at syllable level for the remainder experiments.

Table 2: *Syllable- vs. Word-based "Multilingual A-stabil"*

	Amount	SyllER	Rel. Gain
Word-based	0.75h	51.54%	
Syllable-based	0.93h	52.83%	+24%

5. Experimental Results

We investigated the impact of the number of source languages used for transcription generation to the performance of the final Vietnamese ASR system. We applied our training framework for two (EN, SP), four (EN, SP, FR, GE), and six languages (EN, SP, FR, GE, BG, PL).

5.1. Iterative Automatic Generation of Transcriptions

We started by applying cross-language transfer based on English (EN), French (FR), German (GE), Spanish (SP), Bulgarian (BG) and Polish (PL) acoustic models without any retraining in order to recognize the Vietnamese development set. The SyllER is very high with 90.93% for EN, 92.81% for FR, 93.49% for GE, 89.72% for SP, for 88.49% BG and 86.58% for PL which indicates the challenges of building a Vietnamese ASR system from scratch without any transcriptions. With these initial models we decoded the Vietnamese training data and selected appropriate adaptation data using the "multilingual A-stabil" confidence scores. As we observed in [1] [2], the SyllER drops rapidly when we select those words which are voted for by at least two languages. To reflect this with two, four, and six languages, we chose 0.6, 0.3, and 0.2 as confidence score thresholds. We terminate the process after four iterations as gains seem to saturate. Figure 4 displays the amount of selected data over the iterations given in percentage of all untranscribed words. It shows the resulting transcription quality in terms of SyllER by using two (EN, SP), four (EN, SP, FR and GE) and six source languages (EN, SP, FR, GE, BG and PL) that cover 26, 27, and 28 of the 39 Vietnamese phonemes. The results indicate a close relation between the amount of extracted data and the number of languages respectively the phone coverage. The more target languages we use in our training framework, the more phoneme can we cover from the target language and thereby the more data we are able to select. However, Figure 4 also indicates that the quality of the automatic transcriptions is getting slightly worse when using more languages.

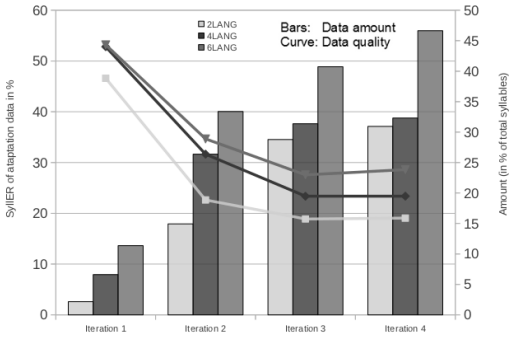

Figure 4: *Amount of selected data given in percentage of all syllables and the corresponding resulting transcription quality in terms of SyllER*

5.2. Cross-language Bootstrapping

We used the selected Vietnamese acoustic training data with the automatic transcriptions from the initial step to train the Vietnamese acoustic model in this final step. First, we trained the multilingual inventory with all existing data from the source languages by applying an IPA-based phone merging [16]. The closest match is derived manually according to IPA similarity. Table 3 summarizes the performance of multilingual acoustic models MM2 (EN, SP), MM4 (EN, FR, GE, and SP), and MM6 (EN, SP, FR, GE, BG, and PL) after cross-language transfer on the development set. The results indicate that a larger number of source languages used to train the multilingual acoustic models improves the cross-language transfer performance on Vietnamese development set. Afterwards, an initial state alignment

Table 3: *Cross-language transfer performance (on dev set) of multilingual acoustic model MM2 (EN, SP), MM4 (EN, SP, FR and GE) and MM6 (EN, SP, FR, GE, BG and PL)*

Systems	SyllER	Rel. Delta
MM2	87.54%	
MM4	82.35%	+5.9%
MM6	76.45%	+7.2%

for the Vietnamese training data is produced by determining the closest matching acoustic models from the multilingual inventory as seeds. Then the Vietnamese 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 in which the number of Gaussian mixtures per state is determined using merge&split training with maximum of 64 mixtures per state. We then applied a global Semi-Tied Covariance (STC) matrix [18] to all the acoustic models followed by three iterations of Viterbi training. To increase the amount of selected acoustic training data, we again decoded the training data. This time the acoustic models from the previous iteration were applied together with data selected applying "multilingual A-stabil" scores. For second iteration we used the acoustic model from the first iteration to generate the state alignments and then trained the system with the same parameters as in iteration 1. Figure 5 summarizes the performance of our Vietnamese ASR system after the first and second iteration and after applying Confusion Network Combination in terms of SyllER by using two (EN, SP), four (EN, SP, FR and GE), and six source languages (EN, SP, FR, GE, BG and PL). The resulting best system achieves up to 16.8% SyllER on the Vietnamese development set and up to 16.1% SyllER 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. Furthermore, using more different languages for our multilingual unsupervised training framework results in better performance of the final Vietnamese ASR system. The final result on the development set is close to the baseline system, trained on about 22 hours transcribed data which achieves 14.3%. But it is still worse than our best system where we applied various language specific optimization steps which achieves 11.8% SyllER.

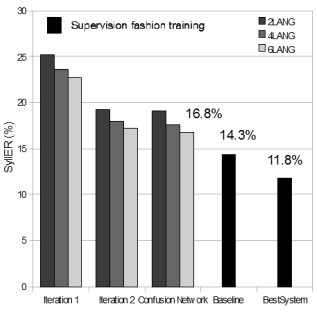

Figure 5: *Cross-language bootstrapping for Vietnamese by using two (EN, SP), four (EN, SP, GE, FR) and all six languages*

6. SUMMARY

In this paper we applied our multilingual unsupervised training framework to Vietnamese, an under-resourced Asian language. We developed a Vietnamese ASR system without any transcribed training data using ASR systems from different languages namely English, French, German, Spanish, Bulgarian, and Polish. A combination of cross-language transfer and unsupervised training was applied by using the "multilingual A-stabil" confidence measure. Our results show that by applying "multilingual A-stabil" at syllable level we could select more data with similar transcription quality compared to "multilingual A-stabil" at word level. Furthermore, the more source languages we used for our multilingual unsupervised training framework, the better is the ASR system for the target language. The best final Vietnamese ASR trained by our framework has a SyllER of 16.8% on the development set and 16.1% on the evaluation set. These results are quite close to the baseline system, trained on about 22 hours transcribed data which achieves 14.3%. But it is still worse than our best system where we applied various language specific optimization steps which achieves 11.8% SyllER.

7. 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] N. T. Vu, F. Kraus and T. Schultz. Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil. In International Conference on Acoustics, Speech and Signal Processing, ICASSP 2011, Prague, Czech Republic, 22-27 Mai, 2011.

[3] 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.

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

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

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

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

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

[9] 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.

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

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

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

[13] 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.

[14] 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.

[15] N. T. Vu, T. Schultz. Vietnamese Large Vocabulary Continuous Speech Recognition. In Automatic Speech Recognition and Understanding, ASRU 2009, Merano, Italy, 13.12.2009.

[16] Handbook, IPA: Handbook of the International Phonetic Association, 1999.

[17] L. Mangu, E. Brill and A. Stolcke. Finding Consensus Among-Words: Lattice-Based Word Error Minimization. In Proc. of EUROSPEECH99, Budapest, Hungary.

[18] M. Gales, "Semi-tied covariance matrices for hidden Markov models", IEEE Transactions Speech and Audio Processing, vol. 7, pp. 272-281, 1999.