MODIFIED POLYPHONE DECISION TREE SPECIALIZATION FOR PORTING MULTILINGUAL GRAPHEME BASED ASR SYSTEMS TO NEW LANGUAGES

Sebastian Stüker

Institut für Theoretische Informatik
Universität Karlsruhe (TH)
Karlsruhe, Germany
stueker@ira.uka.de

ABSTRACT

Automatic speech recognition (ASR) systems have been developed only for a very limited number of the estimated 7,000 languages in the world. In order to avoid the evolvement of a digital divide between languages for which ASR systems exist and those without one, it is necessary to be able to rapidly create ASR systems for new languages in a cost efficient way. Grapheme based systems, which eliminate the costly need for a pronunciation dictionary, have been shown to work for a variety of languages. They are thus destined for porting ASR systems to new languages. This paper studies the use of multilingual grapheme based models for rapidly bootstrapping acoustic models in new languages. The cross language performance of a standard, multilingual (ML) acoustic model on a new language is improved by introducing a new, modified version of polyphone decision tree specialization that improves the performance of the ML models by up to 15.5% relative.

Index Terms— Automatic Speech Recognition, Grapheme based acoustic models, Rapid Porting of ASR systems, Multilingual ASR

1. INTRODUCTION

1.1. A Digital Language Divide Emerges

Linguists estimate the number of currently existing languages to be between 5,000 and 7,000. The fifteenth edition of the Ethnologue [1] list 7,299 languages. Only for a small fraction of these languages automatic speech recognition (ASR) systems have been developed so far. Languages addressed are mainly those with either a large population of speakers, with sufficient economic funding, or with high political impact. The fact that applications using ASR only address a small fraction of the world's languages bears the danger of creating a digital divide between those languages for which ASR systems exist and those without one.

Languages are frequently disappearing. In [2] Janson estimates that in a few generations at least 1,000 of today's languages will have disappeared and that, if the trend holds, in as little as one hundred years half of today's languages will be extinct. Janson attributes this vanishing of languages to a frequently occurring switch to more prevalent languages. The creation of a digital divide as mentioned above is very likely to contribute to this kind of extinction of languages, might even accelerate it. In order to be able to preserve a high language diversity and cultural richness that comes with it, it is thus necessary to create methods for rapidly porting speech recognition systems to new languages, with possibly few resources for development, either in terms of money or available data and knowledge.

The author would like to thank the reviewers for their helpful remarks.

1.2. Grapheme Based ASR

In large vocabulary ASR systems, phonemes are traditionally used as modeling units. They require as a central component a pronunciation dictionary that maps the textual representation of the words to be recognized to their phonemic sequence. The creation of the pronunciation dictionary can be very costly in terms of time and money. It often requires the help of a phonetic expert in the targeted language. Though automatic or semi-automatic methods for phoneme-to-grapheme conversion exist, they still require a training phase and training material, and sometimes manual post-processing by an expert in order to achieve good results. Therefore, the creation of a suitable pronunciation dictionary can easily become either too expensive, may require too much time, or might even be impossible due to the lack of an expert, especially for less prevalent languages.

[3], [4], [5], [6], and [7] have shown that the use of graphemes as modeling units, instead of phonemes, can be a suitable approach for a range of languages. Whether this approach is successful or not depends on the grapheme to phoneme relationship of the particular language. [8] and [5] also conducted first experiments in building multilingual acoustic models based on graphemes, and [5] very briefly reported on porting grapheme models to a new language in a rudimentary way and under the assumption that a large amount of training data in the new language is available.

1.3. Porting Grapheme Based Recognizers to New Languages

While previous work has focused on multilingual grapheme based acoustic modeling, this paper examines the problem of porting grapheme based recognition systems to new languages. In this, the behavior of multilingual acoustic models based on graphemes is examined in more detail than in any prior work. Also, the porting experiments unlike [5] assume only a limited amount of adaptation material in the target language as given.

Polyphone decision tree specialization (PDTS) [9] is a technique for porting the decision tree of a recognition system to a new language and has been applied to phoneme based recognizers in the past. This work applies PDTS to grapheme models and refines the model initialization of the PDTS models. PDTS is further modified and improved by combining it for the first time with a decision tree pruning technique such as described in [10].

The rest of the paper is structured as follows. Section 2 introduces the task and data on which the experiments were performed. Section 3 then briefly describes the monolingual grapheme based ASR systems which were derived from earlier work and which give a ceiling for the porting performance, while Section 4 discusses the baseline multilingual ASR systems used for the porting experiments. Section 5 then describes the naive way of porting grapheme based recognition systems to new languages, while, finally, Section 6 describes how to improve the naive approach by first using PDTS and then improving porting performance even further by applying the modified version of PDTS introduced in this paper.

2. CORPUS AND TASK

The experiments in this paper were conducted on a selection of languages from the GlobalPhone [9] corpus. GlobalPhone is an ongoing data collection effort that now provides transcribed speech data that was collected in an uniform way in 18 languages. The corpus contains newspaper articles read by native speakers and is modeled after the Wall Street Journal 0 (WSJ0) corpus.

For the work presented, the four languages English (EN), German (GE), Russian (RU), and Spanish (SP) were used. Since English is not part of GlobalPhone, the WSJ0 corpus was used. For every language three data sets are available: one for acoustic model training (train), one for development work (dev) such as finding the correct language model weight, and one for evaluation (eval). All three sets are speaker disjunct. When it comes to the porting experiments German receives the role of the new target language to which to port, while for the other languages good models are assumed to be already available. For German it was assumed that only a limited set of 2h of adaptation data (adapt) is available.

		EN	GE	RU	SP
train	hours	15.0	16.0	17.0	17.6
dev	hours	0.4	0.4	1.3	2.1
eval	hours	0.4	0.4	1.6	1.7
adapt	hours	–	2	–	–

Table 1. Size of the data sets in hours

3. MONOLINGUAL GRAPHEME BASED SPEECH RECOGNITION

The performance of the multilingual systems and rapidly ported systems in this paper is best compared against the performance of monolingual grapheme based recognition systems that were trained on their respective language only. From the experience with phonemic recognition systems it can be expected that their performance serves as an upper bound of the performance of the multilingual and ported systems. The monolingual recognizers in this paper are similar to the ones described in [5], [11], and [6]. The preprocessing and training procedures were only slightly modified and harmonized over all languages involved. All acoustic models are left to right Hidden Markov Models (HMM) with three sub-states per grapheme. All experiments in this work were performed with the help of the Janus Recognition Toolkit (JRTk) that features the Ibis single pass decoder [12].

The preprocessing is the same as in [6] and is based on mel scaled cepstral coefficients that are transformed with the help of a linear discriminant analysis (LDA). Training was done with the help of forced alignments obtained with the systems trained in [5], [11], and [6]. First the LDA matrix is estimated, after that random samples for every model were extracted in order to initialize the models with the help of the k-means algorithm. Then these models were refined by eight iterations of label training along the forced alignments. Context-independent (CI) as well as context-dependent (CD) models were trained in this way. The polyphone decision trees for the context-dependent models were taken from the systems from which the forced alignments were obtained, and contained 3000 models each. The decision trees only ask so called 'singleton' questions [5], that is, which graphemes are to the left or the right of the center grapheme of a polygrapheme.

Table 2 shows the word error rates of the context-dependent and context-independent models for every language on their respective development and evaluation sets. The trigram language models used for decoding were unchanged from the previous experiments in [5] and [6]. The differences among the different languages reflect their suitability for the grapheme based approach as well as inherent differences in the respective languages.

		EN	RU	SP	GE
CI	dev	61.2%	52.9%	49.5%	46.0%
	eval	60.2%	56.5%	36.3%	47.1%
CD	dev	17.0%	36.4%	27.0%	14.7%
	eval	18.5%	39.0%	18.3%	15.4%

Table 2. WER of the monolingual grapheme based ASR systems

4. MULTILINGUAL GRAPHEME BASED SPEECH RECOGNITION USING ML-MIX

For training the multilingual grapheme based ASR system the technique ML-Mix [13] was applied. When using ML-Mix, graphemes that are common to one language share the same model and are treated as identical in the rest of the system, e.g. in the polyphone decision tree. All information about which language a grapheme belongs to, is discarded in the system and the data from all languages for this grapheme is used for training it. Since Russian uses a Cyrillic script instead of a Latin based one, as the other three languages involved do, the Cyrillic graphemes were mapped to a romanized representation [6].

First, a context-independent ML-Mix recognizer (ML3-Mix-CI) on the languages that we assume as given — English, Russian, and Spanish — was trained. Then a polygrapheme decision tree with three thousand models was clustered and trained on these languages (ML3-Mix-CD). Table 3 gives the word error rates of the resulting models on the dev and eval sets of the individual languages that were used for training. One can see from the results that for the languages English and Russian there is a clearly visible performance degradation compared to the monolingual recognizers. The degradation for English is larger than for Russian which is to be expected, since English has a more complex grapheme-to-phoneme relation than Russian. Also, Russian contains many graphemes that are not common to the other two languages, so that their models are not broadened by the training material coming from the other languages. For Spanish a high degradation is only visible for the context-independent models. The context-dependent models show only a small degradation on the development data and no degradation on the evaluation data. This is due to the fact, that the, in comparison simple, grapheme-to-phoneme relation for Spanish can be captured by the polyphone decision tree, and no significant tainting of the shared models seems to take place by the sharing of training material.

4.1. Influence of the Multilingual LDA Transformation

[14] has shown for phoneme based models that an LDA matrix that has been trained on many languages performs either equally well or only slightly worse than a monolingual LDA matrix. In order to verify this result for grapheme based models the monolingual ASR systems for English, Russian, and Spanish were retrained, this time using the LDA matrix from the ML3-Mix models. The results in Table 4 show the same behavior for the grapheme based systems as for the phoneme based systems in [14], that is no or only a slight degradation. Furthermore, when using the LDA matrix trained on English, Russian, and Spanish for the German ASR system, the recognition performance improves slightly. Thus, the multilingual LDA matrix is suited for porting ASR systems to new languages.

		EN	RU	SP
ML3-Mix-CI	dev	77.5%	65.5%	62.5%
	eval	74.7%	69.2%	48.2%
ML3-Mix-CD	dev	24.3%	40.7%	28.9%
	eval	26.7%	43.2%	18.3%

Table 3. WER of the ML3-MIX models on the training languages

		EN	RU	SP	GE
CI	dev	62.0%	52.9%	49.9%	44.8%
	eval	59.9%	56.5%	36.5%	45.8%
CD	dev	16.8%	35.8%	27.4%	14.5%
	eval	18.8%	39.4%	18.1%	15.0%

Table 4. WER of the monolingual models using the ML3-MIX LDA

5. NAIVE PORTING OF ML-MIX TO GERMAN

A first, naive approach for creating a new ASR system for German, is to apply the ML3-Mix models directly to German. This has the advantage that no training material in German is needed. But as known from earlier work one can expect a very low performance. Indeed, as can be seen from the results in Table 5, the performance on German is rather poor. For the context-independent models we can see a relative increase in WER of 96% in comparison to the German, monolingual recognizer. For the context-dependent models the WER rises by 448% relative on the dev set and 435% on the eval set. The, in comparison to the context-independent models, four times higher loss in performance for the context-dependent models, suggests that one of the major sources for the WER increase is the multilingual polyphone decision tree.

	dev	Δ	eval	Δ
ML3-Mix-CI	90.8%	97%	92.2%	96%
ML3-Mix-CD	80.6%	448%	82.4%	435%

Table 5. WER of ML3-Mix on German and increase in WER over the German ASR

5.1. Naive Adaptation

The easiest way to exploit the 2h of German adaptation data for improving the performance of the ML3-Mix model on German is to use it for adapting the model's parameters. To do so, 4 iterations of label training on the German data were performed with ML3-Mix. This reduces the WER of the models to 38.6% on the German dev and 38.6% on the eval data (see also Table 7).

5.2. Influence of the Polyphone Decision Tree on Porting Performance

In order to determine the influence of the polyphone decision tree on the porting performance, the monolingual, context-dependent ASR systems were retrained, this time using the multilingual LDA matrix and the multilingual polyphone decision tree from the ML3-Mix system. The word error rates of the resulting systems, as well as the relative increases in WER, compared to the purely monolingual recognizers are listed in Table 6. For English, Russian, and Spanish there is only a moderate increase in WER that is always well below 10% relative. However, for German the increase is massive. This enormous increase is soley due to the multilingual polyphone decision tree which only poorly fits the German data. Therefore, the issue of better fitting the polyphone decision tree to the target language needs to be addressed.

	EN	RU	SP	GE
dev	18.6%	38.6%	27.5%	25.4%
Δ	9.4%	6.0%	1.9%	72.8%
eval	21.0%	41.9%	17.6%	24.6%
Δ	13.5%	5.1%	4.0%	59.7%

Table 6. WER of the monolingual models using the ML3-MIX LDA and decision tree as well as increase in WER

6. ADAPTING THE POLYPHONE DECISION TREE

6.1. PDTS

In order to improve the porting performance and address the issue of the poorly fit multilingual polyphone decision tree, the tree is adapted using the 2h of German data. [9] introduced polyphone decision tree specialization (PDTS) as an approach for adapting a multilingual polyphone tree to new languages. PDTS uses the fact that some of the leaves in the multilingual decision tree are not specialized enough for the new language. Thus, the tree clustering procedure is restarted on the adaptation material in the new language and, depending on the available adaptation material, new, finer grained models are being clustered that fit the target language better. [9] does not describe the way, the new found models are being trained. [15] mentions using MAP to train the models without giving sufficient details, especially on the model initialization which is necessary before applying MAP.

Therefore, for the experiments in this paper a new procedure was developed. The first step in this procedure is to train the models of the newly clustered tree on English, Russian, and Spanish. The LDA matrix was not reestimated, instead the matrix from ML3-Mix was used. For training, random samples using the existing forced alignments were extracted and the models initialized using k-means. Then eight iterations of label training were applied.

As a second step the sample extraction and k-means calculation was also performed on the German adaptation data in parallel to step 1. For the sake of speed the forced alignments used for training the monolingual German system were used, in order to reach the same goal with fewer training iterations. It can now happen that a model in the specialized decision tree, that has been trained in step 1 on the English, Russian, and Spanish data, has not seen enough training

Method	Threshold	dev	eval
Naive	—	38.6%	38.6%
PDTS	—	36.0%	35.2%
mod. PDTS	50	32.8%	32.6%

Table 7. WER of naive adaptation, PDTS, and modified PDTS on German dev and eval data

data because its context was not observed often enough. Because of this, these models can only be poorly adapted to German, no matter whether using MAP as in [15] or applying label training as in this work. To avoid this problem, all models that saw fewer training material on English, Russian, and Spanish in step 1 than on German in step 2, were substituted with the models from step 2. Now it is better possible to adapt these models, in this work by four iterations of label training on the German data.

After applying this procedure the word error rate on the German data improved to 36.0% on the dev set and 35.2% on the eval set (Table 7).

6.2. Modified PDTS

PDTS addresses the problem of model contexts that are not specialized enough for the target language of porting an ASR system, but neglects the inverse problem, that in the ported model contexts exist that are too specialized for the target language. Here, for certain contexts, more general models would be beneficial. In order to address this problem while at the same time keeping the advantages of PDTS, PDTS was combined with a pruning scheme that removes leaves in the decision tree that are underrepresented in the German adaptation data prior to applying PDTS. The scheme used here was described in [10] but has never been used on graphemes and never in combination with PDTS. In order to apply the pruning scheme, the occurrences of polygraphemes in the adaptation material were counted and the leaf in the decision tree determined to which each polygrapheme belongs. In that way one gets a rough estimate on how much training data a model in the decision tree receives. This way of estimating the counts of the models on the German data has the advantage that it does not need any acoustic data and can be applied on text only. Therefore, the polyphone tree can also be pruned if no acoustic adaptation material is available. The pruning was done in an iterative way. The leaf with the lowest count was first removed and its counts were distributed over the remaining leaves according to the pruned distribution tree. Then the next leaf with new lowest count was removed and so on, until no more leaves had counts below a given threshold. The optimal count threshold was determined empirically by trying out a series of thresholds on the development data.

After pruning the ML3-Mix tree, PDTS was applied to it, as described above, and the new tree and its models were trained as before. Table 7 shows the performance of the resulting ASR systems compared to the naive adaptation approach without PDTS, and the traditional PDTS. It can be seen that in the optimal case the WER compared to the standard PDTS models is reduced by 8.9% rel. on the development set, when using a threshold of 50 for pruning the tree, and by 7.4% on the evaluation set.

7. CONCLUSION

The experiments presented in this paper examine the use of multilingual acoustic models for porting grapheme based ASR systems to new languages and describe the aspects of multilingual ASR systems in more detail than in previous work. Different techniques for porting a multilingual, grapheme based model to German, when only 2h of adaptation material are available, were compared. The naive approach of training the multilingual model on the German adaptation data yields a WER of 38.6%, which is 23.2% absolute worse than the performance of the German monolingual recognizer. Applying the known technique of PDTS reduces the WER by 8.8% relative down to 35.2%. However, the new modified PDTS procedure introduced in this paper reduces the WER even further down to 32.6%, thus outperforming conventional PDTS by 7.4% rel.

8. REFERENCES

[1] R. G. Gordon Jr., Ed., *Ethnologue, Languages of the World*, SIL International, fifteenth edition, 2005.

[2] T. Janson, *Speak – A Short History of Languages*, Oxford University Press, 2002.

[3] C. Schillo, G. A. Fink, and F. Kummert, "Grapheme Based Speech Recognition for Large Vocabularies," in *ICSLP*, Beijing, China, 2000.

[4] S. Kanthak and H. Ney, "Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition," in *ICASSP*, Orlando, Florida, 2002.

[5] M. Killer, S. Stüker, and T. Schultz, "Grapheme Based Speech Recognition," in *EUROSPEECH*, Geneva, Switzerland, 2003.

[6] S. Stüker and T. Schultz, "A Grapheme based Speech Recognition System for Russian," in *SPECOM*, St. Petersburg, Russia, 2004.

[7] P. Charoenpornsawat, S. Hewavitharana, and T. Schultz, "Thai Grapheme-Based Speech Recognition," in *HLT-NAACL*, New York, NY, USA, 2006.

[8] S. Kanthak and H. Ney, "Multilingual Acoustic Modeling Using Graphems," in *EUROSPEECH*, Geneva, Switzerland, 2003.

[9] T. Schultz and A. Waibel, "Polyphone Decision Tree Specialization for Language Adaptation," in *ICASSP*, Istanbul, Turkey, June 2000.

[10] R. Wolff, "Adaption von Kontextenscheidungsbäumen auf neue Sprachen," Studienarbeit, Universität Karlsruhe (TH), 1999.

[11] B. Mimer, S. Stüker, and T. Schultz, "Flexible Decision Trees for Grapheme Based Speech Recognition," in *ESSV*, Cottbus, Germany, 2004.

[12] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment," in *ASRU*, Madonna di Campiglio Trento, Italy, December 2001.

[13] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition," *Speech Communication*, vol. 35, August 2001.

[14] T. Schultz, *Multilinguale Spracherkennung - Kombination akustischer Modelle zur Portierung auf neue Sprachen*, Ph.D. thesis, Universität Karlsruhe (TH), Juli 2000.

[15] Z. Wang and T. Schultz, "Non-Native Spontaneous Speech Recognition through Polyphone Decision Tree Specialization," in *EUROSPEECH*, Geneva, Switzerland, 2003.