INTEGRATING THAI GRAPHEME BASED ACOUSTIC MODELS INTO THE ML-MIX FRAMEWORK - FOR LANGUAGE INDEPENDENT AND CROSS-LANGUAGE ASR

Sebastian Stüker

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

ABSTRACT

Grapheme based speech recognition is a powerful tool for rapidly creating automatic speech recognition (ASR) systems in new languages. For purposes of language independent or cross language speech recognition it is necessary to identify similar models in the different languages involved. For phoneme based multilingual ASR systems this is usually achieved with the help of a language independent phoneme set and the corresponding phoneme identities in the different languages. For grapheme based multilingual ASR systems this is only possible when there is an overlap in graphemes of the different scripts involved. Often this is not the case, as for example for Thai which graphemes does not have any overlap with the graphemes of the languages that we used for multilingual grapheme based ASR in the past. In order to be able to apply our multilingual grapheme model to Thai, and in order to incorporate Thai into our multilingual recognizer, we examined and evaluated a number of data driven distance measures between the multilingual grapheme models. For our purposes distance measures that rely directly on the parameters of the models, such as the Kullback-Leibler and the Bhatthacharya distance yield the best performance.

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

1. INTRODUCTION

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. Here the descendants of speakers of a smaller language will start to speak a different, more common and wider spread language instead, without learning the language of their parents. He cites Gaelic as an example of a language that is currently in the process of being replaced in such a way, in this case by English. 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 pronunciation dictionary is a central component of large vocabulary continuous speech recognition systems. Its creation often involves large amounts of manual labor and requires the help of an expert in the targeted language. This makes it an expensive and difficult to create resource, especially for under-resourced languages. The use of graphemes instead of phonemes as modeling units is one way of trying to solve this problem. However, grapheme based speech recognition has the problem that it can only be applied to languages with a suitable script with a reasonably close grapheme to phoneme relation. Also, the creation of language independent acoustic models becomes more difficult, since the alphabets from many languages only have a minor or often no overlap, which hinders the sharing of parameters and training material.

The rest of the paper is structured as follows. Section 2 briefly reviews the topic of grapheme based speech recognition, while Section 3 introduces the corpus and task on which our experiments were performed. Section 4 then describes the monolingual grapheme based ASR systems which were derived from earlier work and which serve as a baseline and comparison for our experiments. Section 5 discusses the multilingual ASR systems used for our experiments and Section 6 introduces the distance measures that we used. Section 7 then describes how we used these distance measures to apply the multilingual recognizer to the Thai data, while in Section 8 we describe how we used the distance measures to integrate the Thai data into the multilingual recognizer.

2. GRAPHEME BASED ASR

In large vocabulary, HMM based ASR systems words are usually divided into sub-word units which are used as modeling units in the HMM. Very often phonemes, or sub-phonetic units, are used as such modeling units. These ASR systems therefore require as a central component a pronunciation dictionary that maps the textual representation of the words to be recognized to their phonetic manifestation. The creation of that 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 and is usually very time intensive. Therefore, the creation of a suitable pronunciation dictionary for under-resourced languages can easily become either too expensive, may require too much time, or might even be impossible due to the lack of an expert.

One possible solution to avoiding the need of a pronunciation dictionary is the use of graphemes instead of phonemes as modeling units. In that way the mapping from the orthography of a word to its sub-word units that are used as HMM states becomes trivial.

Past research has demonstrated that the use of graphemes as modeling units, instead of phonemes, can be a suitable approach to ASR in a wide range of languages. [3], [4], [5], and [6], for example, showed the feasibility of this approach for the languages English, German, Russian, Spanish, and Thai.

When writing about grapheme based speech recognition systems instead of phoneme based ones the terminology changes accordingly. Instead of triphones we now talk about trigraphemes, instead of a polyphone decision tree we use a polygrapheme decision tree, instead of sub-phonemes we talk about sub-graphemes etc. However, the general setup of the grapheme based recognition systems usually stays the same as for the phoneme based ones.

Context dependent ASR systems often utilize a decision tree for HMM state tying. These decision trees require a set of questions to ask about the context of the models. Traditionally, these questions regard the phonetic context of the models. This counteracts the goal of grapheme based ASR which tries to work without any phonetic knowledge. [3] examined different types of questions in the decision tree for HMM state-tying, and found that simply asking for the identities of the graphemes in the context of the models works better than phonetically motivated or automatically derived questions. This type of questions for the identities of neighboring graphemes are called 'singleton questions' in our work.

Sometimes, in order to improve the performance of the grapheme based models, pre-processing steps are applied to the graphemes, such as reordering of the graphemes, or grouping graphemes, to form a separate model. In our work we explicitly do not perform such preprocessing, since we assume as little knowledge as possible given for the target language.

[4] and [3] also conducted first experiments in building multilingual acoustic models based on graphemes, and [3] 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. The experiments were performed using languages with Latin based alphabets that have a larger overlap. While for phonemes the overlap between languages is generally fairly large, for graphemes it can be dramatically worse. The Thai alphabet, for example, does not show any overlap between the Latin based alphabets used in [3].

3. CORPUS AND TASK

The experiments in this paper were conducted on a selection of languages from the GlobalPhone [7] 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 is well suited for research in multilingual speech recognition and rapid deployment of speech processing systems in new languages, because data collection in all languages has been done in an uniform way.

The corpus is modeled after the Wall Street Journal 0 (WSJ0) corpus and contains newspaper articles collected with close talking microphones. The articles were read by native speakers of the respective language.

For the work presented, the four languages English (EN), Russian (RU), Spanish (SP), and Thai (TH) 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.

Further, since Thai also takes the role of the new language for which we want to create a new ASR system, we assume a very limited amount of 30 minutes of adaptation material (adapt) as given. Table 1 shows the size of the individual data sets for the four languages in terms of length in time, number of utterances, and number of speakers.

English and Spanish have a very similar set of graphemes, since their script is Latin based. Russian, however uses Cyrillic script. For Russian we take as given a romanization which in part overlaps with the graphemes of the Latin alphabet and thus with the graphemes from English and Spanish. Table 2 shows the assumed mapping of the Cyrillic graphemes to Latin ones. Note, that each romanized entry in the table is treated as one grapheme and thus modeling unit, so that romanized entries that consist of more than one Latin grapheme are treated as distinct from graphemes in the other languages. Thai on the other hand uses an alphabet that has probably been derived from the Old Khmer Script. The Thai alphabet has no overlap with the English, Spanish or romanized Russian alphabet. Also, since for our experiments we want to assume as little knowledge as possible given about the language to which to port the acoustic models to, we do not use a romanization for Thai and assume it as not given.

		EN	RU	SP	TH
train	hours	15.0	17.0	17.6	24.5
	#utt	7,137	8,170	5,426	12,260
	#spkrs	83	84	82	80
dev	hours	0.4	1.3	2.1	1.3
	#utt	144	898	680	613
	#spkrs	10	6	10	4
eval	hours	0.4	1.6	1.7	1.1
	#utt	152	1,029	564	568
	#spkrs	10	6	8	4
adapt	hours	–	–	–	0.5
	#utt	–	–	–	252
	#spkrs	–	–	–	2

Table 1. Size of the data sets in hours

4. MONOLINGUAL RECOGNIZERS

As an initial baseline for our experiments serves the performance of grapheme based recognition systems that were trained on their respective language only. These recognizers are similar to the ones described in [3], [8], and [5], but the preprocessing and training procedures were slightly modified and harmonized over all languages involved. All acoustic models are left to right Hidden Markov Models (HMM) with three substates 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 [9].

4.1. Preprocessing

The 16kHz, 16 bit audio data was preprocessed by calculating mel scaled cepstral coefficients, liftering, and concatenation of 6 neighboring feature vectors. The resulting 91 dimensional vector was reduced to 32 dimensions with the use of *linear discriminant analy-*

Graphemes	Romanized	Graphemes	Romanized
а	a	р	r
б	b	с	s
в	w	т	t
г	g	у	u
д	d	ф	f
е	ye	х	h
ё	yo	ц	tS
ж	jscH	ч	scH
з	z	ш	sch
и	i	щ	schTsch
й	j	ъ	Q
к	k	ы	i2
л	l	ь	˜
м	m	э	e
н	n	ю	yu
о	o	я	ya
п	p		

Table 2. The Cyrillic graphemes and their romanized form

sis (LDA). The mean of the cepstral coefficients was subtracted and their variance normalized on a per utterance basis. During decoding *incremental feature space constrained MLLR* (cMLLR) [10] and incremental cepstral mean subtraction and variance normalization on a per speaker basis was performed.

4.2. Training

Training was done with the help of forced alignments obtained with the systems trained in [3], [8], and [5]. Initial forced alignments for Thai were obtained by the flat start training procedure described in [3] and [5]. For training the acoustic models, first the LDA matrix was 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 six iterations of label training along the forced alignments and 4 iterations of EM training. The resulting models were used to obtain new forced alignments and the training procedure was iterated until minimum WER on the development set was reached. *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 obtained by a top-down clustering procedure that uses entropy gain as distance measure, in the same way as it was done in [3]. As explained before, the clustering procedure must be able to ask questions about the phonetic context of a polyphone. For our experiments we used the singleton questions described in Section 2.

4.3. Results

Table 3 shows the word accuracies (WA) of the context-dependent and context-independent models for every language on their respective development and evaluation sets. The trigram language models used for English, Russian, and Spanish were unchanged from the previous experiments in [3] and [5]. The trigram language model that was used for Thai was created with the help of the SRI Language Model Toolkit [11] and is an interpolation of a trigram model trained on 3.3 million words of newspaper texts and a trigram model trained on the transcriptions of the training data. The interpolation weight was chosen by minimizing the perplexity of the language model on the development set.The word accuracies are similar to the ones reported in previous work. The differences among the different languages are not only due to their suitability for the grapheme based approach, but also due to inherent differences in the respective languages and ASR system development in general, and can be observed on phoneme based recognizers as well.

		EN	RU	SP	TH
CI	dev	45.8%	48.1%	55.7%	70.8
	eval	46.5%	44.2%	68.6%	71.3
CD	dev	84.4%	64.3%	78.0%	87.3
	eval	82.7%	60.7%	85.9%	86.0

Table 3. WA of the monolingual grapheme based ASR systems on the dev and eval sets of their respective language

5. MULTILINGUAL ACOUSTIC MODEL

On the languages English, Russian, and Spanish a multilingual, grapheme based ASR system was trained using the technique ML-Mix [12]. 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 as described above.

First, a context-independent ML-Mix recognizer (ML-3Mix-CI) was trained. Then a polygrapheme decision tree with three thousand models was clustered and trained on these languages (ML-3Mix-CD). Table 4 gives the word accuracies 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.

6. DISTANCE MEASURES BETWEEN MODELS

Since Thai uses a completely disjunct alphabet from the other languages involved, it is not possible to map the grapheme models of the Thai recognizer to the models of the ML-3Mix recognizer by

		EN	RU	SP
ML-3Mix-CI	dev	27.6%	38.5%	44.5%
	eval	29.2%	33.8%	58.5%
ML-3Mix-CD	dev	78.2%	60.5%	74.7%
	eval	75.9%	44.2%	83.7%

Table 4. WA of the ML-3Mix models on the training languages

simply using the grapheme identity. Such a mapping, however is of interest, e.g. for applying the ML-3Mix model to Thai, or by extending it with the Thai data while at the same time sharing the Thai data with the other models.

Therefore, it is necessary to resort to a data driven mapping that does not rely on the grapheme identity. [13] constructed such a mapping by running a phoneme recognition pass of a multilingual acoustic model on the target language and taking the frame-wise phoneme confusion of the resulting decodings as distance measure. In this work we apply this approach to graphemes, but also examine other distance measures that do no rely on grapheme confusion, but are based on the model parameters itself. Some of these distance functions are defined for probability functions in general, some only for Gaussian distributions. All distance functions have a closed form solution for Gaussian distributions.

6.1. Framewise Grapheme Confusion

[13] used a framewise confusion to establish a mapping between the models of a multilingual recognizer and a target language. For this the multilingual, context independent recognizer was used as a phoneme recognizer to decode the adaptation material in the target language, for which a phoneme reference existed. Then the normalized, framewise phoneme confusion was calculated and used as a distance measure: the higher the confusion, the closer the phonemes. This measure can be applied to graphemes in the same way.

6.2. Grapheme Confusion

This distance measure is a modification of the framewise grapheme confusion. Instead of calculating the confusion between two graphemes on a per frame basis, a mapping between the hypothesized and the reference graphemes is established using the Levenshtein distance. With that the confusion between hypothesized and reference phonemes is calculated and used as the distance measure.

6.3. Euclidean Distance

Given two Gaussian distributions

$$\Gamma_1(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_1|}} \exp^{-\frac{1}{2}(x-\mu_1)\Sigma_1(x-\mu_1)}$$

and

$$\Gamma_2(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma_2|}} \exp^{-\frac{1}{2}(x-\mu_2)\Sigma_2(x-\mu_2)}$$

where d is the dimension of the input vector x, μ_1 and μ_2 are the means of the Gaussian distributions, and Σ_1 and Σ_2 their covariance matrices, it is possible to calculate the distance between Γ_1 and Γ_2 by simply taking the Eculidean distance between their two mean vectors μ_1 and μ_2:

$$d_{eucl}(\Gamma_1, \Gamma_2) = \sqrt{(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T} \tag{1}$$

This distance measure completely ignores the covariance matrices of the two distributions. It therefore makes sense to apply this measure in situations where no or only little information about the similarity of two distributions is expected to be contained in the covariance matrices.

6.4. Extended Mahalanobis Distance

The Mahalanobis distance can be used to measure the distance of a vector x to a set of samples that are distributed with a mean of μ and a covariance of Σ:

$$d_{Mhn}(x) = \sqrt{(x-\mu)^T \Sigma^{-1} (x-\mu)} \tag{2}$$

The Mahalanobis distance can be extended to a distance measure between two distributions by combining the covariance matrices of the distributions:

$$d_{extMhn}(\Gamma_1, \Gamma_2) = \sqrt{(\mu_1 - \mu_2)^T \left(\Sigma_1 + \Sigma_2\right)^{-1} (\mu_1 - \mu_2)} \tag{3}$$

Compared to the Euclidean distance the Extended Mahalanobis distance has the advantage that it also considers the covariance matrices of the distributions in addition to the mean vectors.

6.5. Kullback-Leibler Distance

The Kullback-Leibler divergence between two probability functions P_1 and P_2 is defined as [14]:

$$d_{kl}(P_1, P_2) = \int P_1(x) \log \frac{P_1(x)}{P_2(x)} \tag{4}$$

The Kullback-Leibler divergence can bee seen as a dissimilarity measure between two probability functions. However it is not symmetric and does not obey the triangle inequality and is thus not a true metric. In order to be able to use it as a distance function, one can make it symmetric by averaging the Kullback-Leibler divergence between P_1 and P_2 with the divergence between P_2 and P_1:

$$d_{kl-sym}(\Gamma_1, \Gamma_2) = d_{kl}(\Gamma_1, \Gamma_2) + d_{kl}(\Gamma_2, \Gamma_1) \tag{5}$$

For the case that P_1 and P_2 are Gauss distributions with diagonal covariance matrices, the symmetric Kullback-Leibler divergence takes the following form:

$$\begin{aligned} & d_{kl-sym}(\Gamma_1, \Gamma_2) \\ & = \frac{1}{2} \sum_{i=1}^{d} \frac{\sigma_{1,i}^2}{\sigma_{2,i}^2} + \frac{\sigma_{2,i}^2}{\sigma_{1,i}^2} - 2 + \left(\frac{1}{\sigma_{1,i}^2} + \frac{1}{\sigma_{2,i}^2} \right) (\mu_{1,i} - \mu_{2,i})^2 \end{aligned} \tag{6}$$

where μ_1, μ_2 are mean values of Γ_1 and Γ_2, while $\sigma_{1,i}$ and $\sigma_{2,i}$ are the ith element of the diagonal of covariance matrix Σ_1 and Σ_2, respectively.

6.6. Bhattacharya Distance

When working in a two class scenario often the Bhattacharya distance is used [15]:

$$d_{bhatt}(P_1, P_2) = -\ln \left(\int_x \sqrt{P_1(x) P_2(x)} \right) \tag{7}$$

The Bhattacharya distance is symmetrical but does not necessarily obey the triangle equation. For the case that Gaussian distributions with diagonal matrices are used as above it takes the form:

$$d_{bhatt}(\Gamma_1,\Gamma_2) = \frac{1}{2}\sum_{i=1}^{d} \ln\left(\frac{\sigma_{1,i}^2+\sigma_{2,i}^2}{2\sqrt{\sigma_{1,i}^2\sigma_{2,i}^2}}\right) + \frac{|\mu_{1,i}-\mu_{2,i}|^2}{2\left(\sigma_{1,i}^2+\sigma_{2,i}^2\right)} \quad (8)$$

7. APPLYING THE ML MIX MODEL TO THAI

For this set of experiments Thai takes the role of a language onto which we want to apply the ML-3Mix model, e.g. in order to initialize the grapheme based acoustic model of a new Thai recognizer. We assume that almost no linguistic or phonetic knowledge about Thai is available to us, especially that no romanization or other manual mapping of the Thai graphemes to the models of the ML-3Mix recognizer is available. We further assume that we only have 30 minutes of training material available that is annotated at the grapheme level. For our experiments we simulated this manual annotation by performing a forced alignment with the best Thai recognizer from Section 4.

In order to apply the ML-3Mix model we establish a mapping between the Thai models and the multilingual models using the distance measures described in Section 6 by mapping each multilingual grapheme model to the closest Thai model according to the respective distance measure. For the grapheme confusion based distance measures we worked directly with the ML-3Mix model from Section 5 on the Thai adaptation data.

In order to be able to calculate the distance measures based on the model parameters, we trained two auxiliary acoustic models, a ML-3Mix and a Thai model that only have one Gaussian per model, instead of the Gaussian mixture models used otherwise. With these helper models we then calculated the distances described in Subsections 6.3 to 6.6.

Table 5 shows the word accuracies of the resulting models on the Thai development and eval set when using the different distance measures for establishing the grapheme mapping. The distance measures based on the model parameters significantly outperform the grapheme confusion based measures. Further, the covariance matrices of the models carry significant knowledge, demonstrated by the gap in performance between the Euclidean distance measure and the other measures, that include the covariance matrices. Bhattacharya and Kullback-Leibler clearly outperform the Extended Mahalanobis distance. On the evaluation set Kullback-Leibler also outperforms the Bhattacharya distance.

	dev	eval
Euclidean	13.9%	15.9%
Ext. Mahalanobis	16.7%	16.9%
Kullback-Leibler	20.8%	19.7%
Bhattacharya	20.8%	19.0%

Table 5. WA on the Thai test data for the different distance measures estimated on the Thai adaption set

8. INCORPORATING THAI INTO THE ML MIX MODEL

Using the distance metrics in the same way as in Section 7 it is also possible to integrate the Thai grapheme based acoustic models into the ML-Mix framework. This time the mapping between the Thai and the ML-3Mix models was used to assign the Thai training data to the models in the ML-3Mix recognizer. For this, the training material that belongs to a specific Thai model is assigned to the ML-3Mix model that is closest according to the distance measure. Then the ML-3Mix models were trained anew on the combination of their original training data and the assigned Thai training data. In that way a complete sharing of the Thai training data with the models in the multilingual recognizer takes place and we obtain a ML-4Mix recognizer that was trained on all four languages.

In order to find the optimal distance measure without having to train the complete recognizer, we repeated the experiment from Section 7. This time the auxiliary Thai models were trained on the complete Thai training data, not only the adaptation data. This time the Bhattacharya distance turned out to be the best distance measure for the mapping, when testing ML-3Mix models on the Thai data according to the mapping.

Using this mapping to integrate the Thai data the ML-4Mix recognizer was trained the same ways as described in Section 5. Table 6 lists the word accuracy of the context independent (CI) and context dependent (CD) models on the training languages.

		EN	RU	SP	TH
ML-4Mix-CI	dev	22.1%	32.5%	38.3%	44.1%
	eval	22.7%	28.1%	50.7%	44.6%
ML-4Mix-CD	dev	73.5%	57.9%	71.9%	68.3%
	eval	72.2%	54.3%	81.5%	68.3%

Table 6. WA of the ML-4Mix models on the training languages

9. CONCLUSION

In this paper we have incorporated Thai data into a multilingual, grapheme based recognizer that was trained on languages with a Latin based script. Since there is no overlap of the scripts of these languages with the Thai script the mapping between the models had to be done with the help of data driven distance measures. We evaluated and compared several measures that are based either on grapheme confusion or directly on the model parameters, and found the Bhattacharya distance to perform best for this purpose. We further tested the suitability of the same distance measures for applying the multilingual model to the Thai data, e.g. for the purpose of initializing a new grapheme based model.

10. 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] M. Killer, S. Stüker, and T. Schultz, “Grapheme Based Speech Recognition,” in *EUROSPEECH*, Geneva, Switzerland, 2003.

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

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

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

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

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

[9] 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.

[10] “Maximum likelihood linear transformations for hmm-based speech recognition,” Tech. Rep., Cambridge University, Engineering Department, May 1997.

[11] A. Stolcke, “SRILM – An Extensible Language Modeling Toolkit,” Denver, Colorado, USA, 2002.

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

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

[14] Andrew R. Runnalls, “A Kullback-Leibler Approach to Gaussian Mixture Reduction,” *IEEE Transactions on Aerospace and Electronic Systems*, vol. 43, no. 3, pp. 989–99, July 2007.

[15] Spyros Liapis and Georgios Tziritas, “Image retrieval by colour and texture using chromaticity histograms and wavelet frames,” in *Advances in Visual Information Systems*, vol. 1929 of *Springer Lecture Notes in Computer Science*. Springer, Heidelberg, 2000.