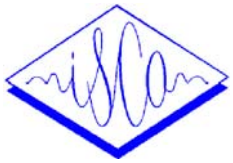

SPECOM'2004: 9th Conference
Speech and Computer
St. Petersburg, Russia
September 20-22, 2004

A Grapheme Based Speech Recognition System for Russian

Sebastian Stüker[1], *Tanja Schultz*[2]

[1]Institut für Logik, Komplexität und Deduktionssysteme
Universität Karlsruhe (TH), Karlsruhe, Germany
[2]Interactive Systems Laboratories
Carnegie Mellon University, Pittsburgh, PA, USA
stueker@ira.uka.de, tanja@cs.cmu.edu

Abstract

With the increasing availability and deployment of speech recognition technology in real world environments fast and affordable adaptation of speech recognition systems to new languages and/or domains becomes more and more important. One of the most expensive components of a recognition system is the pronunciation dictionary that maps the orthography of the words in the search vocabulary onto a sequence of sub-units. Often phonemes act as such sub-units. Human expert knowledge is usually required for crafting the pronunciation dictionary, thus making it an expensive and time consuming task. Even automatic tools for creating such dictionaries often require hand labeled amounts of training material and rely on manual revision. In order to address the problem of creating a dictionary in a time and cost efficient way we have examined recognition systems at our lab that rely soly on graphemes rather than phonemes as subunits. The mapping in the dictionary thus becomes trivial, since now every word is simply segmented into its letters. Therefore no expert knowledge is needed anymore. Our experiments on different languages have shown that the quality of the resulting recognizer significantly depends on the grapheme-to-phoneme relation of the underlying language. Since Russian is a language with an alphabetic script with a fairly close grapheme-to-phoneme relation it is very well suited to be a candidate for this approach. In this paper we present our results on creating a grapheme based Russian recognizer trained on the GlobalPhone corpus that covers fifteen different languages. We compare the performance of the resulting system to a phoneme based recognition system that was trained in the course of the GlobalPhone project, and compare the performance of two grapheme based systems whose context-dependent models were clustered with two different procedures.

1 Introduction

The pronunciation dictionary is a central component of an automatic speech recognition (ASR) system. Its purpose is to map the orthography of the words in the search vocabulary to the units that model the actual acoustic realization of the vocabulary entries. Motivated by linguistics and phonology, phonemes or sub-phonetic units are commonly used in the acoustic model of a speech recognition system. The performance of a speech recognizer is heavily influenced by the quality of the pronunciation dictionary. The dictionary can introduce two kinds of errors. First during training a false mapping between a word and the modeling units will contaminate the acoustic models. The models will not describe the actual acoustic that they represent as accurately as if they were only trained with the correct data. Second, even when the acoustic models are correctly trained, an incorrect mapping will falsify the scoring of a hypothesis by applying the wrong models to the score calculation.

Usually, manually created dictionaries yield the best results. However, their creation requires an expert in the target language and is very time consuming, thus very expensive. For some languages with a large economic impact, such as English, manual creation might be an option. But in today's world there exist an estimated 4000-6000 languages, many of which are only spoken by comparatively few people, and which are not of enough economic relevance to allow for the high costs of manually created dictionaries. Also, in cases where time is of essence, dictionary creation by human experts might not be an option, because it is simply too slow.

So the process has to be at least in part be automatized. Several different methods have been introduced in the past. Most of the time these methods are based on finding rules for the conversion of the written form of a word to a phonetic transcription, either by applying rules [1] or by statistical approaches [2]. Only some of them have been investigated in the field of speech recognition [3, 4].

Recently, the use of graphemes as modeling units, instead of phonemes, has been increasingly studied. Graphemes have the advantage over phonemes that they make the creation of the pronunciation dictionary a trivial task that does not require any linguistic knowledge. However, because of the generally looser relation of graphemes to pronunciation than that of phonemes, the use of context dependent modeling techniques and the sharing of parameters for different models are of central importance. Also prior experiments have shown that the quality of a grapheme based recognizer is highly dependent on the nature of the grapheme-to-phoneme relation of a specific language [5, 6].

Kanthak [4] was one of the first who presented results

for speech recognition systems based on the orthography of a word and the use of decision trees for context dependent modeling. Black et al. [7] sucessfully relied on graphemes for text-to-speech systems in minority languages. [5, 6] both investigated the use of graphemes for languages with phoneme-grapheme relations of differing closeness and in the context of multilingual speech recognition. All these experiments have shown that for certain languages graphemes are suitable modeling units for speech recognition. However, the use of grapheme based pronunciation dictionaries does not yield any pronunciation variants. Therefore variations in pronunciation of the same word have to be modeled implicitly in the parameters of the units used, as it is the case with the differences in pronunciation of the different graphemes depending on their orthographic context.

Lately research in the field of phoneme based speech recognition systems has also turned away from modeling pronunciation variants through explicit variations in the phoneme string but rather explores the possibilities in modeling the variations in pronunciation implicitly, e.g. by the use of single pronunciation dictionaries [8] and sharing of parameters across phonetic models [9]. In that sense, a grapheme based pronunciation dictionary is a single pronunciation dictionary in its purest form.

Traditionally, the variations in pronunciation of phonemes in different contexts are modeled by polyphones, a single phoneme in a specific context. Since the number of different polyphones even for very small context widths is already very large, in fact too large as to have enough training material to estimate the model parameters robustly, the context dependent models are usually clustered into classes. Often this clustering is done by decision tree based state tying [10]. Traditionally, due to early computational and memory constraints, one cluster tree was grown for each substate of each phoneme. In this case, parameter sharing across polyphones with different center phonemes is not possible. The enhanced tree clustering from [11] lifts this constraint.

In this paper we present our work in exploring the possibilities in building a Russian recognizer based on graphemes. We compare the performance of the grapheme based recognizer against the performance of a baseline recognizer based on phonemes. In addition we observe the effects of the enhanced tree clustering presented in [11] on the grapheme based recognition system.

2 Clustering

The representation of words in continous speech as a sequence of phonemes — sometimes called 'beads-on-a-string' [12] — is only a coarse model of reality. Due to the inertia of the human articulators when in motion and due to the sloppiness of speakers, phonemes in different contexts change their attributes. This leads to different acoustic manifestations of phonemes depending on the phonemes surrounding them.

Often these variations in pronunciation of phonemes depending on their contexts are modeled by polyphones, a single phoneme in a specific context [13, 14]. Since the number of different polyphones even for very small context widths is already very large, in fact too large as to have enough training material to estimate the model parameters robustly, the context dependent models are usually clustered into classes using a decision tree based state tying [10].

2.1 CART in Speech Recognition

When using context-dependent models the number of different models already becomes very large for relative small contexts. In general it is not possible to collect sufficient amounts of acoustic material to robustly estimate all the models' parameters. Usually many possible contexts are not even seen in the training material. One solution to deal with this problem is to cluster the models into classes, each representing one model. The clustering scheme has to fulfill the following requirements:

- the resulting number of classes is small enough to robustly estimate parameters for modeling them
- the phonetic contexts clustered into one class are suited to be modeled by a shared set of parameters (e.g. they are acoustically similar)
- phonetic contexts that have not been seen during training can be assigned to a suitable class during recognition.

As a representation of the classes and as means of assigning contexts to classes often classification and regression trees (CART) are used [15, 16]. The number of resulting classes can be controlled by different parameters, and the resulting tree allows to easily classify all possible contexts encountered during decoding. The algorithms for creating the CART in speech recognition can be generally distinguished by the following criteria:

- elements of the classes (e.g. sub-polyphones)
- questions used in the decision tree
- bottom-up or top-down clustering
- measure for determining the distance between classes (e.g. entropy or likelihood based measures).

In speech recognition often a CART for classes of sub-polyphones is trained using an entropy based distance measure. The questions in the nodes of the decision tree usually are about the membership of the phonemes in the polyphone to linguistically motivated classes, e.g. whether the phoneme left of the center phoneme is voiced. Traditionally often several decision trees are grown, e.g. one for every sub-state of every phoneme (thus collecting all polyphones with the same center phoneme in a decision tree of their own). The use of several decision trees speeds up the tree creation and is the result of memory and computational constraints of

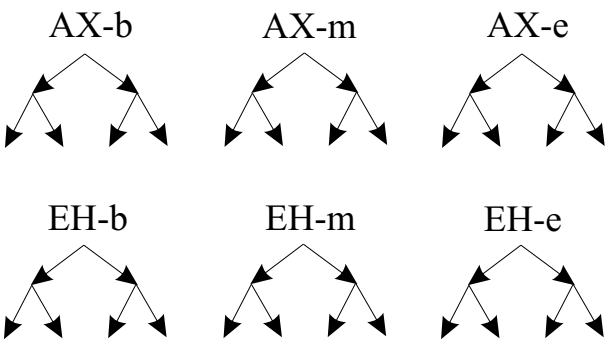

Figure 1: Separate Cluster Trees for Polyphones with different Center-Phonemes

the past. However, at the same time the manual partitioning into several trees limits the ability to model acoustic effects that are common to polyphones with different center phonemes. A possibly beneficial sharing of parameters for such polyphones is therefore not possible. Figure 1 illustrates this approach for the begin (b), middle (m), and end (e) states of two phonemes.

2.2 Enhanced Tree Clustering

[11] presented a new tree clustering approach that lifted the limitations imposed by the growing of separate decision trees for different phonemes. In contrast to the traditional decision tree based state tying, the enhanced tree clustering allows flexible parameter sharing across phonemes. With the enhanced tree clustering one single decision tree is constructed for all the sub-states of all phonemes. The clustering procedure starts with all polyphones at the root. The decision tree can ask questions regarding the identity and phonetic properties of the center phoneme and the neighbouring phonemes plus the sub-state identity. In every node the question to split the polyphones for that node is chosen that gives the highest information gain. This process is repeated until either the number of leaves of the tree reaches a certain size or the amount of training material per leave node exceeds a given threshold.

Our former experiments in [17] showed that separate trees for the begin, middle, and end states and separate trees for graphemes that can be either considered vowels or consonants, can be successfully applied to grapheme based speech recognition.

Figure 2 shows the resulting decision tree for the middle states of our sub-units just before the top-down clustering starts.

2.3 Implicit Pronunciation Modeling through Enhanced Tree Clustering

In sloppy speech people do not differentiate phonemes as much as they do in read speech. Different phonemes might be pronunced very similar. Therefore the enhanced tree clustering is well suited to implicitly capture these

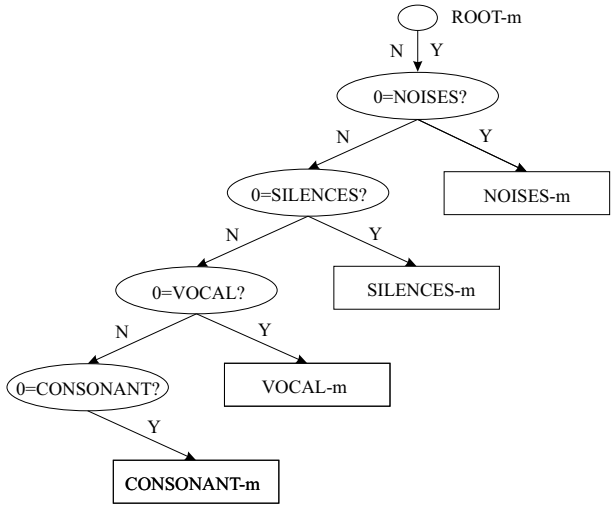

Figure 2: Enhanced Tree Clustering with Vowel and Consonant classes

phenomena by allowing certain polyphones that might be pronounced in the same or a similar way to share the same set of parameters.

Similar effects have to be dealt with in grapheme based speech recognition. Here the dictionary does not capture the fact that (a) the same grapheme might be pronounced in different ways depending on the context and (b) that different graphemes might be pronounced the same way depending on the context. The traditional clustering procedure is able to deal with the effects of (a). But in order to be also able to deal with the implications of effect (b) and at the same time to make the best use of the available training data the enhanced tree clustering is needed.

3 Grapheme Based Speech Recognition

When performing grapheme based speech recognition it often is not sufficient to simply exchange the pronunciation dictionary by one that represents words in terms of graphemes instead of phonemes. In phoneme based recognition systems phonetic knowledge about the modeling units often is also applied when growing the CART used for clustering the context dependent models. The questions being asked in the nodes of the CART often relate to the phonetic properties of the phonemes in the context of the polyphone that the CART is supposed to classify, e.g. whether a phoneme in the context is voiced or not. For graphemes this kind of questions are not given as naturally as it is the case for phonemes. One can also view these questions as a definition of classes. Then a phoneme belongs to a class corresponding to a question, when the answer to the question is positive. We call the collection of such classes a questions set. [6, 18] examined the performance of grapheme based speech recognition systems for four different kinds of questions sets:

- Phoneme-Grapheme Question Generation: Using knowledge about the grapheme-phoneme relation of

#utterance (hours)		
Training	Development	Evaluation
8,170 (17.0)	898 (1.3)	1029 (1.6)

Table 1: Overview over the Russian GlobalPhone corpus

a language, the phonemes in a class in the question set are mapped to graphemes.

- Bottom-Up Entropy: Using an entropy based distance measure the context-independent models are clustered bottom-up. The resulting classes are used as question set.

- Hybrid-Entropy: A hybrid between a bottom-up cluster procedure and an exhaustive search of all possible questions sets proposed by [19]

- Singleton: Only questions regarding the identity of the graphemes in the polygrapheme are allowed (e.g.: "Is the left neighbour an A?")

Our experiments in [6] showed that overall the Singleton question set gave the best performance. We therefore focus our experiments in this work on using the Singleton question set.

4 Experiments

For our experiments we made use of the Janus Recognition Toolkit (JRTk) v5.0 featuring the IBIS single pass decoder [20]. IBIS allows to incorporate full linguistic knowledge at an early stage reducing the amount of necessary search passes from three in our old decoder to only one.

4.1 Database

We trained and tested our recognizers on the Russian portion of the GlobalPhone (GP) corpus [21]. GlobalPhone consists of read newspaper articles with national, international political, and economic topics in fifteen languages. The data were collected under clean acoustic conditions, in 16kHz, 16 bit audio quality. Since the corpus provides uniform conditions across all languages it is possible to concentrate on studying the differences among languages from the view of speech recognition without having to deal with mismatched conditions.

The corpus was divided into three parts, one acoustic training data part (train), one for development, e.g. tuning language model parameters (dev), and one for doing the final evaluation (eval). Table 1 gives an overview of the amount data in the three portions of the corpus.

4.2 Preprocessing

The audio data was preprocessed by calculating mel scaled cepstral coefficients, liftering, and concatenation of 6 neighbouring feature vectors. The resulting 91 dimensional vector was reduced to 42 dimensions with the use of a linear discriminant analysis (LDA). The mean of the cepstral coefficients was subtracted and their variance normalized on a per speaker basis. During decoding the mean and variance of the cepstral features were incrementally estimated for each speaker. Also during decoding an incremental feature space adaptation (FSA) was performed.

4.3 Dictionary

Both the phoneme based and the grapheme based dictionary cover a vocabulary of 25,623 words. The phoneme based dictionary was constructed during the course of the GlobalPhone project in rule based manner. The used phoneme set contains 49 phonemes. After its creation the dictionary was manually postprocessed by a native-speaker, correcting errors in the automatic pronunciation generation and introducing pronunciation variants.

The grapheme based dictionary was constructed by simply segmenting all words into their letters. The Russian alphabet consists of 33 letters, two of which do not have any acoustic manifestation, but influence the pronunciation of the bordering letters. We therefore decided to keep these two letters so that their existence can be used for the context dependent models. Keeping them also coheres with our intention to apply as little phonetic knowledge about the language of our grapheme based recognizers as possible. Table 2 shows the phoneme and grapheme inventory of the two dictionaries. The order in the table gives a rough correspondence between phonemes and graphemes. Phonemes followed by a # refer to the palatalized variants of the phoneme.

4.4 Language Model

For language modeling a statistical trigram language model was trained on 19 million words of newspaper texts collected from the online editions of six newspapers. The articles are from the period of 1997 to 2004. No cut-offs for n-grams were used; the trigram perplexity of the language model on the development set is 1833. The high perplexity of the language model is due to the highly inflective morphology of the Russian language. So far, special techniques for handling this problem have been implemented in our system.

4.5 Training

For the phoneme based as well as for the grapheme based recognizers the models were divided into three sub-states, a begin, middle, and end state, with a left-to-right topology. The emission probabilities of these HMM-states were modeled by a Gaussian mixture model (GMM) with

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

Table 2: The grapheme and phoneme inventories of the recognizers

64 Gaussians, one for every state. These 64 Gaussians per HMM-state are stored in a so-called codebook. First context independent models were initialized by equally dividing the samples of every utterance in the training data over the corresponding models and then running the k-means algorithm (flat-start). After that initial labels (forced alignments) were calculated using a viterbi alignment. Then eight iterations of label training were performed followed by four iterations of viterbi training. This procedure of writing labels and training was repeated three times. With the codebooks from the resulting context independent recognizers 3000 triphone models were clustered using our conventional clustering procedure that grows a CART using entropy gain as optimization criterium. For the phoneme based recognizers we use a question set based on the articulatory properties of the phonemes. For the grapheme based recognizers we use the Singleton question set described above. Using the labels from the context-independent recognizers again eight iterations of label training followed by four iterations of viterbi training were performed. New labels with the context-dependent systems were written and the label and viterbi training repeated.

4.6 Phoneme Baseline

The first row in Table 3 shows the word error rate (WER) of the context-independent recognizers based on the phonetic dictionary on the development and evaluation set. Table 4 shows the respective numbers for the context-dependent system. Since the grapheme-based systems differ from this phoneme based system only by the dictionary, the acoustic model units, and the questions set used in clustering, the word error rates of the grapheme based systems are directly comparable to these numbers.

Since GlobalPhone is a read speech task under clean acoustic condition, we would have expected a higher performance, comparable to those we achieved for various other languages [21]. We hypothesized that the rich morphology of the Russian language is one major source of errors. We verified this by performing a number of cheating experiments in which we artificially improved the language model by adding test material to the training material. Our results indicate that the reason for the lack in performance is indeed the high language model perplexity. A manual review of the errors made by the recognizer revealed that many substitution errors result from errors made in the endings of the words. These inflections are often acoustically confusable.

4.7 Grapheme Based Recognizers

The first grapheme based speech recognizers use the same preprocessing and training procedure as the phoneme based and have the same vocabulary. They only differ in the pronunciation dictionary, the acoustic model, and the question set used for clustering. The second row in Table 3 gives the word error rates of the context-independent recognizers on the development and evaluation set. The second row in Table 4 gives the same numbers for the context-dependent system.

4.8 Enhanced Tree Clustering

In order to observe the effects of the enhanced tree clustering on our recognition system we trained a second grapheme based recognizer that is identical to the first one, only this time using the enhanced clustering procedure.

For our experiments we chose not to grow just one cluster tree, but separate ones for the begin, middle, and end states of the polygraphemes used. Also our experiments in [17] showed that it is beneficial to grow separate trees for graphemes that can be considered vowels, and such that can be considered consonants. Intuitively this make sense, because of the very different nature of consonants with respect to their articulation and therefore their acoustic manifestation. In principal the enhanced tree clustering should be able to find this separation between consonants and vowels by itself. But since the tree clustering algorithm is a greedy algorithm and since the entropy gain criterium might not be an optimal criterium this seems not to happen.

Approach	dev	eval
phoneme	54.3	48.8
grapheme	55.1	51.4

Table 3: Word error rate in % of the context-independent recognizers on the development and evaluation set

Approach	dev	eval
triphones	33.0	33.5
trigraphemes	36.4	37.3
trigraphemes with enh. tree clustering	32.8	35.7

Table 4: Word error rate in % of the context dependent recognizers on the development and evaluation set

In order to be able to grow a cluster tree we need a semi-continuous recognition system in which all polygraphemes that are supposed to belong to the same cluster tree share the same codebook. For our conventional clustering scheme it is possible to use the codebooks from the context-independent system. But for the enhanced clustering it is necessary to train such a semi-continuous system. Since now the number of models that share the same codebook drastically increases, it is necessary to increase the number of Gaussians per codebook. From our experience in [17] we chose to use 1500 Gaussians per codebook for the semi-continuous system. The semi-continuous system was trained by eight iterations of label training along labels written with the best context-independent system.

The third row in Table 4 shows the word error rate of the resulting recognizer.

5 Discussion

When using our traditional clustering procedure the grapheme based systems achieves a word error rate of 37.3% on the evaluation set which is a performance loss of 11.9% relative compared to 33.5% of the phoneme based baseline recognizer. This is a considerable loss in performance indicating that phoneme based dictionaries are superior to grapheme based ones when using the traditional clustering scheme. However, when we apply the enhanced tree clustering procedure to the grapheme based recognizers the word error rate achieves 35.7%, a relative improvement of 4.3%. Thus, the gap between the grapheme based recognizer and the phoneme baseline is closed to 6.6% relative. On the development set the grapheme based recognizer with the enhaced tree clustering even outperforms the phoneme based baseline recognizer. Therefore one can draw the conclusion that Russian is well suited for grapheme based speech recognition and also for implicit pronunciation modeling by sharing training material across polygraphemes with different center-graphemes.

The fact that in the context-independent case the grapheme based speech recognizers are considerably worse than the phoneme based ones emphasizes the importance of the context-dependent modeling when using graphemes, in order to capture the context-dependent nature of graphemes to phonemes that can be found in many languages.

Model	Set of Center Graphemes
VOWEL(|)-b (765)	а ё ю ы
VOWEL(|)-m (477)	ё ю
VOWEL(|)-e (808)	ё е ю ы
CONSONANT(|)-b (348)	j щ ъ
CONSONANT(|)-m (214)	ъ ь
CONSONANT(|)-e (444)	п ф ч щ ъ

Table 5: Examples of models for polygraphemes with different centergraphemes

The relatively low overall performance of the systems for a read speech task, can be attributed to the high language model perplexity, which is due to the highly inflective nature of the Russian language in combination with its relatively flexible word order.

An examination of the cluster tree of the grapheme based recognizer that was trained with the enhanced clustering procedure reveals that 61 models allow the sharing of parameters for polygraphemes with different center graphemes. Table 5 gives examples of center-graphemes for a model from the begin, middle, and end states of models from the VOWEL and CONSONANT classes.

The average depth of a questions node in the cluster tree is 15.1. The average depth of a question node asking for the identity of the center-grapheme is 13.5. This indicates that the identity of the center-grapheme is still an important criterium for partitioning the models. But at the same time the questions for the center-grpapheme are not the first questions asked.

6 Conclusion

In this paper we presented a grapheme based recognition system for read Russian newspaper articles. We compared the performance of a phoneme based baseline system with two different grapheme based systems. The systems only differ in the acoustic model units, the question set and the pronunciation dictionary. One of the grapheme based systems also made use of an enhanced clustering procedure for finding context dependent models. The almost equal performance of the first grapheme based system and the phoneme based baseline proofs that Russian is suited to be acoustically modeled by the use of graphemes. The increase in performance for the second grapheme based system making use of the enhanced tree clustering, shows that it is also possible to implicitly model the grapheme-to-phoneme relation of the Russian language.

The overall results are very encouraging and give

great hope for other languages with a close grapheme-to-phoneme relationship, such as Croatian, Polish, Spanish, Finnish, and Turkish, to name only a few. We also hope to successfully apply this approach to minority languages, for which the writing systems had been developed at later stages according to the pronunciation. Especially in those languages where only limited resources are available, rapid dictionary generation is a major concern and grapheme based dictionaries are a time and cost efficient alternative.

7 Acknowledgement

The authors would like to thank Borislava Mimer for her help in the preparation of the figures in this paper.

This research was partly supported by the European Commission within the project CHIL (Computers in the Human Interaction Loop: http://chil.server.de) under contract no. 506909.

References

[1] A. Black, K. Lenzo, and V. Pagel, "Issues in building general letter to sound rules", in *Proceedings of the ESCA Workshop on Speech Synthesis*, Australia, 1998, p. 7780.

[2] S. Besling, "Heuristical and statistical methods for Grapheme-to-Phoneme conversion", in *Proceedings of Konvens*, Wien, Austria, 1994, pp. 23–31.

[3] R. Singh, B. Raj, and R. M. Stern, "Automatic Generation of Subword Units for Speech Recognition Systems", *IEEE Transactions on Speech and Audio Processing*, vol. 10, pp. 98–99, 2002.

[4] S. Kanthak and H. Ney, "Context-dependent Acoustic Modeling using Graphemes for Large Vocabulary Speech Recognition", in *Proceedings the ICASSP*, Orlando, Florida, 2002, pp. 845–848.

[5] S. Kanthak and H. Ney, "Multilingual Acoustic Modeling Using Graphems", in *Proceedings of European Conference on Speech Communication and Technology*, Geneva, Switzerland, September 2003, vol. 2, pp. 1145–1148.

[6] M. Killer, S. Stüker, and T. Schultz, "Grapheme Based Speech Recognition", in *Proceedings of the EUROSPEECH*, Geneve, Switzerland, 2003, pp. 3141–3144.

[7] A. Black and A. Font Llitjos, "Unit Selection Without a Phoneme Set", in *Proceedings of the IEEE TTS Workshop*, Santa Monica, CA, 2002, p. 7780.

[8] T. Hain, "Implicit pronunciation modelling in ASR", in *ISCA Pronunciation Modeling Workshop*, 2002.

[9] M. Saraçlar, H.J. Nock, and S. Khudanpur, "Pronunciation Modeling By Sharing Gaussian Densities Across Phonetic Models", Computer Speech and Language, vol. 14, pp. 137-160, 2000.

[10] S. Young, J. Odell, and P. Woodland, "Tree-based state tying for high accuracy acoustic modelling", in *Proceedings of the ARPA HLT Workshop*, Princeton, New Jersey, March 1994.

[11] H. Yu and T. Schultz, "Enhanced Tree Clustering with Single Pronunciation Dictionary for Conversational Speech Recognition", in *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech-2003)*, Geneva, Switzerland, September 2003.

[12] M. Ostendorf, "Moving Beyond the 'Beads-On-A-String' Model of Speech", in *Proceedings of the ASRU*, Keystone, Colorado, USA, December 1999.

[13] L. R. Bahl, R. Bakis, P. S. Cohen, A. G. Cole, F. Jelinek, B. L. Lewis, and R. L. Mercer, "Further Results on the Recognition of a Continously Read Natural Corpus", in *Proceedings of the ICASSP*, Australia, 1980, vol. 5, pp. 872–875.

[14] R. M. Schwartz, Y. L. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved Hidden Markov Modeling of Phonemes for Continous Speech Reconition", in *Proceedings of the ICASSP*, 1984, vol. 9, pp. 21–24.

[15] L. Breiman et al., *Classification and Regression Trees*, Wadsworth, Pacific Grove, CA, 1984.

[16] J. R. Quinlan, *Introduction of Decision Trees*, Kluwer Academic Publishers, Boston, MA, 1986.

[17] B. Mimer, S. Stüker, and T. Schultz, "Flexible Decision Trees for Grapheme Based Speech Recognition", in *Proceedings of the 15th Conference Elektronische Sprachsignalverarbeitung (ESSV)*, Cottbus, Germany, 2004.

[18] M. Killer, "Grapheme Based Speech Recognition; Master Thesis at the Swiss Federal Institute of Technology Zurich", 2003.

[19] B. Ray, R. Singh, and R. M. Stern, "Automatic clustering and genaration of textual questions for tied states in hidden markov models", in *Proceedings of the ICASSP*, Phonexi, AZ, USA, 1999.

[20] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A One Pass-Decoder Based on Polymorphic Linguistic Context Assignment", in *Proceedings of the Automatic Speech Recognition and Understanding Workshop*, Madonna di Campiglio Trento, Italy, December 2001.

[21] T. Schultz and A. Waibel, "Language Independent and Language Adaptive Acoustic Modeling", Speech Communication, vol. 35, pp. 31-51, 2001.