FLEXIBLE DECISCION TREES FOR GRAPHEME BASED SPEECH RECOGNITION

Borislava Mimer[1], *Sebastian Stüker*[1] *and Tanja Schultz*[2]

[1]*ILKD, Universität Karlsruhe (TH), Germany*
[2]*ISL, Carnegie Mellon University, Pittsburgh, PA, USA*
mimer@ira.uka.de, stueker@ira.uka.de, tanja@cs.cmu.edu

Abstract: Over the last decades research in the field of automatic speech recognition (ASR) has seen enormous progress. Speech recognition systems are now deployed in real world applications, such as commercial software systems on PCs or Workstations, embedded in consumer devices such as cell phones or car navigation systems, or as part of specialized appliances.

With this increasing economic relevance it has become more and more important to be able to rapidly extend speech recognition systems by new words, or to port them to new, previously unseen languages or domains. Hereby the speed and cost of development are of great importance.

One of the most labor and cost intensive components of a speech recognition system is the pronunciation dictionary. Its creation often requires the application of linguistic knowledge. Even though automatic procedures for the creation of phoneme based dictionaries exist, they often require manual postprocessing by experts. This manual postprocessing is expensive and time intensiv.

Therefore, the use of grapheme based speech recognizers has seen increased research lately [1, 2, 3]. Hereby words are segmented into graphemes instead of phonemes. The use of graphemes as modeling units has the advantage over the use of phonemes that it makes the creation of the pronunciation dictionary a trivial task, saving time and money.

While a phoneme sequence is designed to describe the pronunciation of a word, the relation between the grapheme sequence of a word and its pronunciation is highly dependent on the writing system of the language in question and can be rather loosely coupled. Therefore, the context depending modeling of the units und the sharing of parameters are of central importance. Also, one does not obtain pronunciation variants when using grapheme based pronunciation dictionaries.

However, recent experiments have shown that graphemes are equally well suited as phonemes for specific languages [2, 3].

In this article we investigate the potential of a flexible decision tree clustering scheme for context dependent modeling as proposed by Hua et al. for grapheme based speech recognition. To do so we trained grapheme based speech recognizers in two languages — English and German — and compared the word error rates when using the regular clustering procedure to when using the flexible clustering.

Through the use of the enhanced clustering procedure we were able to reduce the word error rate of the grapheme beased recognizer by up to 9.3% relative, showing that for German and English graphemes are suited as units for implicit pronunciation modeling.

1 Introduction

One of the core components of a speech recognition system is the pronunciation dictionary. 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 units for the acoustic model of a speech recognition system. The performance of a speech recognizer often heavily depends on the quality of the pronunciation dictionary. The dictionary can introduce two kinds of errors. First during the 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.

A handcrafted dictionary usually yields the best results. However, manually created dictionaries require an expert in the target language and their creation are very time consuming, thus also 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 comparativly 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. In addition, as applications become more interactive and adaptive the demand for automatically expanding the dictionary on the fly increases. For example, an application, where the written form is given, and the pronunciation needs to be generated on the fly is voice driven cell phones supporting name dialing.

So the process has to be at least in part be automatized. Several different methods have been introduced in the past. Most of the times 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 [4] or by statistical approaches [5]. Only some of them have been investigated in the field of speech recognition [6, 1].

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 the pronunciation than that of phonemes, the use of context dependent modeling techniques and the sharing of parameters for different models are of central importance.

Kanthak [1] 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. In [2, 3] the use of graphems for languages with phoneme-graphem relations of differing closeness was investigated in the context of multilingual speech recognition. All these experiments have shown that for certain languages graphems are suitable modeling units for speech recognition.

Black et al. [7] showed that graphemes can be sucessfully applied for text-to-speech systems. However, TTS does not suffer from the fact that the use of grapheme based pronunciation dictionaries does not yield any pronunciation variants.

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 using a 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. Therefore, parameter sharing across polyphones with different center phonemes is not possible. The enhanced tree clustering from [11] lifts this constraint.

In this work we present the application of the enhanced tree clustering scheme from [11] to grapheme based speech recognition systems for the languages German and English. We work under the assumption that in this way it is possible to capture the fact that different graphemes may be pronounced in a smilar matter depending on their context.

2 Clustering

2.1 CART in Speech Recognition

When using context-dendent 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 now 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 [12, 13]. 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 distrance measure. The questions in the nodes of the decision tree often 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. for every sub-state for 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 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.

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 crosses a given threshold.

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 phenomena by allowing certain polyphones that might be pronunced 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 pronunced 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 Experiments

In order to examine the suitability of the enhanced tree clustering for grapheme based speech recognition we performed a couple of experiments. We trained phoneme based speech recognizers for the languages German and English, as an absolute baseline against which all the other grapheme based systems can be compared. As a grapheme based baseline act a German and an English grapheme based recognizer that were trained using our conventional clustering scheme. In order to see the effects of the application of the enhanced tree clustering we performed the same training procedure for two new grapheme recognizers, only this time using the enhanced tree clustering.

All experiments were performed with the use of the Janus Recognition Toolkit (JRTk) v5.0 featuring the Ibis single pass decoder [14].

3.1 Database

The English recognizers were trained on the Wall Street Journal 0 (WSJ0) corpus, the German recognizers on the German GlobalPhone (GP) corpus. Just like WSJ, GP consists of read news-

paper texts in fifteen languages, recorded under clean conditions with a sampling rate of 16kHz and a resolution of 16 Bit. The recordings for all languages were done under equal conditions so that the corpus is very well suited for examining differences in speech recognition between different languages. Table 1 gives an overview of the amount of acoustic data for every language, the partioning into training, development, and test data, as well as the vocabulary sizes of the two recognition systems.

	#utterance (hours)			#words
Language	Training	Development	Evaluation	Size of the Dict.
EN	7,137 (15.0)	144 (0.4)	152 (0.4)	9,461
GE	9,259 (16.9)	199 (0.4)	250 (0.4)	24,000

Table 1 - Overview over the GlobalPhone corpus

The statistical trigram language models for the English recognizer were trained on the ngram counts provided by the WSJ corpus, for German the trigram model was trained on roughly 40 million words of newspaper texts downloaded from the Interenet editions of the German newspapers “Süddeutsche Zeitung” and “Frankfurter Allgemeine Zeitung”.

3.2 Preprocessing

The 16kHz, 8 bit audio data was preprocessed by calculating mel scaled cepstral coefficients, liftering, and concatenation of 6 neigbhbouring feature vectors. The resulting 91 dimensional vector was reduced to 32 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.

3.3 Training

The recognizers were intialized with models from earlier grapheme and phoneme based models and initial labels were written. Then context-independent systems were trained, using label training. After that 3000 context dependent models were clustered with the respective clustering methods, label training was performed writing new labels with a context dependent system once.

3.4 Phonetic Baselines

In order to have a general comparison between phoneme and grapheme based recognition systems, we trained a German and English recognition system whose dictionaries were created based on rules and then later manually modified. The training was done with the procedure described above, using the common clustering procedure. The first row in Table 2 gives the word error rates of those systems on the development sets as well as the final evaluation sets.

3.5 Grapheme Baselines

In the same way we trained the phoneme-based recognizer, we developed a grapheme-based recognizer. The only difference lies in the pronunciation dictionary. The results on the evaluation and development sets are shown in the second row of Table 2. When compared to the phoneme based recognizers the word error rate increases considerably for English, and only

Approach	GE		EN	
	dev	eval	dev	eval
phoneme baseline	14.4	15.6	9.7	11.5
grapheme baseline	14.7	14.0	18.1	19.5
grapheme with enhanced tree clustering	14.2	12.7	16.8	18.6

Table 2 - Word accuracy in % of different recognizers on the development set

moderatly for German. The difference in the loss in performance is due to the fact that German has a close grapheme-to-phoneme relation than English. We observed the same effect in [3].

3.6 Enhanced Tree Clustering

For the enhanced tree clustering we performed different experiments with different parameter settings. For the enhanced tree clustering we used the German development set to optimize the parameters. Those optimal settings were then applied to evaluate on the German evaluation data, and transfered to the English evaluation data.

We decided to train separate CART for the three substates of our models. We also run experiments examining whether it makes sense to have separate trees for vowels and consonants. The results showed that manually discriminating between vowels and consonants gives a better performance, so we decided to run our experiments with this distinction.

For the question sets we also performed experiments with different sets and decided to use a singleton question set that can only ask about the identity of different graphemes in the polygrapheme.

In order to apply the entropy criterium for clustering, a semi-continuous system needs to be trained, with one codebook for every root node of the different clustering trees. The mixture weights of the context-dependent models can then be used to calculate the entropy of the cluster in a node of the tree. For the regular clustering approach it is possible to take over the codebooks from a context independent recognizer. For the enhanced tree clustering new codebooks needed to be trained. Since now only a few codebooks remain, it is necessary to increase the number of Gaussians. For our amount of training material 1500 Gaussiens turned out to be a good codebook size.

Applying the enhanced clustering leads to a reduction in error rate for both English and German on their respective development and evaluation sets. For German the word error rate decreased by 9.3% relative and for English by 4.1%.

4 Discussion

The results in Table 2 indicate that the possibility to share training data across context dependent models with different center grapheme is suited to better capture the relation between graphemes and pronunciation. An analysis of the German clustering tree reveals that 68 Models out of the 3000 used are models for polygraphemes with different center graphemes. Besides the ability to share data for different center graphemes, the enhanced tree clustering not only allows to search through a larger space, but also to get on more ways to the same results as the old clustering. But since the tree clustering is a greedy algorithm, and since the entropy gain criterium used is not necessarily an optimal criterium, the gained freedom seems to help in finding a better set of equivalence classes for the polygraphemes. The results also show that a manual partitioning of the models into vowels and consonants is still of advantage. Intuitively this makes sense because

of the quite different acoustic nature of vowels and consonants, but because of the deficiencies of the clustering algorithm mentioned above, it does not seem to be able to find this partition by itself. Table 3 gives examples of some models for the VOWEL and CONSONANT class that share data across polygraphemes with different center graphemes. The first column gives the name of the model, the second column the center graphemes of the polygraphemes that fall into the model classes. For the CONSONANT class we give examples for begin, middle, and end states. Interestingly in the VOWEL class data sharing only happens for the end states of some polygraphemes.

Model	Set of Center Graphemes
VOWEL(|)-e (465)	O ∼O U
VOWEL(|)-e (691)	∼O U
CONSONANT(|)-b (1150)	F H J Q X Y
CONSONANT(|)-b (1522)	Q V X Y
CONSONANT(|)-m (837)	J P Q V W X Y
CONSONANT(|)-m (1276)	C F Q X Y
CONSONANT(|)-e (681)	B G J M Q V W X
CONSONANT(|)-e (417)	C P Q X Y

Table 3 - Examples of models for polygraphemes with different centergraphemes

5 Conclusion

In this paper we investigated the potential of the enhanced decision tree clustering scheme for grapheme based speech recognition. The experiments were done on two languages, German and English, selected due to their different closeness of grapheme-to-phoneme relation. We compared grapheme-based with phoneme-based recognizers, which only differ in the acoustic models, the pronunciation dictionary, and the question set used to build the decision tree. All other aspects such as database, signal preprocessing, or language models were held identical. To train and evaluate our systems, we used the GlobalPhone database.

In both lanuages we see a significant improvement by applying the enhanced clustering scheme, achieving a relative improvement of 9.3% for German and 4.1% for English.

The resulting decision tree together with the relative improvements prove that flexible parameter tying is a successful scheme for grapheme-based speech recognition.

When comparing the grapheme with the phoneme based modeling, the results show that languages with a poor grapheme-to-phoneme relation, such as English, still suffer from a degradation. However, the flexible tree clustering leads to a 4.1% relative improvement compared to the usual clustering approach.

For those languages with an at least reasonable grapheme-to-phoneme relation, such as German, the grapheme-based modeling is a fast and efficient method that avoids the labor and cost intensive manual generation of pronunciation dictionaries.

The overall results are very encouraging and give great hope for those languages with even closer grapheme-to-phoneme relationship, such as Croatian, Polish, Russian, 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.

References

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

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

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

[14] Hagen Soltau, Florian Metze, Christian Fügen, and Alex 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.