Grapheme Based Speech Recognition

Mirjam Killer[1,3], Sebastian Stüker[2], Tanja Schultz[3],

[1]Computer Engineering and Networks Laboratory, ETH Swiss Federal Institute of Technology Zürich
[2]Institut für Logik, Komplexität und Deduktionssysteme, Karlsruhe Universität, Germany
[3]Interactive Systems Laboratories, Carnegie Mellon University, Pittsburgh, PA
mkiller@ee.ethz.ch, stueker@ira.uka.de, tanja@cs.cmu.edu

Abstract

Large vocabulary speech recognition systems traditionally represent words in terms of subword units, usually phonemes. This paper investigates the potential of graphemes acting as subunits. In order to develop context dependent grapheme based speech recognizers several decision tree based clustering procedures are performed and compared to each other. Grapheme based speech recognizers in three languages - English, German, and Spanish - are trained and compared to their phoneme based counterparts. The results show that for languages with a close grapheme-to-phoneme relation, grapheme based modeling is as good as the phoneme based one. Furthermore, multilingual grapheme based recognizers are designed to investigate whether grapheme based information can be successfully shared among languages. Finally, some bootstrapping experiments for Swedish were performed to test the potential for rapid language deployment.

1. Introduction

One of the core components of a speech recognition system is the pronunciation dictionary. It provides a mapping to a sequence of subword units for each entry in the vocabulary. Commonly used subword units are phonemes and polyphones. The performance of a recognition system heavily depends on the accuracy of the pronunciation dictionary. Best results are usually achieved with hand-crafted dictionaries. However, this approach is very time and cost consuming especially for large vocabulary speech recognition. If no language expert knowledge is available or affordable, methods are needed to automate the pronunciation dictionary creation process. Several different methods have been introduced over time. Most of them are based on the conversion of the orthographic transcription to a phonetic one, using either rule based [1] or statistical approaches [2]. Only some of them have been investigated in the context of speech recognition [3, 4]. Kanthak [4] was one of the first who presented results in speech recognition based on the orthographic representation of words and the use of decision trees for context dependent modeling. Black et al. [5] successfully relied on orthographic representations for text-to-speech systems in minority languages.

In this work we built grapheme based speech recognizers and compare their performance to equivalent phoneme based engines. This approach only requires a fully automatic generated question set in addition to the audio training material and transcripts. We implemented four different methods for the creation of decision tree question sets and compare their results. To demonstrate the language independent nature of the grapheme approach, all experiments were performed on English, German, and Spanish. These language were selected because of their different grapheme-to-phoneme relations. English shows the worst correspondence between graphems and phonemes, Spanish shows the best, German lies somewhere in between. Additionally we present results of multilingual grapheme based recognition. Language independent grapheme models were developed resembling work on multilingual acoustic modeling [6]. The potential for rapid adaptation to new languages is investigated by bootstrapping a Swedish recognizer from threelingual grapheme models.

2. Grapheme vs Phoneme based Recognition

All systems and experiments were performed on the GlobalPhone corpus [7], which provides clean read speech data in fifteen different languages. Based on this data, we developed monolingual phoneme based LVCSR systems in the four languages English, German, Spanish, and Swedish, using the Janus Recognition Toolkit [8] featuring the Ibis decoder [9].

2.1. Phoneme based Systems

For each language, the acoustic model of the baseline engines consists of a phonetically tied semi-continuous $3-$state HMM system with 3000 triphone models. A polyphonic decision tree was clustered using a set of linguistically motivated questions. A mixture of 32 Gaussians models each HMM-state. The preprocessing is based on 13 Mel-scale cepstral coefficients with first and second order derivatives, power, and zero crossing rates. After cepstral mean subtraction a linear discriminate analysis reduces the input vector to 32 dimensions [6].

2.2. Grapheme based Systems

The grapheme based recognizers use the same database, preprocessing, HMM-architecture, and language model as their phoneme based counterparts. The only difference lies within the subunits, the pronunciation dictionary, and the question set for creating the context dependent models.

Pronunciation dictionaries for the grapheme based recognizers are built by simply splitting a word into its graphemes. Graphemes with diacritics such as the German umlaut *ü* were treated as an independent grapheme. Digits and numbers were preprocessed by rule-based digit-to-grapheme scripts. As in the case of phonemes, a grapheme is modeled by a $3-$state HMM consisting of a begin, a middle, and an end-state.

2.3. Comparison

Table 1 compares the performance of the phoneme based to the grapheme based recognizers. The grapheme based speech recognizers use phoneme-grapheme questions (see subsection 3.1). The Spanish and German results imply that the grapheme based approach is feasible for languages with a good grapheme-to-phoneme relation. However, for English with its fairly poor grapheme-phoneme correspondence the grapheme based system is significantly outperformed by the phoneme based one.

	WER		
Language	English	German	Spanish
Phoneme	12.7%	17.7%	24.5%
Grapheme	19.1%	17.0%	26.8%

Table 1: Phoneme based vs. Grapheme based recognition using *phoneme-grapheme* question set

Table 2 indicates that the performance of the grapheme based recognition is influenced by the context width of the models. We investigated different context windows: a context width of one (C-1) leading to a trigrapheme system, a context width of two (C-2) yielding a quintgrapheme system, and a context width of three (C-3) resulting in a septgrapheme system. Additionally we developed hybrid systems, in which the question context is larger than the model context. A hybrid trigrapheme system (C-1 Q-2) is a system in which the collected contexts of a grapheme are quintgraphemes, but the final clustered polygraphemes are trigraphemes. For Spanish and German the hybrid system (C-1 Q-2) looks most promising. The authors believe that this is due to the fact that clustering can be done more precisely (distributions are based on quintgraphemes thus enabling more detailed and accurate modeling) without loosing the advantage of good generalization of the final trigrapheme models. Longer context windows do not improve the performance. There are two possible reasons, data sparseness and a suboptimal question set.

	WER				
Language	C-1	C-1 Q-2	C-2	C-2 Q-3	C-3
English	19.1%	19.8%	21.7%	22.4%	23.6%
German	18.1%	17.0%	18.4%	18.7%	18.7%
Spanish	27.0%	26.8%	28.8%	28.2%	31.4%

Table 2: Grapheme based Recognition with different context and question windows using *phoneme-grapheme* questions.

3. Question Generation

Since the set of possible polygraphemes for a standard language in a context dependent speech recognizer is very large, the parameter estimation process often runs into data-insufficiency problems. To reduce the number of free parameters, it is necessary to group the polygraphemes into a limited number of clusters. A good question set is one that prunes the search space without excluding the optimal cluster. The question generation algorithms we investigate in this work are inspired by the work of Singh et al. [3] and Beulen et al. [10].

In this work we compare four different methods for creating decision tree question sets used for the clustering of the polygrapheme models. First, phoneme based linguistic questions are transformed into grapheme based ones serving as a baseline. Second, a bottom-up clustering and third a hybrid clustering procedure based on the entropy distance generate two different question sets. Fourth, a singleton question set is created by taking each grapheme as a single question.

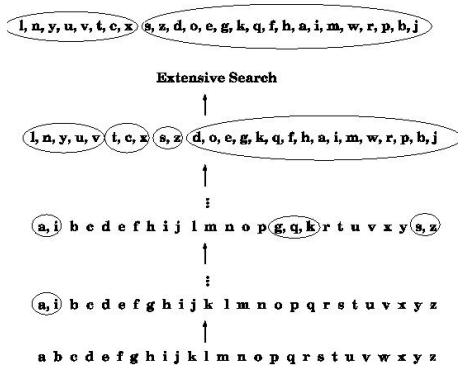

Figure 1: Hybrid Entropy Questions Generation

3.1. Phoneme-Grapheme Questions

To rule out the effect of the question set for the comparison between the phoneme and grapheme based approach, we took the question set as used for the phoneme based recognizer, converted it based on a simple language dependent phoneme-grapheme mapping, and used this for the creation of the context dependent grapheme models. The resulting set is referred to as the phoneme-grapheme question set.

3.2. Bottom-Up Entropy Questions

For the bottom-up entropy question set, we started with a set of monographemes, and clustered them bottom-up using the entropy distance measure until one cluster remains. The nodes of the resulting cluster state tree are taken as the question set.

3.3. Hybrid Entropy Questions

The hybrid entropy method is based on the idea of [3] and illustrated in Figure 1. Starting with a set of monographemes the closest graphemes are clustered together in a bottom-up procedure until the number of partitions can be exhaustively evaluated. This involves the comparison of all possible groupings of clusters resulting in two maximally separated groups. The best partition is chosen as the beginning of the subsequent recursion step. On each resulting subset the bottom-up clustering step is performed again followed by an exhaustive search. If one stores the subsets in each recursion step in a top-down matter they build a tree. The intermediate nodes serve as questions which is equivalent to taking all final partitions resulting after the exhaustive search step as questions.

3.4. Entropy Distance

To calculate the entropy distance between two sets of models for subgraphemes (grapheme-beginning, -middle and -end) a context independent system is trained where all the acoustic models share the same codebook. Let K_1 and K_2 be two sets of distribution models defined by the mixture weights of the Gaussian distributions $\gamma_{1,i}$ and $\gamma_{2,i}$ (they all share the same Gaussians, but differ in their mixture weights). There are n Gaussians to form the Gaussian mixture distribution. K_1 has N and K_2 has

M models in the set. The a priori probability for a set of models is calculated by summing the amount of how many samples of the training data were classified to a certain model. This is equivalent to the sum over all models in K_1 or K_2 of the number of samples ($p_{1,i}$ and $p_{2,i}$) assigned to a polygrapheme model. Now let K_{12} be the union of K_1 and K_2 and :

$$K_{12} = K_1 \cup K_2 \quad (1)$$

$$\gamma_1(k) = \frac{1}{p(K_1)} \cdot \sum_{i=1}^{N} p_{1,i} \cdot \gamma_{1,i}(k)$$

$$\gamma_2(k) = \frac{1}{p(K_2)} \cdot \sum_{i=1}^{M} p_{2,i} \cdot \gamma_{2,i}(k)$$

$$\gamma_{12}(k) = \frac{p(K_1) \cdot \gamma_1(k) + p(K_2) \cdot \gamma_2(k)}{p(K_1) + p(K_2)} \quad (2)$$

The entropy distance between the two sets of models is now:

$$D = (p(K_1) + p(K_1)) \cdot H_{12} - p(K_1) \cdot H_1 - p(K_2) \cdot H_2$$

Where H_i is the entropy of the distribution γ_i. The distance between graphemes is defined as:

$$D_{TOT} = D_b + D_m + D_e \quad (3)$$

For each subgrapheme class (b, m, e) the distance is calculated individually and the sum of the subgrapheme class distances forms the distance between the grapheme models.

3.5. Singletons

Another straightforward idea to generate questions is to simply ask what kind of grapheme the left or right context is. Each question consists of one single grapheme, the resulting question set is called singletons.

3.6. Question Set Evaluation

In the second set of experiments we compared the performance of the different question sets to each other. Table 3 compares the performance of the grapheme based recognizers based on the above question sets. It may seem surprising that the phoneme-grapheme question set does not perform best, but is outperformed by the singletons. Linguistic questions are derived on a phoneme basis, thus characterizing certain sounds that belong to the same sound class, e.g. are pronounced in a somewhat similar way. In the case of graphemes though, the pronunciation of a grapheme depends on its left and right context (e.g. a German "s" with a succeeding "c" and "h" is pronounced very differently than and an "s" followed by an "o"). To cluster classes together such that the acoustic material in one class is close, meaning they represent similar sounds, is in case of graphemes a question of which grapheme is to the right and left, whereas in case of sounds (phonemes) it is a question of to which sound class the phoneme belongs (is it a fricative, a plosive, etc.). This explanation is backed up by the fact that singleton questions perform less good in Spanish and best in English. Because the grapheme-phoneme mapping in Spanish is quite simple, the graphemes in Spanish can be looked at as almost phonemes. In case of English with a loose phoneme-grapheme relation the linguistically motivated questions introduce errors, whereas the singletons are better able to characterize pronunciation and therefore acoustics.

	WER		
Language	English	German	Spanish
Phoneme-Grapheme	23.9%	20.9%	**26.8%**
Bottom-Up Entropy	25.2%	20.2%	27.8%
Hybrid Entropy	22.5%	19.3%	27.5%
Singleton	**21.8%**	**18.6%**	28.7%

Table 3: Word error rate for different question sets

The Hybrid Entropy questions perform generally better than the bottom-up questions, because in case of the bottom-up clustering procedure the acoustic models that are clustered together are similar to each other in the sense of the entropy distance criteria, but the remaining models do not have to fit together at all. The hybrid clustering procedure ensures maximally separated partitions resulting in classes with similarity within a class but larger distance between classes.

4. Multilingual Grapheme based Recognition

In [6] a multilingual framework for phoneme based speech recognition is introduced which is based on the assumption that sounds are similar across languages and therefore can be shared in a language independent, multilingual acoustic model. The main motivation of this approach is to reduce the overall size of model parameters, ease the maintenance of the resulting recognizer, and to provide a good starting point for rapid adaptation to new languages. In this work we investigated if this idea can be transferred to the grapheme based situation.

4.1. Mono- vs. Multilingual Modeling

In order to compare the phoneme based with the grapheme based multilingual engines, we applied two different methods to combine the grapheme based acoustic models of English, German, and Spanish. In the first method (ML3-Mix) the acoustic model of a grapheme that occurs in more than one language is trained by sharing the data across languages. Therefore, the knowledge about the language affiliation of a grapheme is dismissed. The clustering procedure does not differentiate between languages and thus it is possible that a polygrapheme is modeled with contexts from different languages. In the second method (ML3-Tag) the language affiliation of a grapheme is preserved by assigning a language tag. The Gaussian models of a grapheme that occur in more than one language are trained by sharing the data across languages but the Gaussian mixture weights remain language dependent. The phoneme-grapheme questions for the clustering procedure are derived by combining the language specific questions into one language independent question set. Additionally questions for the language itself are added. During clustering the data decide if the language specific context is relevant or not.

As shown in Table 4 the ML3-Mix system performs significantly worse than the monolingual recognizers. In order to rule out that the differences in performance are related to the number of parameters we trained polygrapheme systems with 9000 models, corresponding to three times 3000 models per language. However, the increase in parameters does not improve the performance of the multilingual engines. From these results we conclude that the sharing of models across languages is not appropriate for grapheme based modeling. The results are

Language	WER English	German	Spanish
Monolingual (3x3000)	22.2%	21.9%	26.8%
ML3-Mix (3000)	31.0%	25.9%	34.2%
ML3-Mix (9000)	32.0%	25.6%	34.1%

Table 4: Monolingual vs Multilingual Grapheme based Systems WER [%] using *phoneme-grapheme* questions

not surprising since the graphemic representation does not hold across languages. A Spanish "V" for example is pronounced very differently from an English "V" and the acoustic data is thus not necessarily expected to be similar. Mixing those models without preserving the language information therefore does harm the performance significantly.

Language	WER English	German	Spanish
P-G	27.7%	24.4%	32.2%
Hybrid	29.4%	24.0%	34.9%
Singleton	28.0%	24.3%	34.4%

Table 5: Word error rates for different question sets on ML3-Tag using 3000 models

In order to account for language specific grapheme representations we investigated the ML3-Tag approach for model combination. Table 5 shows the results for the ML3-Tag systems using 3000 models. The results indicate that preserving the language information decreases WER by approximately 2.2% averaged over the languages. The analysis of the decision tree proved that the language questions are heavily used throughout the tree. Additionally we compared the different question sets, showing like in the monolingual case that singleton questions work best for English and the phoneme-grapheme based questions gave best results in Spanish.

4.2. Language Portability

Finally, we investigated if the multilingual grapheme based ML3-Mix recognizer can be applied to the rapid adaptation to new (during training unseen) languages. More specifically, we tested if the usage of ML3-Mix acoustic models outperform a flat start for bootstrapping a grapheme based recognizer. For this purpose the Swedish language acts as the test language. The Swedish recognizer was bootstrapped with the context dependent ML3-Mix system using the 3000 model CD-systems as shown in Table 4. After the first training cycles the Swedish context independent flat start recognizer is significantly outperformed by the multilingual, however after several training iterations the flat start does perform equally well, indicating that the multilingual engine could speeds up the procedure but giving enough training material in the target language a flat start would lead to similar results.

5. Conclusion

In this paper we presented an approach that only requires the orthographic representation of the vocabulary list rather than the pronunciation for each word. We evaluated the generation of question sets for building polyphonic decision trees and showed that this task can be solved without any linguistic knowledge. These results are very promising since they indicate that no further resources other than the audio training material and their transcripts are needed for building an LVCSR system. The resulting grapheme based recognizers perform as well as phoneme based ones tested on language for which the writing systems provide some kind of a grapheme-phoneme relation. However, since only some hundred different writing systems exist in the world and many script have been invented based on the roman alphabet [11], it is reasonable to assume that we can reach a very large number of languages with our approach. Furthermore we investigated the sharing of training data across languages to create language independent grapheme based speech recognizers. From our results we concluded that language specific graphemes significantly outperform language independent ones, which meets our expectation that alphabetic scripts are consistent within a language but not necessarily across languages. Finally, we tested the possibility of using a language independent grapheme based recognizer to bootstrap a new target language and found that it outperformes a flat start procedure only in the first training cycles.

6. References

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

[2] S. Besling, Heuristical and statistical Methods for Grapheme-to-Phoneme Conversion, Proceedings of Konvens, Wien, Austria, p.23-31, 1994.

[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, p. 98-99, 2002.

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

[5] A. Black and A. Font Llitjos, Unit Selection Without a Phoneme Set, Proceedings of the IEEE TTS Workshop, Santa Monica, CA, 2002.

[6] T. Schultz and A. Waibel, Language Independent and Language Adaptive Acoustic Modeling for Speech Recognition, Speech Communication, Vol. 35, August 2001.

[7] T. Schultz, Globalphone: A Multilingual Speech and Text Database Developed at Karlsruhe University, Proceedings of the ICSLP, Denver, CO, 2002.

[8] M. Finke, P. Geutner, H. Hild, T. Kemp, K. Ries, and M. Westphal The Karlsruhe-Verbmobil Speech Recognition Engine, Proceedings of the ICASSP, pp. 8386, Munich, Germany, 1997.

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

[10] K. Beulen and H.Ney, Automatic Question Generation for Decision Tree Based State Tying, Proceedings of the ICASSP, pp- 805-808, Seattle, WA, 1998.

[11] R. Weingarten, http://www.ruediger-weingarten.de /Texte/Latinisierung.pdf, University of Osnabrück, 2003.