CONTEXT-DEPENDENT ACOUSTIC MODELING USING GRAPHEMES FOR LARGE VOCABULARY SPEECH RECOGNITION

S. Kanthak and H. Ney

Lehrstuhl für Informatik VI, Computer Science Department
RWTH Aachen – University of Technology
52056 Aachen, Germany
{kanthak,ney}@informatik.rwth-aachen.de

ABSTRACT

In this paper we propose to use a decision tree based on graphemic acoustic sub-word units together with phonetic questions. We also show that automatic question generation can be used to completely eliminate any manual effort.

We present experimental results on four corpora with different languages, namely the Dutch ARISE corpus, the Italian EUTRANS EVAL00 evaluation corpus, the German VERBMOBIL '00 development corpus and the English North American Business '94 20k and 64k development corpora. For all experiments, the acoustic models are trained from scratch in order not to use any prior phonetic knowledge. Complete training procedures have been iterated to simulate the long optimization history used for the phonemic acoustic models.

With minimal manual effort we show that for the Dutch, German and Italian corpora, the presented approach works surprisingly well and increases the word error rate by not more than 2% relative. On the English NAB task the error rate is about 20% higher compared to experiments using a pronunciation lexicon.

1. INTRODUCTION

In large vocabulary speech recognition, to satisfy the need for scalable vocabularies and to overcome the sparse training data problem, words are most commonly built from acoustic sub-word units. Widely used sub-word units are phonemes [1], polyphones [2] and syllables [3]. All these approaches use pronunciation lexica which provide a mapping from words to sequences of sub-word units. In general, best recognition results are obtained with pronunciation lexica that are manually designed and tuned, which is a time-consuming task. Additionally, context-dependent acoustic sub-word units [4] in combination with decision tree state-tying [5] are used to detail the acoustic modeling and to improve the recognition performance.

Many different methods have been proposed for automatic construction of pronunciation lexica. Most of them convert the orthographic or graphemic transcription of a word to a phonetic transcription. These so-called grapheme-to-phoneme conversion algorithms are based either on deterministic rules [6] or statistics [7, 8, 9]. Almost all methods are based on phonetic pronunciation lexica rather than on acoustic corpora and only some of them [8, 9] have been evaluated in the context of automatic speech recognition. Most rule based methods have the additional drawback that they are not easily adaptable to other languages. In [10] and more recently in [11] the authors presented methods to learn the phoneme inventory for a speech recognizer immediately from acoustic data.

Some of the statistically based grapheme-to-phoneme methods (e.g. [9]) use decision trees in a separate step and most recognizers also use decision trees for efficient state-tying. In this paper we propose to use a single decision tree to do both tasks jointly. With this approach, a recognizer needs only an orthographic lexicon to specify the vocabulary rather than a pronunciation lexicon. To show the portability of our approach, we provide recognition experiments for four different languages.

2. METHOD

As already stated in the introduction, decision trees may be used for grapheme-to-phoneme conversion [9] as well as for context-dependent HMM state-tying [5, 12]. On the one hand, in grapheme-to-phoneme conversion it seems reasonable that the decision tree captures only contextual information. On the other hand, if a decision tree is already used for context-dependent HMM state-tying it can be used to jointly model both mappings.

2.1. Grapheme Sub-Word Units

In the approach presented here we directly apply decision tree based state-tying to the orthographic representation of words. The estimation of decision trees uses the algorithm described in [12] and takes into account the complete acoustic training data as well as a list of possible questions to control splitting of tree nodes. Similar to phonetic sub-word units we now ask questions to graphemes. Contextual information is taken into account automatically by the set of questions.

Decision tree question sets can be generated either manually or automatically [13]:

- Graphemic sets of questions can be easily manually created from phonetic ones. A grapheme is assigned to a phonetic question if the grapheme is part of the phoneme. Nevertheless, for some special cases this task still calls for expert phonetic knowledge.
- Automatic generation of questions is based on a bottom-up clustering of context-independent HMM model states and uses the log-likelihood gain and the observation count as merging criteria [13].

2.2. Training Transcripts

Our method directly uses the orthographic transcription of each word. The training transcripts may also contain symbols that are

Table 1. *Statistics of the different corpora and tasks (* silence portion measured using phonetic alignments).*

	ARISE		EUTRANS		VERBMOBIL		NAB20k		NAB64k
	training	test	training (CD1-41)	test dev00	training WSJ 0+1	test dev94 H1	training dev94 H1	test D1.3c/d	test EVAL00
acoustic data	16.5h	3.1h	7.9h	0.8h	61.5h	1.6h	81.4h	0.8h	0.8h
silence portion*	48%	49%	32%	33%	13%	11%	26%	18%	18%
# speaker	2,364	214	276	25	857	16	284	20	20
# sentences	22,786	4,265	3,187	300	36,705	1,081	37,571	310	310
# running words	74,620	13,676	52,511	5,555	721,904	14,662	649,624	7,387	7,387
# running phonemes	280,943	-	249,802	-	2,484,154	-	2,691,352	-	-
# running graphemes	332,110	-	269,321	-	2,809,822	-	3,186,108	-	-
perplexity (m-gram)	-	bi 13.7	-	tri 28.6	-	tri 57.7	-	tri 124.5	tri 145.9
# tied states	1,501		1,501		2,501		3,001		
vocabulary size	1,106	984	2,807	2,940	11,139	10,810	13,939	19,977	64,735
# lexicon entries	1,106	985	2,807	2,940	15,621	10,946	13,939	22,412	69,970

not orthographic representations of spoken words. Examples of such symbols are: alphanumeric digits, verbalized punctuation marks, and symbols representing non-speech noises. These symbols have to be suitably preprocessed.

3. EXPERIMENTAL RESULTS

For languages where the spelling is close to the orthography (e.g. German, Dutch or Italian) the application of our approach is reasonable. For others (e.g. English) the method is expected not to work very well. In this section we therefore present experimental recognition results performed on four corpora with different languages:

- the Dutch train travel information system corpus ARISE,
- the Italian spontaneous speech evaluation corpus EUTRANS EVAL00,
- the German spontaneous speech development corpus VERBMOBIL '00 and
- the English North American Business (NAB) '94 development, corpus. Recognition experiments on this corpus are carried out with vocabularies sizes of 20k and 64k words.

Table 1 gives an overview of the corpus statistics and compares the most important system parameters. For the recognition tests we used the RWTH continuous Gaussian mixture density speech recognition system which has been described in detail in [14]. The number of tied states was controlled and therefore is equal for all corpora for both the phonetic and graphemic experiments.

No across-word sub-word models were used during the tests. Baseline recognition results for context-dependent phonetic sub-word models are shown in Table 2.

For all tasks we started training of the graphemic acoustic models from scratch to keep them clean from any prior phonetic knowledge. Complete training procedures were iterated 2 to 5 times, because the baseline results using pronunciation lexica with phonetic transcriptions have also been optimized over many years.

3.1. Context-Independent Sub-Word Models

Table 3 compares recognition results on the German VERBMOBIL corpus for phonetic and graphemic sub-word units if no context-dependent state-tying is used. Here, phonetic sub-word units clearly outperform graphemic ones by more than 30% relative. There are two possible explanations:

1. the number of different sub-word units is 30% lower in the case of monographs,
2. the duration of the graphemic HMM state sequences for German sounds like `sch` or `ch` is much longer than a single phone and does not match the true acoustic lengths anymore. This observation is emphasized by the large amount of deletions in Table 3 and the larger number of running graphemes as can be seen in Table 1.

Table 2. *Baseline recognition results using manually designed phonetic pronunciation lexica with variants and state-tying based on context-dependent triphones.*

Task	Model Size [# Dens.]	Word Errors [%] DEL	INS	WER
ARISE	100,152	1.8	3.3	14.7
EUTRANS	119,698	4.3	3.1	16.2
VM	239,775	5.3	3.4	23.5
NAB20k	351,816	1.7	2.0	12.8
NAB64k	351,816	1.9	1.4	10.5

Table 3. *Comparison of recognition results for monophone and monograph acoustic sub-word models (VERBMOBIL only).*

Units	# Units	Model Size [# Dens.]	Word Errors [%] DEL	INS	WER
Monophones	44	20,180	9.5	2.6	33.6
Monographs	29	15,834	13.3	1.8	47.6

3.2. Manually Generated Questions

In the initial experiments for the tasks VERBMOBIL and NAB we used manually created question sets. The questions were based on phonetic features and were taken from [15] for the German VERBMOBIL task and from [5] for the English NAB task. Table 4 shows the recognition results for the four languages.

Table 4. *Comparison of recognition results using pronunciation lexica based on context-dependent phonemes and graphemes with pronunciation variants.*

Task	Units	Model Size	Word Errors [%]		
		[# Dens.]	DEL	INS	WER
ARISE	phon.	100,152	1.8	3.3	14.7
	graph.	135,662	2.4	3.0	15.0
EUTRANS	phon.	119,698	4.3	3.1	16.2
	graph.	97,991	3.3	3.8	16.5
VM	phon.	239,775	5.3	3.4	23.5
	graph.	262,331	6.1	3.1	23.9
NAB20k	phon.	351,816	1.7	2.0	12.8
	graph.	347,042	2.4	1.9	15.2
NAB64k	phon.	351,816	1.9	1.4	10.5
	graph.	347,042	2.7	1.2	13.2

As can be seen from Table 4 the increase in word error rate is smallest for the Dutch ARISE corpus (2.0% relative) and the German VERBMOBIL task (1.7% relative). As expected, the increase in word error rate is largest for the English NAB task (18.8% relative for the 20K vocabulary and 25.7% relative for the 64K vocabulary).

It should be noted here that using graphemic sub-word units does not provide any pronunciation variants. Therefore we also compare the above recognition results without using pronunciation variants. Table 5 shows a comparison of recognition results using pronunciation lexica based on context-dependent phonemes and graphemes without pronunciation variants. It can be seen that the recognition performance for the context-dependent graphemic sub-word units matches the performance for the phonetic sub-word units on the German VERBMOBIL corpus. For the English NAB corpus the relative increase in word error rate is 15% and 21% respectively. As can be seen in Table 1 the phonetic pronunciation lexica for the Dutch ARISE and the Italian EUTRANS corpus do not have pronunciation variants at all, so the word error rates on these two corpora are equal to those of Table 4.

Table 5. *Comparison of recognition results using pronunciation lexica based on context-dependent phonemes and graphemes without pronunciation variants.*

Task	Units	Model Size	Word Errors [%]		
		[# Dens.]	DEL	INS	WER
VM	phon.	239,775	5.0	3.8	23.9
	graph.	262,331	6.1	3.1	23.9
NAB20k	phon.	351,816	1.8	2.0	13.2
	graph.	347,042	2.4	1.9	15.2
NAB64k	phon.	351,816	2.0	1.3	10.9
	graph.	347,042	2.7	1.2	13.2

3.3. Automatically Generated Questions

The baseline approach still calls for expert phonetic knowledge when defining question sets for the estimation of the decision tree. In the following experiments we used the context-independent bottom-up cluster algorithm for automatic generation of questions as described in [13] to learn the questions from the acoustic training data. This completely eliminates the need for phonetic knowledge when building a speech recognition system.

Table 6. *Comparison of recognition results for graphemic sub-word units using automatically and manually generated decision tree questions.*

Task	Auto.	Model Size	Word Errors [%]		
	Quest.	[# Dens.]	DEL	INS	WER
ARISE	no	135,662	2.4	3.0	15.0
	yes	146,798	2.8	3.0	15.3
EUTRANS	no	97,991	3.3	3.8	16.5
	yes	98,177	3.2	3.7	16.5
NAB20k	no	347,042	2.4	1.9	15.2
	yes	346,032	2.4	1.9	15.6
NAB64k	no	347,042	2.7	1.2	13.2
	yes	346,032	2.5	1.4	13.6

Table 6 shows the recognition word error rates using automatically generated questions for the corpora ARISE, EUTRANS and NAB compared to the results from manually generated questions. As can be seen for all corpora, using automatically generated questions additionally increases the word error rate by less than 3% relative.

4. DISCUSSION

As an example, Figure 1 shows the subtree for the first HMM state of the grapheme `h`. The decision tree was estimated on the NAB corpus using the manually designed question set from the experiments in Table 4. Internal tree nodes represent questions that were asked to the left and right context (prefixes *l* and *r*) of a graphemic sub-word model. Branches to the left were followed if the question was answered with *yes*. The leaves contain the indices of the tied states. For example it can be seen from the figure that the decision tree splits the left contexts `t` and `s` and uses separate state models for them. The algorithm seems to learn the phonetic decisions from the acoustic data.

Pronunciation variants cannot yet be generated by the approach. Future investigations will include recognition tests with across-word graphemic sub-word models, variable context lengths and the generation of pronunciation variants.

5. SUMMARY

In this paper, we present a new method to eliminate the need for manual construction of pronunciation lexica for an automatic speech recognition system. We propose to directly use the orthographic transcription of words to map them onto acoustic HMM state models using phonetically motivated decision tree questions. We also show that automatic question generation can be used to further eliminate manual effort in creating a question set.

Experiments carried out on four corpora with different languages have shown the feasibility of our approach. The relative increase in word error rate was below 2% on three of the four corpora and about 20% on the English NAB task.

6. ACKNOWLEDGEMENTS

This work was partially funded by the European Commission under the Human Language Technologies project CORETEX (IST-1999-11876).

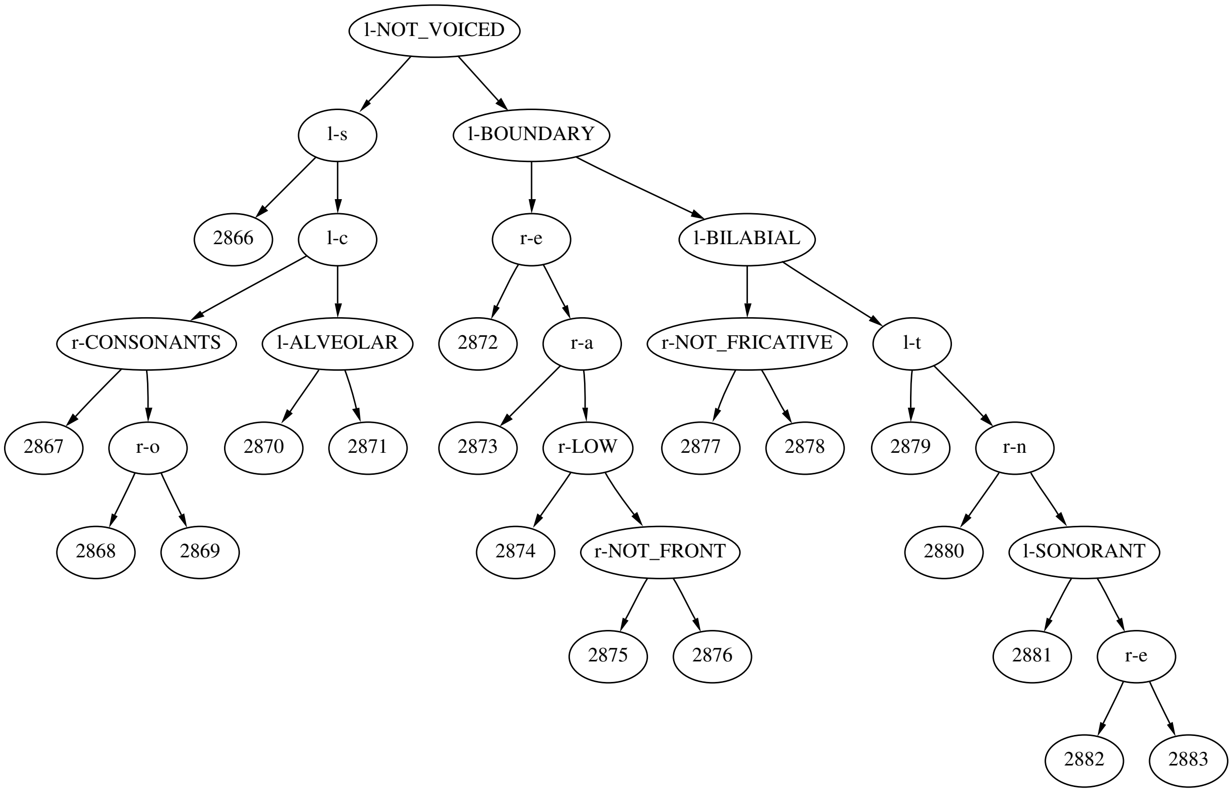

Fig. 1. *Decision subtree for the first HMM state of grapheme sub-word unit* h *(NAB20k). The tree was estimated using the manually designed question set from the experiments in Table 4. Internal tree nodes represent questions that were asked to the left and right context (prefixes l and r) of a graphemic sub-word model. Branches to the left were followed if the question was answered with yes. The leaves contain the indices of the tied states.*

7. REFERENCES

[1] L. R. Bahl, R. Bakis, P. S. Cohen, A. G. Cole, F. Jelinek, B. L. Lewis, and R. L. Mercer, "Continuous parameter acoustic processing for recognition of a natural speech corpus," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Atlanta, GA, Mar. 1981, pp. 1149 – 1152.

[2] E. G. Schukat-Talamazzini, H. Niemann, W. Eckert, T. Kuhn, and S. Rieck, "Automatic speech recognition without phonemes," in *European Conf. on Speech Communication and Technology*, Berlin, Germany, Sep. 1993, pp. 129 – 132.

[3] J. M. Hunt, M. Lenning, and P. Mermelstein, "Experiments in syllable-based recognition of continuous speech," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Denver, CO, April 1980, pp. 880 – 883.

[4] R. M. Schwartz, Y. L. Chow, S. Roucos, M. Krasner, and J. Makhoul, "Improved hidden markow modelling of phonemes for continuous speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, San Diego, CA, Mar. 1984, pp. 35.6.1 – 35.6.4.

[5] H.-W. Hon, *Vocabulary-Independent Speech Recognition: The VOCIND System*, Ph.D. thesis, Carnegie Mellon University, Pittsburgh, PA, 1992.

[6] K. Torkkola, "An efficient way to learn english grapheme-to-phoneme rules automatically," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Minneapolis, MA, April 1993, pp. 199 – 202.

[7] S. Besling, "Heuristical and statistical methods for grapheme-to-phoneme conversion," in *KONVENS*, Wien, Austria, Sep. 1994, pp. 23 – 31.

[8] C. Schillo, G. A. Fink, and F. Kummert, "Grapheme based speech recognition for large vocabularies," in *Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 129 – 132.

[9] J. Suontasuta and J. Häkkinen, "Decision tree based text-to-phoneme mapping for speech recognition," in *Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 199 – 202.

[10] J. M. Lucassen and R. L. Mercer, "Information theoretic approach to the automatic determination of phonemic baseforms," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, San Diego, CA, Mar. 1984, pp. 42.5.1 – 42.5.4.

[11] R. Singh, B. Raj, and R. M. Stern, "Automatic generation of phone sets and lexical transcriptions," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Istanbul, Turkey, June 2000, pp. 1691 – 1694.

[12] K. Beulen, E. Bransch, and H. Ney, "State tying for context dependent phoneme models," in *European Conf. on Speech Communication and Technology*, Rhodos, Greece, Sep. 1997, pp. 1179 – 1182.

[13] K. Beulen and H. Ney, "Automatic question generation for decision tree based state tying," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, pp. 805 – 808.

[14] H. Ney, L. Welling, S. Ortmanns, K. Beulen, and F. Wessel, "The rwth large vocabulary continuous speech recognition system," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, May 1998, pp. 853 – 856.

[15] K. J. Kohler, *Einführung in die Phonetik des Deutschen*, pp. 156 – 191, Erich Schmidt Verlag, Berlin, 1977.