Multilingual Acoustic Modeling Using Graphemes

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 combine grapheme-based sub-word units with multilingual acoustic modeling. We show that a global decision tree together with automatically generated grapheme questions eliminate manual effort completely. We also investigate the effects of additional language questions.

We present experimental results on four corpora with different languages, namely the Dutch and French ARISE corpus, the Italian EUTRANS corpus and the German VERBMOBIL corpus. Graphemes are shown to give good coverage on all four languages and represent a large set of shared sub-word models. For all experiments, the acoustic models are trained from scratch in order not to use any prior phonetic knowledge.

Finally, we show that for the Dutch and German tasks, the presented approach works well and may also help do decrease the word error rate below that obtained by monolingual acoustic models. For all four languages, adding questions about languages to the multilingual decision tree helps to improve the word error rate.

1. Introduction

With the exploration of speech recognition for new languages porting acoustic models becomes more important. As already shown by other groups [1, 2, 3, 4] best results for porting to a new language are obtained when starting with multilingual acoustic models. Other advantages of multilingual acoustic models are:

- in general they are smaller compared to the sum of acoustic models of monolingual systems,
- they cover a broader variety of speakers and acoustic conditions by sharing more acoustic data,
- units with few observations for a particular language may be modelled by data from other languages.

So far, global phoneme sets were most widely used in multilingual acoustic modeling. However, finding a suitable common phoneme set may be challenging and requires phonetic expert knowledge.

As shown in previous work [5] grapheme-based acoustic units in combination with decision tree state-tying may reach the performance of phonemic ones at least on a couple of European languages. The approach is completely driven by the acoustic data and does not require any linguistic or phonetic knowledge. In multilingual acoustic modeling graphemes already provide a globally consistent acoustic unit set by definition.

We evaluate our approach on four European languages namely Dutch, French, German and Italian where the acoustic databases where chosen to have similar conditions.

2. Grapheme-Based Acoustic Sub-Word Units

Context-dependent acoustic modeling using graphemes has been described in detail in [5] and we only give a brief summary here. In that approach we directly apply decision tree based state-tying to the orthographic representation of words. The estimation of decision trees uses the algorithm described in [6] 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.

In all experiments we only use the context of the immediate left and right neighbouring sub-word units. In order to avoid a loss of context we remove multiple successive occurences of consonants from the orthographic scripts of the words.

Questions for the decision tree can be generated manually or automatically. As shown in [5] for manual generation existing phonetic questions can be translated easily. Automatic generation of questions is based on bottom-up clustering of context-independent HMM model states and uses the log-likelihood gain and the observation count as merging criteria [7]. In this paper we focus on automatic generation of questions in order to eliminate any manual effort to train a multilingual acoustic model.

Table 1: *Statistics of the different corpora and tasks (*silence portion measured using grapheme-based alignments).*

	DUTCH		FRENCH		ITALIAN		GERMAN		$\sum$
	training	test	training	test	training	test	training	test	training
acoustic data	16.4h	3.1h	4.2h	0.8h	8.0h	1.6h	14.4h	0.8h	42.9h
silence portion*	41%	-	35%	-	27%	-	26%	-	34%
# speaker	2,364	453	747	102	276	25	446	14	3,833
# sentences	22,786	4,330	4,880	851	3,193	300	10,340	289	41,199
# running words	74,620	13,822	29,319	6,714	55,326	5,555	169,200	5,074	328,465
# tied states	1,501		1,001		1,501		2,501		-
vocabulary size	1,106	984	832	890	2,807	2,934	7,261	10,819	11,873
perplexity (m-gram)	-	tri 16.0	-	tri 7.0	-	tri 28.7	-	tri 35.8	-

3. Multilingual Acoustic Database

Multilingual acoustic modeling requires a consistent multilingual speech database like GlobalPhone [8]. As the GlobalPhone speech database is not publicably available yet, we compiled our own corpus. The corpus covers four European languages and the subcorpora are derived from:

- the Dutch and French train travel information system task ARISE,
- the Italian spontaneous speech task EUTRANS and
- the narrow-band portion of the German spontaneous speech task VERBMOBIL.

The acoustic data used was recorded from 8kHz telephone speech. Both the acoustic conditions and the language domain for the four tasks are very similar, but the amount of acoustic data varies. Table 1 gives a detailed overview of the corpus statistics.

4. Multilingual Acoustic Modeling

Although multilingual acoustic modeling using phonetic pronunciation lexica has been successfully investigated by others we propose to use a more data-driven method. When using phonetic acoustic sub-word units, similarities between languages may be expressed by global phoneme sets like Sampa, Worldbet or IPA [2]. With context-dependent grapheme-based sub-word units there is no need to find a common set of acoustic sub-word units. The common set of symbols shared between the words of the four languages are the characters of the words.

Table 2 summarizes the usage of graphemes across the four languages. Most of the 26 characters of the Latin alphabet are shared by all languages, only the grapheme ’q’ is missing in words of the Dutch vocabulary. French and German are the only languages in our multilingual setup that provide additional graphemes which are not used by any other language. Compared to the large amount of unique phonemes [1], graphemes seem to give a good coverage across the four languages.

Table 2: *Usage of graphemes across languages.*

Graphemes	DU	FR	IT	GE	$\sum$
a,b,c,d,e,f,g,h,i,j,k,l,m, n,o,p,r,s,t,u,v,w,x,y,z	x	x	x	x	25
q		x	x	x	1
à,â,ç,è,é,ê,ë,î,ô,û		x			10
ä,ö,ü,ß				x	4
Monolingual $\sum = 117$	25	36	26	30	
Multilingual					40

Another motivation for using context-dependent grapheme-based sub-word units for multilingual acoustic modeling is given by the distribution of the graphemes. Figure 1 compares the relative frequencies of all graphemes across the four languages. In contrast to the relative frequencies of phonemes [1], the distributions of the most frequently occuring graphemes are similar across the four languages.

5. Experimental Results

For recognition tests we use the RWTH continuous Gaussian mixture density speech recognition system which has been described in detail in [9]. The preprocessing is based on 12 Mel-frequency cepstral coefficients with first order derivatives and the second order derivative of the first coefficient. After cepstral mean substraction we apply a linear discriminant transformation with an overall window length of 3 feature vectors.

The HMM topology for graphemes and phonemes is 3 states with loop and forward transitions resulting in longer words on average when using graphemes. However, transition probabilities are estimated which should compensate for this. Besides the grapheme sub-word models task-specific noise and silence models are used.

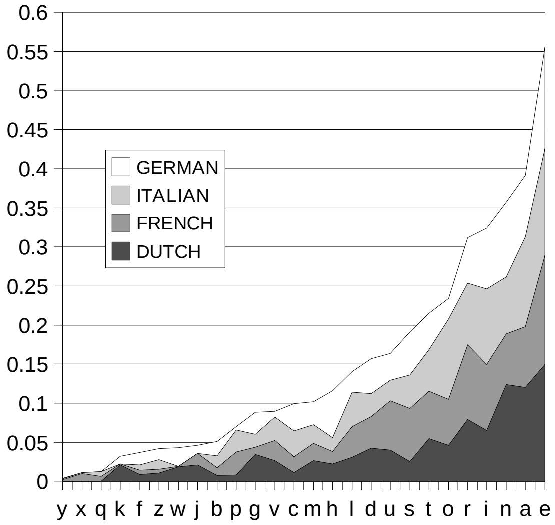

Figure 1: *Relative frequencies of the 26 shared graphemes. The distributions of the most frequently occuring graphemes are similar across the four languages.*

The bottom of Table 1 summarizes the most important remaining system parameters. The number of tied states is controlled and therefore is equal for all corpora for both the phonetic and grapheme-based experiments.

For all grapheme-based acoustic models we start training from scratch to keep the models clean from any prior phonetic knowledge. Complete training procedures were iterated 2 to 5 times as the baseline results using pronunciation lexica with phonetic transcriptions have also been optimized over many years. No across-word sub-word models are used during the tests.

5.1. Monolingual Baseline Systems

Baseline recognition results for the best monolingual systems are shown in Table 3. For comparison, Table 3 also contains results obtained with phonetic pronunciation lexica. As can be seen from the table the word error rate decreases for the Dutch and Italian tasks. For the French task the word error rate increases from 10.2% to 10.8% (6% relative) and for the German task the word error rate increases from 28.0% to 28.8% (2% relative).

5.2. Multilingual Acoustic Models

Based on the common grapheme set from Table 2 we train two different multilingual acoustic models. The acoustic model *ML-MIX* uses only questions about the grapheme, its context and questions about the state of the HMM. The acoustic model *ML-TAG* is being trained using additional questions about the language. Currently, the automatic generation of question sets only clusters questions about the grapheme and its context. The HMM state and language questions are added afterwards.

Table 3: *Baseline recognition results using monolingual acoustic models. For comparison, results using phonetic pronunciation lexica are also given.*

Lang.	Units	Model	Word Errors [%]		
		[# Dens.]	DEL	INS	WER
Dutch	phon.	121,529	1.2	2.3	8.6
	graph.	117,205	1.2	2.2	8.5
French	phon.	47,857	2.3	2.4	10.2
	graph.	55,980	3.1	1.8	10.8
Italian	phon.	96,469	3.7	3.8	16.8
	graph.	97,394	3.4	3.9	16.5
German	phon.	169,865	5.9	4.8	28.0
	graph.	161,663	6.5	5.1	28.8

Figure 2 shows the cumulative likelihood gain during the estimation of the decision tree. It can be seen that the language questions have a large contribution to the overall gain. Questions about three of the four languages are distributed almost equally over the whole splitting process. Obviously, the question for the language French has a smaller contribution to the overall gain. This may be explained by the small corpus size on the one hand and the larger amount of graphemes that are not shared with other languages on the other hand.

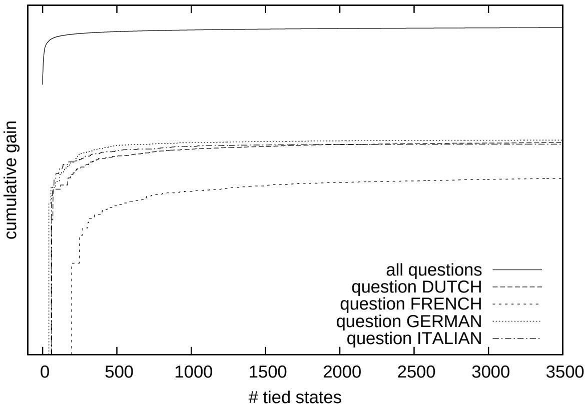

Figure 2: *Cumulative likelihood gain during estimation of the multilingual decision tree. The total cumulative gain after 3500 leafs is* $4.78191e + 07$.

Recognition results for the different acoustic models and the four languages are compared in Figure 3. The multilingual acoustic models using the *ML-MIX* setup give the worst results. The relative distance in word error rate between the monolingual and the *ML-MIX* acoustic models ranges from 11% relative for the German task to 50% for the Italian task.

The *ML-TAG* acoustic model gives much better results. For example, the word error rate for the French task increases from 10.8% when using the monolingual acoustic models to only 12.8% for the multilingual acoustic models (15% relative improvement to *ML-MIX*). For the German task the word error rate even decreases from 28.8% to 28.3%. This is possibly due to similarities between Dutch and German and therefore parts of the Dutch training corpus virtually enlarge the German training corpus. For the Italian task the difference in word error rate is still highest among the four languages (16.5% to 22.8%).

However, adding questions about languages to the decision tree improves the word error rates obtained with the multilingual acoustic models significantly.

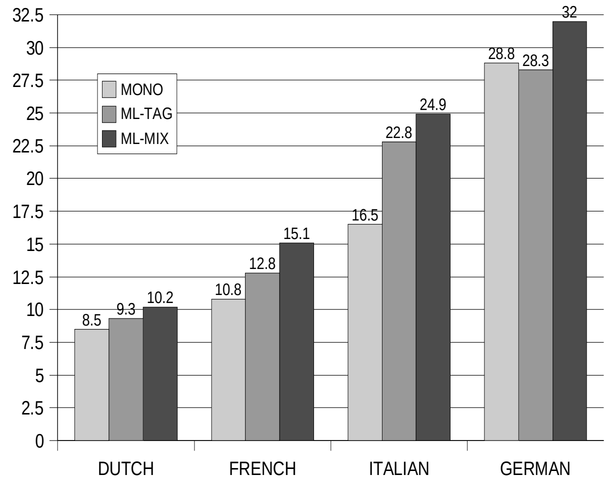

Figure 3: *Comparison of recognition results for monolingual and multilingual acoustic models. For the German task the word error rate even decreases when using multilingual acoustic models.*

6. Discussion

We found that the differences in sizes of the subcorpora can have a negative effect on the multilingual acoustic model. For example, the word error rate for the Dutch and the Italian tasks using the multilingual acoustic model would have been much better for smaller model sizes, i.e. with less mixture densities. However, the numbers in Figure 3 are chosen with respect to the lowest overall word error rate across all languages.

The French subcorpus is even smaller although the word error rates seem not to suffer that much from the size of the training corpus. Nevertheless, it does have an effect on the likelihood gain during the construction of the decision tree.

We think that a more consistent multilingual database would lead to more consistent results and plan to use the GlobalPhone database for future experiments.

7. Summary

In this paper, we apply grapheme-based acoustic sub-word units together with automatic generation of questions for decision tree state-tying to multilingual acoustic modeling. This reduces the effort to find a common set of acoustic sub-word units which in case of phonemes requires phonetic expert knowledge.

Experiments carried out on four corpora with different languages have shown the feasibility of our approach. The relative increase in word error rate between monolingual and multilingual acoustic models was below 20% for three of four languages and about 40% on the Italian task. On the German task the multilingual acoustic model decreases the word error rate by almost 2% relative.

8. Acknowledgements

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

9. References

[1] T. Schultz and A. Waibel, "Experiments towards a multi-language LVCSR interface," in *Int. Conf. on Spoken Language Processing*, Beijing, China, Oct. 2000, pp. 129 – 132.

[2] T. Schultz and A. Waibel, "Language independent and language adaptive large vocabulary speech recognition," in *Int. Conf. on Spoken Language Processing*, Sydney, Australia, Nov. 1998.

[3] T. Schultz and A. Waibel, "Fast bootstrapping of LVCSR systems with multilingual phoneme sets," in *European Conf. on Speech Communication and Technology*, Rhodes, Greece, Sep. 1997, pp. 371 – 374.

[4] J. Köhler, "Language adaptation of multilingual phone models for vocabulary independent speech recognition tasks," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Seattle, WA, 1998, pp. 417 – 420.

[5] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Orlando, FL, May 2002, pp. 845 – 848.

[6] 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.

[7] 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.

[8] T. Schultz, M. Westphal, and A. Waibel, "The GlobalPhone Project: Multilingual LVCSR with JANUS-3," in *Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop*, Plzen, Czech Republic, April 1997, pp. 20 – 27.

[9] 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.