MAXIMUM LIKELIHOOD PRONUNCIATION MODELLING OF NORWEGIAN NATURAL NUMBERS FOR AUTOMATIC SPEECH RECOGNITION

Ingunn Amdal(1), Trym Holter (2) and Torbjørn Svendsen (1)

(1) Department of Telecommunications, NTNU, N-7491 Trondheim, Norway
(2) SINTEF Telecom and Informatics, N-7465 Trondheim, Norway

E-mail: amdal@tele.ntnu.no

ABSTRACT

This paper addresses the problem of optimizing the pronunciation lexicon for use in speaker independent automatic speech recognition. The method investigated in this paper utilizes sample utterances of the vocabulary words in a data-driven manner. The optimization procedure is based on the maximum likelihood (ML) criterion, and might generate multiple baseforms for each word in order to model pronunciation variation.

The experiments show that incorporating ML pronunciation variation modelling in the natural number recognizer, improved the relative word error rate (WER) performance 6–19%, dependent on the complexity of the subword acoustic models.

1. INTRODUCTION

Modelling pronunciation variation is an important issue in today's research on automatic speech recognition. To make dialogue systems more user-friendly, spontaneous speech, noise and varying environments must be handled. Such variation in the user's input is difficult for the current state-of-the-art recognizers. Much of the effort used in modelling speaker variation has, until recently, been put into the acoustic modelling, thereby removing effects due to individual differences among speakers, e.g. pitch variation. Some of the variation in pronunciation is caused by speaking style, speaking rate and dialect, and may be better handled by careful design of the pronunciation dictionary.

The most common way of dealing with pronunciation variation is to add several pronunciation models for each word in the recognizer's lexicon. This must be done with care. By adding entries to the lexicon, the number of similar variants, and thereby the acoustic confusability, will increase. This may in turn decrease the recognition performance. This problem is partly solved in this paper by the ML based optimization procedure.

Connected natural number recognition over telephone lines is chosen as the task for the experiments for several reasons; recognition of natural numbers is important in most dialogue systems (which is the intended application), the words demand sophisticated modelling because they are short and similar and because frequently used words, e.g. numbers, are often pronounced with more variation [1]. One final reason for choosing natural numbers; the vocabulary is limited, making error analysis feasible. Earlier experiments [2, 3] have shown the need for careful pronunciation modelling for this task.

2. PRONUNCIATION VARIATION MODELLING

A subword based recognizer consists of acoustic models representing subword speech units (like phones) and a pronunciation lexicon which states the correspondence between the speech units and the words we wish to recognize. These pronunciation models are often called baseforms. One way of dealing with pronunciation variation is to introduce multiple baseforms for each word in the lexicon. There are two problems:

1. How do we chose the baseforms? The knowledge of how people really talk is not always sufficient. Besides, the acoustic models used in a recognizer does not necessarily correspond to the abstract linguistic units used to describe the pronunciation of a language.

2. Multiple baseforms for each word gives the recognizer more options to chose from and we may experience an increased error rate as an extra baseform for one word may be too similar to the baseform for a different word. Rarely used baseforms may introduce many errors compared

to the ones corrected. This can be helped by adding costs to the alternate baseforms [4].

There are two main approaches to lexicon design:

1. Ruled-based methods:
   a set of pronunciation rules are generated from phonetic and linguistic knowledge.

2. Data-driven methods:
   databases of real speech are employed in order to design the desired baseforms.

In this paper we use the data-driven approach as this employs objective criteria for lexicon design.

2.1. Baseform optimization

Usually all parts of the recognizer except the lexicon are carefully optimized with respect to objective criteria. In this paper we have utilized the ML criterion for the lexicon design as well as for subword hidden Markov model (HMM) training. Several methods for baseform optimization are described in [5]. The optimal baseform for each word is defined as the baseform B' that maximizes the likelihood of a set of sample utterances $\mathcal{T}$ of the word, given a set of valid baseforms $\mathcal{B}$ and an HMM set defined by its parameters θ:

$$B' = \underset{B \in \mathcal{B}}{\operatorname{argmax}} \{P(\mathcal{T} \mid B, \theta)\} \tag{1}$$

This criterion can be used with any grammar defining $\mathcal{B}$. However, for maximum flexibility, no constraints should be put on the search space. This is achieved with the phone-loop grammar, i.e., a grammar in which any number of phones in any order may constitute a baseform. The modified tree-trellis algorithm proposed in [6] is an effective search algorithm which offers a solution to the optimization problem. This procedure is described in larger detail in [7].

The EMCM algorithm reduces the search complexity by constraining the candidate baseforms $\mathcal{B}$ to the N most likely phone-strings obtained by a phone-loop recognition of each of the sample utterances (in this experiment 200). Results in [5] indicate that the suboptimal EMCM algorithm performs similar to the unconstrained ML solution for $N > 5$. In this paper $N = 10$ is chosen.

2.2. ML pronunciation variation modelling

In principle, the algorithms described in Section 2.1 are capable of finding the M most likely baseforms, given all training utterances. However, in order to model pronunciation variation, different baseforms should model different subsets of the training samples. A two-stage ML based algorithm for this task is described in [5]. In stage 1, 1 to M ($M = 4$ in our experiments) candidate baseforms are found independently for each word by a ML k-means clustering procedure. Each candidate baseform represents one cluster of the available sample utterances. In stage 2, the lexicon is composed on basis of the baseforms found during stage 1. The main idea is to start with a lexicon containing a single baseform for each word, and then increment the total number of baseforms by one in each step, in a manner that guarantees a maximum increase of the total likelihood. This process is repeated until the total number of baseforms in the lexicon reaches a predetermined limit.

	# speakers	*# sentences*	*# words*
Training set	382	4475	26680
Validation set	200	2338	13846
Test set	200	2330	13821

Table 1: Partitioning of TABU.0.

3. THE SPEECH DATABASES USED

3.1. TABU.0

The experiments are based on the Norwegian 1000 speakers TABU.0 telephone speech database [8]. 10 different manuscripts were used, i.e. 100 speakers would use each manuscript. Each manuscript contained, among other items, 12 different telephone numbers, where each 8-digit telephone number was listed as 4 number-pairs.

3.2. The Norwegian part of SpeechDat

The SpeechDat project [9] has recorded speech databases in 21 different European languages. The Norwegian part consists of 1000 speakers recorded via the fixed telephone network (as TABU.0). The contents of the database are digits, natural numbers (among them 1 telephone number), letters, command words, phonetically rich isolated words, names and sentences. All speakers had different manuscripts.

3.3. The use of the databases in these experiments

In this experiment the TABU.0 database was used as sample utterances for the baseform optimization (training set), validation set and test set, see Table 1. There were at least 200 utterances of each word present in the training set except "og" which occurred 123 times.

To examine dialect effects, the test set was partitioned into 5 regions; north, middle, west, southwest and southeast. The southeast region has approx-

word	*orth. trans.*	*phon. trans.*	*word*	*orth. trans.*	*phon. trans.*
0	null	n } l	17	sytten	s 2 t n
1	en	e: n	17	sytten	s y t n
1	ein	{i n	17	sytten	s 2 t e n
2	to	t u:	17	sytten	s y t e n
3	tre	t r e:	18	atten	A t n
4	fire	f i: r e	18	atten	A t e n
5	fem	f e m	19	nitten	n i t n
6	seks	s e k s	19	nitten	n i t e n
7	sju	S }:	20	tjue	C }: e
7	syv	s y: v	20	tyve	t y: v e
8	åtte	O t e	30	tretti	t r e t i
9	ni	n i:	30	tredve	t r e d v e
10	ti	t i:	40	førti	f 2 rt i
11	elleve	e l v e	40	førr	f 2 r
12	tolv	t O l	50	femti	f e m t i
13	tretten	t r e t n	60	seksti	s e k s t i
13	tretten	t r e t e n	70	sytti	s 2 t i
14	fjorten	f j u rt rn	70	sytti	s y t i
14	fjorten	f j u rt e n	80	åtti	O t i
15	femten	f e m t n	90	nitti	n i t i
15	femten	f e m t e n		og	O:
16	seksten	s {i s t n			
16	seksten	s {i s n			
16	seksten	s e k s t n			
16	seksten	s {i s t e n			

Table 2: The baseline lexicon for the natural number vocabulary.

imately twice as many speakers as each of the other regions, which are of approximately equal size. There is a tradeoff in partitioning the test set between homogeneous dialect regions and the number of speakers in each region needed for significant results. The chosen 5 regions each contain several dialects, but some major differences should still be present.

The Norwegian SpeechDat database was used for training of the HMMs used both for baseform optimization and testing. This database has approximately the same dialect distribution as TABU.0. The baseline lexicon was also extracted from the SpeechDat lexicon. For the 29 words needed to utter Norwegian natural numbers 0–99, 2 of the words had 4 pronunciations and 11 of the words had 2 pronunciations. This adds up to a total of 46 baseforms, i.e. 1.6 baseforms per word on average, see Table 2 (in SAMPA[1] transcription). With these entries the SpeechDat number lexicon is believed to cover the major dialect variations.

Alternatives for the four numbers "7", "20", "30" and "40" were analysed in [2] and the pronunciations were found to be distributed as shown in Table 3.

[1] http://www.phon.ucl.ac.uk/home/sampa/norweg.htm

word	*phon. trans.*	*distr. [%]*
7	S }:	70.4
	s y: v	29.6
20	C }: e	82.5
	t y: v e	17.5
30	t r e t i	68.8
	t r e d v e	31.2
40	f 2 rt i	97.5
	f 2 r	2.5

Table 3: Distribution in the TABU.0 database of pronunciation for four of the numbers.

4. EXPERIMENTS

The HMMs were trained on SpeechDat according to the recipe "refrec093" made by the "COST 249 SpeechDat task force"[2]. This recognizer is based on the *Hidden Markov Model Toolkit* (HTK V.2.1) [10]. Experiments made on an earlier version of this reference recognizer is reported in [11]. The main difference is that the more recent "refrec093" is suitable for real-time applications, as the 0th cepstral coefficient is used instead of normalized energy, and cepstral normalization is not applied. The HMMs were used both for finding the ML optimized baseforms as described in Section 2 and for comparing the performance of the new lexica to baseline through recognition experiments. The observation probability density functions (pdfs) for each of the 3 states in the HMMs were modelled as Gaussian mixtures. The experiments were performed for varying acoustic details in the models by using from 1 to 16 mixture components.

A word-loop grammar with uniform transition probabilities was used instead of the number grammar employed in [2] and [3], thus avoiding effects of the grammar constraints in the tests. The comparison between the lexica should still be valid although the resulting performance will be worse without the grammar constraints. An approximated two-sided 95% confidence interval for the tests is 1.7 for 50% WER and 1.5 for 30% WER.

5. RESULTS

5.1. Number of baseforms chosen

The first experiments were performed on the validation set to find the optimal number of baseforms in the new lexica generated by ML optimization. The performance of the new lexica showed a small increase in performance when adding more baseforms up to a certain point, when the performance would start to decrease somewhat. For the tests with more complex

[2] http://www.elis.rug.ac.be/ELISgroups/speech/cost249/

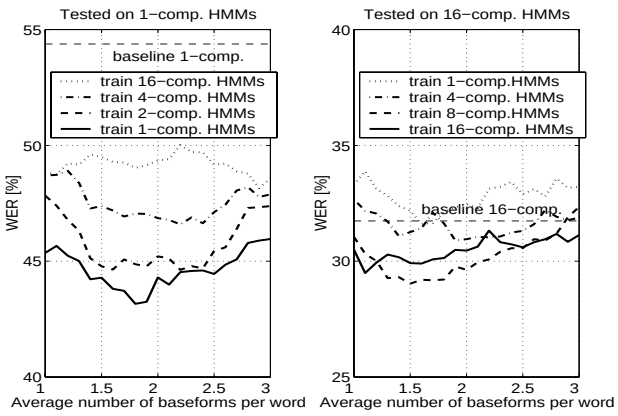

Figure 1: ML optimized baseforms with mismatch in training and test regarding HMM complexity.

HMMs (a larger number of mixture components) the maximum performance was achieved with a smaller lexicon. The validation set tests also indicated improvement over baseline, although less relative improvement for more complex HMMs.

More detailed testing was performed for 1.0, 1.6, and 1.8 baseforms per word on average. As the baseline had 1.6 baseforms per word we wanted to test the two systems with equal complexity. We also wanted to test the optimal lexica derived with the ML optimization without this constraint as the new lexica might model other effects than the baseline lexicon. To find the optimal lexicon we could have found individual number of baseforms to use for the different HMMs used to train the lexicon. As the performance was similar for a large region of lexicon sizes we chose the average of 1.8 baseforms per word. This was the mean minimum and close to minimum for all configurations in the validation set tests. In the following, the ML optimized lexica with averages of 1.0, 1.6, and 1.8 baseforms per word are denoted ML_1.0, ML_1.6, and ML_1.8, respectively.

5.2. Recognition performance

We performed some experiments on the test set with varying number of baseforms to examine mismatch between training and testing HMM complexity, i.e. a different number of components was used in the HMM pdfs during baseform optimization and testing. The ML optimized lexica outperformed the baseline in these tests (as in the validation set tests). Not surprisingly, we found that the best performance was achieved when the same number of mixture components was used in the pdf mixtures for training and testing, see Figure 1. This effect is mainly experienced for low complexity HMMs and it seems that using simpler HMMs during baseform optimization will give only minor, if any degradation in performance.

HMM comp.	*lexicon*	*WER [%]*	*rel. impr. [%]*
1	baseline 1.0	51.4	
1	ML_1.0	45.4	11.8
2	baseline 1.0	47.6	
2	ML_1.0	44.1	7.3
4	baseline 1.0	42.0	
4	ML_1.0	39.5	6.0
8	baseline 1.0	36.7	
8	ML_1.0	34.3	6.6
16	baseline 1.0	32.2	
16	ML_1.0	30.5	5.2

Table 4: ML optimized lexica compared to baseline, both with one baseform per word.

HMM comp.	*lexicon*	*WER [%]*	*rel. impr. to baseline [%]*	*rel. impr. to corr. 1.0-lex [%]*
1	baseline 1.6	54.4		-5.8
1	ML_1.6	43.8	19.4	3.4
1	ML_1.8	43.2	20.6	4.9
2	baseline 1.6	49.7		-4.3
2	ML_1.6	41.5	16.6	6.1
2	ML_1.8	41.5	16.5	6.0
4	baseline 1.6	42.8		-1.8
4	ML_1.6	38.5	9.9	2.4
4	ML_1.8	38.7	9.5	1.9
8	baseline 1.6	36.6		0.2
8	ML_1.6	32.7	10.7	4.6
8	ML_1.8	32.2	12.0	6.0
16	baseline 1.6	31.7		1.3
16	ML_1.6	29.9	5.8	2.0
16	ML_1.8	30.1	5.0	1.2

Table 5: ML optimized lexica compared to baseline, both with multiple baseforms per word.

A baseline lexicon containing a single baseform per word was made using the assumed most common pronunciation. These baseforms are given as the first baseform for each word in Table 2. Table 4 compares the results achieved with this lexicon to the ML_1.0 lexica. To our surprise this baseline performed similarly to the 1.6 baseform per word-lexicon, as shown in Table 5. The last column in Table 5 shows the relative improvement for each of the multiple baseform lexica compared to the corresponding single baseform lexica. The improvement with the ML optimized lexica over baseline is substantially higher with the multiple baseform lexica ML_1.6 and ML_1.8, except for the 16-component HMMs; this is partly due to the decrease in baseline performance when adding more baseforms. The improvement for the 1-component HMMs is as high as a 19% relative decrease in WER, whereas for

the more complex 16-component HMMs the improvement is 6% and the same as for one baseform per word. The new lexica ML_1.6 and ML_1.8 performed similarly except for the 8-component HMMs where the ML_1.8 lexicon was better.

5.3. Lexicon analysis

We examined the details of the iterative algorithm used to add multiple baseforms in the ML optimized lexica. Which words that first are assigned alternative multiple baseforms varies using different number of components in the HMM mixture pdfs. Some words are still more frequently chosen: "7", "13", "14", "16", and "30". These words are thus least sufficiently modelled by a single baseform, according to the ML criterion. All of these words also have multiple baseforms in the baseline lexicon. The last words which are assigned multiple baseforms also varies; the two least frequently chosen words are "9" and "10", and these words have only one baseform in the baseline lexicon as well. The data-driven lexicon generation algorithm agrees with linguistic knowledge on which words need more careful lexical modelling.

When the 1-component HMMs were used in training of single baseform lexica, 6 baseforms were identical to baseforms in the baseline lexicon. For the 16-component HMMs this number increased to 18. Surprisingly, this number did not increase further when ML_1.6 was compared to baseline. The equal baseforms were (with one exception) in the single baseform baseline lexicon.

We also examined the distribution of the different baseforms for each word in order to investigate if some baseforms were more frequently used in specific regions. This was done by forced alignment, inspecting the baseforms that gave 10% or more deviation in one or more regions compared to the baseform distribution for the whole test set. For the baseline the results were similar to the previous reports on TABU.0. The optimized lexica gave more baseforms with skewed distribution, e.g. for the word "5" the form /f {: m/[3] were used in 80–90% of the occurrences in the north, middle and west regions whereas only in 50% for the southeast region where the form /f e rn/ was equally frequent. It seams that some dialect variations are better modelled than in the baseline lexica.

5.4. Dialect aspects

In Norway dialects are accepted when communicating orally in any situation, both formal and informal. This means that dialogue systems that are able to accept dialect variation will be if not demanded, certainly preferred. We examined the results for each

[3] ML_1.6 trained with 16-component pdfs

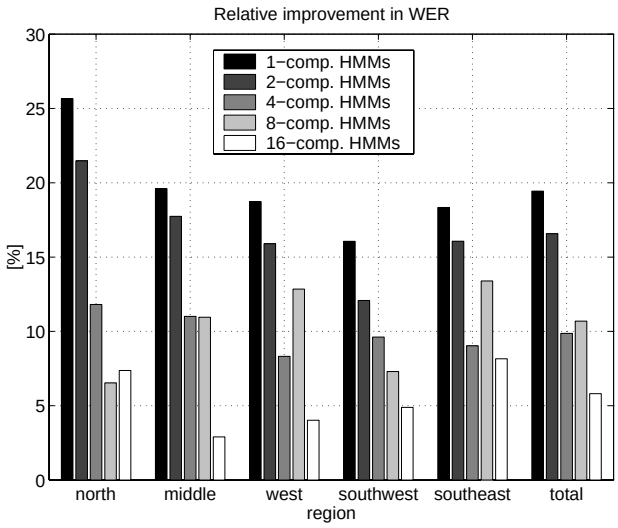

Figure 2: ML optimized lexicon with 1.8 baseforms per word compared to baseline for different regions

of the 5 regions in the test set to see if any dialect modelling effect of the ML optimized lexica was visible. The improvements for the 5 regions differ, but it seems that which region has the largest improvement is dependent on the number of components in the HMM pdfs used for baseform optimization.

For the models with low acoustic detail, a large proportion of the lexical entries may be used for modelling of allophonic differences, as this is not captured by the HMMs. We also believe that these lexica capture dialect variation used in larger regions. With the 1-, 2-, and 4-component HMMs, the largest performance gain is achieved in the north region. As more detailed acoustic models are employed, the effort in the baseform optimisation seems to be moved towards modelling of phonemic variations. This hypothesis is supported by the fact that the largest improvement with 8- and 16-component HMMs is achieved in the southeast region.

When inspecting the 1-component ML_1.6 lexicon we see that some baseforms may be interpreted as more or less allophonic; /e: rn n/ and /{: {i m/ for "1" and /s 2 r t i/ and /s 2 y t i/ for "70". The two variants for "7"; /S }: }/ and /s y:/, may on the other hand be interpreted as the frequently used two forms of "7" also present in the baseline lexicon. In the 16-component ML_1.6 lexicon we still find allophonic variants like /f {: m/ and /f e rn/ for "5", but we also have more phonemic variants like /s {i s t n/ and /s e k s t } rn/ for "16" and /C }: e/ and /C y: v 2:/ for "20" not present in the 1-component lexicon. A more detailed analysis is necessary to gain more insight about which effects the ML optimized lexica model better than the baseline lexica.

6. DISCUSSION AND CONCLUSIONS

The optimized lexicon resulted in significant WER improvement compared to the baseline lexicon. We also note that further improvement is achieved when multiple baseforms are incorporated for each word. This effect was not experienced for the manually generated lexica, thus suggesting that the ML based optimization scheme is particularly well suited to model variation in pronunciation by adding the correct extra baseforms. When adding multiple baseforms, the number of baseforms in the optimized lexica which were identical to the baseforms found in the baseline lexicon did not increase.

Some expected results were experienced when the results were analysed; there was less improvement for more complex HMMs and the results were better with matching HMMs complexity in training and test, although this was of less importance for more complex HMMs. With more complex HMMs, a larger proportion of the optimized baseforms were identical to the corresponding entries in the baseline lexicon. This is not surprising, as the more complex HMMs are able to model a larger degree of allophonic variations resulting in acoustic HMMs which correspond more closely to phonemes.

Earlier results [3] on the same task reports much better figures of WER[4], but using quite different and more sophisticated HMMs (context dependent and vocabulary tuned) and a more constrained language model.

The main result is that it is possible to find baseforms from sample utterances that outperforms tailor made baseforms. Although the improvement is less for more complex HMMs, this is still an useful algorithm for finding a lexicon covering the variations found in the vocabulary. To find out exactly what is modelled better by the ML optimized baseforms, a more thorough analysis of the corrected errors are needed.

7. REFERENCES

[1] S. Greenberg, "Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation," in *Proc. Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, (Rolduc, the Netherlands), pp. 47–56, 1998.

[2] K. Kvale, "Norwegian numerals: A challenge to automatic speech recognition," in *Proc. ICSLP-96*, (Philadelphia, USA), pp. 2028–2031, 1996.

[3] K. Kvale and I. Amdal, "Improved automatic recognition of Norwegian natural numbers by incorporating phonetic knowledge," in *Proc. ICASSP-97*, (Munich, Germany), pp. 1763–1766, 1997.

[4] W. J. Byrne, M. Finke, S. P. Khudanpur, J. McDouough, H. Nock, M. D. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos, "Pronunciation modelling using a hand-labelled corpus for conversational speech recognition," in *Proc. ICASSP-98*, (Seattle, USA), pp. 313–316, 1998.

[5] T. Holter, *Maximum likelihood modelling of pronunciation in automatic speech recognition*. PhD thesis, NTNU (Norwegian University of Science and Technology), 1997.

[6] T. Svendsen, F. K. Soong, and H. Purnhagen, "Optimizing baseforms for HMM-based speech recognition," in *Proc. EUROSPEECH-95*, (Madrid, Spain), pp. 783–786, 1995.

[7] T. Holter and T. Svendsen, "Maximum likelihood modelling of pronunciation variation," *Speech Communication*, Accepted for publication, 1999.

[8] I. Amdal and H. Ljøen, "TABU.0 – en norsk telefontaledatabase," tech. rep., Telenor R&D, Report 40/95, 1995.

[9] H. Höge, H. S. Tropf, R. Winsky, H. van den Heuvel, R. Haeb-Umbach, and K. Choukri, "European speech databases for telephone applications," in *Proc. ICASSP-97*, (Munich, Germany), pp. 1771–1774, 1997.

[10] S. Young, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book (Version 2.1)*. Entropic Cambridge Research Laboratory, 1997.

[11] F. T. Johansen, "Phoneme-based recognition for the Norwegian SpeechDat(II) database," in *Proc. ICSLP-98*, (Sydney, Australia), pp. 333–336, 1998.

[4] The baseline results was 8.8% WER. For the improved recognizer testing on only non-noise sentences and the South-East region the performance was 3.9% WER