Results vs complexity for context independent and dependent models

Ingunn Amdal

Telenor Research and Development, Instituttveien 23, N-2007 Kjeller, Norway
E-mail: Ingunn.Amdal@fou.telenor.no

Presentation for the COST 249 meeting in Zürich October 17–18 1996

ABSTRACT

This paper addresses the problem of speaker-independent connected natural number recognition over telephone lines. Increasing the vocabulary from digits (0–9) to natural numbers (0–99) introduces many new problems, some of them language-specific. Our first step was to use standard methods to achieve the best results before manual inspection of the remaining errors. We have compared context independent phoneme and word models to context dependent phoneme models with both word internal and word external context. For real-time applications complexity is important, the comparisons are therfore done regarding complexity.

1. INTRODUCTION

The number of services based on automatic speech recognition (ASR) over telephone lines has increased over the last few years, and many of these applications are based on isolated or connected digit recognition [1]. However, in many languages long numbers are normally memorised and pronounced as natural numbers. Thus, to offer more user-friendly services the recognisers have to cope with all the numerals from 0 to 99 or even 9999.

This paper describes automatic telephone number recognition. In Norway, 8-digit telephone numbers are listed as number-pairs in the telephone directory e.g. 22 34 56 78, and are therefore usually pronounced as natural numbers. We restricted the task to recognition of exactly 4 pairs of numbers, where a number-pair may be two digits or a natural number 00–99. This is a reasonable task constraint in Norwegian [3].

In this paper section 2 describes the task defining the grammar network for the Norwegian natural number vocabulary. In section 3 the recogniser is described, and section 4 summarises the results. Discussion and further work to improve the recognition performance is described in section 5.

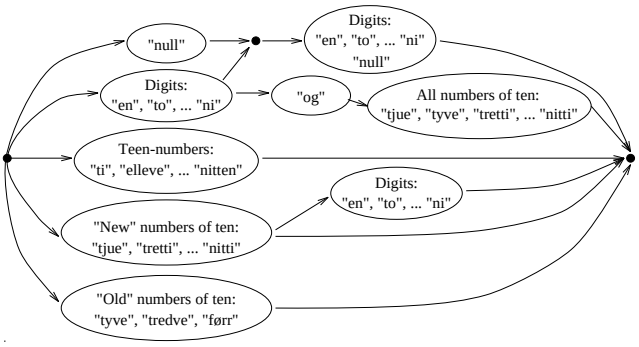

Figure 1. Grammar network for Norwegian number-pairs 00–99

2. THE TASK

This investigation is based on the Norwegian 1000 speaker TABU.0 speech database [6] and [4], which is a part of the Scandinavian 3000 speaker Rafael.0 database [5]. The speakers were called up by interviewers and asked to read telephone numbers from a manuscript in the same way as speaking to an automatic service. 10 different manuscripts were used, each containing 12 different telephone numbers. The manuscripts were designed to give enough samples for training of each word in a natural number vocabulary. Uniform distribution of natural numbers 0–99 was not possible in achieving this. The database has been manually inspected and about 40 % of the speech material was perceived with considerable channel or background noise. Mispronunciations with respect to the manuscript were not discarded as long as the resulting utterance still consisted of Norwegian natural numbers.

A particular problem in Norwegian is the two ways of pronouncing natural numbers, e.g. 52 may be pronounced both as “femti to” (fifty two), as in English and Swedish, or as “to og femti” (two and fifty), as in German and Dutch. In 1951 it was decided that the first way of pronunciation should be used, this last way of pronouncing numerals is therefore called the “old” counting style.

The vocabulary of the Norwegian natural numbers 0–99 are built from 33 different whole-word units. This includes the word "og" (and) used for the "old" counting style and two transcriptions of 7, 20, 30 and 40. Figure 1 shows the grammar network for Norwegian number-pairs. All telephone numbers in the database were automatically transcribed with the phonotypical transcription of south-eastern Norwegian. A phoneme based recogniser was used to choose the pronunciation of 7, 20, 30 and 40.

3. THE RECOGNISERS

The recogniser is based on the Hidden Markov Model Toolkit (HTK) [8]. Each 10 ms speech frame was represented by 12 mel frequency cepstral coefficients plus normalised log-energy together with their corresponding first and second order regression coefficients. Cepstral mean subtraction was applied for each sentence file. Each sentence file consisted of one 8-digit telephone number. The models were trained using k-means clustering and Baum-Welch reestimation. Different numbers of states were used, but all models were modelled as left-to-right continuous density hidden Markov models with no skip transition and with diagonal covariance matrices.

For this 34 word vocabulary (33 words and silence) the most natural choice of acoustical unit is whole-word units. Similar experiments on Danish [2] showed that context-dependent phoneme models outperformed whole-word models. Hence, we wanted to compare whole-word units with different kinds of context dependent and independent phoneme units. For the context independent whole-word models (CIWW) we used a number of states per model depending on the number of phonemes in each word. In this way the complexity of the CIWW models will be the same as for word-dependent phoneme models.

28 context independent phoneme models (CIP) (out of 47 in Norwegian) and a silence model are needed for this vocabulary. For all the phoneme based models we used three-state HMMs.

99 context dependent phoneme models with word internal context (CDWIP) were estimated by cloning the context independent phoneme models and then reestimated using triphone transcription. Clustering gives less complex systems and better trained models for rare phonemes. Training of rare CDWIP models is not a problem in this experiment, as all words and phonemes occur in sufficient numbers. Nevertheless we wanted to investigate the performance keeping the complexity constant and increasing the number of mixtures by state-clustering. The clustered word internal context models (CDWIPC) reduced the number of states from 297 needed for CDWIP models to 130.

# mix	# parameters	correct words	word accuracy	correct sentences
1	7000	80.5 %	73.0 %	27.7 %
3	21000	86.6 %	81.2 %	41.3 %
5	35000	87.9 %	82.9 %	45.0 %
7	49000	89.4 %	85.0 %	50.0 %
9	63000	90.2 %	86.3 %	52.4 %
11	77000	90.9 %	87.4 %	55.3 %
13	91000	91.4 %	88.0 %	56.9 %
15	105000	91.7 %	88.4 %	58.2 %
17	119000	92.1 %	88.9 %	60.6 %
19	133000	92.2 %	89.0 %	60.9 %
21	147000	92.6 %	89.6 %	61.6 %
23	161000	92.8 %	89.8 %	62.2 %

Table 1. Results for CIP models

Modelling context dependent phoneme models with word external context (CDWEP) was far more complex, and lack of relevant training material became a problem as not all contexts appeared in sufficient numbers for training, and some contexts were completely lacking. The 34 word vocabulary resulted in 654 word external contexts, of which 551 appeared in our database. We applied state-clustering to train rare contexts, to include the non-occurring contexts, and to keep the complexity at a reasonable level. The CDWEP models were also trained from the CIP models. To kinds of clustering were investigated: data driven clustering and knowledge based clustering. For the data driven clustering (CDWEPC2) the number of states were reduced to 183 from the 1650 needed for unclustered models. The number of states for the knowledge based clustering (CDWEPC1) was 581.

The different models were trained on 580 speakers and tested on 200, giving a total of 12520 sentences (telephone numbers) for training and 2168 for test. There are 120 different telephone numbers in the manuscripts used. Speech material based on the same manuscripts were used for both training and testing.

4. RESULTS

Since the recogniser we wanted to find by this experiment, is intended to be used in a practical application, we had to constrain the number

# mix	# parameters	correct words	word accuracy	correct sentences
1	28500	90.4 %	85.5 %	51.0 %
2	57000	92.3 %	87.6 %	57.3 %
3	57000	93.3 %	89.2 %	61.2 %
4	85500	94.0 %	90.4 %	64.9 %
5	114000	94.5 %	91.1 %	67.2 %

Table 2. Results for CIWW models

# mix	# parameters	correct words	word accuracy	correct sentences
1	23500	90.3 %	84.6 %	50.1 %
3	70500	93.6 %	89.6 %	63.9 %
5	117500	94.5 %	91.2 %	68.5 %
7	164500	94.9 %	91.7 %	69.9 %
9	211500	95.1 %	92.1 %	71.2 %
11	258500	95.3 %	92.5 %	72.0 %
13	305500	95.4 %	92.6 %	72.7 %
15	352500	95.6 %	92.7 %	72.8 %

Table 3. Results for CDWIP models

of parameters in our design. All the comparisons are therefore done with respect to the number of parameters. Tables 1 to 3 show the results as percentage correct words and accuracy. The number of mixtures are listed in the tables as well as percentage correct sentences.

There is a scoring problem for natural numbers, "words" are not the same unit as "digits". Digits may be represented by 1–3 words when represented by natural numbers. Some word substitutions are irrelevant. I.e. a telephone number will still be correct if "femti to" (fifty two) is subsituted by "fem to" (five two). Different words representing the same digit may be interesting to investigate in an error analysis, but will not give errors for pratical purposes. We have included these kind of errors. Figures 2 to 4 show recognition results as percentage correct sentences, as we found these most comparable across results for the different recognisers.

As expected the CIP models were not able to compete with the context dependent CDWIP models, see figure 2. Clustering did not give better performance for CDWIP models. For CDWEP models, clustering is neccessary and the data driven method performed best, see figure 2. Figure 4 shows that the context-dependent CDWIP and CDWEPC2 achieved approximately the same scores. The CDWIP models performed slightly better than CDWEPC2. However, a network

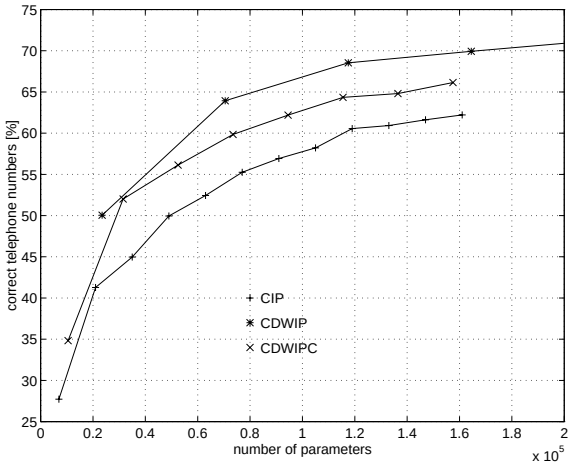

Figure 2. Context independent vs word-internal context dependent phoneme models

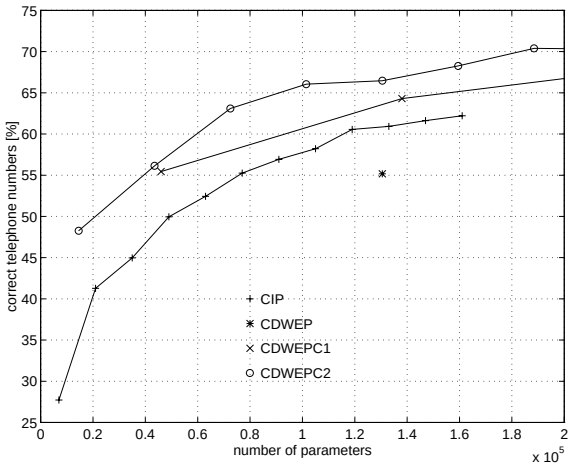

Figure 3. Context independent vs word-external context dependent phoneme models

based on CDWEP models is more complex, making such models less suitable for practical applications. The CDWIP models performed slightly better then CIWW models for low numbers of parameters, see figure 4.

5. DISCUSSION AND FURTHER WORK

Context-dependent phoneme models performed better than whole word models. They give better duration modelling, and this may be the reason for better performance with the same complexity. Using word external context did not give the expected increase in performance. Although our speech material consits of only 120 different telephone numbers, most contexts are present.

The CDWIP models seem to be most promising for further refinements. Since the performance with these models saturated at about 5 mixtures

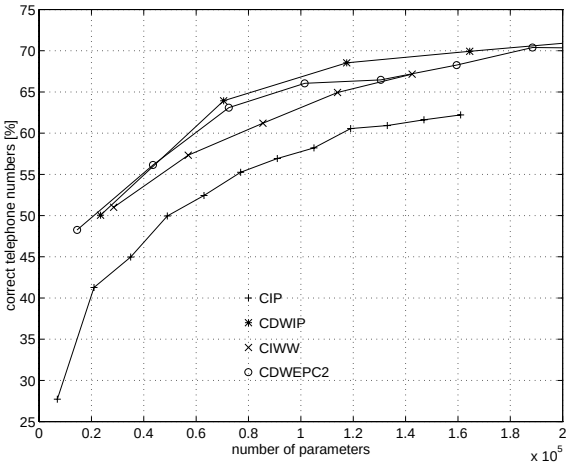

Figure 4. Context dependent phoneme models vs whole word models

grammar	correct sentences
word loop (34 words)	46.3 %
natural number loop	53.4 %
number pair loop	58.9 %
natural number loop with grammar scaling	64.3 %
4 number pairs	68.5 %

Table 4. Results for different grammar networks

we took a closer look at these recognition results. A more complete error analysis is reported in [3].

These models gave 68.5 % correct sentences which is a too high error-rate for practical purposes. We used manual inspection of the errors to tell us where to put the effort to achieve better performance. There is a scoring problem for natural numbers as explained in section 4. For 5 mixtures CDWIP models the result on digit level is 94.0 % compared to 91.2 % on word level. The percentage correct sentences when counting digit errors was 70 %, this will be the percentage correct telephone numbers in an application.

Table 4 shows different kinds of grammar networks testet on the 5 mixture CDWIP-models. This shows that task constraints like the one we have used give an improved performance which is necessary for pratical purposes. Grammar scaling using word punishment may help when such a strong task constraint is impossible.

First we analysed the errors with respect to noise, dialects, sex and age. Speech signals disturbed by noise reduced the recognition accuracy dramatically. Surprisingly, the recogniser performed significantly worse on women than men. The main reason for this was background noise. Typically, when women talked on the telephone children cried or shouted in the background, whereas this never happened with men. More robust features together with more advanced silence and noise models may handle the noisy utterances better.

Words which differed by only one or two distinctive phonetic features were more misrecognised in the noisy part of the testset. Discriminative training is expected to take better care of these kinds of problems. Discriminative trained models have been shown to give better performance at low complexity [9].

A more complete transcription including different dialectal pronunciations may alleviate some of the misrecognitions found in other dialect regions than the south-eastern part. Especially in western parts of Norway some of the natural numbers are pronounced differently from this transcription. New segmentation using several pronunciations will be tried. More advanced whole-word modeling using more models and different topology may give better performance. Investigating results on the trainingset suggests that better performance for CIWW is possible.

The investigation of misrecognised telephone numbers showed that the numbers of ten were most prone to errors. Uttered in isolation the final vowel was prolonged and an extra word was inserted. Uttered in a number-pair with the "new" pronunciation the number of ten became unstressed and reduced. In order to alleviate this problem we modelled the two different pronunciations separately by adding 38 extra CDWIP models. The number of parameters for the 5 mixture models then became approximately the same as for the 7 mixture "single" CDWIP models but the performance was better: 74.2 % correct sentences.

Summary of possible improvements:

- more robust features
- more advenced silence and noise models
- discriminative training
- new transcribation and segmentation using several pronunciations
- several models for the same word in different contexts
- several different topologies

REFERENCES

[1] B. H. Juang, J. G. Wilpon, "Recent Technology Developments in Connected Digit Speech Recognition". *Proc. ICSLP-94*, pp. 2135–2138, Yokohama, September 1994.

[2] C. N. Jacobsen, J. G. Wilpon, "Automatic recognition of Danish natural numbers for telephone applications". *Proc. ICASSP-96*, pp. 459–462, Atlanta, May 1996.

[3] K. Kvale, "Norwegian numerals: A challenge to automatic speech recognition". *Proc. ICSLP-96*, Philadelphia, October 1996.

[4] H. Ljøen, I. Amdal, F. T. Johansen, "Norwegian Speech Recognition for Telephone Applications". *Proc. NORSIG-94*, Ålesund, June 1994.

[5] P. Rosenbeck et al, "The Design and Efficient Recording of a 3000 Speaker Scandinavian Telephone Speech Database: Rafael.0". *Proc. ICSLP-94*, pp. 1807–1810, Yokohama, September 1994.

[6] I. Amdal, H. Ljøen, *TABU.0 – en norsk telefontaledatabase*. Technical Report 40/95, Telenor, June 1995.

[7] T. F. Pedersen, *Automatisk gjenkjenning av tallord fra norske dialekter*. Diploma thesis, Norwegian University of Science and Technology, December 1995.

[8] S. Young et al, *HTK – Hidden Markov Model Toolkit V2.0.*. Entropic Cambridge Research Laboratory, November 1995.

[9] F. T. Johansen, *Global discriminative modelling for automatic speech recognition*. PhD thesis, Norwegian University of Science and Technology, May 1996.