IMPROVED AUTOMATIC RECOGNITION OF NORWEGIAN NATURAL NUMBERS BY INCORPORATING PHONETIC KNOWLEDGE

Knut Kvale and Ingunn Amdal

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

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) opens for more user-friendly services, but also introduces many new, language-specific problems, such as more similar sounding words, a more complex grammar network, and more ambiguities due to segmentation problems of connected natural numbers. The paper shows that incorporating phonetic knowledge into a Norwegian natural number recogniser, improved the recognition performance from 70.6 % to 76.3 % correctly recognised 8-digits telephone numbers in noisy conditions.

1. INTRODUCTION

The number of services based on automatic speech recognition (ASR) over telephone lines has increased vastly over the last few years, and many of these applications are based on isolated or connected digit recognition. However, in many languages long numbers are normally memorised and pronounced as natural numbers such as “fifty two” rather than single digits, “five two”. Thus, in order to offer more user-friendly services the recognisers have to cope with all the numerals from 0 to 99, or even up to 9999.

We have investigated ASR of Norwegian natural numbers in the context of automatic telephone number recognition. In Norway, most telephone numbers are listed as four number-pairs in the phone directories e.g. 22 34 56 78, and are therefore usually pronounced as connected natural numbers. However, if a number-pair begins with 0, it has to be pronounced as two digits. Some speakers, especially young people, tend to read long numbers with single digits (which is normal e.g. in Swedish).

We therefore restricted the ASR-task to recognition of exactly 4 pairs of numbers, where a number-pair may be *two digits* or a *natural number*. This constraint resolved some of the inherent ambiguities in connected Norwegian natural numbers which may be very difficult to model, e.g. both 50 2 and 52 are pronounced /**femti tu:**/.

2. NORWEGIAN NATURAL NUMBERS

2.1. The speech database

The experiments are based on the Norwegian 1000 speakers TABU.0 speech database [1], which is a part of the Scandinavian 3000 speakers Rafael.0 database [2]. The speakers were phoned 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, where each 8-digit telephone number was listed as 4 number-pairs.

The manuscripts were designed to provide enough samples for training of each *word* in the natural number vocabulary, see table 1, yielding an over-representation of the “-teen” numbers (13–19) and the numbers of ten.

The database has been manually inspected and about 40 % of the speech material was perceived as having considerable channel or background noise. Deviations from the manuscript were not discarded as long as the resulting utterance still consisted of Norwegian natural numbers.

Nb.	*Ortho.*	*Phono.*	*Nb.*	*Ortho.*	*Phono.*
0	null	**nʉl**	15	femten	**femtn**
1	en	**e:n**	16	seksten	**sæistn**
2	to	**tu:**	17	sytten	**søtn**
3	tre	**tre:**	18	atten	**atn**
4	fire	**fi:r**ə	19	nitten	**nitn**
5	fem	**fem**	20	tjue	çʉ:ə
6	seks	**seks**	20	tyve	**ty:v**ə
7	syv	**sy:v**	30	tretti	**treti**
7	sju	ʃʉ:	30	tredve	**tredv**ə
8	åtte	ɔ**t**ə	40	førti	**føʈi**
9	ni	**ni:**	40	førr	**før**
10	ti	**ti:**	50	femti	**femti**
11	elleve	**elv**ə	60	seksti	**seksti**
12	tolv	**t**ɔ**l**	70	sytti	**søti**
13	tretten	**tretn**	80	åtti	ɔ**ti**
14	fjorten	**fjuʈɳ**	90	nitti	**niti**
				og	ɔ:

Table 1. Orthographical and phonotypical transcription (South-Eastern Norwegian) of the whole word units in the Norwegian numerals 0–99

2.2. Vocabulary and grammar

A particular problem in Norwegian is the two morphologically different ways of pronouncing natural numbers, e.g. 52 may be read both from left to right, “femti to”, (fifty two), as in English and Swedish, or the other way round, “to og femti” (two and fifty), as in German and Dutch.

Earlier the German-Dutch way of reading numbers was

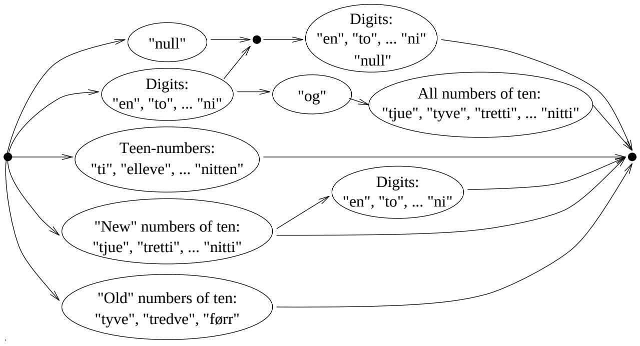

Figure 1. Grammar network for Norwegian number-pairs 00–99

used, but in 1951 the parliament decided that from then on numerals should be read from left to right in Norwegian. However, today both ways of counting are used, especially in informal everyday speech. In formal speech, e.g. reading telephone numbers from a manuscript, people are less likely to use the "old" pronunciation. In our database, 7.8 % of the speakers used the "old" pronunciation, though most of them mixed the two pronunciations.

The 1951 reform also established /**ʃʉ:**/, /**çʉ:**ə/, /**treti**/ and /**føʈi**/, as the standard pronunciations of 7, 20, 30 and 40 respectively, removing the alternative pronunciations /**sy:v**/, /**ty:v**ə/, /**tredv**ə/ and /**før**/. However, in our database the "old" forms of these numerals were still used in 23.1 % of the cases. For more details, see [3].

When including both counting styles and alternative pronunciations of 7, 20, 30 and 40, the vocabulary of the Norwegian natural numbers 0–99 can be built from the 33 different whole-word units listed in table 1, where the word "og" means and. All the telephone numbers in the database were automatically transcribed with the phonotypical transcription of South-Eastern Norwegian shown in table 1. A phoneme based recogniser was used to identify the actual pronunciation of 7, 20, 30 and 40.

Figure 1 shows the grammar network for the Norwegian number-pairs, including the two different counting styles.

3. ACOUSTIC MODELLING

For this 34 word vocabulary (33 words and silence) the most natural choice of acoustical unit is the whole-word units listed in table 1. However, similar experiments on Danish natural numbers have shown that context-dependent phoneme models outperformed context-independent whole-word models [4]. Hence, we wanted to compare whole-word units with different kinds of context dependent and context independent phoneme units.

The recogniser was based on the *Hidden Markov Model Toolkit* (HTK v.2.0) [5]. 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 8-digit telephone number.

The different acoustical units were modelled as left-to-right continuous density hidden Markov models (HMMs) with no skip transition and with diagonal covariance matrices. The models were trained using k-means clustering and Baum-Welch re-estimation. The phoneme-based models were three-state HMMs, whereas in the context independent whole-word (CIWW) models the number of states per model depended on the number of phonemes in each word. In this way the complexity of the CIWW-models was the same as for word-dependent phoneme models.

For this vocabulary, only 28 context independent phoneme (CIP) models and a silence model are needed, (out of a total of about 45 phonemes in Norwegian). The 99 context dependent phoneme models with word internal context (CDWIP) were estimated by cloning the context independent phoneme models and then re-estimating them using triphone transcription.

Context dependent phoneme models with word external context (CDWEP) were far more complex to model, and lack of relevant training material became a problem since not all contexts were represented in sufficient numbers for training, and some contexts were not represented at all. In this 34 word vocabulary there are 653 word external contexts, of which 550 appeared in our database with 120 different telephone numbers. We applied state-clustering to train rare contexts, to include the non-occurring contexts, and to keep the complexity at a reasonable level. The data driven clustering reduced the number of states from 1650 needed for unclustered models, to 183 states. The CDWEP-models were also trained from the CIP-models.

The various models were trained on 580 speakers and tested on 200, giving a total of 12520 telephone numbers for training and 2168 for test. Since speech material based on the same manuscripts was used for both training and testing ("text overlap"), we did not estimate statistical language models.

4. RESULTS

There is a scoring problem when evaluating ASR of natural numbers, because there is no one-to-one correspondence between the word units and the numerals. In our testset there were 6.1 words per telephone number on average. For instance, the two digits 52 may be represented with three words as "to og femti" (two and fifty), or two words as "femti to" (fifty two) or "fem to" (five two), and 50 may be one word "femti" (fifty) or two words "fem null" (five zero). Hence, the string of digits in a telephone number may be correctly recognised although some words are substituted. Different words representing the same digit, e.g. both "sju" and "syv" represent 7, may be interesting to investigate in an error analysis, but will not give errors in a practical application. However, in this paper we have included these kinds of errors in the recognition results.

Thus, *string accuracy* expresses percentage correctly recognised telephone numbers regarded as a *string of words*; not as a string of digits. For instance the CDWIP-models with 5 mixtures achieved a *word* accuracy of 91.2 % (see table 2) which gave 94.0 % *digit* accuracy. Similarly, the string accuracy for these models was 68.5 %, yielding 70.6 % correctly recognised 8-digits telephone numbers.

4.1. Comparing models

For practical applications the complexity of the models has to be constrained. Figure 2 compares the different models as a function of the number of parame-

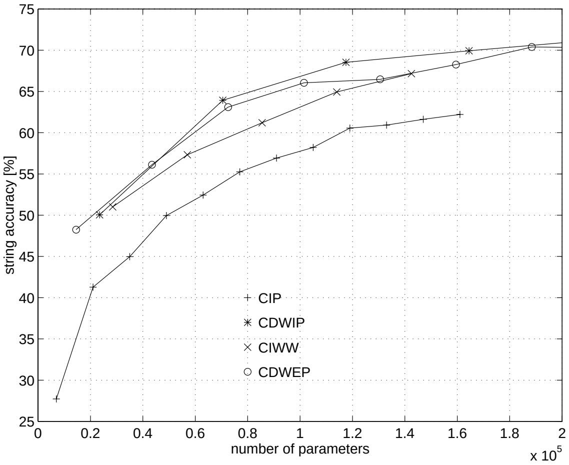

Figure 2. String accuracies for the different models as a function of the number of parameters

Model	*# Mix*	*# Parameters*	*Word accuracy*	*String accuracy*
CIP	15	105000	88.4 %	58.2 %
CIWW	4	114000	90.4 %	64.9 %
CDWEP	7	101500	90.8 %	66.1 %
CDWIP	5	117500	91.2 %	68.5 %

Table 2. Word and string accuracies for the different models with approximately the same number of parameters

ters. As expected the context-dependent phoneme models, CDWIP and CDWEP, outperformed the context-independent phoneme CIP-models. Probably due to better duration modelling, the CDWIP and CDWEP-models also achieved higher string accuracies than the context-independent whole-word CIWW-models.

Modelling word external contexts did not improve the string accuracy, although most of these contexts were present in our material. Actually, the CDWIP-models performed slightly better than the CDWEP-models. In addition, a network based on CDWEP-models is more complex, making such models less suitable for practical applications.

Hence, the CDWIP-models seemed most promising for further refinements. Since the performance with these models saturated at about 5 mixtures, we examined these recognition results more carefully to find where to concentrate our effort on improving the recogniser. Table 2 lists the word and string accuracies for the different models with approximately 110000 parameters.

4.2. The 5 mixtures CDWIP-models

First, we analysed the errors with respect to noise, dialects, sex and age, as shown in table 3, where males and females represent people older than 12 years of age. Speech signals disturbed by noise reduced the recognition accuracy. Especially words which differed by only one or two distinctive phonetic features were more often misrecognised in the noisy part of the testset.

Part of testset		*# Telephone numbers*	*String accuracy*
All	All	2168	68.5 %
Noise	Noise	923	61.9 %
	Non-noise	1245	73.5 %
Dialect	East	763	72.1 %
	West	669	65.5 %
	North	736	67.7 %
Sex	Male	961	72.1 %
	Female	1024	67.3 %
Age	8–12	183	56.8 %
	13–18	358	72.6 %
	19–34	763	71.4 %
	35–59	723	69.3 %
	60+	141	53.9 %

Table 3. Recognition performances with 5 mixtures CDWIP-models at different parts of the testset

Surprisingly, the recogniser performed significantly worse on women than men. The main reason for this, however, was background noise. Typically, when women talked on the telephone children cried or shouted in the background, whereas this never happened with men.

As seen in table 3, the recogniser performed significantly worse than average on 8–12 years old children and people older than 60 years, which is consistent with findings from the Danish database [6]. In addition to physiological differences it seemed that these people spoke with either too little or too much intensity and that they hesitated more and produced more non-speech sounds.

The main reason for the differences between the three dialect regions of Norway shown in table 3, is that only phonotypical transcription of a south-eastern Norwegian dialect was used for training and testing the models. Especially in western parts of Norway some of the natural numbers are pronounced differently from this transcription.

A careful analysis of the misrecognised telephone numbers for the 5 mixtures CDWIP-models [3], revealed that the numbers of ten (20, 30,..., 90) were particularly prone to errors:

- When the numbers of ten were pronounced in isolation the final phonemic short vowel was often prolonged and realised as a schwa towards the end. Thus, our recogniser aligned the number of ten correctly, but it inserted an extra digit at the end of the prolonged final vowel. For instance 50-/**femti**/ was often recognised as /**femti e:n**/-51, or /**femti tre:**/-53.
- With the "new" pronunciation, the natural numbers 21–99 are commonly pronounced with an *iambic* or *anapaestic* stress pattern, i.e. only the last digit is stressed. Since the number of ten is unstressed, it is realised shorter, with less intensity and more reduced than in stressed position. This made the numbers of ten prone to errors in this position.
- With the "old" pronunciation the first syllable in both digits of the numbers were stressed, and therefore much easier to recognise correctly.

5. APPLYING PHONETIC KNOWLEDGE

5.1. Alternative transcriptions

Some of the recognition errors were due to dialectal pronunciations which differed from the selected phonotypical transcription. Hence, we expanded the transcription for some of the most error prone numerals and tested these expanded transcriptions in the regions of Norway where these particular pronunciations are commonly used. As the "single" 5-mixtures CDWIP-models from the original words were applied, new training was not needed for this experiment.

- 16 is commonly pronounced as **/sekstn/** in major parts of southern-Norway, except for the Oslo-area. Testing on this region without any modifications of the transcription gave 67.4 % string accuracy. Applying the two different pronunciations of 16 improved this score to 70.4 % for this region and to 69.2 % for the whole country. The word **/sekstn/** was build up from the CDWIP-models of **/seks/** from 6 and 60, and **/tn/** from all the "-teen"-numerals.
- 90 may be pronounced **/neti/** in some parts of mid-Norway (north-Trøndelag). In this region the original models achieved 69.8 % string accuracy, whereas the alternative pronunciation gave 71.2 % for this region and 68.9 % for the whole country. We modelled **/neti/** rather crudely by **/ret/** from 13, and **/ti/** from all the numbers of ten.
- 1 is pronounced **/æin/** in several parts of south-Norway, except for the Oslo-area. However, expanding the transcription with **/æin/** did not improve the string accuracy. For this region the string accuracy without any modification was 63.3 %, with the alternative pronunciation 63.5 %. This result may be due to the crude models for **/æin/** which were taken from **/sæistn/**.

40 is rarely pronounced **/før/** in our database (less than 3 % of all examples). Removing this alternative pronunciation did not alter the recognition accuracy.

5.2. Numbers of ten

The investigation of misrecognised telephone numbers showed that the numbers of ten were most prone to errors and that these numbers were realised both stressed and unstressed. Therefore, the numbers of ten were retrained with two different models: One for the *stressed* pronunciation (i.e. uttered isolated or in a number-pair pronounced with the "old" counting style), and one for the *unstressed* pronunciation (i.e. uttered in a number-pair with the "new" pronunciation). We removed the **/før/** transcription of 40 before retraining the models.

With two sets of models for the numbers of ten we had a total of 126 CDWIP-models. This increased the number of parameters for the 5 mixture models to 150 000, which was still less than for the 7 mixtures "single" CDWIP-models (165 000). With fewer parameters the "double" models achieved 74.2 % string accuracy whereas the 7-mixtures models obtained 69.9 %.

In addition, when the two transcriptions of 16 were used, the string accuracy for these "double" CDWIP-models increased to 74.8 %.

5.3. Special models for the different regions

Since the recognition performance differed for the three regions of Norway, we retrained by embedded re-estimation the 5-mixture CDWIP-models which were trained on speech samples from the whole country, on the training set for each region. Testing these "regional" models on their own region only, yielded 76.9 % string accuracy for East, 69.4 % for West and 68.6 % for North-Norwegian.

The improvements for the East region is probably due to a smaller variation in the realisations of the words, and therefore better correspondence with the given transcription. The West region included all dialects with dorsal realisations of **/r/** for which we now have a better model.

With respect to complexity, these results can be compared with the 15 mixtures CDWIP-models trained on the whole country, which obtained 72.8 % string accuracy for the whole country and 75.5 % for East, 70.5 % for West and 71.9 % for North-Norwegian.

6. CONCLUSIONS

In this paper we have applied phonetic knowledge to improve a low complexity recogniser for Norwegian natural numbers. The most noticeable effect was achieved by training two different models for the numbers of ten, one for stressed and one for unstressed pronunciation. For 5 mixtures CDWIP-models, this modelling increased the string accuracy from 68.5 % to 74.2 %, which gave 76.3 % correctly recognised 8-digits telephone numbers. Alternative pronunciations of certain numbers also improved the recognition accuracy; especially in the regions where the alternative pronunciation is commonly used. The alternative pronunciations may be used to label the training set.

The best result with 150 000 parameters was 85.9 % string accuracy, obtained with two pronunciations of 16 and "double" 5-mixtures CDWIP-models tested on the South-Eastern region of Norway and non-noise sentences only. This 85.9 % string accuracy yielded 87.8 % correctly recognised 8-digits telephone numbers, or 96.1 % *word* recognition accuracy and 97.4 % *digit* accuracy.

REFERENCES

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

[2] 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.

[3] K. Kvale, "Norwegian numerals: A challenge to automatic speech recognition". *Proc. ICSLP-96*, pp. 2028–2031, Philadelphia, October 1996.

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

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

[6] J.G. Wilpon, C.N. Jacobsen, "A study of speech recognition for children and the elderly". *Proc. ICASSP-96*, pp. 349-352, Atlanta, May 1996.