NORWEGIAN NUMERALS: A CHALLENGE TO AUTOMATIC SPEECH RECOGNITION

Knut Kvale

Telenor Research and Development
P.O.Box 83
N-2007 Kjeller, NORWAY
Email: knut.kvale@fou.telenor.no

ABSTRACT

This paper addresses the problem of speaker-independent connected numeral recognition over telephone lines. Increasing the vocabulary from digits (0-9) to numerals (0-99) opens for more user-friendly services, but it also introduces many new, language-specific problems. This paper investigates morphological, phonemic and allophonic variations in the pronunciation of numerals in Norwegian. If improvements in recognition performance are to be achieved these language-specific issues have to be considered.

1. INTRODUCTION

The number of services based on automatic speech recognition (ASR) over telephone lines has increased tremendously over the last few years. Many of these applications are based on connected digit recognition, e.g. credit card and account number validation, catalogue ordering, reverse directory services and voice dialling by spoken digits.

However, for some applications such as digit dialling by voice, e.g. Norwegians normally do not pronounce the phone numbers as single digits, but group them as pairs of numbers, e.g. 22 34 56 78, as they are also listed in the phone directories. If a number-pair begins with 0, it is pronounced as single digits. This may influence the speaker to read the whole phone number with single digits. Especially young people tend to read numbers with single digits (which is normal e.g. in Swedish). Thus, the recognizer has to cope with all the natural numbers from 0 to 99.

In the following sections we will investigate how language specific factors influence the automatic speech recognition of phone numbers.

2. RECOGNISING PHONE NUMBERS

2.1 Speech database

This investigation is based on the TABU.0 speech database [1], [2], which consists of 1000 speakers from all over Norway. The speakers were called up by interviewers and asked to read phone numbers from a manuscript as they would have spoken to an automatic service. The 8 digit phone numbers were grouped as 4 pairs of numbers. The database was designed for training with a uniform distribution of *words*. Therefore, the "-teen" numbers (13-19) and the numbers of ten are over-represented.

2.2 Phonotypical transcription

For ASR over telephone lines it is of crucial importance to classify the stressed vowels correctly. The stressed vowels are most clearly articulated and have most intensity and are thus less influenced by signal distortions.

First stressed vowel	Phonotypical transcription (South-Eastern Norwegian pronunciation)
iː	4-/fiːrə/, 9-/niː/, 10-/tiː/
i	19-/nitn/, 90-/niti/
yː	20-/tyːvə/, 7-/syːv/
eː	1-/eːn/, 3-/treː/
e	5-/fem/, 50-/femti/, 15-/femtn/, 13-/tretn/, 30-/treti/, 6-/seks/, 60-/seksti/, 11-/elvə/, 30-/tredvə/
ø	17-/søtn/, 70-/søti/, 40-/føʈi/, /før/
ɑ	18-/ɑtn/
ɔ	8-/ɔtə/, 80-/ɔti/, 12-/tɔl/
u	14-/fjuʈɳ/
uː	2-/tuː/
ʉː	7-/ʃʉː/, 20-/çʉːə/
ʉ	0-/nʉl/
æi	16-/sæistn/

Table 1: The Norwegian numerals 0-99 grouped with respect to the first stressed vowel.

Note the two different pronunciations of 7, 20, 30 and 40 in table 1, giving a total of 32 different words, (see also section 3.1). However, only 29 phonemes were needed for transcribing the numerals (out of about 50 phonemes in natural spoken Norwegian).

2.3 The recognizer

For this task we have developed a recognizer based on the Hidden Markov Model Tool Kit (HTK), [3]. Each 10 ms speech frame is represented by 12 mel frequency cepstral coefficients plus normalised log-energy together with their corresponding first and second order regression coefficients. Cepstral mean subtraction (CMS) is applied for each phone number. The phoneme models were trained on the phonotypical transcription, see table 1. We estimated 99 word-internal triphones by cloning

the context-free phonemes and then re-estimating using triphone transcription. The triphones were modelled as three-state left-to-right continuous density hidden Markov models (CDHMM) with no skip transition. We applied a diagonal covariance matrix, 5 mixtures and no triphone or state clustering, yielding about 117500 parameters for the word-internal triphone models. The recognizer was trained on 580 speakers and tested on 200.

2.4 Results

We restricted the task to recognition *of exactly 4 pairs of numbers*. However, there may be more or less than 8 *words* in the strings. For instance the two digits 28 may be pronounced with three words "eight-and-twenty" or two words "twenty-eight", and the two digits in 20 as one word, "twenty".

For this task our recognizer obtained a word error rate of 8,2 % yielding 70 % correct recognised phone numbers. However, the recognizer performed significantly worse than average on children (8-12 years) and people older than 60. The reasons for this may be that these people spoke with either too little or too much intensity and that they hesitated more and produced more non-speech sounds, e.g. clicks and breath noise.

Surprisingly, the recognizer 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. About 7 % of the recognition errors were caused by background noise, and women were highly over-represented.

3. ERROR ANALYSIS

Measurements of ASR performance with dynamic programming (DP)-based string alignments and confusion matrices may give false impressions of the recognition errors. When e.g. 15-/femtn/ is recognised as /fem e:n/-5 1, the DP-alignment shows that /fem/ is inserted and that /femtn/ is substituted with /e:n/, whereas a manual analysis shows that the word /fem/ is correctly recognised and that /tn/ is substituted with /e:n/.

We therefore analysed manually all the 649 phone numbers which were wrongly recognised, and tried to classify the errors.

This analysis of recognition errors showed that numerals with identical first stressed vowel (table 1) were most frequently confused with each other. This means that the stressed vowels were recognised correctly but the consonants and unstressed vowels were prone to errors.

Classifying the ASR-errors is difficult because most of the errors are due to many co-occuring factors, such as huge differences in signal level, speaking rate and phoneme realisations both within a single speaker and across speakers, telephone-bandwidth speech, signal distortion by transducer, channel variations and background noise, (both extraneous speech and acoustic events). In addition, when reading phone numbers the last pair of numbers is often pronounced with creaky voice, less intensity and final lengthening, making them more prone to recognition errors. In spite of this we have categorised the errors into five classes: Morphology, dialect, the numbers of ten, the "-teen" numbers and connected numerals.

3.1 Morphology

In 1951 it was decided that numerals should from then on be read from left to right in Norwegian, e.g. 52 as /femti tu:/ and not as /tu: ɔ femti/ which was common at that time. The main argument for the change was that it is easier to process numerals if they are pronounced as they appear in texts, from left to right. Also a growing use of the "new" pronunciation (as in Swedish and English) in the defence forces and among switchboard operators caused mix-ups (for people using the "old" pronunciation).

The optimists forecast that the reform would be accepted by the public within a five year period. But there were misgivings as well, since the trochaic or dactylic stress pattern of the "old" pronunciation agreed with the normal stress pattern in Norwegian, while the "new" pronunciation gave an iambic or anapaestic stress pattern which normally only occurs in Norwegian in words of foreign origin.

Especially in non-formal everyday speech the "old" pronunciation is frequently used, both by old and young people. In formal speech, e.g. reading phone numbers from a manuscript, people are less likely to use the "old" pronunciation. However, in our database 336 (3 %) of totally 10922 numerals which could be pronounced in both ways, were pronounced in the "old" way. Of 780 speakers, 61 (7,8%) used "old" pronunciation, though most of them mixed the two pronunciations. Thus, 45 years after the reform we have two ways of pronouncing such numbers in Norwegian, and we never know for sure which one will be used.

The 1951 reform also established /ʃʉ:/, /çʉ:ə/, /treti/ and /føɻi/ as the standard pronunciation of 7, 20, 30 and 40, removing the alternative pronunciations /sy:v/, /ty:və/, /tredvə/ and /før/. Our analysis of read aloud phone numbers shows that the "new" forms of these numbers were used in 76.9 % of the cases, with 70,4 % for /ʃʉ:/, 82,5 % for /çʉ:ə/, 68,8 % for /treti/ and 97,5 % for /føɻi/. This shows that the reform has only been successful for the /føɻi/ pronunciation.

Accordingly, we have to expand the grammar to accept the two morphologically different ways of pronouncing numerals in Norwegian, include an extra model for the word "og" /ɔ/ and include two transcriptions of 7, 20, 30 and 40. With more alternatives to choose among, the recognizers will be more prone to errors.

Recognition errors due to the "old" number pronunciation

Due to the trochaic or dactylic stress patterns of the "old" pronunciation the first syllable in both digits of the numbers are stressed. The stressed syllables have more intensity and are more clearly articulated than unstressed syllables, and are thus easier to recognise. Therefore, 94 (98,9 %) of the 95 numerals read with "old" pronunciation in the testset were correctly detected. In 6 of these numerals one of the digits was wrongly identified.

On the other hand, "old" pronunciation was *inserted* 32 times. A typical error occurred when a pair of numbers was followed by a number of ten, as 34 50 pronounced /treti fi:rə femti/, but recognised as /treti fi:rə ɔ femti/, which is 30 54. The only phonemic difference between these pronunciations is the extra /ɔ/ for the function word "og" (which is normally realised as a short, reduced, centralised vowel). The "new" pronunciation of 34, with stress on 4, may result in a longer /ə/ than normal and makes it possible for the recognizer to split the schwa into /ə/ + /ɔ/.

Also hesitations, repeated starts and extraneous speech may mislead the recognizer to insert an extra "old" pronunciation, e.g. 90 pronounced /æh niti/ is recognised /fem ɔ niti/, 95.

Confusions between "old" and "new" forms of 7, 20, 30 and 40 do not lead to recognition errors of numbers. However, augmenting the lexicon with different pronunciations may lead to other confusions such as 30-/tredvə/ substituted by 11-/elvə/ and 7-/sy:v/ confused with 20-/ty:və/ or 4-/fi:ɾə/.

Recognition errors due to the "new" number pronunciation

With the "new" pronunciation, the numerals 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 makes the numbers of ten prone to errors in this position. For instance 20-/çʉ:ə/ is likely to be confused with 7-/ʃʉ:/ even when uttered in isolation. With the "new" pronunciation, e.g. 22-/çʉ:ə tu:/, the schwa in the unstressed number of ten is often elided, giving /çʉ: tu:/, which is even more like 7. In addition, there is a growing trend among young Norwegians to pronounce /ç/ as /ʃ/. Thus 22 may be realised /ʃʉ: tu:/, and the confusion with 7 2, /ʃʉ: tu:/, is complete.

Of 2684 numerals pronounced with the "new" pronunciation, 128 (5 %) were wrongly identified because of the unstressed number of ten.

Ambiguity is another problem with the "new" pronunciation. The only difference between e.g. 40 2 and 42 is the stress on the number of ten. Thus, although the recognizer identifies the phoneme sequence correctly, /føɽi tu:/, it may lead to the wrong phone number.

3.2 Dialects

There is no widely accepted standard pronunciation of Norwegian. In fact Norwegians use their own dialect in most situations. Of the numerals between 0 and 99, only 2, 9, and 10 are pronounced fairly uniformly all over the country. For the rest of the numerals the pronunciation varies widely. In spite of this, only 114 (0,86 %) of a total of 13252 words in the testset were wrongly recognised because of dialectal pronunciations.

There are two main reasons for this: (i) The informants normalised when reading phone numbers from a manuscript, and (ii) well known dialectal sound changes in natural speech were not so prominent for the numerals.

Due to normalisation, dialectal pronunciations as 11-/æɽəvə/ or /øløv/, 12-/tɔɽv/ or /tøɽv/, 17-/søyçɑ/, /sɑɯçɑn/, /sytn/ or /sætn/, 18-/ɔ:çɑ/ or /ɑ:çɑn/, 19-/ni:çɑ/ or /ni:çɑn/, 20-/çʉ:gə/, /tjʉ:ə/ or /cçʉ:/ and 70-/ʃʉti/, /ʃøti/ or /syti/, were relatively rare in the testset. If such forms are not included in the lexicon, the recognizer will probably err, but we expect that these forms also will be used rarely in practical services.

As regards dialectal variations of certain phonemes, one special problem in Norwegian is caused by the pronunciation of /r/ which occurs in several numerals. Depending on the speaker's dialect, /r/ is produced as an apical tap or trill, a uvular tap or trill, an alveolar, post-palatal, velar, or uvular approximant or fricative [4]. Although these /r/-realisations vary acoustically, and may cause problems for recognising natural spoken language, only the approximant realisations seem to cause serious problems for our recognizer. The approximant realisation of /r/ is normal in the South-Western parts of Norway and occurs intervocalically, i.e. in 4-/fi:ɣə/. In the waveform the /i:ɣə/ has no closure phase and looks like one long vowel. In all the other numerals with /r/ it is realised voiceless in a /tr/- or /rt/-clusters. Thus, in these numerals the difference between apical and dorsal /r/ is small. Some dialects with dorsal /r/ may also change the vowel quality in 3 to /tɣi:/, which may cause confusions with e.g. 7-/sy:v/ and 10-/ti:/, and changing 4 to /fi:ɣɑ/ and 8 to /ɔtɑ/.

Other dialectal variants which did not cause any problem for the recognizer are:

- Alveolar sounds, which are often palatalised in Mid-Norway, but for numerals only the /l/ in 12 may be palatalised. Very few pronounced it like this and this palatalization did not lead to confusion with other numbers.
- Velarization of /l/, e.g. 11 pronounced /ælˠvə/ (which is common in the Lake Mjøsa area north of Oslo).
- The typical vowel deletion in Mid-Norwegian dialects was also less common than expected, and occurred only in 4-/fi:r/ and 8-/ɔt/, which seldom lead to recognition errors.
- Lack of retroflexing of e.g. /rt/-clusters in Western parts of Norway only occurs in 14 and 40 pronounced /fjurtn/ and /førti/. Most often the /r/ was realised voiceless in these words, resulting in a minimal difference with the retroflex realisations /fjuɽɳ/ and /føɽi/.

The numbers which caused most recognition errors due to dialectal pronunciation were:

- 1 pronounced /æ:n/ and /æin/, /et/ or /eit/, where e.g. /æ:n/ was often confused with /fem/.
- 3 pronounced /tɣe:/ or /tɣi:/ with voiced and voiceless /ɣ/. With voiceless /ɣ/, /tɣe:/ was substituted with /fem/, and /tɣi:/ was substituted with /sy:v/ and /fi:rə/.
- 4 pronounced /fi:r/, /fi:ə/, /fi:ɣə/ or /fi:ɣɑ/, where e.g. /fi:ɣə/ was often confused with /tre:/.
- 16 pronounced /sekstn/, which was confused with e.g. /seksti/-60, /seks e:n/-6 1, /seks fem/-6 5 or /seks tre:/-6 3.
- 17 pronounced /sytn/ and 70 pronounced /syti/ which may be confused with 19-/nitn/ and 7 3 -/sy:v tre:/ respectively.
- 19 pronounced /netn/ and 90 pronounced /neti/, (in some dialects in Mid-Norway), led to confusion with 13-/tretn/ and 30-/treti/ respectively.

The only error due to dialect variation in the three Northern counties of Norway was 80-/ɔːti/ (long vowel).

The stress pattern of the "new" number pronunciation makes it possible to recognise the digits correctly in spite of large dialectal variations. On the other hand, dialectal variations in the number of tens are more prone to errors with the "new" pronunciation. For instance, 40 /føʈi/ is often confused with 70 /søti/ over the telephone lines. When 40 is realised without retroflex /ʈ/ and with a voiceless dorsal /r/, these numbers become even more similar acoustically. When in addition the number of ten is unstressed due to "new" pronunciation of e.g. 42, confusion with 72 is very likely.

3.3 Numbers of ten

When the numbers of ten (20, 30, ... 90) are pronounced in isolation the final phonemic short vowel (/e/ for 20 and /i/ for the rest) is often prolonged and realised as a schwa at the end. Thus, our recognizer aligns the number of ten correctly, but it inserts an extra digit at the end of /i/ because of final lengthening. For instance 50-/femti/ was often recognised as /femti eːn/-51, or /femti treː/-53. In the testset there were 1455 numbers of ten uttered in isolation, of which 135 (9,3%) were recognised with an extra digit.

Another typical error (6,3%) for the numbers of ten was that e.g. 50-/femti/ was recognised as /fem treː/ or /fem eːn/. This error may occur because the closure phase of the plosive /t/ was recognised as a word boundary (silence).

Surprisingly, only 11 (0,8%) of the numbers of ten were substituted with their "-teen" counterparts, e.g. 50-/femti/ recognised as /femtn/-15.

3.4 The “-teen” numbers (13-19)

The plosive /t/ in the numerals 13-19 is normally released as a nasal plosion and the final nasal is syllabic. Especially 16 was often realised without a plosive at all; /seisn/. However, when people speak very clearly they may pronounce the last syllable of these numbers as -/ten/. This "over-careful" pronunciation led to most errors for these numbers, where the first part of the number was recognised as the corresponding digit, i.e. /fem/ in the case of 15, and the last /ten/ was confused by e.g. 1-/eːn/ or 3-/treː/. Of 1905 “-teen” numbers 110 (5,8 %) were wrongly recognised due to this type of confusion.

3.5 Connected numerals

When pairs of numbers are read as a part of an 8 digit phone number, coarticulation effects between the pairs of numbers may cause problems for the recognizer. One typical error was when a number of ten was followed by a 0, such as 20 0 -/çʉːə nʉl/. Here the recognizer often inserted /eːn/ as the last part of /ə/ and the first part of /n/. Also, when "-teen" numbers were followed by a 0, such as 19 0 -/nitn nʉl/, the geminate nasals were impossible to separate leading to the confusion error /niː nʉl/-9 0.

At least 15 word errors were due this segmentation problem.

4. CONCLUSIONS

We have shown that morphological, phonemic and allophonic variations in Norwegian put an extra load on the ASR-system. The numerals with identical first stressed vowel were most frequently confused with each other.

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. Modelling the two different pronunciations with separate, whole word models may improve the recognition performance.

We are relieved to conclude that for the numerals 0-99 surprisingly few errors were due to the many and varying dialectal pronunciations of Norwegian. On the other hand, transcribing different pronunciations of e.g. 16 in the lexicon does not rule out the possibility of ASR errors.

We conclude that ASR of numerals in Norwegian is a particularly hard task, and we will not be able to show satisfactory recognition results until language-specific knowledge is applied.

ACKNOWLEDGEMENT

I wish to thank Ingunn Amdal for developing the recognizer, and Arne Kjell Foldvik for helping me with phonetic and language-specific problems.

REFERENCES

1. Amdal, I. and Ljøen, H., *The Norwegian telephone speech database TABU.0*, Scientific Report 40/95, Telenor 1995 (in Norwegian).

2. Ljøen, H., Amdal, I. and Johansen, F.T., "Norwegian speech recognition for telephone applications", *Proc. Norsig-94*, pp. 121-125, 1994.

3. Young, S., et.al, *The HTK Book*, HTK V2.0, 1995.

4. Foldvik, A.K., "The change from apical to dorsal r in Norwegian", *Proc. 11th International Congress of Phonetic Sciences*, Vol.1, pp. 177-178, 1987.