Norwegian Speech Recognition for Telephone Applications

Harald Ljøen *Ingunn Amdal* *Finn Tore Johansen*

Norwegian Telecom Research
N-2007 Kjeller, Norway

ABSTRACT

In this paper we present a Norwegian telephone speech database, TABU.0. We discuss the database design specification and some experiences with recording and labelling of the database. We also present some preliminary results with a word-based recogniser trained on a subset of the database.

1. INTRODUCTION

Automatic speech recognition has now reached a level of development where it is rapidly moving from the laboratories into real-world applications, many of them over telephone lines. The most important single factor for successful speech recognition seems to be the speech database used to train the recogniser. Therefore, a Norwegian speech database is necessary in order to perform any serious work on speech recognition for Norwegian users. Furthermore, for Norwegian Telecom, the natural speech recognition applications are services in or over the public telephone network.

General public services in the telephone network present high challenges to speech recognisers due to variations in voice characteristics, dialect, background noise and speaker behaviour. Additionally, for telephone speech, there are also substantial variations in line and handset characteristics. All this variability should in some way be incorporated into the speech recogniser, i.e. it must be included in the speech database used for training, as the recogniser cannot generally be expected to cover more variability than what is spanned by the training data.

A particular problem with Norwegian speech recognition for the public, is the common use of dialects. We assume that a potential telephone service must be able to handle most Norwegian dialect pronunciations to achieve wide acceptance. There are two official written language variants of Norwegian (*bokmål* and *nynorsk*), but neither of these have any "official" pronunciation (except for in the national broadcasting, NRK).

In this paper we first present our Norwegian telephone speech database. We describe the specification (speaker distribution over dialect areas, age, etc), the vocabulary (numbers, control words, phonetically balanced text) and some experiences from the data collection and labelling phases. Finally, we present some preliminary results with a speech recogniser trained and tested on data from the database.

2. DATABASE DESIGN

The TABU.0 database consists of 1000 speakers and has been recorded over real telephone lines. PCs with ISDN cards were used to capture the A-law PCM signal, just like it would appear in a real public application.

The database design has been made in close cooperation with Jydsk Telefon, using their experiences from speech recognition over the telephone line [1], as well as experiences from the design of a similar Danish database, described elsewhere in these proceedings.

As shown in Table 1, TABU.0 contains speakers from a wide age range, from 8 years and up. The speakers are also distributed among all the main Norwegian dialect groups.

The vocabulary of the database consists of:

- **General numbers,** designed to cover all subwords between 0 and 9999. These numbers can be used for various services requiring number input like database inquiries, banking, etc.
- **Telephone numbers,** in various contexts, including a natural command phrase. This is useful for reverse directory services, voice dialling etc.
- **Command words** for some typical telephone services like list processing, etc. These words are listed in Table 3.
- **Countinous speech.** These are sentences from a school book, assembled to produce phonemic coverage. Histogram statistics (estimated from phonotypical transctiptions of the manuscript) are shown in Table 2. The additional Norwegian diphthongs /Ai/, /Oy/,

	Region / Age group	8-12		13-18		19-34		35-59		60+		
		k	m	k	m	k	m	k	m	k	m	sum
1	Finnmark nord	1	1	2	2	3	3	3	3	1	1	20
2	Finnmark sør	1	1	2	2	3	3	3	3	1	1	20
3	Troms	2	2	4	4	6	6	6	6	2	2	40
4	Narvikområdet	2	1	2	3	4	3	4	4	1	1	25
5	Bodøområdet	1	1	2	3	4	4	4	3	1	2	25
6	Mo i Rana området	1	1	2	2	3	3	3	3	1	1	20
7	Brønnøysundområdet	1	1	2	2	3	3	3	3	1	1	20
8	Ytre Trøndelag	2	3	5	5	7	8	8	7	3	2	50
9	Indre Trøndelag	4	3	6	7	10	10	10	9	3	3	65
10	Søndre Trøndelag	1	1	2	2	3	3	3	3	1	1	20
11	Moldeområdet	1	1	2	2	3	3	3	3	1	1	20
12	Ålesundområdet	2	2	4	4	6	6	6	6	2	2	40
13	Ytre Sogn og Fjordane	1	1	2	2	3	3	3	3	1	1	20
14	Indre Sogn og Fjordane	1	1	2	2	3	3	3	3	1	1	20
15	Vossområdet	1	1	2	2	3	3	3	3	1	1	20
16	Hordaland	1	1	2	2	3	3	3	3	1	1	20
17	Bergensområdet	3	3	6	7	10	9	10	10	3	4	65
18	Stavangerområdet	3	3	6	7	10	9	10	10	3	4	65
19	Vest-Agder	1	1	2	2	3	3	3	3	1	1	20
20	Aust-Agder	3	3	6	6	9	9	9	9	3	3	60
21	Gjøvikområdet	5	5	10	10	15	15	15	15	5	5	100
22	Osloområdet	7	8	15	14	22	22	21	22	7	7	145
23	Øst og Vestfold-området	5	5	10	10	15	15	15	15	5	5	100
	Total:	50	50	99	101	150	150	151	149	50	50	1000

Table 1. Speaker distribution in database TABU.0 (planned)

/}i/, and /ui/ are not present in the vocabulary.

The speakers were called up by interviewers and asked to read from the written text, trying to pronounce the vocabulary as naturally as possible. Real spontaneous speech would of course have been preferable, but is very difficult to record without an expensive and time consuming "Wizard of Oz" type of data collection. By using text read from a manuscript, we hoped to minimise influence from the interviewer.

3. RECORDING EXPERIENCES

For some categories of the database distribution, the speakers were hard to find. This was especially the case for the youngest and oldest speaker groups. The real distribution will therefore be slightly different from the one in Table 1.

The speakers were not paid, and the recording, including information and guiding, normally took between 10 and 15 minutes. The manuscript was 7 pages long, with both *bokmål* and *nynorsk* versions on the same page. Some of the speakers who at first had been willing to contribute, changed their mind when they saw the amount of text. Most were surprised, however, that the vocabulary only took about 3 to 5 minutes to read. Another manuscript layout might have been less frightening. Some kind of reward could also have been presented to the speakers to increase their motivation.

As expected, children and elderly people often made mistakes and had to repeat part of the sentences. For these groups as many as half of the sentences contain wrong words, restarts and stuttering. This requires careful manual transcription.

We also experienced the well known list effects, as the pronunciation of the command words is highly dependent of the place in the list. The last word tends to having falling pitch, the rest varying. This was however compensated for to some degree in the design, by having ten different list orderings.

Many speakers omitted the pauses we had asked for between the words. We also observed an important effect of layout on the pronunciation. In the leading text, people were instructed to read the numbers line by line, but many read these numbers column by column. In one section the speakers were asked to read the telephone numbers in pairs, and we had printed them that way. Children, however, seemed to prefer to read them digit by digit, especially when the first telephone number on the list was "01 00 03 15".

4. OBSERVED PRONUNCIATION VARIATIONS

We have made some preliminary observations about pronunciation variations in the database.

vowels		consonants	
/A:/	381	/p/	354
/A/	931	/b/	293
/{:/	168	/t/	793
/{/	35	/d/	418
/O:/	474	/k/	423
/O/	369	/g/	318
/e:/	404	/f/	200
/e/	1522	/v/	313
/i:/	209	/s/	702
/i/	385	/S/	86
/2:/	89	/C/	63
/2/	30	/j/	138
/u:/	223	/h/	471
/u/	187	/m/	514
/y:/	57	/n/	889
/y/	42	/N/	177
/}:/	173	/l/	651
/}/	104	/r/	923
diphthongs		retroflexes	
/{i/	195	/rt/	108
/2y/	29	/rd/	47
/A}/	40	/rn/	48
		/rl/	45
		/rL/	56

Table 2. Phoneme distribution (SAMPA notation) for continuous speech section of TABU.0

However, since we have only listened to a small subset ourselves (about 50 speakers), these may not cover all the variations present in the database.

All but one of the interviewers were from the Oslo-region. In spite of this, and the unnatural recording situation, speakers tend to use their dialect quite freely, without fear of not being understood. This strenghtens our belief that Norwegian speech recognition for telephone applications must understand large dialect variations to be accepted in real life.

We have collected the different subwords used to produce the numbers 0 to 9999, and have so far found 85 different subwords. We have not counted differences in nonstressed vowel, different "r"s or different "l"s. Some of the fricatives have been grouped. "20" can be said in at least 6 different ways: "kjue", "sjue", "tjue", "kjuge", "kju" and "tyve". ("sjue" is particularly common among children.) The first 3 of these pronunciations are counted as one as the fricative will be hard to distinguish.

We have also found some pronunciations that are so rare that people from other parts of the country might have problems understanding them. Of the three pronunciations of "17" we have found: "sytn", "søtn" and "saukjan", the last may be of this class.

The single subword which has most pronunciation variations is "100" (13 variations). Depending on its subword context, both the beginning and the end may and may not be omitted, and in addition there are variations in the vowel.

With the 85 subwords found so far, we should be able to cover numbers from 0 to 999 999 fairly well. We have however not yet labelled the complete database, so there might still be variations left to discover.

There are several ways of grouping the digits of an 8-digit telephone number when it is uttered spontaneously (without manuscript). Most common are pairs and one by one. Not surprisingly, however, odd groupings occur when the telephone number was originally learnt as an old 5- or 6-digit number, with a 2 or 3-digit "area code" added in front.

Two examples of grouping are: "12 345 6 78" and "12 3 45 6 7 8". The most unexpected pronunciation of a telephone number we have encountered is "34 56 78 with 12 first". We will need a robust dialogue system to take care of utterances like this one.

In Norway there are two ways of saying the numbers 20 to 99, the "old" counting system ("femog-tjue") and the "new" ("tjue-fem"). The majority of speakers seem to use the "new" system, especially children. Very few use the "old" way of counting consistently. Even when pronuncing a familiar telephone number in context the majority use "new" counting.

5. LABELLING

Manual labelling of speech is generally very time consuming, and should be kept to a minimum, with as much as possible left to automatic procedures. We therefore specified labelling of the database only at a "sentence" level for telephone numbers, isolated digits and continuous text. The numbers from 100 to 9999 and the control words are segmented at the word level.

For both these levels, short inter-word silence is equally divided between neighbouring speech, while long silent periods get their own label. We classify out-of-vocabulary sounds in a fairly detailed manner, which can be summarised into three main groups:

- Speaker-generated noise. This group is further divided into cough, laughter etc.
- Background noise. Divided into silence, background speech, line noise and other background noise.
- Other utterances from the speaker. Here we try to make a distinction between "useful" garbage speech (words typically occuring in spontaneous communication with a telephone service) and "useless" (such as communication with the interviewer, restarts, etc.). Un-

fortunately, most of these utterances in the database ended up defined as "useless".

Labelling also includes orthographic transcription, i.e. modification of the manuscript to what has actually been said. This can be a substantial part of the job for some speakers.

Some variations in pronunciation are "orthographic", meaning that a different set of subwords than the "standard" are used to construct the word. This applies to the "old" way of counting described above, as well counting using hundreds instead of thousands ("trettenhundre" instead of "ettusentrehundre"). Such variations are identified in the label.

Short out-of-vocabulary sounds inside a sentence are indicated in the sentence label. This is done to be able to test the recogniser for "non-perfect" sentences. For longer out-of-vocabulary sounds, we have divided the sentences into usable and non usable parts as described above. Many of these sub-sentences are not comprehensible.

A subset of the speakers will also be labelled at a detailed word level for the isolated digits (0 to 9). These labels do not include inter-word silence and has a detailed pronunciation (corresponding to the 85 subwords described above) indicated.

To do the labelling we have used hired personel with no special background, and the Waves+ program from Entropic Research Inc. The labelling was planned to take 600 hours for the complete database, and at the time of writing, this seems to be a fair guess.

6. WORD BASED RECOGNITION

As a first preliminary experiment with speech recognition using the TABU.0 database, we have trained word models for the 16 control words listed in Table 3. We used a simple, continuous density HMM for each word. The number of states was roughly equal to three times the number of phonemes, plus two states to include pre- and post-word silence. Observation probabilities were modeled by single mixture Gaussians with diagonal covariances. The features were identical to the COST 232 reference recogniser [2], i.e. LPC-based cepstra and log energy with deltas. The training used Baum-Welch reestimation in isolated word mode, with segment boundaries provided from the ordinary word-level labelling described above.

A silence/background noise model and a garbage speech model, both with 3 HMM states, were also included in the training. This allowed us to perform recognition on continuous input recordings, without explicit endpointing.

Training and testing was accomplished with the Hidden Markov Model Toolkit (HTK) from Cambridge University Engineering Department [3].

Words	States	Words	States
hjelp	14	stopp	14
ja	8	feil	14
firkant	23	ring opp	20
nummer	17	nei	11
igjen	14	neste	20
avslutt	20	jo	8
telefonnummer	35	gjenta	17
stjerne	20	hold	11

Table 3. Vocabulary and HMM size

We divided the first 40 sentences available from labelling into a training set of 30 speakers, and a testset of 10. All of these speakers used the *bokmål* vocabulary. Speakers from all age groups were represented in both training and test set, and the distribution among gender and dialect is shown in Table 4.

Region		Trainset		Testset	
		M	F	M	F
2	Finnmark sør	1	0	0	0
13	Ytre Sogn og Fjordane	0	1	0	0
17	Bergensområdet	1	1	0	0
19	Vest-Agder	2	1	0	0
20	Aust-Agder	0	4	0	0
22	Osloområdet	1	1	4	3
23	Øst- og Vestfold	7	10	1	2
	Sum	12	18	5	5

Table 4. Distribution of training and testing speakers. M indicates male and F female speakers.

Using the HMMs for continuous mode recognition, we obtained the results shown in Table 5. Recognition of silence and garbage speech is ignored in the statistics. However, the presence of out-of-vocabulary speech in the data accounts for a large part of the insertions obtained. This indicates that our garbage models are not sufficient for garbage rejection. This was to be expected, however, with such a small amount of training data.

7. FUTURE WORK

Since our final project goal is to have a recogniser for a yet unknown vocabulary, training word models as above is not very useful in the end. What we need is a vocabulary independent recogniser, that can be adapted to a new vocabulary without too much additional training.

We intend to use the continuous speech portions

Data	W	C	I	D	S	%Corr.	%Acc.
Train	477	476	12	0	1	99.97	97.27
Test	162	156	8	0	6	96.30	91.36

Table 5. Recognition results for 16 word vocabulary. W, C, I, D and S indicate total number of words, number of correctly recognised words, and number of insertions, deletions and substitutions, respectively.

of the TABU.0 database to train vocabulary independent phoneme models for this purpose. A set of initial models obtained from the Norwegian EUROM.0 and EUROM.1 recordings [4] will probably be used to do a semi-automatic phonetic segmentation for a small part of the database [5].

The phonetic segmentation will allow us to identify the most common word pronunciations in the continuous text passages. Improved models can then be found by embedded reestimation (using multiple pronunciations) on the larger TABU.0 database.

With the improved vocabulary-independent recogniser, we plan to test simple dialogue design for a particular application. Finally, we will attempt to collect a new (and smaller) database with spontaneuous speech for a chosen dialogue. This can be done by "Wizard of Oz" techniques, but we hope to be able to use the vocabulary independent recogniser and do it automatically instead.

8. CONCLUSIONS

In this paper we have presented some experiences from a large database collection effort. The database, TABU.0, seems to have captured Norwegian dialect variations well, although the subjects were asked to read from a manuscript. We believe there is sufficient speech material in the database to build robust recognisers for the specified number and control word vocabulary. As a first experiment, we have presented a very simple speech recogniser using whole-word models trained from the database. On a 16 word vocabulary, we obtained 96% correct recognition in continuous recognition mode, which we feel is a promising starting point. In the future, we will try to build a vocabulary-independent recogniser, based on the continuous speech parts of the database.

REFERENCES

[1] Per Rosenbeck, Bo Baungaard. *Experiences from a real-world telephone application: teleDialogue.* Proc. ICSLP'92 , pp 1585-1588.

[2] F.T.Johansen. *The COST 232 Reference Recogniser, Version 1.1.* To appear in final report from the COST 232 project.

[3] *HTK - Hidden Markov Model Toolkit V1.5.* Cambridge University Engineering Department and Entropic Research Labs., Sept. 1993.

[4] Knut Kvale. *The Norwegian EUROM.1 Database.* Technical report TF-Rxx/94, Norwegian Telecom Research, May 1994. (To appear)

[5] Knut Kvale. *Segmentation and Labelling of Speech.* Dr.ing. thesis 1993:126, Norwegian Institute of Technology, Dept. of Telecommunication, 1993.