The Norwegian part of SpeechDat: A European Speech Database for Creation of Voice Driven Teleservices

Finn Tore Johansen *Ingunn Amdal* *Knut Kvale*

Telenor Research and Development
N-2007 Kjeller, Norway

ABSTRACT

In this paper we describe the Norwegian part of a European telephone speech database, SpeechDat. We describe the database content, the recruitment and recording procedure and the annotation specification and procedure. We also report some preliminary results obtained by testing our telephone number recognizer on the database.

1. PROJECT OVERVIEW

The development of automatic speech recognition is highly dependent upon large amounts of recorded speech for training and testing. The EU project "SpeechDat – Speech Databases for Creation of Voice Driven Teleservices" (LE2-4001) was initiated to create reusable speech resources for 21 European languages and language variants [1, 2].

In this project, three types of databases were defined:

- FDB: A *fixed network* database, with many speakers recorded over the ordinary telephone network
- MDB: A *mobile* database, recorded over analogue and digital (GSM) mobile networks
- SDB: A database intended for *speaker verification*, recorded with many repetitions in various networks and conditions

Common minimum standards has been defined for content, annotation and documentation of these databases. An FDB is recorded for all participating languages, whereas only a few partners record MDBs and SDBs. All databases resulting from the project will be validated centrally and distributed through ELRA (European Language Resources Association) [3].

Telenor is responsible for the Norwegian part of the project, and has recorded a 1016 speaker FDB. Section 2 of this paper gives an overview of the Norwegian FDB design. Section 3 describes how speakers were recruited and recorded. Section 4 summarises the principles of annotation, while Section 5 present statistics computed from the annotations. A first recognition experiment with the database is described in Section 6, and our conclusion is given in Section 7.

2. DATABASE DESIGN

For each speaker in the database, 45 utterances were recorded. Seven of these were produced as spontaneous answers to questions asked, while the rest were read from a manuscript sheet. The content is summarised in Table 1.

Spontaneous items	**Codes**
1 given name of caller	O1
1 spelled given name of caller	L1
1 city of call	O2
1 date of birth	D1
1 time of day	T1
2 yes/no questions	Q1–2
Read items	
1 isolated digit	I1
1 sequence of ten digits, with pauses	B1
1 eight digit prompt sheet number	C1
1 eight digit telephone number	C2
1 sixteen digit credit card number (SDB)	C3
1 six digit PIN code (SDB)	C4
2 natural numbers	N1–2
1 currency amount	M1
1 date (e.g. *27. februar 1997*)	D2
1 relative date phrase (e.g. *i morgen*)	D3
1 analogue time (e.g. *halv 12*)	T2
6 application words (30 different)	A1–6
1 application word embedded in sentence	E1
2 city names (1000 largest in Norway)	O3–4
1 proper name (SDB)	O7
1 spelled given name	L2
1 spelled letter sequence	L3
4 phonetically rich words	W1–4
9 phonetically rich sentences	S1–9
1 sentence (SDB)	S0

Table 1. Content of the Norwegian SpeechDat FDB

Proper names were selected from the ONOMASTICA database [9]. This database contains spellings and pronunciations for the most frequent entries in the telephone directory. This results in more male names than female.

Phonetically rich words were selected from an available pronunciation lexicon used for speech synthesis [8]. In order to be readable, only words between 3 and 15 characters long were used. A maximum of 5 repetitions of each word can occur in the database. With these words, there will be a minimum of 200 repetitions of each common Norwegian phoneme. Rare phonemes, such as the diphtongs "ui" "oi" and "oy", were not taken into account in the design. In our definition, the number of common phonemes was 45.

The phonetically rich sentences were taken from news agency texts (NTB 1994). For readability, limits were put on the number of words and characters. All sentences were also manually checked for readability and possibly offending content. In total we have approximately 12000 sentences. About 15% of the sentences were translated to form manuscript sheets in Nynorsk. The phonetically rich sentences gave a minimum of 500 repetitions for each common phoneme.

All items marked by SDB in Table 1 were designed to facilitate imposter testing in speaker verification. These items were taken from a limited set, containing 150 different utterances. The utterances should correspond to recorded items in a Norwegian SDB, which has not yet been planned.

3. SPEAKER RECRUITMENT AND RECORDING

The method of speaker recruitment is different among the SpeechDat participants. In Norway we chose to contact potential callers by direct mail. People were randomly selected from the telephone directory, and sent a letter explaining the project. The letter included a manuscript sheet and a reply form. Nynorsk and Bokmål versions of the letter were sent according to the form officially selected in the municipality.

The database was recorded over real telephone lines to a digital (ISDN-based) recording platform. The recording platform consisted of two PCs running Windows 95 and recording software (ADA) from the Polytechnical University of Catalonia.

The recordings were constantly monitored. This allowed us to adjust minor problems in the recording process and with the received caller distribution. About 20% of the people mailed actually completed a call, more than our most optimistic predictions. The distribution of speakers selected for the final database is given in Table 2. The entire recording was done in nine weeks.

4. ANNOTATION METHOD

The database contains 45720 utterances. To keep the amount of manual work at a reasonable level, annotation was only done on the *orthographic level*. The main principle was to *write down all the words and sounds heard*. This principle implied that restarts, repetitions and talk which did not appear in the manuscript also should be transcribed.

Sex	**Callers**	**%**
Female	498	49.0
Male	518	51.0
Sum	1016	100
Age		
8-15	3	0.3
16-30	300	29.5
31-45	364	35.8
46-60	195	19.2
61-	137	13.5
Unknown	17	1.7
Sum	1016	100
Dialect region		
01 Finnmark nord	11	1.1
02 Finnmark sør	7	0.7
03 Troms	28	2.8
04 Narvik-området	16	1.6
05 Bodø-området	25	2.5
06 Mo i Rana-området	12	1.2
07 Brønnøysund-området	14	1.4
08 Ytre Trøndelag	36	3.5
09 Indre Trøndelag	75	7.4
10 Søndre Trøndelag	11	1.1
11 Molde-omrødet	11	1.1
12 Ålesund-området	38	3.7
13 Ytre Sogn og Fjordane	12	1.2
14 Indre Sogn og Fjordane	10	1.0
15 Voss-området	7	0.7
16 Hordaland	27	2.7
17 Bergens-området	74	7.3
18 Stavanger-området	85	8.4
19 Vest-Agder	34	3.3
20 Aust-Agder	25	2.5
21 Indre Østlandet	92	9.1
22 Oslo-området	219	21.6
23 Øst- og Vestfold-området	137	13.5
24 Foreign background	10	1.0
Sum	1016	100

Table 2. Distribution of speakers

The official Bokmål and Nynorsk dictionaries formed the basic annotation symbol set, with a few additions to cover frequently used "unofficial" words such as `SYV`. Special symbols were used for mispronounced words, unintelligible speech, truncated recordings and non-speech acoustic events (see Table 3).

All normal dialectal and stylistic pronunciation variations were regarded as correct. For instance the word "hodet" may be pronounced as differently as "hode, hodet, hue, huggu, huvvu, haue" and "haude", but should be transcribed as `HODET` or `HOVUDET`. Hence the pronunciation "hu slo seg i haue" should be transcribed `HUN SLO SEG I HODET`. Sentence context should be taken into

Symbol	Description	# Symb.
WORD	Words occuring in the lexicon	226 072
*	Mispronounced words	646
~	Truncated words	451
**	Unintelligible stretches of speech	692
[spk]	Speaker noise markers	28 343
[fil]	Filled pause markers	742
[spk]	Stationary noise markers	14 441
[int]	Intermittent noise markers	18 539
	Sum, symbols	289 926

Table 3. Content of annotation files

Annotation job per utterance	# Utt.	%
Suggested annotation accepted	8 700	19
Only noise markers added	26 101	57
Other modifications performed	7 826	17
Annotation created from scratch	3 093	7
Sum, utterances	45 720	100

Table 4. Annotation work overview

account when selecting the dictionary form. The word "bar'n" in the utterance "ække du i bar'n a" should thus be transcribed BAREN (the bar) not BARN (child).

Four categories of non-speech acoustic events were transcribed: Two originate from the speaker; *filled pauses* such as "øøh, æææ, mmm", and *speaker noise* such as lipsmack, breath and throat clear. The other two categories originate from other sources. We distinguished between *stationary* background noise such as road noise, channel noise and voice babble, and *intermittent noise* such as door slam, phone ringing, cross talk and baby crying.

Annotation was done by a semi-automatic procedure. The annotators were first presented a suggested transcription, generated partly from the manuscript, partly from the speaker database (spontaneous questions) and partly by a natural number recognizer. The recognizer was used to select between different realisations of telephone and credit card numbers.

The annotators then listened to the signal, made necessary modifications to the transcription and added special symbols. All this was done using a WWW-based annotation tool, designed so that several people could work simultaneously on the same centralised database, without any other tool than a WWW-browser and audio capabilities. The annotation was done part-time by eleven students at NTNU and took about seven weeks.

5. DATABASE STATISTICS

In total, 19728 different words were encountered in the transcriptions. 453 of these were added by the annotators. A pronunciation lexicon for all these words will be generated and included on the database CD-ROM.

Classification of utterance quality	# Utt.	%
Total utterances	45 720	100
Utterances without speaker errors	44 049	96
No speaker errors or [int]'s	28 257	62
No speaker errors, [int]'s or [sta]	19 666	43

Table 5. Quality of utterances

The 45720 annotation files contain 289926 annotation symbols, as specified in Table 3. Speaker noise, intermittent noise and stationary noise markers together account for 21% of the annotation symbols, whereas the other special symbols only add up 0.9 %.

In order to get a view on the amount of manual work needed to annotate the database, Table 4 classifies the utterances according to the annotation work performed. We can see that adding the noise markers is the most common operation needed. Without noise markers, only 24 % of the suggested annotations would have had to be changed.

The quality of the database for training and testing of speech recognizers can also be evaluated by looking at the annotations. When testing speech recognizers, it is customary to exclude mispronunciations, unintelligible speech and truncated utterances, since these are normally associated with speaker errors, not recognition errors. In speech recognizer training, one would often like to exclude utterances with intermittent noise, corresponds to the [int] symbol, in addition to speaker errors. If one would like to train or test on "clean" speech, i.e. without any background noise at all, both [int]'s and [sta]'s should be discarded.

From Table 5 we see that 96 % of the utterances can be used for testing. This number however varies from 91 % for the I1 digit item to more than 99 % for applications word items A1–6. Discarding noisy utterances however reduces the available data significantly.

6. RECOGNITION EXPERIMENT

As a first example application of the database, a Norwegian telephone number recognizer [6] was tested. This recognizer is based on continuous density hidden Markov models [7], with word-internal triphones, and has been trained on the TABU.0 database [4, 5]. A fixed length (8 digits) grammar was used. This grammar only allows telephone numbers uttered in four pairs (xx xx xx xx).

Two tests were performed on the telephone number item (C2) of the SpeechDat database. One was done on the full database, and one on a 200 speaker randomly selected testset. In both cases, telephone numbers coming from prompt sheets with the alternative spacing pattern (xxx

Test database	Number of speakers in testset	Number of utterances in test	Word accuracy	String accuracy	Correct telephone numbers
TABU.0 testset [6]	200	2168	93.1 %	74.8%	76.3 %
SpeechDat	1016	741	94.6 %	74.9%	77.5 %
SpeechDat testset	200	140	95.3 %	77.1%	80.0 %

Table 6. Telephone number recognition results on TABU.0 and SpeechDat

xx xxx) were left out, along with utterances containing pronunciation errors or extraneous speech. This reduced the number of test utterances from 1016/200 to 741/140 for the two testsets, respectively.

Results are reported in Table 6. In the word and string accuracy numbers, confusions between "sju/syv", "tjue/tyve" and "tretti/tredve" were all counted as errors. None of these will generate errors in the telephone number. In that context, errors such as "seksti en" being recognized as "seks en" are also ignored.

As we can see, the TABU.0 and SpeechDat databases give very similar results. This indicates that they are both representative for Norwegian telephone speech that can be expected in real applications.

7. EXPERIENCES AND CONCLUSION

In this paper we have presented the design, recording and annotation of the Norwegian SpeechDat database for fixed networks (FDB).

From this project we have gained a number of experiences. First of all, there were a lot of details to be worked out in the specification and design phase. The common European scope of the project was valuable since we could share common experiences from similar projects to a large extent. On the other hand, the many participants made the project management a challenging task, and the specification phase took longer time than expected. After this, both recording and annotation of the Norwegian database went fairly smooth. Most subjects seemed very positive to participate in the project, as reflected by the response rate of 20%. The resulting quality and usability of the database also seems very high.

In the immediate future, we foresee several applications of the database. It can be used to test a commercial isolated digit recognizer and to compare it with our in-house recognition models. Furthermore, the database can be used to improve our existing natural number recognizer, by providing balanced test and training material. We also intend to use the database to develop a flexible vocabulary recognizer which can be used in a wide range of applications.

REFERENCES

[1] SpeechDat Home Page *http://www.phonetik.uni-muenchen.de/SpeechDat.html*

[2] H. Höge, H. Tropf, R. Winski, H. van den Heuvel, R. Haeb-Umbach and K. Choukri *European Speech Databases for Telephone Applications* Proc. ICASSP-97, München, April, 1997.

[3] ELRA Home Page *http://www.icp.grenet.fr/ELRA/home.html*

[4] Harald Ljøen, Ingunn Amdal and Finn Tore Johansen *Norwegian Speech Recognition for Telephone Applications* Proc. NORSIG-94, Ålesund, pp. 121-125

[5] Ingunn Amdal and Harald Ljøen *TABU.0 – en norsk telefontaledatabase* Norwegian Telecom Research TF R 40/95

[6] Knut Kvale and Ingunn Amdal *Improved automatic recognition of Norwegian natural numbers by incorporating phonetic knowledge* Proc. ICASSP-97, München, April, 1997, pp. 1763-1766

[7] Steve Young et al. *The HTK Book* Cambridge University Engineering Department and Entropic Research Labs.

[8] Jon Emil Natvig and Per Olav Heggtveit*En sanntids demonstrator for norsk tekst-til-talesyntese* Norwegian Telecom Research TF R 15/93

[9] ONOMASTICA Final report