Acoustic databases for Norwegian

About this translation

This is the English translation of the official documentation of the NST acoustic databases for Norwegian. The Norwegian documentation describes the resource as it was when the document was written in 2011, and this translation renders the Norwegian version faithfully.
Språkbanken, January 2020

About these databases

The acoustic databases described below were developed by the firm Nordisk språkteknologi holding AS (NST), which went bankrupt in 2003. In 2006, a consortium consisting of the University of Oslo, the University of Bergen, the Norwegian University of Science and Technology, the Norwegian Language Council and IBM bought the bankruptcy estate of NST, in order to ensure that the language resources developed by NST were preserved. In 2009, the Norwegian Ministry of Culture charged the National Library of Norway with the task of creating a Norwegian language bank, which they initiated in 2010.

The resources from NST were transferred to the National Library in May 2011, and are now made available in Språkbanken, for the time being without any further modification. Språkbanken is open for feedback from users about how the resources can be improved, and we are also interested in improved versions of the databases that users wish to share with other users. Please send response and feedback to sprakbanken@nb.no.

The text in the description that follows is written in its entirety by Gisle Andersen, and is taken from the report *Gjennomgang og evaluering av språkressurser fra NSTs konkursbo*[1], written in 2005 before the consortium bought NST's bankruptcy estate (see above), which is a technical review of the resources. The text is adapted to the state of affairs today. Gisle Andersen has granted Språkbanken the right to use the text. The mentioned report and other information are available for download at the following link: http://www.nb.no/sbfil/dok/dok.tar.gz.

The National Library of Norway, June 2011

[1] *Review and Evaluation of Language Resources from NST's Bankruptcy Estate*

1. Acoustic databases for speech recognition

The acoustic databases comprise two subtypes: databases made for speech recognition/dictation and databases made for speech synthesis. The former category is definitely the more comprehensive.

Section 2 contains a general description of NST's databases for speech recognition/dictation. Subsequently follows a language-specific resource overview for Norwegian/Swedish/Danish in section 3, with a description of the size, formats and degree of validation of the resources. Section 4 provides a description of the methods and standards behind the validation of the recordings. Section 5 and 6 cover specific linguistic issues and dialect areas. After that, a qualitative evaluation of the acoustic resources follows in section 7. Acoustic databases for speech synthesis are described in section 8.

2. General overview of NST's acoustic databases

The material is divided into different categories defined by the purpose of the recording. These are described below.

It is generally true that the database is collected in its entirety and validated by NST itself. The exceptions from this are the subdatabases SpeechDat and Telia, which are purchased and part of the so-called "in kind" resources that NST acquired through share issue. SpeechDat consists of data for mobile and landline telecommunication for Norwegian, Swedish and Danish. NST does not have property rights over this material, and it is consequently not described in the next section. This resource is well documented elsewhere. The Telia data consists of recordings in an office environment for speech recognition conducted in the Stockholm area.

The main part of the acoustic resources are recorded and validated by means of software from L&H.[2] The recording software DSDR (Desktop Speech Digital Recorder) has been used unless otherwise specified below. All validation has been done by means of the validation software DSVS (Desktop Speech Validation Station).

Access to this proprietary software is, however, not necessary for using the data for future purposes. The data themselves are available in generally usable formats: audio files in PCM/wav, and spl log files in plain text format. (More detailed descriptions are found in the following text.) The audio files are stored in uncompressed form (without use of zip or similar tools).

A quantitative and qualitative description of the content of the database is given below.

[2] Lernout & Hauspie (translator's comment)

3. Acoustic databases for Norwegian

3.1 Recordings for speech recognition (ASR/dictation), 16 kHz

The files in this subdatabase are available for download at the following links. Due to the large volume, the material is distributed over multiple files:

- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-1.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-2.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-3.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0463-4.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.0464-testing.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dyf.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dym.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.validering.tar.gz

The subdatabase ADB_OD_Nor.NOR is aimed at technology for acoustic modelling for PC/Multimedia speech recognition and for automatic dictation (Office ASR and Dictation). The recordings took place in a closed office environment and are based on phonetically balanced manuscripts, produced on the basis of sentences from NST's Norwegian corpus. The database consists of a training and a test set, where the former is used to train the acoustic model itself, while the latter is used for testing. The distribution of the recordings is shown in the table below:

Purpose	Manuscript	Lines	Individuals	Recordings	Size (GB)
Training	nor0463	312	900	280800	97.5
Testing	nor0464	987	80	78960	26.9

The recordings are stored as one audiofile per manuscript line, which corresponds to one recorded entity, i.e. most often a sentence or in some cases a phrase or a single word. The database is organized according to a specific catalogue structure, given below.

Audiofiles:	D:\adb_0464\speech\scr0464\23\04642301\r4640007
Annotation file:	D:\adb_0464\data\scr0464\23\04642301\r4640007.spl
List of annotation files:	D:\adb_0464\doc\Spl.lst
Instructions to the informant:	D:\adb_0464\doc\nor464.scr

The naming conventions for the spl files in the catalogue *data* are as follows: .../manuscript_number/station_number/group_number/log_file. Here follows the technical information of this subdatabase:

Signal coding:	linear PCM
File format:	headerless raw
Sampling frequency:	16 kHz
Resolution:	16 bit
Format:	Intel PCM
Channels	2 (stereo)

This format is in accordance with the requirements of L&H. A part of the data is converted from this to a format required for production of acoustic models based on IBM technology. These recordings can be found in the following archives:

- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dyf.tar.gz (women)
- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.lh-ibm.dym.tar.gz (men)

This subdatabase consists of the 312 recordings from the training set from 415 men and 485 women. These are the technical specifications of this material:

Signal coding:	linear PCM
File format:	headerless raw
Sampling frequency:	16 kHz
Resolution:	16 bit
Format:	Motorola PCM
Channels	1 (mono)

The recording manuscript on which the training data are based, has a dictation part and an ASR part. The former part consists of ordinary corpus-extracted sentences required for general dictation purposes, and comprises the first 222 entities (sentences). The latter part comprises the last 90 entities and consists of phrases with personal names, names of locations, individual words, acronyms etc. required for specific speech recognition (ASR) purposes. Punctuation is read explicitly.

The recording manuscript on which the test data are based have a similar division in a dictation part and an ASR part. The dictation part comprises the first 750 entities in the manuscript, while the ASR part comprises the last 237 entities.

All the data have been validated according to the criteria and methods described below. The validation files can be found at the following URL:[3]

[3] This URL is currently not working (translator's comment)

- http://www.nb.no/sbfil/talegjenkjenning/16kHz/no.16khz.validering.tar.gz

3.2 Recordings for dictation, 22 kHz

This subdatabase can be found at the following links:

- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.mnt.data.nstdata.ambe.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.mnt.data.nstdata.norskdiktering.ambe.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.skript-doc.tar.gz
- http://www.nb.no/sbfil/talegjenkjenning/22kHz/no.22khz.tar.gz

The subdatabase ADB_D_IBM-N is aimed at development of acoustic modelling technology for automatic dictation (desktop). Unlike in the preceding resource, the recordings are made using the IBM software ObjectRexx. The recordings were conducted at the start of the collaboration between NST and IBM as part of the training period for the NST employees. The database consists of three sets recorded for different purposes: a test set, a training set and a modelling set. The distribution of the recordings is shown in the table.

Purpose	Manuscript	Lines	Individuals	Recordings	Size (GB)
Modelling	mod	260	104	27040	6.24
Testing	test	160	20	3200	0.90
Enrollment	enroll	156	20	3120	0.90

These are the technical specifications of this subdatabase:

Signal coding:	linear PCM
File format:	headerless raw
Sampling frequency:	22 kHz
Resolution:	16 bit
Format:	Motorola PCM
Channels	1 (mono)

The recordings are conducted in a closed office environment, and are based on a phonetically balanced manuscript, produced from newspaper text from Aftenposten from 1996. The recordings are stored as one audiofile per manuscript line, which corresponds to a recorded entity (sentence, phrase, unique word, sequence of numbers, sequence of letters).

This subdatabase is not validated. The documentation is therefore limited. The microphone used is an Andrea NC-61, and the sound card is Turtle Beach Montego II.

3.3 Recordings for telecommunication

This subdatabase is not freely accessible. Please contact sprakbanken@nb.no if you are interested in this material. (Translator's comment.)

The subdatabase ADB_T_Nor.NOR contains recordings for telecommunication, divided into landline and mobile telecommunication. These are suitable for making speech recognition technology for telecommunication. The data are not partitioned into a test and a training set. NST has followed the general SpeechDat II procedures when recording. The recordings are in part done using L&H's software, in part by means of UMS Diginform.

The informants have received a phone number which they could call and read sentences over the phone. The recordings contain 17 utterances of spontaneous speech in the form of answers to questions, and 40 utterances with sentences from a manuscript which are read out loud. The manuscript nor0531.scr was used both for landline and mobile recordings. In connection with the development of an ASR application for the Norwegian State Railways (NSB) names of Norwegian railway stations were also recorded. These are found in the manuscript nor0666.scr. The table shows how these recordings are distributed.

Purpose	Manuscript	Lines	Individuals	Recordings	Size (GB)
Landline	nor0531.scr	57	6231	355167	27.2
Mobile	nor0531.scr	57	2018	115026	
Landline	nor0666.scr	101	65	6565	0.4
Mobile	nor0666.scr	101	37	3737	

These are the technical specifications for this subdatabase:

Signal coding:	mu-Law
File format:	wav
Sampling frequency:	8 kHz
Resolution:	16 bit
Format:	8-bit mu-Law Compressed
Channels	1 (mono)

The subdatabase for the NSB project is also found here. The data are stored in IBM compatible telecommunication format. These are the technical specifications for the NSB data:

Signal coding:	A-Law
File format:	wav
Sampling frequency:	8 kHz
Resolution:	16 bit
Format:	8-bit A-Law Compressed
Channels	1 (mono)

The telecommunication subdatabase is only partly validated. 3108 landline and 1596 mobile recordings have been validated.

During validation, a number of files have been discarded due to insufficient quality. These are found in the subfolder ...\forkastede opptak. A number of additional recordings are discarded for other reasons than low quality, e.g. to ensure an optimal distribution of informants. These are validated and can be found in the subfolder ...\Opptak på vent. Files which are validated can be found in the subfolder ...\Validering. In addition to this, there are approx. 1000 recordings which haven’t been through the validation process at all.

The entire NSB part of the database has been validated.

3.4 Database of recorded hesitations

This subdatabase can be downloaded via the following link:

- http://www.nb.no/sbfil/talegjenkjenning/nor.nolelyder.tar.gz

A subdatabase was built for specific models for hesitation, i.e. non-verbal sound which are uttered when a speaker hesitates between words. The hesitations comprise a nasal and a vocalic sound transcribed as <*mmm*> and <*eeeh*> respectively in the manuscripts. This material is an addition which is used when producing general dictation systems. The hesitations were recorded together with the database ADB_OD_Nor.NOR described in section 3.1. The technical specifications are the same as for that subset. Similarly to the rest of ADB_OD_Nor.NOR, hesitations database consists of a test and training set as shown in the table.

Purpose	Manuscript	Lines	Individuals	Recordings	Size (GB)
Training	nor4631	28	10	200	0.6

Testing	nor4641	51	2	100	

The training manuscript consists of 20 regular sentences with two hesitations in each sentence, as well as four isolated repetitions of each hesitation. The test manuscript consists of 50 regular sentences with two hesitations in each sentence, as well as four isolated repetitions of each hesitation.

4. Validation

The term *validation* is sentral in NST's work on acoustic resources. Common validation procedures are used for the data collection projects mentioned above, and these are compliant with the procedures agreed upon by L&H. Language assistants conducted the validation, collaborating closely in groups, which were coordinated by a group leader.

Validation involves listening to every recording, marking the duration of the speech, control of the correspondence of orthography and pronunciation, marking of non-verbal events and background noise, marking of mispronunciation and dialectal forms, and indicating the technical and linguistic quality of the recording.

Non-verbal events are indicated by means of a finite set of codes and markers, defined as follows:

SPK: Indicates distinct sounds which are not speech, made by the speaker, e.g. coughing, clearing of the throat, breathing, spit sounds etc. Such sounds are not marked word-medially.
FRA: Used for mispronunciation, repetition, or for nonsense words. The marker *[FRA]* is placed in front of the word which is mispronounced.
INT: Used for time-limited sounds such as music, other voices, sirens etc. *[INT]* is placed where the sound occurs, or in front of a word in the case of a word-medial occurrence.
STA: Used for continuous sounds, e.g. sirens. The marker *[STA]* is placed where the sound is heard for the first time.
TRC: Used for truncated sentences, either at the beginning or the end of the signal. This can occur if the informant starts reading too early, or if the caller hangs up too early.
FIL: Used for filled breaks, i.e. hesitations. The marker *[FIL]* is placed where the hesitation occurs in the sentence.
DST: Is used for telephone disturbance. The marker *[DST]* is placed in front of the distorted word.
DIT: Is only used when a whistle tone is heard at the beginning of the signal. This is an indication to the informant that they should begin to speak, and should not be part of the recording.
Due to these annotations, when models are made from the material, it is possible to make sure that the data which are part of the acoustic model only contain human speech and not disturbing signals, and thereby increase the quality of the acoustic model.

The quality marking of the data is based on observations of both technical and linguistic issues. A scale from A to E is used, where A indicates the best quality and E indicates discarded recordings. The quality criteria are given below.

A: If nothing is heard apart from the speech, the mark *A* is used.
B: If non-verbal events are observed in the background (other people talking, spit sounds, background noise etc.), the mark *B* is used.
C: If these background noises are very distinct, but do not drown the speech, the mark *C* is used. *C* is also used if there is less than 100 ms between the beginning or the end of the speech signal and the end of the recording.
D: If the background noise drowns the speech such that it is difficult to hear what the informant says, the mark *D* is used. *D* is also used in cases of mispronunciation or hesitations. If there is 0 ms between the beginning or the end of the speech signal and the beginning or the end of the recording, the mark *D* is also used.
E: Recordings with the mark *E* are discarded. This mark is used in cases where it is not possible to understand what the speaker says, where the speaker reads the wrong words, or where nothing is said. *E* is also used if the recording is truncated.

For every set of recordings, the corresponding annotation is collected in a Speech Logging File. This is a plain text file with the file extension *.spl. It contains comprehensive metainformation about recording equipment, the speaker, the manuscript and the individual files. An example of the type of information in such a file is given below.

Ex. of metainformation in an spl-file and information about the 4 first recordings (of 320)

```
[System]
Delimiter=>-<
Version=0001_1
CharacterSet=ANSI
ByteFormat=01
Script=463
Channels=2
Board=2;NI DSP2200
Frequency=16000
Coding=PCM;Linear
DOS Codepage=850
ANSI Codepage=1252
Memo=Kontor##1,5m, 2,5m, 1,5m, 2,5m##Shure bordmikrofon##Shur
[Info states]
1=Speaker ID>-<001>-<
2=Name>-<**** ****>-<
3=Age>-<57>-<
4=Sex>-<Female>-<
5=Region of Birth>-<Voss og omland>-<
6=Region of Youth>-<Voss og omland>-<
7=Remarks>-< frå hardanger>-<
[Session]
Directory=c:\adb_0463\data\scr0463\10\04631001\r4630001
Imported sheet file=c:\adb\dsdr\scripts\nor463\nor463.psh
Record session=1
Sheet number=1
RecDate=02 jul 1999
RecTime=08:52:13
Record duration=75' 36"
Number of recordings=312
[Record states]
1=2>-<>-<(...Vær stille under dette opptaket...)>-<1024>-<257024>-
<u0001001.wav>-<>-<1024>-<257024>-<bISa1>-<bISa1
2=2>-<>-<Tester en to tre fire fem seks sju åtte>-<1024>-<561024>-
<u0001002.wav>-<>-<257024>-<817024>-<tISa1>-<tISa1
3=2>-<>-<Blåbærturen ute på landet var en rein fornøyelse og flere av
turgåerene hadde kilosvis med bær.>-<1024>-<593024>-<u0001003.wav>-
<>-<817024>-<1409024>-<cISa1>-<cISa1
4=2>-<>-<Piloten hadde sitt svare strev med å få landet flyet i uvær
og svart natt.>-<1024>-<529024>-<u0001004.wav>-<>-<1409024>-
<1937024>-<cISa2>-<cISa2
5=2>-<>-<Det kan virke helt overveldende ute ved havet når den salte
skumsprøyten slår innover holmer og skjær.>-<1024>-<577024>-
<u0001005.wav>-<>-<1937024>-<2513024>-<cISa3>-<cISa3
```

**** marks anonymization made for this report

5. Dialect areas

This section contains an overview of the division into dialect areas in NST's acoustic databases. The data collection of all the subdatabases described above is based on this division. The speakers are in the 18-70 age range, and both genders are represented. No comprehensive statistics has been found of the distribution within each of these groups, but sporadic findings in archived reports suggest that the distribution is relatively even. This is confirmed by Kolbjørn Slethei at the Unit for Humanistic Informatics of the University of Bergen, who participated in working out the dialect division. In any case, the information about the distribution of informants can be extracted from the spl files.

The Norwegian database is divided into speakers from the following 11 dialect areas:

- Hedmark and Oppland
- The Oslo area
- Outer Oslofjord
- Southern Norway
- Southwestern Norway
- Bergen and outer parts of Western Norway
- The Voss area
- Sunnmøre
- Trøndelag
- Nordland
- Troms

There is no documentation which lists the criteria for this division, but Kolbjørn Slethei informs that it is mainly based on linguistic criteria and secondarily on statistical and socioeconomic conditions. An instruction from L&H stated that the maximal number of dialect areas should be five, and NST had to negotiate to be allowed to have a higher number of dialects and, consequently, a higher number of informants than was the case for similar data collections for other European languages.

6. Linguistic considerations

The following text contains comments on some linguistic considerations and choices which have been made during the work on the Norwegian part of the database.

For Norwegian (and Swedish), the number of dialect areas and informants is relatively high compared to similar data collection projects for other languages in L&H's portfolio. NST have had to justify this choice to their technology partners based on the large dialectal variation in Norwegian.

NST has intentionally chosen to focus on Norwegian Bokmål and exclude Norwegian Nynorsk. As a result, the recording manuscripts only contain Bokmål text. If a speaker reads a Bokmål word as Nynorsk, e.g. if they pronounce *skole*, 'school', as *skule* or *åttende*, '8th', as *åttande*, the orthography of the word is changed during validation so that it corresponds to the pronunciation. In such cases, the word is marked as Nynorsk in comments field of the annotation file. This is not considered an error, and such cases are evaluated according to the quality criterion *A*, provided that the orthographic form is compliant with the Nynorsk written norm. A similar strategy is chosen in cases where a dialectal form replaces a Bokmål form, such as when *venner*, 'friends', is pronounced *venna*. If the dialectal form is relatively common and does not deviate substantially from the written norm, quality criterion *B* is used.

A prerequisite for working with acoustic databases is a corresponding pronunciation lexicon with information about the orthography and pronunciation of the words. A description of NST's lexical databases can be found on Språkbanken's website, where it is also possible to download these databases. This lexicon reflects pronunciations that occur in the acoustic database. To a certain degree, variation in the acoustic database has been considered when creating the pronunciation lexicon. The lexicon contains pronunciation variants in cases where they occur naturally in the recordings and in speech in general. For example, *vende*, 'to turn', has two pronunciation variants: *[^{2}vɛ.nə]* and *[^{2}vɛn.də]*[4]. *Morgen*, 'morning', *måned*, 'month', *tredje*, 'third', *sytti*, 'seventy', are other examples of words with multiple pronunciations in the lexicon. The actual pronunciation of the speaker is not given in phonetic representation in the spl file, but it is given in the lexicon.

Phonetic reduction is handled in different ways depending on the context where it occurs. Some pronunciation variants will receive the quality mark *A*. This is the case when reduction almost always occurs, as in *meteorolog*, 'meteorologist', and *amerikaner*, 'American', pronounced as *[ˌmɛ.trʊ.ˈloːg]* and *[ˌɑm.rɪ.ˈkɑː.nər]* respectively. This is considered the correct pronunciation of these words, and in such cases, the orthophonic pronunciation will actually be considered overarticulated and consequently get the quality mark *B*. Quality mark *B* due to overarticulation is also used in cases of unnatural geminates (*telefonnummer*, 'telephone number', pronounced with two *n*-s) or neuter forms where the ending is pronounced as a plosive (*hodet*, 'the head', pronounced as *[^{2}hu:dət]*). In cases of non-obligatory, but common reductions, quality mark *B* may be used, e.g. *forutsette*, 'assume', pronounced *[ˈfɔr.ʉ.ˌsɛ.tə]* or *Sarpsborg* pronounced without *[p]*. The same is true, e.g., when *Hurdal* is pronounced *[ˈhʉ.ˌɖɑɳ]*, in which case also the orthography of the annotation is changed to *Hurdalen*, in accordance with the pronunciation.

In the case of rarer and unwanted reductions, quality marks *C* to *E* are used, depending on degree. As an example, the dialect forms *saukjan*, *akjan* and *nikkjan* of the numbers 17, 18 and 19 will receive the quality mark *E*, and will consequently be discarded.

Phonetic/allophonic variation occurs naturally in the data and is annotated according to similar quality criteria. In most cases, a set of variants will be accepted as compliant with quality mark *A*. This is the case, e.g., for variation such as */çiː.nʊ/* vs. */ʃiː.nʊ/* for *kino*, 'cinema', */ˈkɔɰʈ/* vs. */ˈkɔʈ/* for *kort*, 'short' and */ˈɔp/* vs. */ˈʊp/* for *opp*, 'short'. Normally, the variation will be represented as phonemically different variants in the pronunciation lexicon, although not in the case of *kino*, which is only represented by the pronunciation with *[ç]*. As a

[4] The superscript *2* denotes tone 2 [translator's comment]

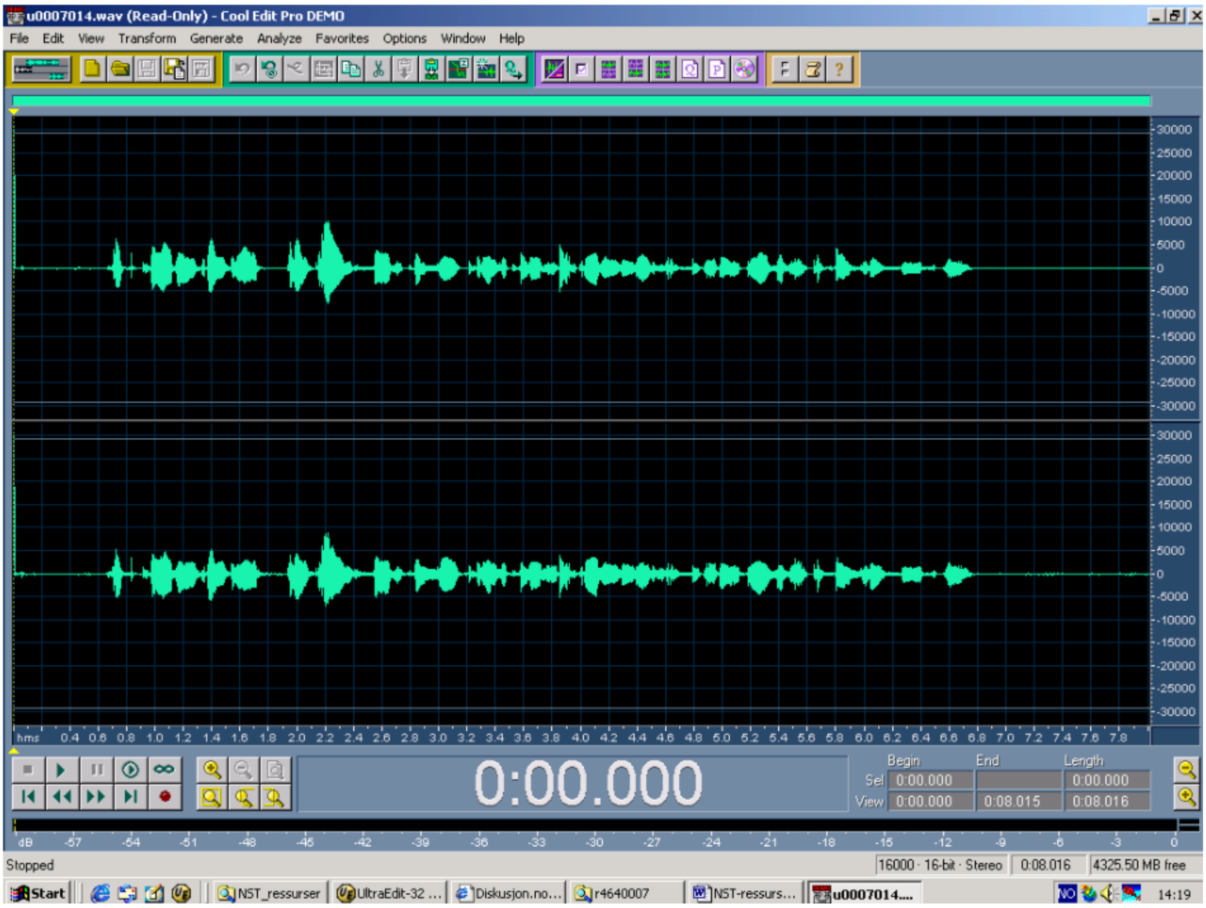
consequence of this, the two realizations will become allophonic variants of the phoneme *[ç]* in the language models.

In more dialect-specific cases, variants may be marked with other marks than *A*. For example, */b d g/* for *[p t k]* (so-called *soft consonants*) in Southern Norway are annotated with quality mark *B*, while */ʃ/* in the verbal form *ser*, ‘sees’, in Trøndelag dialects is not accepted, receiving quality mark *E*.

It is worth noting that the database is not specifically made to represent foreign accents, but there may be informants with a foreign origin in the data. The instructions in this area are not unambiguous, but it seems that pronunciations due to foreign accents are accepted as Norwegian pronunciation variants as long as they are part of a consistent pattern, for example */iː/* for *[yː]* (as in *lyd*, ‘sound’) or */uː/* for *[ʉː]* (as in *du*, ‘you’). Such pronunciations will, however, not receive the quality mark *A*.

7. Quality assessment

The spot check which was conducted as part of the present investigation suggests that the sound data are generally of high quality and validated in a thorough and precise manner. The sound quality is on the whole good, with minimal amounts of noise. The recordings are cut automatically by the recording software if the speech exceeds a certain sound level. In the spot check, no recordings were found where this occurs. This would be noticeable on a spectrogram, where automatically cut recordings would have a flat curve on the top instead of having natural peaks (cf. the figure below).


u0007014.wav (Read-Only) - Cool Edit Pro DEMO
File Edit View Transform Generate Analyze Favorites Options Window Help
30000 25000 20000 15000 10000 5000 0 -5000 -10000 -15000 -20000 -25000 -30000
30000 25000 20000 15000 10000 5000 0 -5000 -10000 -15000 -20000 -25000 -30000
hms 0.4 0.6 0.8 1.0 1.2 1.4 1.6 1.8 2.0 2.2 2.4 2.6 2.8 3.0 3.2 3.4 3.6 3.8 4.0 4.2 4.4 4.6 4.8 5.0 5.2 5.4 5.6 5.8 6.0 6.2 6.4 6.6 6.8 7.0 7.2 7.4 7.6 7.8
0:00.000
dB -57 -54 -51 -48 -45 -42 -39 -36 -33 -30 -27 -24 -21 -18 -15 -12 -9 -6 -3 0
Start NST_ressurser UltraEdit-32 ... Diskusjon.no... r4640007 NST-ressurs... u0007014....
14:19

	Begin	End	Length
Sel	0:00.000		0:00.000
View	0:00.000	0:08.015	0:08.016

Stopped 16000 · 16-bit · Stereo 0:08.016 4325.50 MB free

Example of stereo recording (8 kHz, office environment)

As mentioned above, a 100-200 millisecond gap was added between the beginning and the end of the speech signal and the beginning and the end of the recording. During the spot check, a few telecommunication recordings were found which were cut at the end, where the last part of the sentence was missing. The extent of this problem is unknown, but it seems to be small. From the office environment recordings, no cut recordings were found. In any case, such cut recordings will be removed by means of the annotation during validation.

The quality control of the data through the validation seems to have been conducted in a thorough manner. The validation appears to be consistent, which has to be due to good documentation and consistent supervision of the language assistants. The speech of the informants is distinctly annotated by means of markers, it can be distinguished clearly from unwanted sounds such as inhalation, exhalation, spit sounds and other kinds of noise. Due to this, it is possible to produce acoustic models which do not contain noise disturbances. As mentioned before, the validation contains detailed information about possible mispronunciations by the speakers.

It must be mentioned that creaky voice occurs in the data without it being explicitly marked or removed through annotation. This is, however, not necessarily a weakness, but can be included as part of the modelling as it also occurs in natural speech. During recordings for speech synthesis, segments with creaky voice are removed or re-recorded.

In addition to the comprehensive validation of the date, systematic spot checks were performed. first by NST's group leaders, then by L&H's personnel. Around 10% of the data are spot check controlled by NST, while 5 percent are spot check controlled in Belgium after delivery. This manual control therefore exceeds the 5 percent required by the ELRA standard. An internal meeting reports states the following concerning quality and spot checks:

N.N. explained a bit about the dictation evaluation. NST is praised for the amount of data we have, i.e. 900 recordings for which the validation is completed. In Belgium, they have spot-checked 150 dictation recordings, and these are of very high quality. Belgium has, therefore, decided to drop further internal spot checks, …

The resources are well documented, and the naming of files and directory structure is consistent.

The validation groups had frequent meetings and were supervised by group leaders with linguistic training. The validation work was coordinated between the three languages, and the language resources have been made in parallel with the valuation. From the frequent meetings within and across the language-specific groups, it is clear that NST has put a substantial effort into coordinating the work and develop a common standard for validation. Due to this, the validation has a high consistency.

It must be noted, however, that parts of the documentation only exit in Norwegian, which must be changed if the resources are to be part of an international resource database (cf. ELRA's specification).

Developing acoustic databases is a resource-demanding process, and NST has invested a lot of work in making acoustic databases consisting of a total of nearly 2 million recordings. Just the planning phase of such a project involves a substantial amount of tasks, such as mapping out the dialect areas, writing manuscripts, project management, hiring staff for the data collection, buying recording equipment, transporting equipment and staff to recording locations, recruiting informants, writing contracts, handling logistics etc. The work with the recordings comes in the next phase, and finally, validation.

A natural objection is that the division into dialect areas does not always follow dialect boundaries as they are documented in the literature. It seems like the areas for data collection have been chosen based on market concerns and practical concerns rather than on real dialectal differences. It is appropriate to ask why the Voss dialect is included, while other regions with distinct dialects such as Sogn, Sunnfjord etc. are not.[5] It is, however, difficult to evaluate what consequences this choice will have on the actual recognition rate, if any, and whether it is a meaningful goal to represent every single dialect in a complex region like the Nordic countries. It is more important to recognize that a broad geographic area is covered in the databases for the three languages.

[5] The NST office was located in Voss (translator's comment).

Another relevant objection, which only concerns Norwegian, is that Norwegian Nynorsk is not covered. Both main manuscripts only contain text in Norwegian Bokmål. The reason is, presumably, market considerations, but this can be problematic from a language policy point of view. The data collection does indeed cover a large geographic area which includes Nynorsk areas, but it may be necessary to supplement the data with Nynorsk-based recordings in the future. Besides, NST has chosen a reasonable handling of Nynorsk forms in the acoustic database.

Despite these objections, the conclusion is that NST's acoustic database is compliant with the current ELRA standard concerning its technical and linguistic quality, degree of validation, spot check control, consistent methods and documentation.

8. Acoustic databases for speech synthesis

The subdatabase described below is available for download here:

- http://www.nb.no/sbfil/talesyntese/no.ibm.talesyntese.tar.gz

NST has multiple times made recordings for the production of speech synthesis, firstly during the development of the L&H software RealSpeak for the Scandinavian languages, still available from the company Nuance, secondly for the development of IBM's speech synthesis, which didn't reach the marked before NST's bankruptcy. Both syntheses are concatinative systems. The former is a diphone synthesis, while the latter is a data-driven nit-selection synthesis. In addition to this, NST has developed IBM-based test systems for domain-specific synthesis, called Phrase Splicing, and made recordings for some specific customer applications.

8.1 RealSpeak

The recordings which were made for the development of RealSpeak were conducted entirely in L&H's recording studio in Ieper, Belgium. These recordings are not part of the bankruptcy estate of NST.

8.2 IBM's speech synthesis

During the development of IBM's speech synthesis, professional voices were recruited, i.e. one male voice per language. The recordings were made with IBM software in a recording studio in Voss, but proprietary software is not an obstacle to future use, as the recordings are available in the usable PCM format. These are the technical specifications for the three subdatabases:

Signal coding:	Linear PCM
File format:	headerless raw

Sampling frequency:	44 kHz
Resolution:	16 bit
Format:	Motorola PCM
Channels	2 (stereo) + laryngograph

The stereo recordings have the speech signal in one channel and the signal from a laryngograph in the second channel. The recording manuscripts are based on NST's corpus. An optimized set of sentences is produced by means of IBM's software OptScript. The manuscripts are phonemically optimized in order to obtain a broad coverage of possible diphone combinations. The manuscripts are not prosodically balanced, but they are nevertheless distributed over categories which give some prosodic variation, such as narrative sentences, wh-questions, yes/no-questions and listings. A smaller subset of the manuscripts contain sentences and number formats needed for the development of specialized speech applications in the banking domain.

The resources are distributed as follows:

Total number of recordings:	5363
Banking domain subset	417

8.3 IBM phrase splicing

In addition to the aforementioned recordings aimed at commercial applications, several datasets were recorded for the development of test and demonstration systems for IBM's software for phrase splicing, i.e. a system which is a hybrid of application-specific recordings and normal, concatenating speech synthesis. The systems were developed for use in the banking domain. These are the technical specifications:

Signal coding:	Linear PCM
File format:	headerless raw
Sampling frequency:	44 kHz
Resolution:	16 bit
Format:	Motorola PCM
Channels	2 (stereo) + laryngograph

The technical specifications are the same as for the data in 8.2. However, the scope and quality of the material is substantially different. Professional voices have not been used in these recordings. The informants are entirely recruited from NST's own staff. There are also fewer manuscripts and number of recordings than for the data sets above.

8.4 Quality assessment

The recordings aimed at the development of IBM's speech synthesis, mentioned in section 8.2, have been made in a sound studio using recording equipment and a laryngograph approved by IBM for developing professional speech synthesis systems. It is good to emphasize the high sampling frequency, the high quality of the recordings themselves, and the fact that the informants are professional actors.The recordings have been controlled by NST's speech synthesis product developers. These acoustic databases have been segmented and annotated using IBM's annotation tool. The material has first been annotated manually, and the annotation has subsequently been checked manually by the developers.

The material complies with current standards for acoustic resources and represents the state of the art for speech synthesis development. It is therefore recommended that it remains a part of a continuation of NST's acoustic resources. It would also be beneficial if the annotation giving the segmentation and the alignment of sound and text were made available. For this, a dedicated agreement with IBM would probably be necessary.

The recordings made for phrase splicing mentioned in section 8.3 have been made as part of the training of NST staff in developing speech synthesis technology at IBM's offices in Heidelberg, Hursley and Paris. The recordings are made in office environments that are sometimes noisy, not in a recording studio. Due to this, they are not very suitable for future use. Also, these databases do not have the same phonetic coverage as those described in section 8.2.