FonDat1: A Speech Synthesis Corpus for Norwegian

Ingunn Amdal and Torbjørn Svendsen

Department of Electronics and Telecommunications
Norwegian University of Science and Technology,
N-7491 Trondheim, Norway
{ingunn.amdal,torbjorn}@iet.ntnu.no

Abstract

This paper describes the Norwegian speech database FonDat1 designed for development and assessment of Norwegian unit selection speech synthesis. The quality of unit selection speech synthesis systems depends highly on the database used. The database should contain sufficient phonemic and prosodic coverage. High quality unit selection synthesis also requires that the database is annotated with accurate information about identity and position of the units. Traditionally this involves much manual work, either by hand labeling the entire database or by correcting automatic annotations. We are working on methods for a complete automation of the annotation process. To validate these methods a realistic unit selection synthesis database is needed. In addition to serve as a testbed for annotation tools and synthesis experiments, the process of producing the database using automatic methods is in itself an important result. FonDat1 contains studio recordings of approximately 2000 sentences read by two professional speakers, one male and one female. 10% of the database is manually annotated.

1. Introduction

State-of-the-art text-to-speech synthesis (TTS) systems are based on unit selection speech synthesis. This methodology relies on searching an annotated database of pre-recorded speech for the unit sequence which best matches a set of desired features, predicted by the TTS front-end. The quality is thus highly dependent on the database, which must be annotated with accurate information about identity and position of the units, traditionally time consuming manual process (Möbius, 2000). Because of the high development cost, only a limited number of voices are available.

The project Fonema[1] is a research project within the KUNSTI (Knowledge generation for Norwegian language technology) framework (Maegaard et al., 2006). The motivation for the Fonema project is a need to bring forth the necessary basis for state-of-the-art unit selection speech synthesis in Norwegian.

More automatic procedures in TTS development will make rapid and cost efficient deployment of new voices possible. The goal is to make the process as automatic as possible but still achieve good quality. One of the most important issues for the Fonema project is therefore high quality automation by developing tools for automatic phonemic annotation (labeling and segmentation) and prosodic labeling of speech corpora for use in unit selection synthesis systems. The tools need to be robust with respect to variations in voices as well as speaking styles.

The development of automatic annotation tools in the Fonema project requires evaluation experiments. For these experiments the only available Norwegian database has been ProsData, (Natvig and Heggtveit, 2000), which is a single speaker database with approximately 500 sentences. In the development process of the annotation tools, (Heggtveit and Natvig, 2004) and (Meen et al., 2005), there is a need for a larger speech corpus for verification and tuning. The database FonDat1 was collected to serve as a testbed for the annotation tool development. In addition, the database will be used to conduct experiments with unit selection synthesis, e.g. (Bjørkan et al., 2005). The database is a crucial part of a unit selection synthesis system, and the process of developing such a corpus is in itself an important part of the project. FonDat1 is the first of two databases planned within the project.

For FonDat1 we have chosen a standard approach based on speech community practice. Experiences from both the BITS (Ellbogen et al., 2004) and CGN (Schuurman et al., 2004) projects have been useful as well as the guide from Festvox (Black and Lenzo, 2003).

2. Speech Synthesis Corpus Development

2.1. Specification

Two of the main factors in database design are the content selection and the annotation of the recorded database (Black and Campbell, 1995). One of the first steps in database development is to choose between careful design of a manuscript for a smaller database or less careful design of a manuscript for a larger database which can be pruned after recording. A related approach to the latter is to use pre-recorded databases such as audio books.

Selecting the size of a unit selection database is a trade off between the desired coverage and the time and cost related to development, as well as search time and storage. The speech synthesis evaluation project *The Blizzard challenge*[2] used 1200 sentences in their competition databases, which is regarded as quite small.

Unit selection synthesis will copy the voice quality and speaking style of the "donor". Selection of speakers is therefore important. To make realistic sounding systems, speaker artifacts (like creaky voice) should be preserved. Unfortunately the automatic annotation methods used often fails when encountering such phenomena.

[1] http://www.iet.ntnu.no/projects/fonema/

[2] http://festvox.org/blizzard/

Speaker profiles	
Speakers	2 speakers: one male and one female
Language	Native Norwegian
Dialect	South-East Norwegian
Age	At least 21 years old
Occupation	Professional speakers e.g. actors
Contents	
Task	No task specified
Domain	Text taken from newspapers
Phonemic content	All Norwegian phonemes and realistic diphones present in the original text
Prosodic content	Include yes/no questions and multiple (2) clause sentences
Vocabulary	A lexicon of all the words in the database should be produced: - Orthographic form - POS code - Pronunciation in SAMPA format (including lexical word accents)
Speech material per speaker	- Approximately 2000 different sentences - 40 prosodically marked sentences
Speaking style	
Style	Read speech: An "expressive" speaking style is desired, in the sense that prosodic events should be clearly realized
Recording setup	
Acoustical environment	Studio recording
Script	Speakers reading prompts from a CRT display in their native language
Microphone	Desk mounted studio microphone
Technical specifications	
Sampling rate	16 kHz
Sample type	Linear, not compressed
Number of channels	Two channel recording: Speech and EGG
Signal file format	WAV
Annotation file format	Praat TextGrid files for a test set of 10%
Meta data file format	XML files with predicted phonemic and prosodic content for all files given in prosXML format (Natvig, 2003)
Lexicon format	Three-column plain text file: orthographic form, pronunciation and POS

Table 1: Corpus specification

The choices we have made for FonDat1 are summarized in the specification in Table 1.

2.2. Manuscript Selection

We have used texts from "The Oslo Corpus of Tagged Norwegian Texts"[3] provided by the Text Laboratory at the Faculty of Arts, University of Oslo, as a basis for the manuscript selection. This corpus contains news, novels and factual prose. The factual prose was excluded due to low readability (using standard ways of calculating readability based on sentence and word length). The news texts were chosen because there are lots available and the different sub-genres like feature articles, sport, and culture have their own characteristics.

The provided 32 MB of newspaper texts were first split into sentences. Several filtering steps using ad hoc rules (remove formatting, numbers, special characters etc.) were performed to clean up the text. In order to be regarded as "well formed", the sentences were required to have an inflected verb and words present in the "bokmål" version of the Norwegian computational lexicon NorKompLeks (Nordgård, 2000). (Norwegian has two written standards: "bokmål" and "nynorsk". The vast majority of speakers in the South-East dialect chosen for the database use the "bokmål" variant and this was therefore the natural choice for the manuscript.)

For readability we put several limits on the sentence length:

- Maximum word length: 15 characters
- Maximum sentence length: 100 characters
- Maximum number of words in a sentence: 28

The limits are important because longer sentences will be

[3] http://www.tekstlab.uio.no/norsk/bokmaal/english.html

preferred in the greedy search since they contain more phonemes. To elicit boundary tones we wanted to include two clause sentences even if they are longer.

The manuscript selection parameters were based on a statistical analysis of the candidate sentences left after the clean up (as well as the time and resources available). Phonemic and prosodic prediction was provided by the pre-processor of Telenor Talsmann ©. From a basis of 75,000 "well formed" sentences extracted, only diphone coverage was chosen as the selection criterion. A greedy search was performed to select 2092 sentences using a threshold of at least 6 instances of each diphone. The sentences were proofread before being used in the recordings.

10 sentences were selected to experiment with the possibility of using text formatting to guide prosodic realization. Each of these sentences were repeated in 4 versions (in addition to the "non-formatted") with different word(s) emphasized using capital letters.

2.3. Speaker Selection and Recording

The best way to select speakers is to test how good the resulting synthesis will be. Screening of speaker candidates by letting them read a smaller text and synthesize utterances for listening tests is therefore usual. We had no automatic Norwegian unit selection synthesis available and a manual process was deemed to costly for this first database. We therefore chose to use audio book actors as speakers as they already are "screened" for a different, yet similar task. The actors, one male and one female speaker, were chosen in cooperation with *Lydbokforlaget*[4] mostly based on their ability to read consistently and accurately. Both actors read the same manuscript of 2132 sentences giving a total of 4264 sentences in the database.

A sound laboratory at NTNU was used for the recordings. The studio contained a sound-proofed recording room and a control room for supervising the recording process, see Figure 1. The recording setup was made by one of the project members with a control program implemented as a Perl script. The script utilizes Praat[5] for display and playback of the recorded speech and Total Recorder v. 4.4 from High Criteria[6] for data capture.

A desk mounted high quality microphone and the EGG in the recording studio was connected to a DAT sampling the stereo signal (one channel each for speech and EGG) at 48 kHz. The DAT signal was transmitted by optical fiber to a sound unit, re-sampling the signal to the resulting 16 kHz sampling frequency. Details are given in Table 2

The instructions for reading were to use "normalized pronunciation" and a distinct way of speaking without overarticulating. Which pronunciation to use was not always self-evident (e.g. mnemonics, abbreviations, and numerals), and the two speakers made different choices for some words. The manuscript was read one sentence at the time.

The manuscript was given in a text file, where each line contained an ASCII identifier and the orthographic text of one sentence. The ASCII identifier was used as the base filename, and the speech, EGG, and text files were given the extensions .wav, .lar and .txt respectively.

Equipment	Manufacturer
Microphone	Milab LSR 1000
EGG	Laryngograph Ltd
DAT	Fostex D10 Digital Master Recorder
Sound unit	Creative Studios Sound Blaster Live 5.1 Platinum

Table 2: Recording equipment

Prompts were communicated to a separate PC with two displays (one in the control room and one in the recording room). The sentence length restriction ensured that each sentence only needed one line (Figure 1 is misleading in this aspect). The color of the text would change to show recording cues. The recording supervisor had visual and auditive control of the recorded speech and could accept or reject recordings, i.e. initiate re-recordings or move on to the next sentence in the manuscript. The supervisor monitored both noise, truncations as well as pronunciation and mis-readings.

The recordings were done in 2-4 hour sessions with one break every hour. The female speaker needed shorter sessions as the EGG necklace was annoying. The total time spent in studio was about 20 hours per speaker and the amount of recorded speech was 4–5 hours per speaker including rather long silence segments, cf. section3.1..

2.4. Manual Annotation

A test set defined to be approximately 10% of the corpus, i.e. 200 sentences per speaker, was manually annotated using Praat. The annotation of ProsData (Natvig, 2000) and the CGN[7] and BITS[8] projects was used as a basis for the annotation specification. The annotation was performed in several steps, with a mix of automatic and manual steps:

1. Manuscript conversion to a format suitable for automatic transcription
2. Automatic phonemic segmentation based on speech recognition
3. Manual correction of phonemic annotation
4. Manual prosodic labeling

The phoneme prediction in step 1 was provided by Telenor Talsmann ©. The speech segmentation in step 2 was chosen to be rather rudimentary to avoid bias in the manual annotation towards the system used on phonemic annotation experiments (Meen et al., 2005). Steps 3 and 4 were performed in separate sessions.

Two phonetics students performed the manual annotation. They were given an initial instruction course, which included annotating 10 sentences that were corrected by the course leaders to ensure a common annotation practice. For the rest of the sentences the annotators corrected each

[4] http://www.lydbokforlaget.no/

[5] http://www.praat.org/

[6] http://www.highcriteria.com/

[7] http://lands.let.kun.nl/cgn/ehome.htm

[8] http://www.phonetik.uni-muenchen.de/Forschung/BITS/

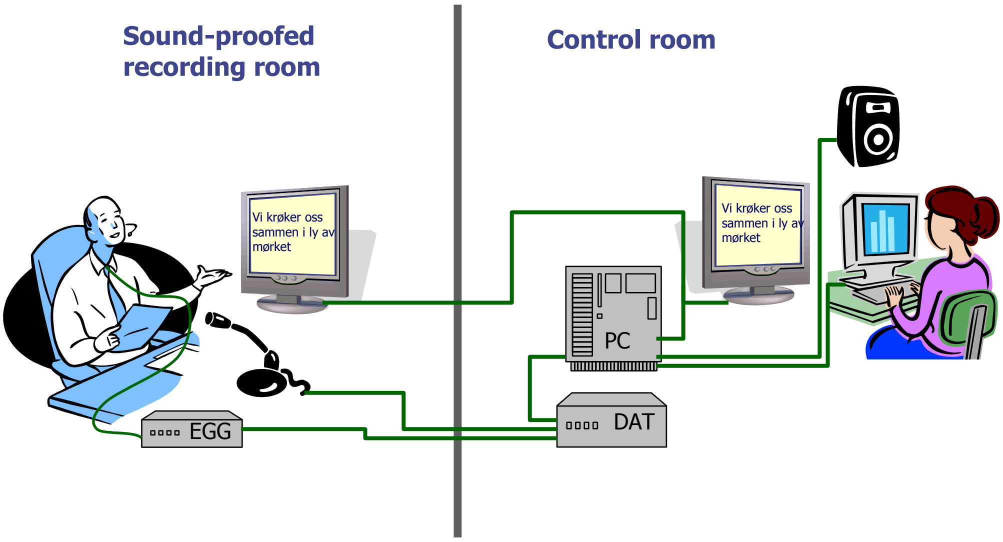

Sound-proofed recording room
Control room
Vi krøker oss sammen i ly av mørket
Vi krøker oss sammen i ly av mørket
PC
DAT
EGG

Figure 1: Recording setup.

other's work, the original annotator being responsible for the final annotation decision. A snapshot of the phonemic annotation setup is shown in Figure 2. The annotation was done during 10 weeks and took a total of 400 hours.

2.4.1. Phonemic Annotation

The main principle is that the manual phonemic annotation should reflect the perceived phonemic content (in contrast to phonotypical transcription). Both labels and timing of the automatic proposal should be corrected. The phoneme symbol set consisted of the Norwegian SAMPA[9] augmented with English phonemes (/aU/, /@U/, /dZ/, /T/, /D/, /z/, /Z/, and /w/) and pause labels (silence, breath sound, filled pause, and epenthetic pause).
4 tiers were presented to the annotators:

- Sentence tier
- Word tier
- Phonemic annotation tier with automatic proposal
- Phoneme segment comment tier

In addition the speech waveform and spectrogram were displayed. The annotators were free to use F0 and/or formant estimates. The timing of the word tier and phoneme comment tier should be changed to be in agreement with the phoneme tier. In the control phase a fifth tier was added for correction comments, this tier was removed when finalizing the sentence.
There will always be comments to the annotation and keeping them in a searchable format will increase the usability of them. The phoneme comment tier had a set of predefined codes for some common sources of uncertainty (uncertain phoneme identity, uncertain segmentation, voiced/unvoiced phoneme with unvoiced/voiced region etc.). In addition, each annotator wrote a free format log for other comments.

Prominence	**Label**
Unaccented	-
Accented	+
Focal accent	++
Word boundaries	**Label**
Normal boundary	-
Marked utterance internal boundary	+
Utterance boundary	++

Table 3: Prosody labels

2.4.2. Prosodic Labeling

For the prosodic labeling, we used 3 levels of prominence and 3 levels of word boundaries, see Table 3. The annotators should only add labels, and not change timing at this step. In addition non-normalized stress or word tone and erroneous splitting of word compounds should be marked using pre-defined codes. No automatic suggestion for prosodic labeling were given to the annotators. 5 tiers were presented to the annotators:

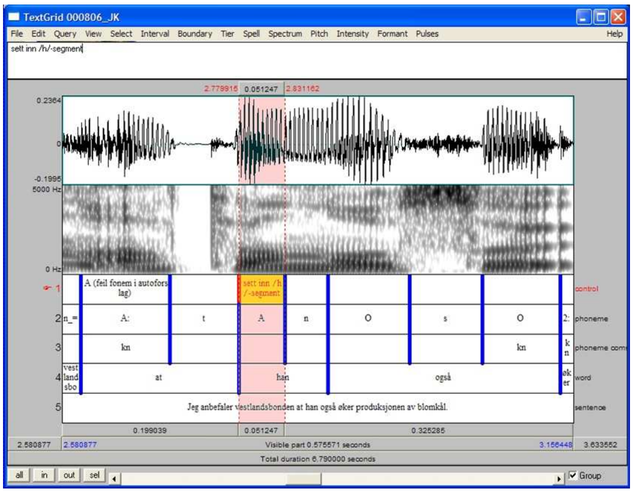
- Sentence tier
- Word tier
- Prominence tier (interval tier)
- Word boundary tier (point tier)
- Word segment comment tier

3. Post-processing

3.1. Speech Files

Due to variable network latency and a small variation in the start-up time of the recording program, it was occasionally problematic for the speakers to ”hit” the recording time window. The recording window was calculated from

[9] http://www.phon.ucl.ac.uk/home/sampa/norweg.htm


TextGrid 000806_JK
File Edit Query View Select Interval Boundary Tier Spell Spectrum Pitch Intensity Formant Pulses Help
2.779915 0.051247 2.831162
0.2364
0
-0.1995
5000 Hz
0 Hz
1 A (feil fonem i autofors lag) sett inn /h /-segment control
2 n_= A: t A n O s O 2: phoneme
3 kn kn k n phoneme com
4 vest land sbo at han også øk er word
5 sentence
0.199039 0.051247 0.325285
2.580877 2.580877 Visible part 0.575571 seconds 3.156448 3.633552
Total duration 6.790000 seconds
all in out sel Group

sett inn /h/-segment

Jeg anbefaler vestlandsbonden at han også øker produksjonen av blomkål.

Figure 2: Snapshot of phonemic annotation setup. The highlighted segment shows a comment from the controller to add an /h/ in the word "han".

the number of phones plus a buffer and had to be set quite generous to avoid truncation. Speaker specific window calculation parameters were needed as the two speakers had quite different speaking rates.

The utterance files therefore include long leading and trailing silence parts. Using the speech segmentation system these segments were identified. Silence segments in the beginning and end were set to a maximum of 300 ms. Silence segments inside sentences where left unchanged (these are up to 1 sec long). All discarded segments were "proof listened" to avoid erroneous segmentation. For the male speaker 1.8 hours of silence were removed resulting in 3.1 hours of speech. For the female speaker 1.1 hours of silence were removed resulting in 2.7 hours of speech.

A label for signal truncation was added to the symbol set used for phonemic annotation as two of the manually annotated files (i.e. 0.5%) were truncated (both for the male speaker).

3.2. Transcription and Phoneme Prediction

There are several reasons for deviations between the phoneme prediction and the actual pronunciation:

- Transcription errors
- Lexicon errors
- Parsing errors (POS tagging errors)
- Reading errors

The starting point for the annotation is an orthographic transcription whose quality depends on the text clean-up. These steps are usually performed using ad hoc rules and may introduce (or fail to correct) errors that affect the quality of the resulting database. About 4% error rate in transcriptions is reported in (Huang et al., 1996). Careful monitoring during the recording phase can reduce the number of errors, but the speaker may also introduce new errors when requested to repeat utterances. We will always encounter divergence between what is predicted from the manuscript and what is actually said (Saikachi, 2003), and manual corrections have been inevitable.

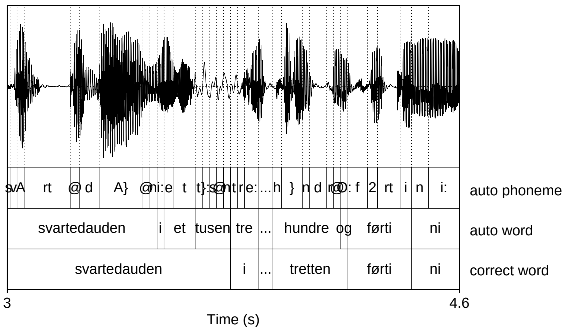


Figure 3: Mismatch between number pronunciation and prediction for the phrase "Black death in 1349". Automatic word alignment in bottom tier, corrected word sequence in second tier and automatic phoneme alignment in top tier.

Numeral expressions are for example notoriously difficult to predict. An example is shown in Figure 3 of a Norwegian sentence containing the phrase "...svartedauden i 1349..." ("... the Black death in 1349..."). The text normalization fails to predict 1349 as a year and suggests "one thousand three hundred and forty nine" instead of "thirteen forty nine" which is spoken, causing a severe misalignment of the phone and word positions.

419 sentences contained numerals and acronyms (all capital letters) and were manually checked on an orthographic level for both speakers. About 10% of these 419 sentences contained errors in phoneme prediction.

Some "nynorsk" sentences had slipped through the manuscript filtering (cf. the manuscript selection in section 2.2.). They were identified during recording. A check on the phoneme prediction for these sentences revealed that several of them were predicted wrongly. These sentences were put on a "blacklist". Some sentences contained words not in lexicon (due to proofreading corrections). They were also put on the blacklist to be able to ignore them when building the TTS system.

3.2.1. Automatic Phoneme Transcription Verification

The project has produced a new annotation assessment method using log likelihood ratio based utterance verification on the recorded database. The utterance verification is applied to detect utterances where there is a likely mismatch between the predicted pronunciation and what is actually spoken, or where an automated procedure for phonemic labeling misaligns the phone labels and the acoustic content. Further details were presented in (Amdal and Svendsen, 2005).

3.3. Manual Phonemic Annotation

The phonemic annotation was controlled by letting the two annotators correct each other while keeping one annotator in charge for each sentence. A sample control revealed several suspicious annotations. A manual control of the annotation was therefore performed, checking all instances of deviation between the automatic and manual annotation for

the two speakers for the same word. Corrections were documented by using CVS on the annotation files.
Epenthetic sounds were difficult to label consistently. Especially the male speaker used epenthetic sounds to mark word boundaries. We therefore decided to add two labels, one for word internal anaptyxis and one for anaptyxes between words.

4. Discussion

The database is already in use in the Fonema project. First of all we have a running unit selection system based on Festival using an entirely automatic annotation. With this system we are able to perform experiments on unit selection synthesis using listening tests. We are also using FonDat1 for HMM-synthesis experiments.
With FonDat1 the project has been through the entire process for an automatic unit selection synthesis development. We are now able to test the various tools made in the project and can focus our work to the most critical parts. The corpus development process will be put to the test when we record the planned production database.

5. Conclusions

The purpose of FonDat1 is mainly to serve as a reference database for the annotation tools developed in the project. An additional purpose is to conduct initial experiments with unit selection synthesis. Experiences from designing, producing and using the database will be exploited for a forthcoming production database. FonDat1 is planned to be made available for non-commercial use through a Norwegian language resources project *Norsk språkbank*.

6. Further Work

The project plans to record a production database during 2006 based on the lessons learned from FonDat1:

- Have an audition of voice talents by building a limited TTS system using our fully automated process
- Use a stricter text normalization discarding problematic sentences
- Synthesize prompts in advance to check phoneme prediction
- Present expected pronunciation to reader, either by synthesized prompt or in text
- Consider continuous reading of paragraphs rather than sentence by sentence.
- Use a head mounted microphone to control volume

Our prosody model is based on South-East Norwegian and this is the main reason for working with only this dialect. We would like to make tools for other dialects, but this requires more knowledge on Norwegian intonation.

7. Acknowledgments

The authors wish to thank project members Jon Emil Natvig and Per Olav Heggtveit at Telenor R&D for their substantial contribution in the making of FonDat1. This work has been funded by the Research Council of Norway as a part of the KUNSTI program.

8. References

Amdal, I. and Svendsen, T. (2005). Unit selection synthesis database development using utterance verification. In *Proc. Eurospeech 2005*, pages 2553–2556, Lisboa, Portugal.

Bjørkan, I., Svendsen, T., and Farner, S. (2005). Comparing spectral distance measures for join cost optimization in concatenative speech synthesis. In *Proc. Eurospeech 2005*, pages 2577–2580, Lisboa, Portugal.

Black, A. W. and Campbell, N. (1995). Optimising selection of units from speech databases for concatenative synthesis. In *Eurospeech95*, Madrid, Spain.

Black, A. W. and Lenzo, K. A. (2003). *Building Synthetic Voices*. [online description]. [cited 2006-02-09]. URL: http://www.festvox.org/bsv/.

Ellbogen, T., Schiel, F., and Steffen, A. (2004). The BITS speech synthesis corpus for German. In *Proc. LREC 2004*, pages 2091–2094, Lisboa, Portugal.

Heggtveit, P. O. and Natvig, J. E. (2004). Automatic prosody labelling of read Norwegian. In *Proc. ICSLP 2004*, Jeju island, Korea.

Huang, X., Acero, A., Adcock, J., Hon, H.-W., Goldsmith, J., Liu, J., and Plumpe, M. (1996). Whistler: A trainable text-to-speech system. In *Proc. ICSLP 1996*, Philadelphia (PA), USA.

Maegaard, B., Fenstad, J.-E., Ahrenberg, L., Kvale, K., Mühlenbock, K., and Heid, B.-E. (2006). KUNSTI – Knowledge generation for Norwegian language technology. In *Proc. LREC 2006*, Genoa, Italy.

Meen, D., Svendsen, T., and Natvig, J. E. (2005). Improving phone label alignment by utilizing voicing information. In *Proc. SPECOM 2005*, pages 683–686, Patras, Greece.

Möbius, B. (2000). Corpus-based speech synthesis: Methods and Challenges. Technical report, University of Stuttgart, AIMS 6 (4).

Natvig, J. E. (2003). prosXML – en datamodell for Fonemadata. Technical report, Fonema working paper 2. in Norwegian.

Natvig, J. E. and Heggtveit, P. O. (2000). PROSDATA version 2.0. A speech database for studies of norwegian prosody. Technical report, Telenor R&D, N 20/2000.

Nordgård, T. (2000). NorKompLeks: A Norwegian computational lexicon. In *Proc. COMLEX 2000*, pages 89–92, Patras, Greece.

Saikachi, Y. (2003). Building a unit selection voice for Festival. Master's thesis, University of Edinburgh.

Schuurman, I., Goedertier, W., Hoekstra, H., Oostdijk, N., Piepenbrock, R., and Schouppe, M. (2004). Linguistic annotation of the Spoken Dutch Corpus: If we had to do it all over again ... In *Proc. LREC 2004*, pages 57–60, Lisboa, Portugal.