ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2004 - ICSLP
8th International Conference on Spoken
Language Processing
ICC Jeju, Jeju Island, Korea
October 4-8, 2004

Automatic Prosody Labelling of read Norwegian

Per Olav Heggtveit, Jon Emil Natvig

Telenor Research and Development
Snarøyveien 30, N-1330 Fornebu Norway
per-olav.heggtveit@telenor.com
jon-emil.natvig@telenor.com

10.21437/Interspeech.2004-662

Abstract

In this paper we present initial work on a method for automatic stress and boundary labelling of read East-Norwegian. The context of this work is automatic corpus annotation for unit selection speech synthesis.

A phonological model of Norwegian prosody is described. The identification of syllable stress and major intonational boundaries are key prosodic events for building a prosodic description of a Norwegian utterance according to this model.

A CART based method for automatic classification of syllable stress is presented. Initial experiments show that the method is capable of classifying syllables as unaccented or accented with high accuracy. 92.1 % of the unaccented syllables and 90.4 % of the accented syllables were correctly classified.

1 Introduction

State of the art text to speech synthesis relies on the concatenation of pre-recorded units selected from a large database of annotated speech. Unit selection speech synthesis can be split into two main steps:

1. Text analysis giving a phonological description of the input text, consisting of the phoneme sequences augmented with prosodic information.
2. Selection and concatenation of speech segments. The task is to search a large database of natural speech for occurrences of candidate speech segments according to the specification from the first step. Given a set of alternative segments, the concatenation module selects an optimal sequence according to a set of cost functions to produce the best possible synthesis of the input text.

Prior to the runtime operation of the synthesis system, a database must be recorded and annotated with segmental and prosodic features used in the selection process.

In our study we explore an approach to prosodic annotation similar to the one presented in [1]. Stored speech units are classified using a small set of prosodic categories. These categories define the basic intonation units in a phonological description of Norwegian as described in Section 2. A main advantage of this approach is that there is less need to produce detailed information of the realisations at the acoustic level in terms of fundamental frequency (F0) and segment durations.

The challenge is to be able to retrieve suitable units from the database at synthesis runtime, using only high-level information. Compared to traditional speech synthesis, the problem is shifted to the definition of descriptive tags that ensure the selection of the best suitable acoustic units in the required context. Though simpler at first glance, this task needs to be based on knowledge about the factors that are believed to influence the segmental and prosodic properties of the units at the acoustic level.

Manual labelling of complex prosodic events is problematic due to low transcriber agreement [2]. In this paper we describe an initial experiment with an automatic method for identification of two important prosodic events in read Norwegian: word and phrase accents.

We are following earlier studies on automatic prosody labelling in applying feature extraction and classification [1,3,4,5,6]. As in [6] we chose to approach the task as a static classification problem. We use syllable based prosodic feature extraction, and Classification and Regression trees (CART) for classifications of prosodic events.

2 Norwegian Prosody

This section describes a phonological model with factors that are believed to have an effect on the realisation of prosody on the acoustic level.

2.1 Intonation

Our description of Norwegian intonation is based on the *Trondheim Model* (TM) as described in [7]. Its terminology is compatible with most varieties of Norwegian. In a previous work, [8] we showed that this model is a good basis for synthesizing Norwegian prosody.

The Accent Phrase (AP) is the basic unit for analysing the different intonation patterns and is the domain of the lexically based *tonal accent* (word accent) and the *phrase accent.* An AP always starts with a primary stressed (accented) syllable, and most often stretches up to the last syllable before the next primary stressed syllable.

Given the primary stressed syllables in an utterance, the APs can be built algorithmically by starting with an accented syllable and include the syllables up to the next accented syllable or a major boundary. A phrase accent with no following accented syllable signals a major boundary. It should also be noted that the AP boundaries do not always coincide with the word boundaries.

2.1.1 The phrase accent

In East Norwegian the phrase accent is a tonal rise (H) [9] and it is linked to the last syllable in the AP [7]. The tonal rise of the phrase accent signals its degree of prominence. The most prominent AP in an utterance typically has a higher pitch level at the right edge than the less prominent APs. The most prominent AP is termed *focal* and the other APs are termed *non-focal*.

2.1.2 *The tonal accent*

The melodies of tonal accents are (including the phrase accent H): LH for accent 1 and HLH for accent 2, where L and H represents lows (L) and highs (H) in the pitch curve as shown schematically in Figure 1. Basically, in an utterance context, the two tonal accent patterns differ in the timing of the HL transition over the accented syllable.

The AP can include more than one word, but the tonal accent assigned to an AP is usually decided by the lexical tone of the first word in the AP.

2.1.3 *Boundary tones*

The utterance level is associated with boundary tones L% and H%. The symbol L% indicates a low boundary tone and a declarative utterance, whereas the symbol H% indicates a high boundary tone and an interrogative utterance. This is the case for East Norwegian dialects.

Utterance boundaries are almost always associated with pauses. Utterance internal AP boundaries and in particular focal AP boundaries can also be associated with pauses. We include pauses as a phonetic context and mark them as separate segments.

2.1.4 *Declination*

A common property of East Norwegian intonation is the pitch declination in APs after the last focal AP in an utterance (post-focal APs). It is important to be able to identify units that are subjected to this tonal effect to help selecting appropriate units at synthesis time.

2.2 Duration

The segment duration varies depending on a number of linguistic and non-linguistic factors. Fortunately, many of these happen to coincide with intonation factors, such as accentuation, prominence and position. Some of the contextual factors reported in the literature are:

- *Phrasal position.* At the word and phrase levels, the segment duration is found to vary depending on the position in the phrase. The greatest positional effect is the phenomenon of final lengthening that appears to be of considerable generality as a phonetic phenomenon. A typical encoding for this purpose seems to be initial (possible shortening), medial (unaffected) and final or pre-pausal (lengthening expected).
- *Stress.* Duration is the most reliable correlate of stress [10]. Stressed syllables have typically longer durations. In most analyses of Norwegian stress, three stress levels are assumed: Primary stress (always associated with tonal accentuation), secondary stress and unstressed.
- *Focus.* Focusing a word increases its duration.
- *Word length.* There is a tendency that syllables in longer words are pronounced more rapidly than in shorter words under the same condition.

3 Method

On the basis of the prosodic model description given above, we can conclude that accented syllables and major intonational boundaries are key prosodic events for building a phonological description of a Norwegian utterance.

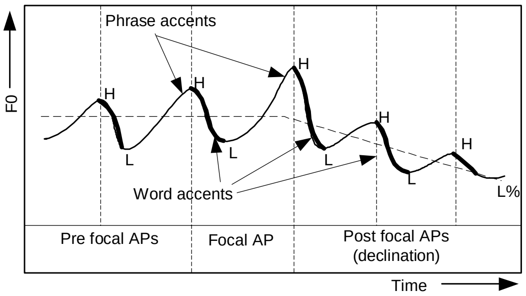

Figure 1. Stylised Norwegian intonation according to the trondheim Model.

We propose a four-step procedure for an automatic prosody labeller for Norwegian:

1) Detect accented syllables
2) Detect major boundaries (phrase accents not coinciding with AP-AP boundaries)
3) Construct the AP sequence using accented syllables and boundaries as delimiters
4) Analyse the phrase accents to classify APs in pre-focal and post-focal (and downstepped) positions.

The context of this work is prosodic annotation of speech corpora for unit selection speech synthesis. For this purpose and context we can make some helpful assumptions compared to the general approach: We assume that a professional speaker will be able to follow instructions concerning reading style, dialect, pronunciations and prosody. We expect the speaker to read using "normal" East-Norwegian pronunciation and prosody. In this way we can assume that the text is read according to the lexical information available: Only lexically defined stress syllables may be accented, word accent type (1 or 2) will be realised in accordance with lexical information, utterance boundaries will consistently be realised as L% for declarative utterances and H% for interrogative utterances. We also assume that sentences will be read one by one.

3.1 Prosodic events

Stress
As described in Section 2, the position of accented syllables is the key to the prosodic structure of a Norwegian utterance. These events define the start of Accent Phrases. Consequently, we chose as our primary task to predict stress on the syllable level and to assign to each syllable one of the following attributes:

u	*Unstressed*
s	*Secondary Stress*
a	*Accented/Primary stress*

Boundaries
The speech database is read sentence by sentence, and consequently the major boundaries are known.

The task remaining is to detect any utterance internal boundaries not coinciding with an accented syllable, i.e. where an AP would end prior to the next accented syllable. This is the case where an utterance internal AP boundary is tonally marked, but is not ending in the next accented syllable.

3.2 Features

Prosodic input features can be extracted from the speech by acoustic analysis. Linguistic features are extracted from the text by lexical lookup and analysis.

The main acoustic cue for an accented syllable in an utterance context in Norwegian is the HL transition in F0. In addition we expect syllable duration (lengthening) to be helpful.

Normalised F0 features

For F0 features, two kinds of normalisation were applied:

1) Utterance normalisation. F0 values are normalised with respect to the average F0 taken over each utterance.
2) Local normalisation. As discussed in Section 2 and illustrated in Figure 1, post focal APs are subjected to declination, a gradual reduction in F0 levels and amplitudes towards the end of the utterance. This could confuse accent prediction if based on absolute F0 measures [11]. To compensate for this effect, we introduced a normalised F0 defined as the ratio of F0 to the local average over the current, previous and next syllable.

Several features were designed to reflect the F0 trajectory in the vincinity of each syllable (see Section 5 below).

Duration features

All syllable duration values were normalised using z-score values [12] for each phoneme instance *i*: $z_{ij} = (dur_{ij} - \mu_j) / \sigma_j$ where μ_j is the mean duration and σ_j the standard deviation of phoneme j. The syllable z-score duration is computed as the sum of the phoneme z-scores divided by the number of phonemes.

Energy features

The energy feature is the syllable nucleus RMS energy normalised by the mean energy of the nuclei of the complete database: $\hat{E}_i = E_i / \mu$ where μ is the mean energy of all syllable nuclei.

Linguistic/lexical features

- Lexical features included are:
- Word final (0, 1)
- Lexical stress (unstressed, stressed, accent 1, accent 2)
- Punctuation (for boundary detection)

4 The Prosdata Corpus

The speech database PROSDATA [13] used for these experiments is a collection of 502 Norwegian sentences read aloud by a female speaker. The recording was made in a studio, using a sampling rate of 16 kHz. The data have been manually segmented in terms of phonemes, syllables and words. The sentences are read in an informative news style, with a pleasant but neutral reading style.

For the purpose of detecting prosodic events, this database is not ideal. A more lively and prosodically distinct reading could make the task easier.

The database contains the following data:

- Lexicon with a phonotypical pronunciaton of each word, lexical primary and secondary stress syllables.
- F0 computed with ESPS/waves and smoothed by median filtering. F0 is interpolated over unvoiced regions.

The following manually determined prosodic labels are available:

- Accented syllables
- Prominence on the word level
- Boundary levels (break index)

5 Experiments

We applied Classification And Regression Trees (CART). We trained a static classifier where the CART tree is classifying the prosodic categories directly without any language model to limit the possible sequence of labels.

We used the WAGON software from Edinburgh Speech Tools [14]. The CART models were trained stepwise i.e. features are included one by one, and the feature that improves the classification most is included at each step until no further significant improvement is achieved.

80% of the Prosdata database was used for training, 10% for testing during the CART training process, and the last 10% (approx. 1500 syllables) was used in an independent test.

5.1 Stress labelling

A number of experiments were carried out. Table 1 summarizes the best results and the features that were selected by the CART stepwise training process in each case.

	F0 normalisation	
	Utterance mean	**Local mean (Three syllables)**
Selected features	- Lexical stress - Syllable duration - F0 mean current - F0 range previous - Duration difference[1)]	- Lexical stress - Syllable duration - F0 range next - F0 mean current - F0 mean prev
unaccented	92.7%	92.1%
accented	88.1%	90.4%

Table 1 *Results from experiments with two different F0 normalisations. Percentage of labels correctly classified.*

1) Duration difference: duration z-score difference between syllable rhyme and onset.

The best result was achieved with two levels of syllable stress, unaccented (u) and accented (a), and a local 3 syllable F0 normalisation. 92.1% of the unaccented syllables and 90.4% of the accented syllables were correctly classified.

The local three-syllable F0 normalization achieves better results in classifying accented syllables than the normalisation by the utterance F0 mean. This result seems to confirm the assumption that local normalization compensates for the F0 declination in post focal accents. The utterance mean normalization only compensates for F0 variations between utterances.

A model with three levels of syllable stress, unstressed (u), secondary stressed (s) and accented (a), did not perform well. Only 25.4% of the secondary stressed syllables in the test set were correctly classified, while the correct classification of accented syllables was 82,7%. The reason for this poor classification could be a general problem of inconsistent manual labelling of stress in the database (secondary stressed and deaccented syllables).

5.2 Boundary labelling

Initial experiments on a model for detecting utterance internal boundaries not coinciding with AP boundaries showed that 58% of these boundaries were linked to commas. The other boundaries occur at various syntactical boundaries, but the syntactical phrase and clause boundaries alone are not sufficient to reliably detect the remaining boundaries. Features combining syntactical boundaries and acoustic prosodic features will be tested in a continuation of this work.

6 Conclusions and Further work

In this paper we have proposed and tested an automatic prosody labelling procedure for accented syllables in read Norwegian. In experiments on detecting accented versus unaccented syllables we achieved above 90% correct classification. This result is encouraging taking into account that the speech material available for training and testing represented a relatively indistinct reading style and that better results could be expected in a corpus with a natural and "lively" reading style with more pronounced prosodic events.

The work reported here is part of an ongoing project focusing on automatic tools for preparation of speech corpora for unit selection synthesis taking into account both segmental and prosodic information. In a continuation of this work we will establish a more suitable speech database for these studies, including both male and female talkers. The method reported here will be tested and further developed on this material, e.g. to take into account cross-speaker variations. In addition to this, we will extend our methodology for detection of secondary stress, prosodic boundaries and prominence.

7 Acknowledgements

This work has been carried out under the FONEMA project "Metodeutvikling for naturtro norsk talesyntese" ("Methods for realistic speech synthesis in Norwegian" [15]. The project is funded by the Norwegian council for Research and is cooperation between Telenor R&D and Norwegian University of science and technology (NTNU).

8 References

[1] Wightman, C. W. et al., "Perceptually Based Automatic Prosody Labeling and Prosodically Enriched Unit Selection Improve Concatenative Text-To-Speech Synthesis." In *Proc. ICSLP 2000*, Beijing, China.

[2] Syrdal, A.K. and McGory, J., "Inter-transcriber reliability of ToBI prosodic labelling." In *Proc ICSLP,* Bejing, 2000

[3] Wightman, C. W., Ostendorf M., "Automatic Labeling of Prosodic Patterns" In *IEEE Trans. Speech and Audio Processing, vol 2*, 4:469-481 (1994)

[4] Batliner, A. et al., "Boiling down Prosody for the Classification of Boundaries and Accents in German and English." In *Proc. Eurospeech 2001*, Aalborg, Denmark

[5] Shriberg, E. et al, "Prosody-based automatic segmentation of speech into sentences and topics." In *Speech Communication* 2000 32:127-154.

[6] Vereecken, H, et al, "Automatic Prosodic Labeling of 6 Languages", In *Proc ICSLP* 1998, Sidney Australia

[7] Kristoffersen, G., *The Phonology of Norwegian*, Oxford University Press (2000)

[8] Natvig, J.E., Heggtveit, P.O., "Intonation Modelling with a Lexicon Of Natural F0 Contours", In *Proc. Eurospeech 2001*, Aalborg, Denmark

[9] Silverman, K et al, "TOBI: A standard for labelling English prosody". In *Proc. ICSLP 92* pp 867-870, Banff, Alberta, Canada

[10] Fant G., Kruckenberg, A., "Towards an integrated view of stress correlates.*",* In *Proc ESCA Workshop on Prosody*, 1993, Lund, Sweden, pp 41-45.

[11] Chen, et al, "A Maximum Likelihood Prosody Recognizer" In *Proc. ISCA International Conference on Speech Prosody 2004*, Nara, Japan, March 2004.

[12] Campbell, W and . N. and Isard, S. D., "Segment durations in a syllable frame," *Journal of Phonetics*, vol. 19, pp. 37–47, 1991.

[13] Natvig, J.E., Heggtveit, P.O., *PROSDATA – A speech database for study of Norwegian prosody v2.0*, Telenor R&D, N 20/2000, Kjeller 2000

[14] Taylor P. et.al. *Edinburgh Speech Tools Library, System Documentation Edition 1.2*, Centre for Speech Technology, University of Edinburgh, http://www.cstr.ed.ac.uk/projects/speech_tools/manual-1.2.0/

[15] http://www.tele.ntnu.no/projects/fonema/