# Intonation Modelling with a Lexicon of Natural F0 Contours

*Per Olav Heggtveit, Jon Emil Natvig*

Telenor Research and Development
Instituttveien 23, N-2027 Kjeller Norway
per-olav.heggtveit@telenor.com
jon-emil.natvig@telenor.com

## Abstract

We describe a new approach for generating Norwegian intonation in text to speech synthesis. The method is based on a phonological representation of utterances. The overall f0 contour of an utterance is synthesised by concatenation of stored f0 contours corresponding to accent units. Candidate accent units are found by searching a lexicon derived from natural speech and selecting the unit that is the best match with respect to the properties of the target accent units of the utterance to be synthesised.

A formal subjective test confirms that the new approach leads to more natural speech than a former rule based method, but the quality is still inferior to intonation copied from natural speech.

## 1. Introduction

Prosody and in particular intonation plays a key role in the perceived naturalness of synthetic speech [1,2]. In most TTS systems, generation of prosody takes place in two steps: first an abstract description of the sentence prosody is derived on a linguistic level, typically represented as word prominence levels and phrase structures. Given this information, the next step calculates the physical parameters (f0 contour, phone durations, pauses etc) to produce the acoustic description of the intended prosody.

In this paper we report work on generation of f0 contours for Norwegian utterances. Our approach consists of a simple concatenation of stored f0 contours similar to [3,4,5]. Candidate f0 contours are stored in a searchable lexicon. At synthesis time, the utterance is represented as a sequence of accent units. For each unit, a context dependent search is carried out and the candidate unit that is the best match with respect to the properties of the target accent units of the utterance is selected. The pitch contour over the utterance is determined by linear interpolation between a set of target pitch values located around the accented syllables.

The paper is organised as follows: in Section 2 we present a phonological model for Norwegian intonation. Section 3 describes our system, including the generation of the lexicon and the selection process. Section 4 reports on subjective evaluation results comparing the new model with a conventional rule based system and intonation copied from natural speech.

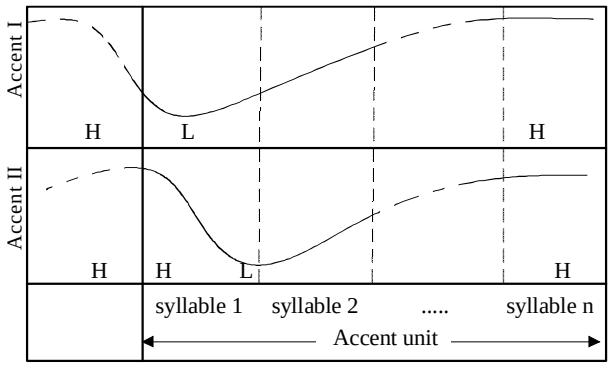


*Figure 1. Characteristic pitch contours for accent I and II in East Norwegian*

## 2. The intonation model

Our model is based on a simplified version of the "Trondheim Model" for phonologic representation of Norwegian intonation [6]. The basic unit of the model is the accent unit (AU). An AU consists of an accented syllable (carrying the word accent) followed by a sequence of unaccented syllables.

The AU is terminated by a phrase accent. In East Norwegian the phrase accent is a tonal rise (H) and the size of the rise signals phrase the degree of prominence. The most prominent unit, the focal AU, normally has the highest rise. Any initisl unaccented syllables, or unaccented syllables following the phrase accent are classified as "AU external".

A characteristic feature of Norwegian intonation is the distinction between two tonal accents. The tonal accents in East Norwegian are typically realised as shown stylised in Figure 1: Accent I is characterised by a low level (L) in the first syllable of the accent unit, whereas accent II is a H-L transition through the first syllable. In a sentence context, accent I can also be seen as a H-L transition from the H level of the previous phrase accent. As shown in Figure 1, the two accents can be regarded as time shifted versions of the same contour.

Accent units are grouped into Intonation phrases (IP). An IP consists of a sequence of non-focal AUs and is terminated by a single focal AU. Accent units following the last IP are subjected to a declination effect which reduces the pitch amplitudes of both tonal accents and phrase accents.

The highest phonological level of the model is the Intonation Utterance (IU), consisting of one or two IPs. In read speech, the IU normally corresponds to a sentence. The

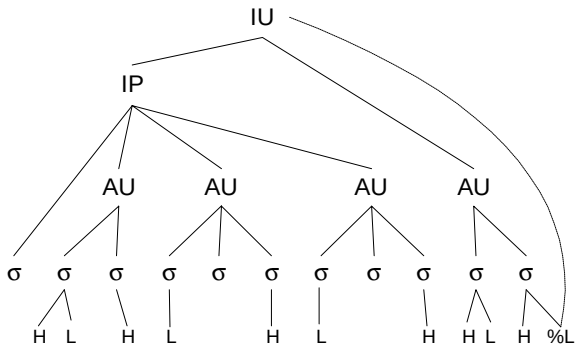


*Figure 2. Example of the tonal hierarchy.*

IU is terminated by a boundary tone %H for interrogative and %L for declarative utterances.
An example of the tonal hierarchy is shown in figure 2 with a syllable layer (σ) and a tone level layer (L/H).

## 3. System description

### 3.1. The corpus

The speech database PROSDATA [7] 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 database was recorded mainly for prosody studies, and not specifically for intonation generation. The sentences are read in an informative news style, with a pleasant but neutral voice.

The database includes word prominence ratings and break indices which have been determined by subjective evaluation. For each sentence a smoothed f0 curve has been determined.

An automatic labelling according to the prosodic model outlined above has been carried out as follows:

1. For each word marked as prominent, the syllables carrying the word accent were determined using a lexicon containing the phonotypical pronunciation.
2. Accent units were then constructed starting at each accented syllable and ending at the syllable before the next accented syllable or at a break index different from “no break”. Accent units were marked as focal or non focal according to the prominence ratings at the word level.
3. Syllables falling outside accent units were grouped into “external units”.

### 3.2. Intonation lexicon

A searchable lexicon of accent units was automatically extracted from the training corpus. For each accent unit we stored

- The unit type: focal, normal or external.
- Tonal accent type: I or II.
- Position code: post focal (Accent units after the *last* focal accent unit of an utterance), initial and final positon in the utterance.
- The number of syllables in the unit
- The phoneme string for each syllablen in the unit.
- The pitch value at syllable start and end, and vowel start and end for the first, second and last syllable of the accent unit.
- The average pitch value of the syllable preceeding the unit.
- The average pitch value of the last syllable of the current unit

### 3.3. Prosody synthesis

#### *3.3.1. Prosodic analysis*

The first step in the prosody generation process is to specify the intended prosodic phrasing and focusing of the input text. In a TTS system this is typically done by text analysis resulting in some kind of symbolic prosodic representation.

We use a symbolic prosodic representation in the following format:

<*unit_code*>(<*phoneme_string*>)
<*unit_code*>(<*phoneme_string*>)
….
<*boundary_tone*>

*unit_code* is one of the following:

*f1*: accent unit with tonal accent I

*F1*: focal accent unit with tonal accent I

*f2*: accent unit with tonal accent II

*F2*: focal accent unit with tonal accent II

*e*: unit consisting of external syllables

*phoneme_string* is a string of Norwegian SAMPA symbols [8] including syllable boundaries ($) and secondary stress (%).

*boundary_tone* is either %H or %L.

Example sentence:

“Men i dette tilfellet så han trolig ikke bilen”.

results in the following six units:

e(men$i)
F2(det$@$%til$fel$@)
f1(sO:$hAn)
f2(tru:$li$%ik$@)
f1(bi:l$@n)
%L

#### *3.3.2. Lexicon search*

Given an input sentence in the prosodic format described in section 3.1.1, the lexicon is searched to find the best matching candidate units. The search is performed by traversing the accent units of the sentence from left to right.

The key used to retrieve candidate units from the lexicon is built as a context dependent key as follows:

*<left_unit, cur_unit, right_unit>*

where:

*left_unit* is unit type and position of preceeding unit

*cur_unit* is unit type, tone and position of current unit

*right_unit* is unit type, tone and position of next unit

The list of candidates found in the search are scored according to the appropriateness of the units compared to the target unit in the local context. The factors taken into account are:

- Syllable structure of the first (accented) syllable
- Number of syllables in the accent unit
- F0 continuity at concatenation point

If no candidate unit is found, the criteria in the unit selection key are gradually relaxed until at least one candidate is found.

The syllable structure of the accented syllable is scored according to the similarity to the target unit. An overall candidate score is calculated by weighting the syllable structure score (0.6) and the unit length score (0.4).

If the average f0 level in the syllable at the concatenation point differs significantly (set ad-hoc to >20%), the candidate is discarded. Monosyllabic candidate units are excluded when the target is polysyllabic, and vice versa.

#### *3.3.3. F0 curve generation*

The result of the database search result in a set of f0 target points located in and around the first syllable of each accent or external unit. The overall f0 contour was determined using linear interpolation. Experiments using spline interplation did not result in any perceivable improvement.

#### *3.3.4. Duration modeling*

A CART model for phone durations was trained using the software tool Wagon from The Edinburgh Speech Tools Library [9].

The model input factors were selected among the factors available in the phonological representation used as input to the intonation model.

The database PROSDATA [7] used for intonation modelling is also used for the duration modelling. The model is trained on 90 % of the database and tested on the other 10 % (every tenth sentence).

The following factors are found important for predicting phone duration in our data and have been used to train the CART model:

- Current phoneme
- Current phoneme class
- Next phoneme class
- Previous phoneme class
- Syllable stress (unstressed, stressed, accented)
- Syllable break index (3 levels)

The position of a syllable in higher level units such as accent unit and length of the higher level units containing the syllable did not contribute significantly to the prediction of phone durations. Linguistic level factors on word or sentence level did not improve the overall performance of the prediction either. This is probably due to data sparsity when applying CART in phone duration modelling. The mean prediction error of the model is 14 ms and the corresponding correlation is 0.773.

## 4. Perceptual evaluation

### 4.1. Stimuli

#### *4.1.1. Sentences*

Ten relatively short sentences from the Prosdata database were selected. These sentences were then excluded from the prosody database when extracting the intonation lexicon described in section 3.2. The sentences were selected to represent factors like position of focal AU (initial, medial, final), tonal accents I and II. The number of syllables in AUs varied from 1 to 8.

#### *4.1.2. Prosodic conditions*

Three different prosody models were applied to the ten sentences. This resulted in lists of phonemes, durations and pitch values, which were fed into the same concatenative speech generator taken from the Talsmann TTS [10]. In all cases, the pitch contour is therefore degraded by the f0 strategy of the speech generator (only 2 pitch values per phoneme).

The following three models were compared

*Copy synthesis form the original (CPY)*
The original prosody (intonation and durations) was copied from the recording of each sentence and shifted down in pitch to match the male voice of the synthesiser.

*The new prosody model (NEW)*
Each of the ten sentences were manually analysed in terms of accent units, and the result fed into the intonation and duration models described in section 3, to produce intonation contours and durations.

*The Talsmann TTS model (TTS)*
The input to the standard TTS system Talsmann [10] was the sentence texts with emphasis tags on the focal words. The Talsmann TTS uses rule-based intonation modeling and a simple regression model for durations trained on a limited corpus of short declarative sentences.

### 4.2. Experimental procedure

We used a computer-based set up where the subjects could perform the test on their own PCs. For each sentence, a window was presented offering a play button corresponding to each prosody model. The identity of the models was unknown to the subjects. The subjects could listen and compare the three samples by clicking the corresponding play button as much as they wanted.

The task was to rank the three samples from perceived as "Most Natural" to "Least Natural". The subjects were forced to give rank orders in all cases. Two initial sentences were presented as familiarisation to the test and were not included in the analysis.

The presentation sequence of sentences and the placement of the play buttons corresponding to each prosody model on the screen was randomised.

### 4.3. Results

Table 1 shows the percentage of judgements where a prosody model was rated better than another.

| | Preferred prosody model | | |
|---|---|---|---|
| | TTS | NEW | CPY |
| TTS | - | 71.4 | 83.3 |
| NEW | 28.6 | - | 76.7 |
| ORIG | 16.7 | 23.3 | - |

*Table 1. Percentages of preference for one prosody model over another.*

All preferences in Table 1 are statistically significant at the 0.001 level resulting in a very clear rank order:

CPY >NEW > TTS.

In Figure 3 we present the average rank order for each of the ten sentences as well as the average over all sentences. We observe that the above rank order is achieved in 8 out of 10 sentences. For sentence 6, the NEW model is judged slightly better on the average than the CPY prosody. For sentence 7, the TTS prosody was judged to be slightly better than the NEW model.

In two cases, (sentences 3 and 6), the NEW model performs close to the CPY model.

## 5. Conclusions

In the study reported here, we compared a new corpus based method for generating prosody in text to speech. We have shown that this approach leads to a clear improvement over the rule based model used in Talsmann TTS [10]. According to our findings, context dependent accent units seem to be appropriate units of natural f0 contours for concatenation in Norwegian.

Further work should focus on design and fast building of new databases of different speaking styles and dialects. Lexica of different prosodic styles could then be applied to change the speaking style of a TTS voice.

An extension of our work would be to improve the unit selection to optimise the overall quality. Another extension could be to integrate the phonological prosody model as part of the segmental unit selection system.

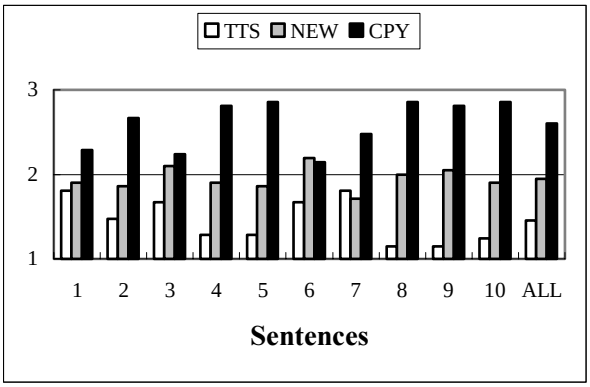


*Figure 3.* Average rank order for sentences

## 6. References

[1] Bunnell. H.T., Hoskins, S.R., Yarrington, D., Prosodic vs. Segmental Contributions to Naturalness in a Diphone Synthesizer. *Proceedings of the third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan, Autralia, Nov. 25-29, 1998

[2] Plumpe, M., Meredith, S. Which is more Important in a Concatenative Text To Speech System – Pitch, Duration or Spectral Discontinuity?, *Proceedings of the third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan, Autralia, Nov. 25-29, 1998

[3] Aubergé, V Developing a structured lexicon for synthesis of prosody, in *Talking Machines*, Elsevier 1992, p307-321

[4] Morlec, Y, Bailly, G. Aubergé, V., Synthesis and evaluation of intonation with a superposition model, *Proc. EUROSPEECH'95*, Madrid Sept.1995, p 2043-2046.

[5] Malfrère, F., Dutois, T., Mertens, P., Fully Automatic prosody generator for Text-To-Speech, *Proc. ICSLP '99*

[6] Fretheim, T. Themehood, Rhemehood and Norwegian Focus Structure, in *Folia Linguistica XXVI/1-2*, Mouton/de Gruyter, Berlin.

[7] 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

[8] Wells, J., SAMPA – Computer readable phonetic alphabet, http://www.phon.ucl.ac.uk/home/sampa/norweg.htm

[9] 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/

[10] Talsmann, A text-to-speech system for Norwegian, http://www.fou.telenor.no/prosjekter/taletek /talsmann