IMPLEMENTATION AND EVALUATION OF A MODEL FOR SYNTHESIS OF SWEDISH INTONATION

Merle Horne and Marcus Filipsson

Dept. of Linguistics and Phonetics, Lund University, Helgonabacken 12, S-223 62 Lund, Sweden

ABSTRACT

An abstract prosodic structure is implemented in terms of acoustic parameters realizing underlying word accents and boundary markers. The system is further evaluated in order to determine whether listeners prefer intonation contours produced with 1) default focal accent placement ('sentence stress') on the last content word in each prosodic phrase vs focal accent assigned to the last 'new' content word in each prosodic phrase, 2) minimal prosodic boundary signalling at commas and full stops vs more detailed prosodic boundary signalling generated using an underlying prosodic structure.

1. INTRODUCTION

In a number of previous articles, we have presented the structure of the different components in a linguistic preprocessor to a Swedish text-to-speech system which has been developed within the project 'Intonation in Restricted Texts: Modelling and Synthesis'.

The first component developed was a referent tracker. This is felt to be a crucial component for any text-to-speech system since generation of natural prosody is dependent on being able to extract knowledge on information structure. The referent tracker identifies coreference or cospecification relations between lexical words on the basis of morphological identity as well as lexical semantic identity-of-sense relations (hyponymy, synonymy, meronymy/partonymy) [1]. These identity relations are modelled in a computerized lexicon [2] and the tracking procedure works within an adjustable text window. The output of the referent tracker is a text where all lexical words are specified as either contextually 'new' (N) or 'given' (G). This information can then be used in the F0 generating component in order to appropriately assign 'focal' vs. 'non-focal' word accents (in Swedish, the 2 lexical word accents are followed by a H tone if they are 'focal', i.e. associated with a word constituting 'new' information [3]).

Another goal of the project has been to generate an abstract prosodic structure which can be used in the text-to-speech system in order to better model prosodic boundary signalling. In [4-5], a prosodic structure is proposed containing three hierarchically ordered levels: the Prosodic Word (PW), the Prosodic Phrase (PPh) and the Prosodic Utterance (PU). The PPh is the central constituent on the basis of which the other constituents are defined. Prosodically, it is characterized by a boundary tone (H% or L%), a degree of Final Lengthening and a Silent Interval (breath pause) [6-7]. It corresponds syntactically very often with the clause; however, syllable count also plays a role in determining the position of PPh boundaries: a number of clauses can be grouped together in a PPh if they consist of a limited number of syllables (e.g. elliptic clauses). Even the number of focussed constituents in a clause can influence PPh boundary assignment. For example, an optional PPh boundary was observed in our stock market data (radio speech) between two focussed (new) Predicate Complements (e.g. a Direct Object and a prepositional phrase functioning as an Adjunct).

Within the PPhs, PWs are defined. Lexically, these are composed of a Content Word (CW) followed by any Function Words (FW) up to the next Content Word. The Prosodic Word is a rhythmical unit and is characterized by a word accent and a boundary tone which is H# if the word does not have a focal word accent and L# if the word does have a focal accent. These boundary tones function to create the transitions between word accents. PUs correspond textually to paragraph boundaries and correlate with discourse topic shifts. Prosodically, they are marked by a greater degree of Final Lengthening and Silent Interval duration than those associated with PPhs.

2. IMPLEMENTING THE PROSODIC STRUCTURE

Using the information obtained from the referent tracker and the prosodic parser, we are currently involved in developing a rule component for generating intonation contours. A set of rules associate the underlying word accent representations and prosodic boundaries with acoustic parameters (F0, duration (Final Lengthening), Silent Intervals) has now partially been implemented. Syllable boundaries, word accent type and stress are further indicated in the lexical entries.

The rules make reference to the new/given status of words when the word accent form is assigned. For example, if a word is associated with the label N(ew) and is marked as Accent 1 in the lexicon, it will be realized phonetically with one of the tonal patterns HL*H$^-$, L*H$^-$ or H$^-$ depending on the number of syllables it contains: if it contains a prestress syllable, it will be associated with all three tones in the representation HL*H$^-$, where the first H is associated with a point in the prestress syllable, the L* with the beginning of the stressed Vowel, and the final H$^-$ with some point in the syllable after that containing the stressed vowel as in Figure 1 (a). If there is no prestress syllable, the word will be associated with the two rightmost tones L*H$^-$ in Figure 1 (b), and if the word is a monosyllable, then priority is given to the focal H$^-$ which is the only underlying tonal component realized in the speech style being modelled (i.e. professional read speech) (Figure 1 (c)):

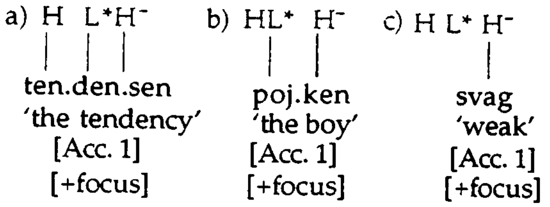


Figure 1: 'Focal' Word Accent 1 in Swedish (HL*H⁻) showing its realization in relation to the number of syllables in the lexical entry. (Tones not associated with segments are not realized phonetically.)

If the word is G(iven), the final H⁻ tonal component associated with focus is not present in the word accent representation. The rules thus generate a number of contextual variants of the underlying word accent representations.

The new/given status of words also conditions the amount of accentual prominence assigned to a word. A 'new' word, for example, is assigned more prominence in terms of F0 peak values than a contextually given word. Function words are further assigned less prominence than given content words. Furthermore, there are rules for downstepping of word accents after the last focal accent in a PPh (see [8] for a discussion of this phenomenon in Swedish). These rules for downstepping must also make reference to the new/given status of words.

Reference to prosodic structure is needed first of all in order to determine the scope of the focal H⁻ and the location of the PW-boundary tones (H# or L#) which constitute the transitions to a following word accent. For example, there is a rule which says that the focal H⁻ spreads from a point at the end of the stressed syllable to the syllable preceding the PW boundary as in Figure 2 where (.) represents syllable boundary:

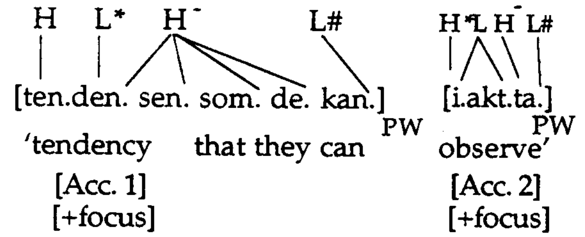


Figure 2: Tonal association showing the spreading of the focal H⁻ up to the last syllable in the first PW.

Reference to prosodic structure is further needed in order to be able to model how speakers chunk up speech into PPhs and PUs. In our data-base of professsional read radio speech, clause boundaries were associated with PPh boundaries 67% of the time. PPhs consisted, furthermore on the average of 24 syllables at a speech rate of 5 syllables per second (they were never longer than 63 syllables and never shorter than 7 syllables). Thus information on clause boundaries as well as syllable count is important in constructing a prosodic structure that can control the insertion of PPh boundaries in text-to-speech conversion.

2.1. Methodology

As speech data for implementation purposes and testing of our rules for F0 generation, we have used recorded speech in order to obtain an optimal segmental quality. After some practice, a male speaker (the second author) was able to produce utterances with a more or less flat F0 contour, and with minimal variation in intensity and duration (using a monotonous, robot-like speaking style). This was desirable in order to be able to test the effect of moving the location of accents while avoiding secondary influences from duration differences. Intonation contours were then generated using an implementation of the PSOLA technique [9].

A system for creating F0-files from the pre-recorded sentences and the associated prosodic structure was developed. The pre-recorded original sentence was labeled with (1) an orthographic transcription of each word, (2) syllable boundaries and (3) vowel onset time for the primary stressed syllable. If the last word ended in (a) voiceless sound(s), a label for the (i.e.) actual end of voicing was added in order to be able to correctly time the final phrase accent.

The system creates a text file from the label file with words which is then analysed by the referent tracker and prosodic parser [4-5]. The parser looks up each word in the lexicon and constructs a prosodic structure for a sentence in terms of PWs, PPhs and PUs. The labels 'New' or 'Given' are also assigned to each word based on the referent tracker mentioned above. From the lexicon the system also derives word accent type and the number of syllables for each word. The system contains a large set of rules for transforming the linguistically parsed sentence to a label file consisting of a sequence of F0 values. Durations and timing were taken from the syllable boundaries and vowel onset times which were manually labeled. In the top label tier in Figure 3, just below the F0 contours, is an example of such a sequence of F0 values expressed in Hertz.

The final part of the system takes the F0 values, interpolates them linearly and produces an F0 file which is then used as input to the re-synthesis algorithm. Examples of two such F0 contours can be found in Figure 3. Figure 4 shows an example of a prosodically parsed sentence from which the lower F0 curve in Figure 3 is derived.

3. EVALUATION OF THE SYSTEM

Having implemented a considerable number of rules for the generation of F0 contours, we felt it important at this stage to compare and test their output against an output generated without the extra analysis that our system involves i.e. referent tracking and prosodic structure generating. Thus, we decided to develop a test to determine 1) whether listeners prefer the output of a system with referent tracking to one without and 2) whether listeners prefer the output of a system with prosodic parsing to one without.

In order to test 1), it is necessary to compare segmentally identical utterances with default focal accent placement on the last content word in each phrase (systems without referent

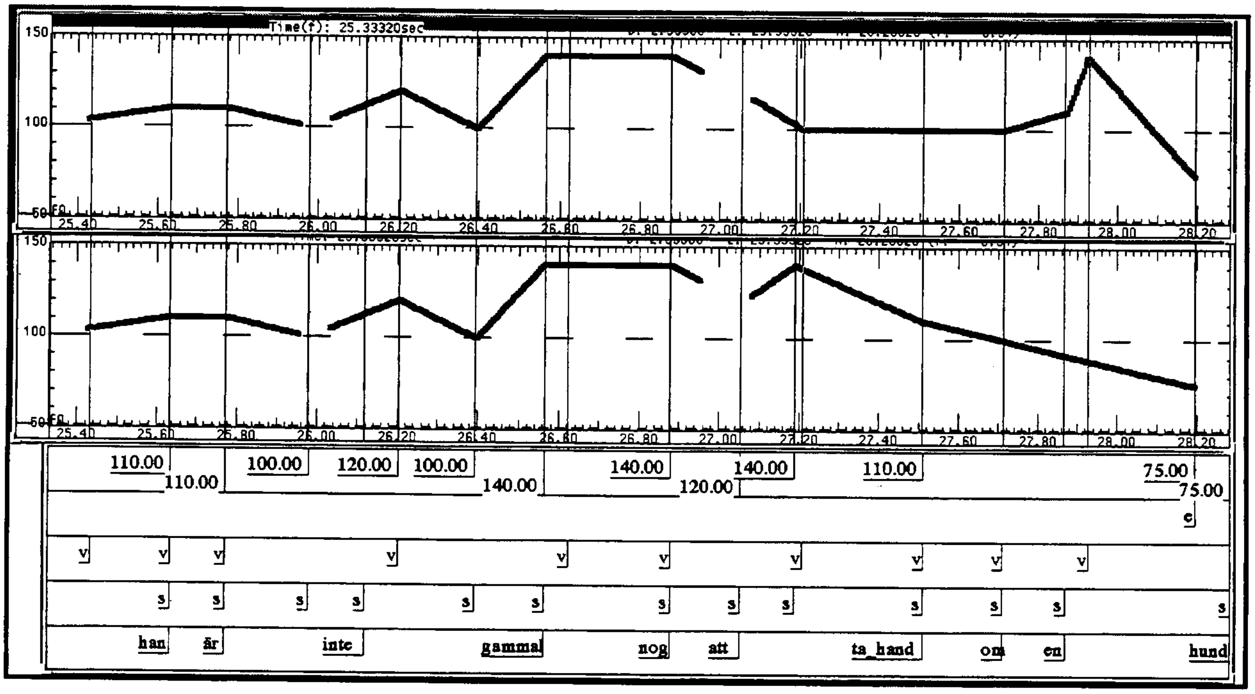


Figure 3. F0 contours generated for the sentence *Han är inte gammal nog att ta hand om en hund* 'He is not old enough to take care of a dog'. This sentence is used for the test of focal accent placement. The top version has focal accents on *gammal* 'old' as well as on the last content word *hund* 'dog'. This differs from the lower version where the last focal accent does not fall on *hund* but rather on *ta **hand*** 'take **care**'. At the bottom are label tiers for words, syllable boundaries (s), vowel onset (v), end of voicing (e), and the automatically generated sequence of F0 values (for the lower contour).

tracking) vs focal accent placed on the last new content word (our system). This can be done using a functional test as in [10] where listeners are asked which of a pair of synthetic stimuli constitutes the most appropriate answer to a preceding question. In order to test 2) presence vs. absence of a prosodic constituent hierarchy segmentally identical test stimuli with and without prosodic boundaries (PW and PPh) can be presented to listeners who are asked to evaluate the second version of the pairs of synthetic stimuli as being better, equal or worse than the first with respect to naturalness.

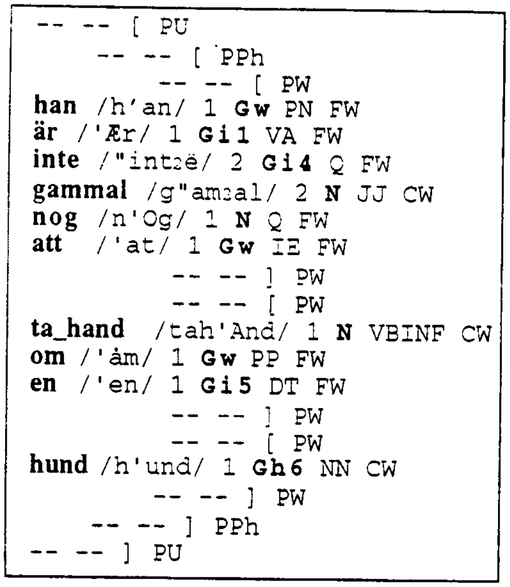
In the functional test on the placement of focal accent, the test sentences are being presented as answers to two different possible questions (Q). An example of these question/answer pairs is given in (1). As the most appropriate answer to the first question (Q1) in (1), one would expect the answer in a) with the words after **hand** deaccented since they are given in the context (**hund** 'dog' is a superordinate term with respect to **tax** 'dachshund' and is therefore marked as Given by the referent tracker). As the most appropriate answer to the second question (Q2) in (1) on the other hand, one would expect that listeners would choose b) with the final content word **hund** 'dog' accented since it is new information in the context and would sound inappropriate if it were not accented as in a).

Two different realizations of each sentence have been generated with different locations of the final focal accent. Each of these two sentences has been spliced together with each of the two questions in two orders of presentation, thus yielding four question-answer pairs. Listeners are being presented with a question followed by the two answers and are asked to indicate which of the alternates (a) or (b) constitutes the most appropriate answer to the preceding question.

(1) (Words written in bold represent focally accented words)
Q1: *Varför köper du inte en **tax** till din son?*
'Why don't you buy a **dachshund** for your son?'
a) *Han är inte **gammal** nog att ta **hand** om en hund.*
'He's not **old** enough to take **care** of a dog'
b) *Han är inte **gammal** nog att ta hand om en **hund**.*
'He's not **old** enough to take care of a **dog**'

Q2: *Varför köper du inte en **häst** till din son?*
'Why don't you buy a **horse** for your son?'
a) *Han är inte ens **gammal** nog att ta **hand** om en hund*
'He's not even **old** enough to take **care** of a dog'
b) *Han är inte ens **gammal** nog att ta hand om en **hund***
'He's not even **old** enough to take care of a **dog**'

For the second test on the absence vs. the presence of a prosodic constituent hierarchy we have constructed two segmentally identical versions of a short text (see (2)) which are associated with different intonation contours. From one version of the text, a prosodic structure is derived as shown in (2) which is used in generating boundary tones, Final



```
-- -- [ PU
    -- -- [ PPh
        -- -- [ PW
han /h'an/ 1 Gw PN FW
är /'Ær/ 1 Gi1 VA FW
inte /"intɜë/ 2 Gi4 Q FW
gammal /g"amɜal/ 2 N JJ CW
nog /n'Og/ 1 N Q FW
att  /'at/ 1 Gw IE FW
          -- -- ] PW
          -- -- [ PW
ta_hand  /tah'And/ 1 N VBINF CW
om /'åm/ 1 Gw PP FW
en /'en/ 1 Gi5 DT FW
          -- -- ] PW
          -- -- [ PW
hund /h'und/ 1 Gh6 NN CW
        -- -- ] PW
    -- -- ] PPh
-- -- ] PU
```

Figure 4: The prosodically parsed sentence in (1a). Associated with each word is a phonetic transcription, accent type (1 or 2), referent status (N(ew) or G(iven)), word class and CW/FW-status. The prosodic structure consists of PWs, PPhs and PUs. PN=Pronoun, VA=Aux. Verb, Q=Quantifier, JJ=Adjective, IE=Inf. Marker, VBINF=Infinitive Verb, PP=Preposition, DT=Determiner, NN=Noun. The sequence *ta hand* 'take care' is treated as a lexicalized phrase *ta_hand*.

Lengthening and Silent Intervals. Both PW's and PPh's are associated with boundary tones ((H#/L#) and (H%/L%), respectively). PPh boundaries are further associated with Final Lengthening and a Silent Interval [7]. In the other version of the text, there is no such underlying prosodic structure assumed, i.e. there are no groupings of content words and function words into units corresponding to PWs in our system. Moreover, there are no divisions of words into PPhs other than at commas and full stops. (In most current text-to-speech systems, commas are associated with slight 'continuation rises' and full stops with falls.) Thus we are testing the importance for listeners of taking into consideration linguistic (lexical, syntactic, and semantic) information when building up prosodic structure. Listeners are being presented with the two versions of text and asked to indicate whether the second version sounds better, worse or equal to the first with respect to aspects of rhythm and naturalness.

The listening tests are currently underway and the results will be presented at the conference.

ACKNOWLEDGEMENTS

This research was supported by a grant from the Swedish Language Technology Program (HSFR-NUTEK).

2)

```
[[Efter en vikande]PW [inledande]PW [handel på]PW
 After a receding     opening       trade at
[torsdagens]PW [StockholmsbörsPW]PPh, [[så fick de]PW
 Thursday's   Stockholm's stock-exchange made the
[mycket positiva]PW [delårsrapporterna från]PW [AGA
 very positive      semi-annual reports from    AGA
och]PW [Astra]PW [kursutvecklingen att]PW [vända]PW ]PPh.
and     Astra     the rate development to  turn.
 [[Tillbakagången under]PW [förmiddagen]PW [återhämtades
   The decline    during    the morning          recovered
sedan]PW [successivt]PW [tack vare]PW [stigande]PW [kurser
later      gradually     thanks to     increasing     rates
i]PW [marknadsledande]PW [papper]PW]PPh. [[Omsättningen
in    market-leading       shares              Sales
under]PW [sessionen var ]PW [liten]PW]PPh [[och gick]PW
during    the session were   small          and went
[endast upp till knappt]PW [234]PW [miljoner]PW
only    up  to  not quite   234     million
[kroner]PW]PPh, [[varav nära 50]PW [procent]PW
crowns           of which nearly 50  per cent
[utgjorde]PW [handel i]PW [Astra]PW, [Ericsson och]PW
represented   sales in     Astra       Ericsson and
[Bilspedition]PW ]PPh.
Bilspedition
```

REFERENCES

1. Horne, M., Filipsson, M., Ljungqvist, M. and Lindström, A. "Referent tracking in restricted texts using a lemmatized lexicon: implications for generation of prosody," *Proc. Eurospeech '93* (Berlin) Vol. 3: 2011-2014, 1993.
2. Hedelin, P., Jonsson, A. and Lindblad, P. "Svenskt uttalslexikon: 3 ed. Tech. Report, Chalmer's Univ. of Technology, 1987.
3. Bruce, G., *Swedish Word Accents in Sentence Perspective*, Gleerups, Lund, 1977.
4. Horne, M. and Filipsson, M. "Generating prosodic structure for Swedish text-to-speech," *Proc. ICSLP 94* (Yokohama), Vol. 2: 711-714, 1994.
5. Horne, M. and Filipsson, M. "Computational extraction of lexico-grammatical information for generation of Swedish intonation". *Progress in speech synthesis*, ed. J. van Santen, R. Sproat, J. Olive, and J. Hirschberg. New York; Springer, In press.
6. Horne, M. and Filipsson, M. "Computational modelling and generation of prosodic structure in Swedish," *Proc. XIIIth ICPhS, Stockholm*, Vol. 4: 364-367, 1995.
7. Horne, M., Strangert, E. and Heldner, M. "Prosodic boundary strength in Swedish: Final Lengthening and Silent Interval duration," *Proc. XIIIth ICPhS, Stockholm*, Vol. 1: 170-173, 1995.
8. Bruce, G. "Developing the Swedish intonation model," *Working Papers* 22 (Dept. of Ling., U. of Lund): 51-116, 1982.
9. Möhler, G. and Dogil, G. "Test environment for the two level model of Germanic prominence," *Proc.Eurospeech'95* (Madrid), Vol. 2: 1019-1022, 1995.
10. Pols, L. Voice quality of synthetic speech:representation and evaluation. *Proc. ICSLP 94* (Yokohama), Vol. 3, 1443-6, 1994.