INTONATION MODELLING IN A TEXT GENERATION PROGRAM

Carlos Gussenhoven, Toni Rietveld and Lex Elich

Department of English & Department of Language and Speech
University of Nijmegen/The Netherlands

ABSTRACT

In this contribution we describe the implementation of an autosegmental description of intonation in Dutch.

1. INTRODUCTION

We present a brief description of the implementation of an autosegmental description of the intonation of Dutch in an allophone synthesis system. Our longer-term aim is for this program to be part of a text-generation program. At present, we assume the following as given:

1. A phoneme string, with word and syllable boundaries; the latter are symbolized by -.
2. Accents, symbolized * before accented syllables.
3. Association Domain (AD) boundaries, symbolized (...);
4. Scaling Domain (SD) boundaries, symbolized {...}. An SD contains a string of one or more ADs.
5. A phonological transcription of the intonation contour.

In (1) we give the representation of *Marjan* ('girl's name'), without the information under 5.

(1) {(m A r - *j A n)}

Our program translates the phonological transcription into a string of **targets.** A target is a point defined by (a) an F0-value (Hz), and (b) a time value relative to the phoneme string. The intonation contour is obtained by interpolating between all targets. We will first describe the phonological representation of the contours, then given the timing rules, and finally the F0-rules.

2. THE PHONOLOGICAL REPRESENTATION

The representation is built up by inserting intonational morphemes (see also Gussenhoven [2,3].

Pitch Accents. Minimally, each * in the string must be provided with one of three PITCH ACCENTS: H⃰L, L⃰H, or H⃰LH. The choice is free, except that all nonfinal *'s in an AD must have the same Pitch Accent.

In addition, a larger number of optional morphemes are available. These specify AD's, or, in one case, the SD.

Accentual Downstep. An AD may be provided with ACCENTUAL DOWNSTEP. This means that the H⃰ of each non-initial H⃰L will be realized with lower pitch than the preceding H⃰. ACCENTUAL DOWNSTEP is symbolized by placing the diacritic ! before the AD.

Accentual Downstep with Spreading. An AD which has been provided with DOWNSTEP, may additionally be provided with SPREAD. This means that each non-final H⃰ will be realized as a plateau instead of a peak, while the final H⃰ is rewritten L⃰, and the preceding L is deleted. Spreading-cum-downstepping is indicated by means of ~! before the AD.

Narration. NARRATION is applicable to H⃰L and L⃰H, and causes their T⃰ to spread. Thus, while an unnarrated realization of H⃰L will show a peak at the accented syllable, a narrated one will show a high plateau beginning at the accented syllable and ending just before the next *, or the end of the AD. Narration cannot occur in an AD with downstep. It is symbolized by placing & before the AD.

Modifications. Pitch Accents come in a number of variant forms, which are described as 'modifications' of the basic shapes. Two modifications are implemented. The first, DELAY, causes the association of the Pitch Accent to be shifted rightward. DELAY is implemented as as the prefixation of a L tone segment, which is timed like the T⃰ of the undelayed Pitch Accent; it is symbolized by placing @ before the AD. The second modification, HALF COMPLETION, is possible only for the last Pitch Accent of an AD. It causes the contour to end on mid pitch. It does not combine with either NARRATION or DOWNSTEP. It is symbolized by placing = before the AD.

Phrasal Downstep. An SD may be provided with PHRASAL DOWNSTEP. This means that each non-initial AD will be realized with lower pitch than the preceding AD. PHRASAL DOWNSTEP is symbolized by placing ! before the SD.

Adjustment rules:

Boundary tones. Before the first Pitch Accent of an AD, an initial tone segment L is inserted, except when the first Pitch Accent is a narrated L⃰H, in which case H is inserted. This initial tone segment is referred to as %T. Moreover, H⃰LH rewrites as H⃰LH% in AD-final position. After the last Pitch Accent of the SD, L% is inserted after H⃰L, and H% after L⃰H, unless the AD has NARRATION or HALFCOMPLETION. The representations are transformed into 'narrower' transcriptions before they are processed by the timing and F0-rules. Domain-specific information is shown locally in these narrower representations. The adjustment rules change the representation in (2), for example, into (3). Observe that the boundary tone segments have been added, and that DOWNSTEP, SPREAD labels on the second AD have here been translated into the spreading of the first H⃰, and a rewriting of the second H⃰ as L⃰. Also, the DOWNSTEP label on the SD has been translated as the phrasal downstepping of the second AD.

(2) !{(L⃰H L⃰H H⃰LH) ~!(H⃰L H⃰LL%)}
(3) {(%L L⃰H L⃰H H⃰LH%) !(%L ~H⃰ L⃰LL%)}

3. IMPLEMENTATION: Timing of Targets

The domain for the timing rules is the AD. That is, the first and the last frame of the AD act like firm walls, and rules cannot locate targets beyond them. Accented syllables provide AD-internal reference points. The notation T is used for tone segments other than %T, T⃰, or T%. Each tone segment yields either one or two targets. The %T, the last T of an AD-final Pitch Accent, and a spread T⃰ translate into two targets. All other tone segments translate into single targets. Where a tone segment yields two targets, the earlier one is called Target1 and the later one Target2. For ease of exposition, we here distinguish three levels of representation. First, there is the segmental string, with *'s added before the accented syllables, and (...) and {...} in place. Second, the temporal string, consisting of a series of frame numbers associated with the successive segments. (We here assume for the sake of convenience that a frame represents 10 ms.) The third representation is the tonal string.

```
(4)                    {( m    A     r  -  *j    A    n )}
- temporal string:     123456 1234  123456 123456 1234 12345
- tonal string:        %L H⃰LH%
```

We show the timing of targets with the help of association lines between tone segments and the frames. Our rules contain a number of timing parameters, whose values can be manipulated in order to explore the perceptual effects of different timings of tone segments. We give default values which have informally been observed to give reasonable results.

- T⃰: locate Target at a distance of STARTIME of the vowel duration, counting from the beginning of the vowel. (Default STARTIME = 50%)

- ~T⃰: a spreading T⃰ (for SPREAD or NARRATION) has two targets:
  Target1 as above

Locate Target2 at TOTIME from frame associated with next T*. (Default TOTIME=100ms.) If AD is narrated, locate Target2 at (SPREADTIME)*(TOTIME) from next * or). (Default SPREADTIME = 1.3)

- %T: associate Target1 with first frame of AD;
  Locate Target2 at TOTIME before *. (Default TOTIME=100ms)

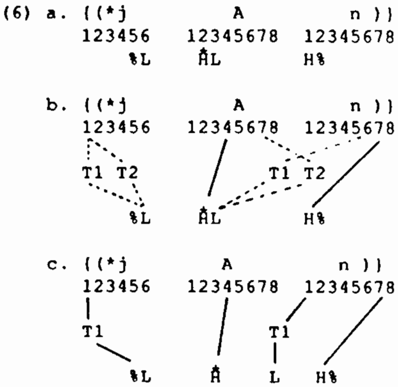
For non-final Pitch Accents:
Each T is located at FROMTIME from the preceding target. If the distance between preceding target and * is less than FROMTIME + TOTIME, locate Target midway between preceding target and *. The last T is located at TOTIME from next *. If the distance between preceding target and following * is less than FROMTIME + TOTIME, position Target midway between preceding target and *. (Default FROMTIME=100ms)

For final Pitch Accents:
All T's except the last as above. The last T receives two targets:
Locate Target1 FROMTIME after *.
Locate Target2 at TOTIME before the end of AD, if T% follows. If no T% follows, associate Target2 with last frame.

- T%: associate Target with last frame

Where the space provided by the segmental string is less than FROMTIME + TOTIME, Target2 may inappropriately be timed earlier than Target1. In such a case, OVERLAP and SIMPLIFY apply so as to associate Target1 with the frame that lies midway between them, and to delete Target2.

In (6), we illustrate a situation in which the available time is less than FROMTIME and TOTIME. Representation (6a) results from applying the first timing rule (STARTIME). In (6b), we see the result of the other timing rules without OVERLAP. Target2 of %L was positioned by going TOTIME leftward from *, and hitting the lefthand boundary of AD (cf the dotted association line). It is thus associated with the same frame as Target1. For Target1 of L, we count FROMTIME from *. For Target2, we count TOTIME back from the righthand boundary. Notice than the two targets overlap, as shown in the added target tier. Their associations are given as dotted lines to indicate their provisional status. Representation (6c) gives the state of affairs after the application of OVERLAP and SIMPLIFY. This representation is ready to go to the F0-rules.



4. IMPLEMENTATION: F0

The calculation of F0-values is performed by an implementation model described in Van den Berg et al. [1]). This model is a modified version of that proposed in Ladd [5]. Briefly, it provides a high reference value (which equals that of the first H*) and a low reference value (which equals that of L*), together defining a register, whose width is is referred to as TRANGE (i.e. the distance between H* and L*). The starting values are determined by three parameters that are intended to model speaker-to-speaker variation in general pitch height, and different degrees of prominence and liveliness. Their settings remain in force throughout the SD. An Accentual Downstep factor **da** determines the lowering of H* targets in an AD with accentual downstep. The distance between L* and the most recent F0-value for a (downstepped or undownstepped) H*-target is referred to as !TRANGE. For targets after undownstepped H*, !TRANGE equals TRANGE.

A Phrasal Downstep factor **dp** determines the lowering of AD's in an SD with phrasal downstep.

For targets other that those of T*, we can be flexible in the sense that not only the high and low reference values will be used, but any intermediate value. That is, we refer to values around the reference values by means of percentages, in the manner of Horne [4].

4,2. F0-rules

- %T: Target1: L = STARTSINK of TRANGE (Default STARTSINK=35%);
  H = H*
  Target2 = (STARTSLOPE)*Target1 (Default STARTSLOPE= .9)

- H* (= high reference)
  L* (= low reference)

- !H* F0 as given by the Accentual Downstep factor **da**.

- L in final Pitch Accent = L* (Target1 and Target2). If HALF-COMPLETION is in effect, delete Target1 and scale Target2 at HALF of TRANGE (Default HALF = 60%).
  L in non-final Pitch Accent = SAG of !TRANGE (Default SAG = 25%)

- L% = ENDSINK of TRANGE (Default ENDSINK = -10%)
  H% = previous Target + (ENDRISE of TRANGE) (Default ENDRISE = 30%).

4.3. The F0(m,n)-module

The implementation model F0(m,n) is given below. It calculates the F0-value for the *n*th T* in the *m*th AD.

$$F0(m,n) = Fr * N dp^{Sp*(m-1)} * w^{T} * da^{0.5*Sa*(1+T)*(n-1)}$$

Parameters:

Sp = +1 if Phrasal Downstep, 0 if not;
Sa = +1 if Accentual Downstep, 0 if not;
T = +1 for H*, and -1 for L*;

Fr = Reference line at the bottom of the speaker's range (default: 50 Hz for men and 100 Hz for women)

N = Defines the range, or the mean starting value above Fr (Default: 2.1)

W = Determines the distance between H* and L*. (Default: 1.6)

da = Downstep factor for downstepping H* targets within the AD. ("Accentual Downstep". Default: .80 if Sp = 1, and .70 if Sp = 0)

dp = Downstep factor for downstepping AD's in the SD. ("Phrasal Downstep". Default: .90).

5. INTERPOLATION

Interpolation between targets is by means of a 2nd order spline function. Future research involves the evaluation of different measures that can be taken if the time provided by the segmental string is insufficient to produce interpolations with slopes that remain within a pre-set speed limit. One measure might be UNDERSHOOT. Targets other than those provided by T* and T% would be undershot. Another approach would be to create more space by adjusting the position of *, thus creating more space (SHIFT). A third might be STRETCH, which would increase the time available by lengthening the segments concerned.

REFERENCES

[1] Berg, R. van den, Gussenhoven, C. & Rietveld, A.C.M. (1991), "Downstep in Dutch: Implications for a model", To appear in G. Docherty & D.R. Ladd (eds.) *Papers in Laboratory Phonology II*. Cambridge: Cambridge University Press.

[2] Gussenhoven, C. (1988), "Adequacy in Intonation Analysis: The Case of Dutch". In H. van der Hulst & N. Smith (eds) *Autosegmental Studies on Pitch Accent*. Dordrecht: Foris. 95-121.

[3] Gussenhoven, C. (1991), "Tone segments in the intonation of Dutch", In Th.F. Shannon & J.P. Snapper (eds.) *The Berkeley Conference on Dutch Linguistics 1989*. Lanham (MD): University Press of America.

[4] Horne, Merle A. (1988), "Towards a quantized, focus-based model for English sentence intonation",. *Lingua*, 75, 25-54.

[5] Ladd, D.R. (1987), "A phonological model of intonation for use in speech synthesis", In J. Laver & M. Jack (eds). *Proceedings of the European Conference on Speech Technology*. Vol. 2. Edinburgh: CEP Associates. 21-24.