Synthesizing Intonation using the RFC Model

Paul Taylor
ATR Interpreting Telecommunication Laboratories, 2-2 Hikaridai,
Seika-cho, Soraku-gun, Kyoto, JAPAN.
(email: paul@itl.atr.co.jp)

ABSTRACT

This paper describes an implementation of the rise/fall/connection (RFC) model of intonation for use in the speech synthesis component of a machine interpretation system. The phonological input is described, as is the algorithm which converts this to a rise/fall/connection description, and eventually to an F_0 contour. As we wish to model different speakers' intonational characteristics, some details are described on the method for adapting the model's parameters to match the behaviour of a given speaker.

1 INTRODUCTION

The English speech synthesis system currently under development at ATR is part of a machine interpretation system. In the design of the phonetic intonation module of this system (which converts phonological intonation descriptions to F_0 contours) we had two basic goals: comprehensive coverage of the intonational phenomena of English and natural sounding intonation that reflects an individual's speaking style.

These design criteria dictate that our system has a different structure and operation to intonation systems found in most text-to-speech (TTS) systems (e.g. [1]). The system has to synthesize a wide range of intonational effects as we want to be able to ask different types of questions and statements. It is not simply enough to produce a "neutral declarative" style of intonation. However, as we were dealing with machine generated utterances we have a richer, more accurate linguistic description of that utterance than would be typically be available from a text-input system. Thus our system has to perform more complex tasks than most TTS intonation systems, but to balance that we have access to more detailed information about the utterance.

In addition to synthesizing all the intonational effects of English we want to capture particular speakers' phonetic realisation characteristics. This requires that we must have some method of deriving a speaker's intonational characteristics from a database of that speaker's voice. Thus our system has to be capable of analysing F_0 contours, extracting intonational characteristics and using these characteristics for synthesis.

2 BASIC SYNTHESIS SCHEME

The synthesis algorithms are based on the *rise/fall/connection* ("RFC") model of intonation [8]. This model was designed to be formal in its workings so that all its operations are explicitly defined and therefore easy to implement on computer. The model was designed for both analysis and synthesis: the synthesis mapping takes a phonological input and produces an F_0 contour, the analysis mapping takes a F_0 contour and produces a phonological description. Algorithms have been developed which can automatically label F_0 contours, and it has been shown that contours generated using the model are very similar to naturally occurring ones [8], [9]. This paper focuses on the practical requirements needed to use this model in a speech synthesis system.

There are several theoretical advantages [8] this model has over others but the principle reason for its use here concerns the analysis capability of the system. As precise, detailed F_0 contour labelling is straightforward, we can easily use large quantities of data to determine the settings and parameters of the model for a particular speaker, and thus imitate that speaker's phonetic realisation behaviour.

2.1 INPUT

The input to the system is a specification of the intonational phonological description of the utterance. This is made up from four things: *tune association, tune type, pitch range* and *phrasing*.

Tune association simply indicates where the phonological elements occur. Pitch accents are associated with single syllables, and boundary elements (see below) are associated with phrase boundaries.

Tune type refers to which *elements* and*features* in the HLCB system are being used. The HLCB tune description system uses four basic elements. Pitch accents are of either type **H** (high) or **L** (low), phonologically relevant connection elements (see below) are marked **C**, and rapid rises at phrase boundaries are marked **B**. Within these classes, features are used for sub-classification. **H** accents are the most common type found in utterances and have three features , "downstepped" ($\mathbf{H}_d$, "late" ($\mathbf{H}_l$) (for accents with peaks late in the syllable, corresponding to the British school rise-fall [7]) and "elevated" ($\mathbf{H}_e$) for extra high accents (high fall in the British school) [3]. The **L** class has a single feature "antecedent" ($\mathbf{L}_a$) which indicates whether the **L** has a steep fall leading into it. **C** elements are normally of no phonological significance, and simply "fill in" the contour between pitch accents and boundary elements, but can be marked with the feature "rising" to distinguish them from the usual (level or falling) behaviour. $\mathbf{C}_r$ elements are often found after the nuclear accent, e.g. $\mathbf{H}_d$ $\mathbf{C}_r$ is similar to a fall-rise accent in the British school. **B** elements can occur at the starts of phrases (often termed "declination resets"), is which case they are marked as being "initial", or at the ends of phrases where they are unmarked.

Phrasing structure is in principle recursive in that there ,are an arbitrary number of phrase types and each phrase may include any other phrase [4]. In practice, we define a small number of phrase types (currently four) and only include daughter phrases of the same or a lower type in a parent phrase. The phrase information is mainly for use with the duration module, which determines phrase-final lengthening and pause durations [2]. The intonation module itself only makes use of the phrase boundary locations and is not concerned with the actual type of phrase.

Each phrase can take a number of parameters which are constant for the duration of that phrase. The most important parameter for the intonation module is *pitch range.* A (variable) number of pitch range levels are defined, each with a name, a maximum and a minimum F_0 value. The pitch range of a phrase determines the starting F_0 value of that phrase and the maximum and minimum values that the F_0 contour can vary between. A phrase may be marked as lying in a particular pitch range or be left marked. If marked, that pitch range type specifies the starting pitch of the phrase, and the minimum and maximum values that the contour can very between. If left unmarked, the starting F_0 of the phrase is the same as the last F_0 of the previous phrase. In the unmarked case the maximum and minimum values are determined by moving up through the phrase tree until a parent phrase is found with a marked pitch range, and that pitch range is then used. The top level phrase always has a marked pitch range. This pitch range system is similar to the register system of Ladd [5], but here we only specify one register per phrase rather than one for each accent. We don't have to specify a register setting for each accent as it is thought that our tune description system is powerful enough to account for the prominence variation of accents within the phrase.

2.2 Example Input

```
(Phrase ((Type S) (PitchRange two))
    (Phrase ((Type C) ())
        (okay                  (H (d 1)) (B ()))
    )
    (Phrase ((Type C) ())
        (Phrase ((Type P) ())
            (i'll                           )
            (send                    (H (d)))
            (you                            )
        )
        (Phrase ((Type P) (PitchRange three))
            (a                       (B (i)))
            (form                    (H (d)))
        )
    )
)
```

A typical input to the system is given above. (Only the information relevant to the intonation module is shown, and thus words are given in their text form rather than their phonological form.) The scope of each phrase is shown by its bracketing. The first entry in the phrase is a list of features for that phrase, including its type (mainly used in the duration module) and its pitch range. The names and values of the pitch ranges are pre-defined. Here we simply use the names "one", "two", "three" etc. For example, pitch range "two" is defined as having a minimum F_0 values of 90Hz and a maximum of 220Hz. After each word there is an optional list of HLCB elements, with the features for each specified in brackets.

3 SYSTEM OPERATION

The system creates F_0 contours by taking the HLCB description and producing a RFC description, which is a linear sequence of *rise*, *fall* and *connection* elements, each with a duration and amplitude. From this RFC description equations are used to generate F_0 contours.

The equation for the fall element is given below and the equation for the rise element is the same as this but reflected in the y (F_0) axis. Connection elements are realised as straight lines.

$$\begin{array}{ll} f_0 = A(1 - 2(t/D)^2) & 0 < t < D/2 \\ f_0 = 2A(1 - t/D)^2 & D/2 < t < D \end{array} \tag{1}$$

Where t is time, A is the amplitude and D is the duration of the element.

3.1 HLCB and Pitch Range to RFC

HLCB elements re-write to RFC elements. **H** becomes a rise element followed by a fall element, **L** becomes a fall element followed by a rise element, **B** becomes a rise element, and **C** becomes a connection element.

A pitch accent requires five parameters; the rise amplitude and duration, the fall amplitude and duration, and some indication of how these elements are aligned with the segmental part of the utterance. At present, accent alignment is defined in terms of the distance between the start of the vowel in the syllable and the position of the peak (juncture between rise and fall elements in a **H** accent) or trough (juncture between fall and rise elements in a **L** accent). A *definitions table* (which is read from file at run time) determines basic default parameter values for each element. This table also specifies the modification that the features make on the elements (see section 4).

Most of the phonological tune specification is in terms of **H**, **L** and **B** elements. **C** elements need only be marked if they carry the feature "rising" as a **C** that is unmarked is assumed as the default. After the **H**, **L** and **B** accents are realised, rising **C** elements are added, and then any remaining part of the utterance which is not covered by an element is designated as being a connection element. After this, the entire utterance is specified in terms of RFC elements.

Pitch range information is used to determine the absolute (i.e. with respect to 0 Hz) amplitudes of the elements. The maximum and minimum values for the phrase pitch range are determined by the method described in section 2.1. If a phrase is explicitly marked as being in a particular pitch range, the defined minimum F_0 of that pitch range is used as the absolute amplitude of the start of the first element, otherwise the end amplitude of the previous element is used. In most cases the first element in a phrase is a $\mathbf{B}_i$ which serves to raise the F_0 level from the pitch ranges's minimum to a more medial level.

Once the starting amplitude of the phrase has been decided, the absolute amplitude of each element is calculated by adding its amplitude (which is negative in the case of fall elements) to the end amplitude of the previous element. If this operation results in the F_0 level exceeding the specified maximum or minimum of that phrase, the amplitudes are constrained so as the F_0 contour is kept within the designated pitch range. An additional phrase-final lowering rule (similar to that described in Liberman and Pierrehumbert [6]) rule states that if the last accent in a phrase is a $\mathbf{H}_d$ this always falls to the bottom of the pitch range.

3.2 RFC to F_0

The conversion of an RFC description to an F_0 contour is straightforward. The list of RFC elements is processed left to right and the equations are used to produce a continuous contour. Depending on the F_0 input for the particular synthesizer, this F_0 contour can be further processed. For example, some synthesizers require there to be a zero F_0 during unvoiced segments, and this can easily be achieved simply by masking the contour and setting it to zero in these regions. As yet, no segmental perturbations or micro-prosody effects are incorporated into the system.

4 DEFINING ELEMENTS AND FEATURES

At present, nearly all the operation of the system can be determined at run time by specifying how HLCB descriptions are mapped onto RFC descriptions. These parameters are kept in the definitions file which may be altered at will. The basic re-write rules cannot be varied, but the default sizes of the HLCB elements, and the effect the features have on those elements are all variable. It is also possible to specify what units these parameters are given in, e.g. both Hertz and semitones can be used to described amplitudes. Some typical parameters definitions are given below. Each statement consists of a variable, an operator, and a value. The "=" operator implies a straight assignment, an operator such as "+=" means "the existing value plus the new value" (in a similar syntax to the C programming language).

```
(define Element H                   (define Feature late
      (rise_amp    = 50 Hz)               (rise_amp      += 0 Hz)
      (fall_amp    = 50 Hz)               (fall_amp      += 0 Hz)
      (rise_dur    = 120 ms)              (rise_dur      += 100 ms)
      (fall_dur    = 160 ms)              (fall_dur      += 0 ms)
      (peak_pos    = 0.5 rel)             (peak_pos      *= 1.8 rel)
)                                   )
```

The first list defines the default **H** accent properties. In this example the accent amplitudes are defined in Hz, and the peak position is defined as occurring half way through the syllable, although it is possible to specify peak position in absolute ms terms also. The second list defines the behaviour of the late feature. In this definition, the values defined for the **H** accent are modified, such that everything is left unchanged except for the `rise_dur` which is increased by 100 ms and the `peak_pos` which is shifted back by a factor of 1.8. The feature definitions can also have non-modifying assignments, for example one could define downstepping accents as having no rise component, which would be achieved by setting stating `(rise_amp = 0 Hz) (rise_dur = 0 ms)`.

Current work is concentrated on deriving the definitions automatically by using the RFC analysis system on large sets of F_0 contours [9].

References

[1] J. Allen, S. Hunnicut, and D. Klatt. *From Text to Speech: the MITalk System.* Cambridge University Press, 1987.

[2] W. N. Campbell and S. D. Isard. Segmental durations in a syllable frame. *Journal of Phonetics*, 19:37–47, 1991.

[3] D. Robert Ladd. Phonological features of intonation peaks. *Language*, 59:721–759, 1983.

[4] D. Robert Ladd. Intonational phrasing: the case for recursive prosodic structure. *Phonlogy Yearbook 3*, pages 311–340, 1986.

[5] D. Robert Ladd. A model of intonational phonology for use with speech synthesis by rule. In *European Conference on Speech Technology*. ESCA, 1987.

[6] Mark Liberman and Janet Pierrehumbert. Intonational invariance under changes in pitch range and length. In Mark Aronoff and Richard T. Oehrle, editors, *Language Sound Structure*. MIT Press, 1984.

[7] J. D. O'Connor and G. F. Arnold. *Intonation of Colloquial English*. Longman, 2 edition, 1973.

[8] Paul A. Taylor. *A Phonetic Model of English Intonation*. PhD thesis, University of Edinburgh, 1992.

[9] Paul A. Taylor. Automatic recognition of intonation from F_0 contours using the rise/fall/connection model. In *Proc. Eurospeech '93, Berlin*, 1993.