Intonation Assignment in Text-to-Speech Synthesis

Mark Tatham – University of Essex
Kate Morton – University of Essex
Eric Lewis – University of Bristol

Paper read at the one-day meeting of the Institute of Acoustics, 29th January 1998. This is an early version of the authors' "Assignment of Intonation in a High-level Speech Synthesiser" published in *Proceedings of the Institute of Acoustics* (1998). St Albans: Institute of Acoustics. 255-262

Introduction

The basis of the Essex-Bristol **tts intonation model** lies in both linguistics and acoustics/aerodynamics. Our underlying principle is that in the human being there are physical processes intrinsic to the overall speech mechanism and that some of these processes are open to cognitive representation – such that they are able to enter into the domain of language processing.

Three types of intrinsic physical process are recognised:

- **incidental processes** intrinsic to the physical system which exist notwithstanding language, and which are tolerated by language – e.g. breathing, general mechanical and aerodynamic inertia;
- **intrinsic processes** of the physical system, but which can be *supervised* by cognitive intervention [Tatham 1995] – e.g. the progressive lowering or raising of sub-glottal air pressure, some mechanical and aerodynamic inertia;
- **extrinsic processes** which can be changed at will – at an extreme, reversed – e.g. vocal cord tension (negligible mechanical or aerodynamic inertia).

Most physical processes entering into language will be of type 3, but many will be of type 2 – that is, some intrinsic phenomena can be sufficiently supervised to be reliably included in language. Here there are two general requirements for language use [Tatham 1991]:

- a single sound or prosodic effect must be able to be replicated within production and perceptual constraints – a sound can be repeated and can be perceived as **'same'**;
- any two sounds or prosodic effects must be able to be produced reliably and repeatedly distinctly and perceived as **'different'**.

The physical basis of the model

We classify the progressive long-term raising or lowering of sub-glottal air pressure within type 2. Long-term here means over stretches of speech linguistically classified as greater than a word in length. We use the terms ***inclination*** and ***declination*** respectively to refer to these changes in sub-glottal air pressure. We use the same terms at the symbolic level to show correlation between levels.

- We regard the basic **long term** intrinsic 'direction' of change of rate of vocal cord vibration as being associated in speech with a falling sub-glottal air pressure – that is declination. We regard inclination as successfully supervised declination.
- **Short term** changes of fundamental frequency direction we suppose are brought about by local alterations of vocal cord tension, and thus constitute a modulation of

the current inclination or declination. Such short-term or sub-word length changes are classified within type 3.

- We recognise also a category of **mid-term** change in fundamental frequency direction, often of word length. Falling within type 2, human beings are able to supervise changing sub-glottal air pressure – within its general direction – to produce a mid-term 'push' in either direction. Thus a push can be overlaid to produce a mid-term increase or decrease in either downward or upward *trend* – we call this **turn-down** or **turn-up**.

The cognitive basis of the model

Cognitive processing is usually modelled as symbolic ***– intonation is the symbolic correlate of fundamental frequency change.*** Here there is *association* between cognitive and physical phenomena, and thus there is association between cognitive and physical representations; we are careful to make each plausible in the ears of the other – the associations are essentially principled [Tatham and Lewis 1992].

Speakers and listeners seem to be sensitive to a number of physical phenomena associated with the fundamental frequency of a speech wave – it is these for which we establish symbolic representations. Note that the sensitivity need not be conscious. Among such phenomena we have dealt with in our model are:

- a basic *f0* and intonational domain – the sentence;
- 'breaks' in the general *f0* trend (and in other prosodic phenomena) – end pointing a sub-domain: the intonational phrase;
- local *f0* changes within words;
- *f0* changes within the intonational segment – the syllable.

These are the physical parameters available for association with cognitive representation. For the speaker and listener there is a clear baseline of intonational expectation – a norm or neutral representation which lends itself to modification in 'special' cases for adding emotional or intentional content to the message being conveyed. Categories such as these, though often defined according to linguistic function, are used by many researchers, notably in recent times Pierrehumbert (1981) and Silverman *et al*. (1992).

[*Footnote:* The concept of 'neutral' intonation has been discussed by a number of researchers, notably Monaghan (1989), usually in terms of an acceptable intonation for synthesis which is not overly adventuresome – that is, is constrained in range and rate of change to minimise the impact of error. This is good practice in the design of the prosodic part of a tts system, but we introduce neutrality also on a theoretical basis. We are explicitly modelling the system as a two level process involving a basic neutral intonation and overlays for special effects. For us, neutrality of intonation contour is included less for practical reasons than for expressing a basic theoretical concept.]

So for example, declination must have a representation in terms of a useful symbol set. Since people report high-rate vocal cord vibration as producing sound high in pitch we use the symbol ***H*** for an intonational point which is reported as 'high'. ***L*** is similarly used for a low intonational point. The relationship between ***H*** and ***L*** and fundamental frequency is notional. A transition from ***H*** to ***L*** is thus declination, and a successful reversal of the direction as a transition from ***L*** to ***H*** is inclination. We referred earlier to our use of the word 'declination' for both a physical and a cognitive phenomenon: this is our key association between representations at these two levels.

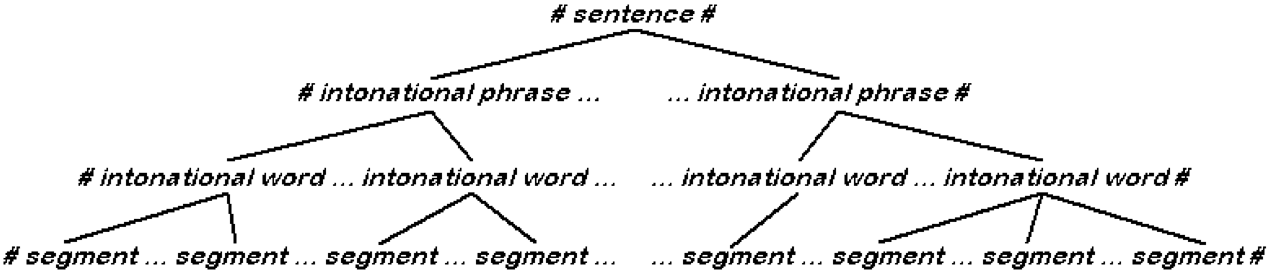
The symbolic representation

The top level domain of the symbolic representation is the ***sentence***. Here we represent ***slope*** – inclination and declination, e.g.

# L[………]H # - inclination

# H[………]L # - declination

Each sentence consists of one or more ***intonational phrases***. In our model intonational phrases are delimited according to sentence syntax. We assume that the sentence has been adequately parsed, using, in our model, a finite state grammar heavily dependent on lexical syntactic category markers attached to words in a dictionary module in the tts system. We also take advantage of the distribution of punctuation marks in the input text [O'Shaughnessy 1990] – though the creators of the texts are notoriously inconsistent in the use of punctuation. We have developed a set of heuristics which are able to assign boundary markers for intonational phrases dependent on the syntactic surface structure of the sentence. For example, a boundary marker is inserted immediately before a conjunction. Our intonational boundary marking is therefore linguistic. This contrasts with the statistical approach adopted by some researchers [Wang and Hirshberg 1991].


# sentence #
# intonational phrase ...
... intonational phrase #
# intonational word ... intonational word ...
... intonational word ... intonational word #
# segment ... segment ... segment ... segment ...
... segment ... segment ... segment ... segment #

Slope is represented here too as modulation of sentence slope, e.g.

# L[………]L H[………]H #
# H[………]H L[………]L #

Within each intonational phrase there are one or more ***intonational words***.

Intonational words comprise one or more ***intonational segments***. Intonational segments, 'syllables', are either ***stressed (S)*** or ***unstressed (U)***. Thus, e.g.

H[U | S U U | U | S U | U U S]L # – 'The furniture would vanish overnight.'

Pierrehumbert (1980) includes two 'tones', ***H*** and ***L***, in her tone sequence theory for assigning intonation in American English – our ***S*** and ***U*** are similar. Mertens (1990), however includes four tones in his model for French, and uses them in a slightly different way.

'Push' or mid-term changes in upward or downward trend in intonation – ***turn-up*** and ***turn-down*** – are symbolised by ***T+*** and ***T-*** respectively. These are phenomena which occur in neutral speech toward the end of intonational phrases. Thus, e.g.

H[S | S | U | S | S | S T-]H L[U | S | S … – 'He wore a pale blue shirt, a dark red …'

The accompanying diagrams [at the end of the paper] show three sentences:

- *'We have to chain the garden furniture down or it would vanish overnight.'*
- *'He wore a pale blue shirt, a dark red tie and light green socks.'*

- *‘Capital initials can, if the typography allows it, be rendered by small capitals.’*

For each of these we show:

1. The **waveform** of a human being pronouncing the sentence with neutral intonation. There is nothing canonical, though, about the pronunciation – variation for ‘same’ sentences is a commonplace.
2. The **measured** *f0* from the waveform. Once again, this is not a canonical version – just one of a number of possibilities.
3. A **symbolic representation** of the mark-up of the text as generated by our tts intonation model – this is *not* a mark-up of the waveform above.
4. The **calculated** *f0* based on the symbolic representation.

One additional symbol is present in the symbolic representation in the diagrams – ***F***. This mark is placed on the ***S*** intonational segment of the word having the greatest claim for assignment of ***focus*** within the sentence domain. Focus is an example of ***overlay*** – a term we use for effects which modulate (both symbolically and physically) the neutral intonation to produce special effects. We assign focus according to the sentence parse arrived at earlier; Sproat (1990) points out the need for such a parse in certain areas of English syntax.

[*Footnote:* In this presentation we do not discuss these overlay effects *per se* – but provision of pathways within the accompanying finite state transition network is represented by the ***[res]*** (reserved) symbol.]

Our fourth diagram shows the possibilities for symbolic representation of the intonation of objects within the intonational domains of sentence and intonational phrase. For purposes of comparison with the phonological model of intonation proposed by Pierrehumbert (on which our representation is a development) we have adapted her finite state transition network [Pierrehumbert 1980] to show the overall representational choices.

The diagram shows five nodes in the network – all five of which are involved in representing the possibilities within the intonational phrase domain. The connection between the final and initial nodes indicates the possibility (unconstrained here) of **sequenced intonational phrases**. The initial node and node 1 are linked by declination markers, as are node 3 and the final node: these outermost connections establish **declination** or its modification to **inclination**. Connections between nodes 2 and 3 determine **turn-up or turn-down**. Intonation representation for intonational words and segments is handled between nodes 1 and 2. This part of the diagram has been expanded separately. Here the top connection establishes the possibility (unconstrained here) of **sequence**, and other connections indicated ***S*** or ***U*** symbols establish **stress** possibilities. The connections labelled ***[res]*** are reserved ‘hooks’ on which we peg further symbols used in the representation of ***pragmatically determined overlays*** [not discussed in this presentation, but see Morton (1992) and Morton and Tatham (1995)].

From symbolic representation to physical representation

In the Essex/Bristol tts intonation model we have a **quasi-abstract physical representation** standing between our symbolic representation and the final *f0*. There are two reasons for this:

- we believe that the transition between the highly symbolic representation and the *f0* to be calculated is eased by this **intermediate representation**;
- we have incorporated this quasi-abstract representation to provide a hook for the rendering of **different voices** by the system – each with its own different *f0* range.

We define therefore an *f0* range for ‘a voice’. The highest *f0* to be expected for a particular voice is assigned a value of 63 and the lowest *f0* for the voice is given a value of 0;

the range is therefore quantised linearly into 64 levels. As an example of how this works we might assign to the first ***S*** segment within an intonational phrase the value 40 and to the last ***S*** segment the value 20. This establishes the declinational baseline for this sentence for this speaker and all ***S*** segments are notionally allocated a value associated with this baseline.

U segments derive their values from their surrounding ***S*** segments (except for phrase-leading and -trailing ones). In an intonational phrase having a declination baseline, for example, a sequence of one or more ***U*** segments drops sharply from the ***S*** preceding it to 'recover' *f0* as the sequence approaches the ***S*** following it. We have introduced a number of rules which deal with how a sequence of ***U*** segments relate to one another *within* this general recovery of *f0*: this removes awkward perceptual effects associated with too linear a movement of *f0*.

T+ and ***T-*** (turn-up and turn-down) are in general given a local domain of a single intonational word. For a good percentage of the time spent on the word unit *f0* is incremented or decremented beyond the normal expectation to produce the special effect. The percentage of the word depends on the ***S*** and ***U*** sequence within the word and on its position within the intonational phrase. In the accompanying diagrams there are examples of ***T-*** occurring finally in intonational phrases.

Finally, the entire quasi-abstract representation of *f0* is smoothed to remove abrupt transitions between values and to minimise the quantisation error introduced by the abstraction. This smoothing is varied for special effect – but in the accompanying diagrams is set to its minimum value throughout. At this point the representation is translated into an actual *f0* contour by defining the appropriate voice range.

Special effects

We have referred often in this paper to 'special effects'. We are using this as a cover term for intonational effects which go *beyond* descriptions of normal utterances to embrace the whole gamut of **pragmatically determined variations.** Intonation is not, of course, the only parameter used in the rendering of these effects – rhythm and stress also play their roles. We have been modelling these effects as overlays upon the neutral contour generation described in this paper; it seems to us that this is a good route toward handling the enormous problem of variability in modelling intonational effects which convey phenomena such as emotion and intention. Here it is enough to say that the basic model has been designed assuming the general overlay concept; we have therefore built in various hooks and other devices which can ensure the extensibility of the model into situations where neutral intonation is inappropriate.

Conclusion

In this paper we have presented the salient properties of the Essex/Bristol text-to-speech model of intonation. The model has a number of features and reflects our general philosophy of factoring out intrinsic and extrinsic physical phenomena to create associations between physical and cognitive representations. The model is linguistically rather than statistically based and is generalisable to assign intonation for many voices rather than being tied to one single voice. The model is transparently extensible to handle variability beyond the neutral rendering of intonation using the concept of overlays to incorporate pragmatically determined intentional and emotional effects.

References

Mertens, P. (1990) Intonation. In C. Blanche-Benveniste *et al*. (eds.) *Le français parlé*. Paris, Editions du CNRS

Monaghan, A.I.C. (1989) Phonological domains for intonation in speech synthesis. *In Proceedings of Eurospeech 89*, pp. 502-506. Paris, ESCA

Morton, P.K.J. (1992) Pragmatic phonetics. In W.A. Ainsworth (ed*.), Advances in Speech, Hearing and Language Processing*, pp. 17-55. London, JAI Press

Morton, P.K.J., with M.A.A. Tatham (1995) Pragmatic effects in speech synthesis. In J. Pardo (ed.), *Proceedings of Eurospeech '95*, pp. 1819-1822. Madrid, ESCA

O'Shaughnessy, D. (1990) Relationships between syntax and prosody for speech synthesis. *In Proceedings of the ESCA Tutorial on Speech Synthesis*, pp. 39-42. Autrans, ESCA

Pierrehumbert, J. (1980) *The Phonology and Phonetics of English Intonation.*. PhD dissertation, MIT, Indiana University Linguistics Club

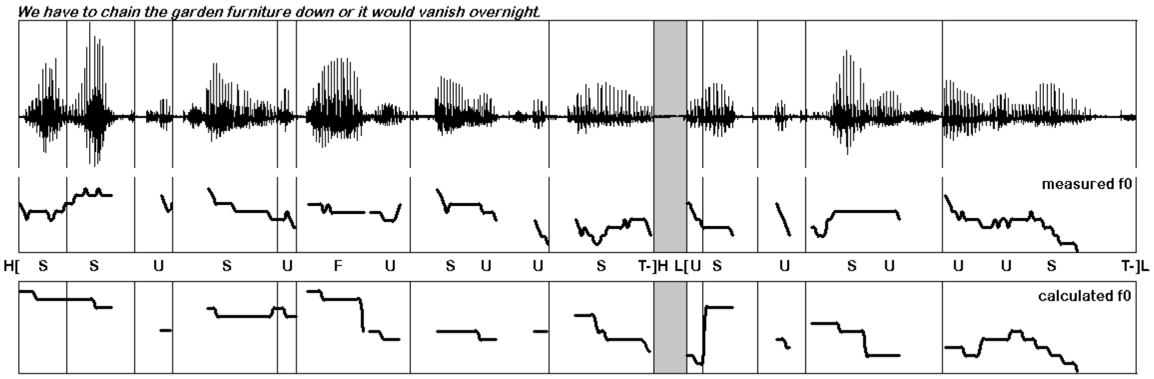
Pierrehumbert, J. (1981) Synthesizing intonation. *Journal of the Acoustical Society of America*, Vol. 70(4), pp. 985-995

Sproat, R. (1990) Stress assignment in complex nominals for English text-to-speech. *In Proceedings of the ESCA Workshop on Speech Synthesis*, pp. 129-132. Autrans, ESCA

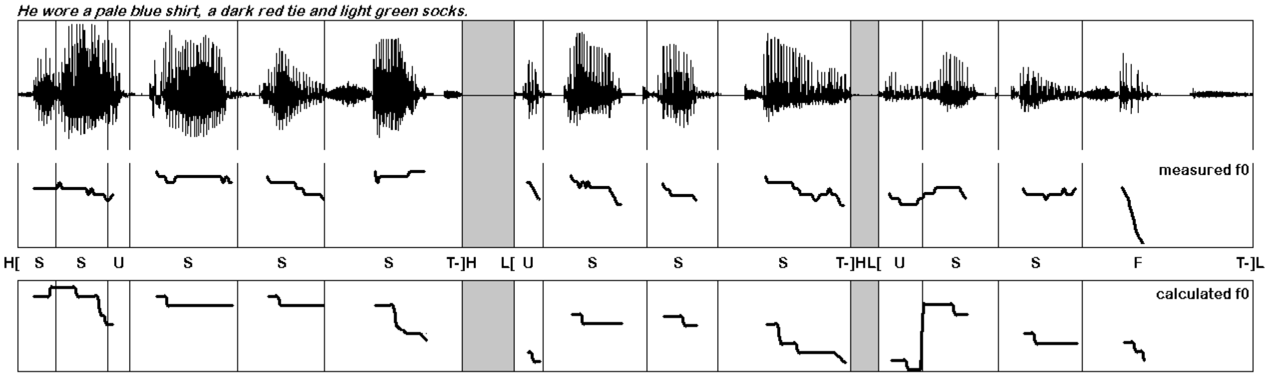
Tatham, M.A.A. (1991) Cognitive Phonetics. In W.A. Ainsworth (ed*.) Advances in Speech, Hearing and Language Processing*, Vol.1, pp. 193-218. London, JAI Press

Tatham, M.A.A. and Lewis, E. (1992) Prosodic assignment in *SPRUCE* text-to-speech synthesis. In R. Lawrence (ed.), *Proceedings of the Institute of Acoustics*, Vol.14. St. Albans, IoA. Tatham, M.A.A. (1995) The supervision of speech production. In C. Sorin *et al*. (eds*.) Levels in Speech Communication*, pp. 115-125. Amsterdam, Elsevier

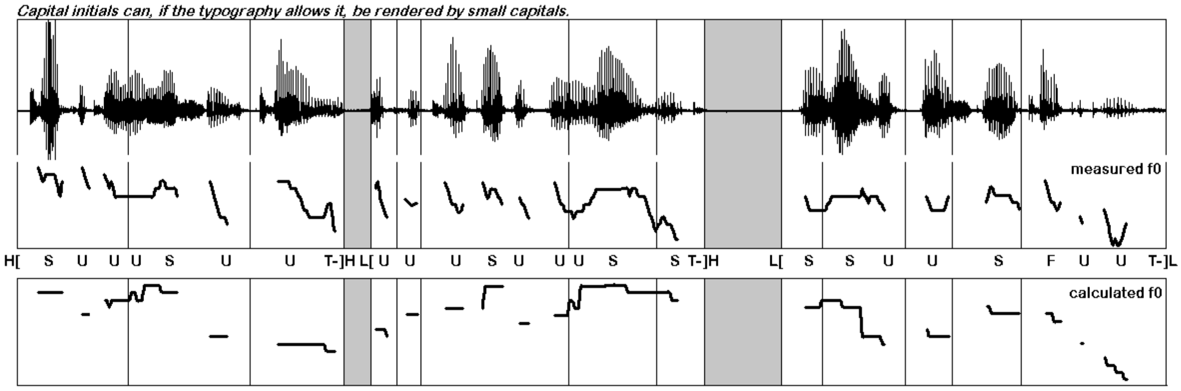
Silverman, K., M. Beckman, J. Pitrelli, M. Ostendorf, C. Whightman, P. Price, J. Pierrehumbert and J. Hirshberg (1992) ToBI: a standard for labeling English prosody. *In Proceedings of the XIIIth International Congress of Phonetic Sciences*, Vol.2, pp. 630-633



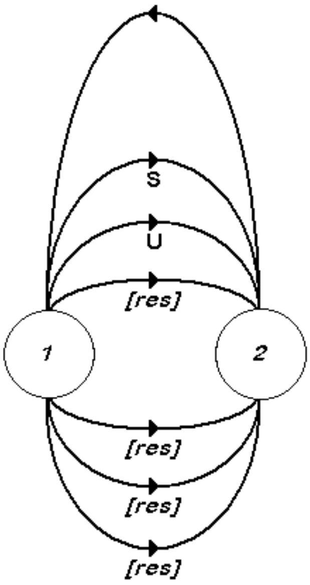
We have to chain the garden furniture down or it would vanish overnight.



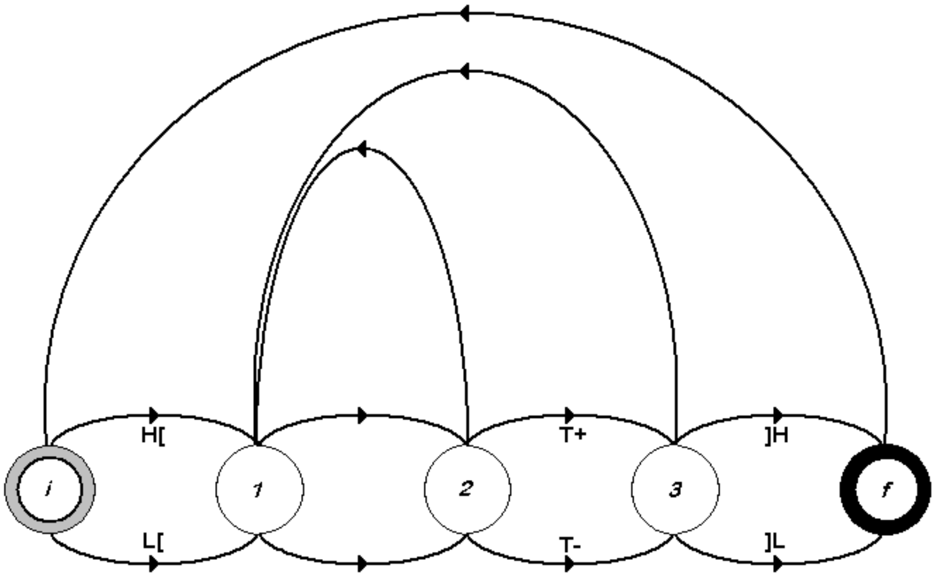
He wore a pale blue shirt, a dark red tie and light green socks.



Capital initials can, if the typography allows it, be rendered by small capitals.



S
U
[res]
1
2
[res]
[res]
[res]



H[
T+
]H
i
1
2
3
f
L[
T-
]L