A NEW INTONATION MODEL FOR TEXT-TO-SPEECH SYNTHESIS

Katherine Morton*, Mark Tatham*, and Eric Lewis†
**University of Essex, UK, †University of Bristol, UK*

ABSTRACT

The text-to-speech intonation model we are developing derives from both linguistics, and the acoustics and aerodynamics of speech production. Our underlying premise is that in human speech production there are physical processes intrinsic to speech production, and that some of these processes can be cognitively represented – they can therefore become part of the domain of language processing. The model is based on our general philosophy of factoring out intrinsic and extrinsic physical phenomena to create associations between physical and cognitive representations. The model is easily extended to handle variability beyond the neutral rendering of intonation, using overlays to add pragmatically determined intentional and emotional effects.

1 INTRODUCTION

The text-to-speech (tts) intonation model described here has both physical and cognitive bases. We begin by identifying in speech production a number of physical processes intrinsic to the speech mechanism, some of which are amenable to cognitive representation – this means they can enter into the symbolic domain of language.

The model defines three types of physical process:

- **Incidental processes** which are intrinsic to the physical system; these do not interfere with language but are not directly involved in any encoding of linguistic content – e.g. breathing, *general* mechanical and aerodynamic inertia.
- **Intrinsic processes** of the physical system, which can be *monitored* and *supervised* by cognitive intervention [1] – e.g. the progressive lowering or raising of sub-glottal air pressure, or *some* mechanical and aerodynamic inertia like *voice onset time* which differs systematically between languages and at the same time is basically an intrinsic phenomenon (processes of this type are the focus of **Cognitive Phonetic Theory** [2]). Supervised intrinsic processes contribute to the phonology of the language.
- **Extrinsic processes** which can be changed at will or even reversed if necessary – e.g. vocal cord tension; such processes are assumed to have *negligible* mechanical or aerodynamic inertia (these processes are the focus of **Classical Phonetic Theory** [3]). Extrinsic processes contribute to the phonology of the language, and any accompanying intrinsic coarticulatory or coproduction processes are disregarded.

Thus the model distinguishes between directly controlled processes not significantly constrained by processes intrinsic to the system (type 3), processes which manipulate existing intrinsic processes to make them significant (type 2) and processes which are largely ignored in language encoding (type 1).

Most physical processes in speech production are type 3, but many are type 2 – that is, some intrinsic phenomena can be sufficiently supervised to be reliably included in language. There are two general requirements for use in language [2]:

- a sound or prosodic effect must be able to be replicated within production and perceptual constraints; this simply means that any one sound must be able to be reliably repeated in such a way as to be perceived as the *same* sound each time it is repeated;
- any two sounds or prosodic effects which are *intended to be different* must be able to be produced reliably and repeatedly distinctly and perceived as *different* sounds.

These two criteria are the basis of phonological speech patterning – a cognitive representational system enabling speakers and hearers to have a shared understanding of which sounds are the same and which are different

2. THE MODEL'S PHYSICAL BASIS

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 longer than a word. 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 higher symbolic level to imply correlation between physical and symbolic representations.

- We regard the basic **long term** intrinsic direction of change of rate of vocal cord vibration as being associated with falling sub-glottal air pressure – that is declination. We regard inclination as successfully supervised declination. In this model sub-glottal air pressure is progressively falling, unless it is actively manipulated to rise.
- **Short term** changes of fundamental frequency direction are brought about by local alterations of vocal cord tension, and thus constitute a modulation of the current inclination or declination.
- We recognise **mid-term** change in fundamental frequency direction, often of word length. 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**.

3. THE MODEL'S COGNITIVE BASIS

Cognitive processing in language is modelled symbolically, and intonation is the *symbolic correlate* of fundamental frequency change at the acoustic level**.** We assume there is *association* between cognitive and physical phenomena, and thus there is association between corresponding cognitive and physical representations. We are careful to make each representation transparently associated with the other, that is, the associations are principled [4]

Speakers and listeners seem to be linguistically sensitive to a number of physical properties of fundamental frequency, and

these must figure in our symbolic representations. Among the properties we have included in the model are:

- a basic *f0* and intonational domain called the **sentence**;
- 'breaks' in the general f0 trend which often serve to end-point subdomains called the intonational phrases;
- local f0 changes within intonational words;
- f0 changes within basic units called intonational segments; these correspond to syllables.

These are the physical parameters available for association with cognitive representations.

For both speakers and listeners there is a clear baseline of expectation for intonation – a norm or neutral representation which can be modified in *special* cases for adding emotional or intentional content to the message being conveyed [5] [6]. Categories such as these, though often defined according to linguistic function rather than in terms of physical parameters, are used by many researchers, notably in recent times Pierrehumbert [7] and Silverman *et al*. [8]. The concept of *neutral intonation* has been discussed by a number of researchers, notably Monaghan [9], usually in terms of an acceptable intonation for synthesis constrained in range and rate of change to minimise the impact of error. This is good practice in the design of tts prosodics. However we introduce the idea of neutrality here *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. Thus we introduce the concept of neutrality not for practical reasons, but as an important part of our theory.

To give an example of how we explicitly relate cognitive and physical representations, take declination – a physical event which must also have a symbolic representation. 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 (after Pierrehumbert [7] and Silverman *et al.* [8]). 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.

4. THE SYMBOLIC REPRESENTATION

The top level domain of the symbolic representation is the ***sentence***. We represent sentence-wide ***slope*** (a generic term) – inclination and declination, e.g.

L[……]H # – *inclination*
H[……]L # – *declination*

In the representation the sentence domain is bounded by #. Since declination and inclination take in the entire sentence their markers L, and H are used to bracket the sentence itself. L goes to H for inclination and H goes to L for declination

Each sentence has one or more ***intonational phrases***. Intonational phrases are defined by the sentence syntax. The sentence is parsed using a finite state grammar heavily dependent on syntactic category markers on words in the tts dictionary module. We also take advantage of the distribution of punctuation marks in the input text [10]. We have developed a set of heuristics which assign boundary markers for intonational phrases depending on the sentence surface syntactic structure. For example, a boundary marker is inserted immediately before a conjunction. Our intonational boundary marking is therefore linguistic in origin. This contrasts with the statistical approach adopted by some researchers [11].

Local 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*** and these comprise one or more ***intonational segments***. Intonational segments, syllables [12], 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.

[For the sake of comparison with other researchers, we can note that Pierrehumbert [13] includes two 'tones', ***H*** and ***L***, in her **tone sequence theory** for assigning intonation in American English – our ***S*** and ***U*** are similar. Mertens [14] 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 (Figs. 1 and 2 – grouped at the end of the text) show two sentences:

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

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

For each of these we show:

- A human **waveform** with neutral intonation (author MT). There is nothing canonical about the pronunciation – the sentences could easily have been spoken differently.
- The waveform's **measured *f0***. Again, not a canonical version – just one possibility our subject happened to use.
- 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 but the way our system assigned a representation.
- The **calculated *f0*** based on the generated 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 which has 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 [15] points out the need for such a parse in some areas of English syntax. In this paper we do not discuss these overlay effects *per se* – but provision of pathways within the finite state transition network in Fig.3 is represented by the ***[res]*** (reserved) symbol.

Fig.3 shows the possibilities for symbolic representation within the intonational domains of sentence and intonational phrase. This diagram is included to enable comparison with the phonological model of intonation proposed by Pierrehumbert (of which our representation is a development) we have adapted her

finite state transition network [13] to show *our* overall representational choices.

The diagram shows five nodes in the network involved in representing possibilities in the intonational phrase domain. The connection between the final and initial nodes indicates the possibility (here unconstrained) 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 words and segments is handled between nodes 1 and 2. This part of the diagram has been expanded separately. In the expansion 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' to peg other symbols used in the representation of ***pragmatically determined overlays*** (see Morton [5] and Morton and Tatham [6]).

5. FROM SYMBOLIC TO PHYSICAL REPRESENTATION

In our tts intonation model we move between the symbolic representation outlined above and the final *f0* by means of a *quasi-abstract physical representation*. There are two reasons:

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

1. We define 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.

2. ***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 sequences of ***U*** segments relate to one another *within* this general recovery of *f0*. This removes any awkward perceptual effects associated with too linear a movement of *f0*.

3. ***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 Figs.1-2 there are examples of ***T-*** occurring finally in intonational phrases.

4. 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 examples it 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.

6. SPECIAL EFFECTS

We have referred several times to *special effects* – a cover term for intonational effects going *beyond* descriptions of normal utterances to embrace the whole gamut of **pragmatically determined variations** [5]. Intonation is not the only parameter used in rendering such effects – the other prosodic phenomena of rhythm and stress are also involved. We have been modelling these effects as overlays on neutral contours generated by the model described here. It seems to us that this is a good route toward handling the variability problem in modelling intonational effects conveying phenomena such as emotion and intention. In this paper we have not dealt with these effects, and the basic model has been designed assuming the general overlay concept. We have built in various hooks and other devices to ensure the *extensibility* of the model into situations where the most basic neutral intonation is inappropriate.

7. CONCLUSION

We have presented the major properties of our tts intonation model. The model has a number of features reflecting our approach of factoring out intrinsic and extrinsic physical phenomena to create associations between physical and cognitive representations. The model is linguistically, not statistically based, and is generalisable to assign intonation for many voices rather than just 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

[1] Tatham, M. 1995. The supervision of speech production. In C. Sorin, J. Mariani, H. Meloni and J. Schoentgen (eds.) *Levels in Speech Communication – Relations and Interactions*, 115–125. Amsterdam: Elsevier

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

[3] Gimson, A.C. 1989. *An Introduction to the Pronunciation of English*. London:Arnold

[4] Tatham, M. and Lewis, E. 1992. Prosodic assignment in *SPRUCE* text-to-speech synthesis. In R. Lawrence (ed.), *Proceedings of the Institute of Acoustics*, 14. St. Albans: Institute of Acoustics

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

[6] Morton, K. and Tatham, M. 1995. Pragmatic effects in speech synthesis. In J. Pardo (ed.), *Proceedings of Eurospeech '95*, 1819-1822. Madrid: ESCA

[7] Pierrehumbert, J. 1981. Synthesizing intonation. *Journal of the Acoustical Society of America*, 70:4, 985-995

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

[9] Monaghan, A.I.C. 1989. Phonological domains for intonation in speech synthesis. In *Proceedings of Eurospeech 89*, 502-506. Paris: ESCA

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

[11] Wang, M.Q. and Hirschberg, J. 1991. Predicting intonational boundaries automatically from text: the ATIS domain. *Proceedings of the DARPA Speech and Natural Language Workshop*, 378-383
[12] Tatham, M. and Lewis, E. 1998. Syllable recovery from polysyllabic words. In *Proceedings of Speech and Hearing 98*. St Albans: Institute of Acoustics
[13] Pierrehumbert, J. 1980. *The Phonology and Phonetics of English Intonation.*. PhD dissertation, MIT, Indiana University Linguistics Club
[14] Mertens, P. 1990. Intonation. In C. Blanche-Benveniste *et al*. (eds.) *Le français parlé*. Paris: Editions du CNRS
[15] Sproat, R. 1990. Stress assignment in complex nominals for English text-to-speech. In *Proceedings of the ESCA Workshop on Speech Synthesis*, 129-132. Autrans: ESCA

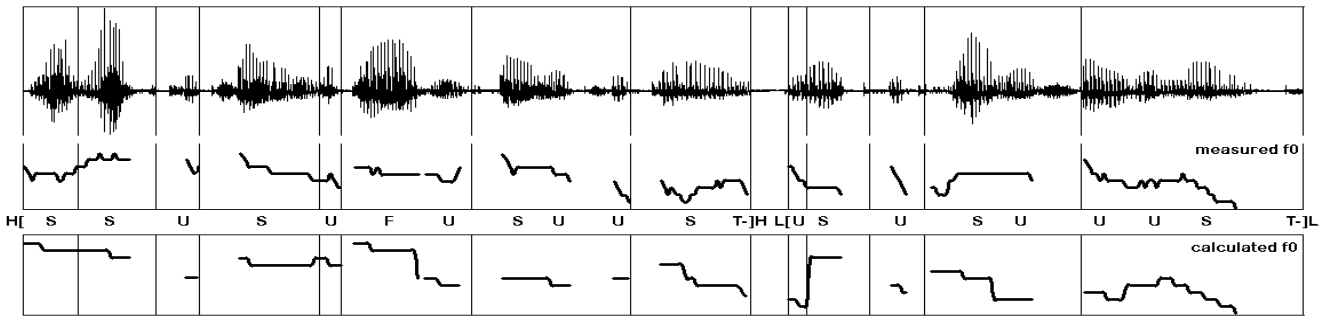

Fig.1 *We have to chain the garden furniture down or it would vanish overnight* – showing a. an example human waveform, b. the measured f0, c. generated text symbolic mark-up, and d. the calculated f0.

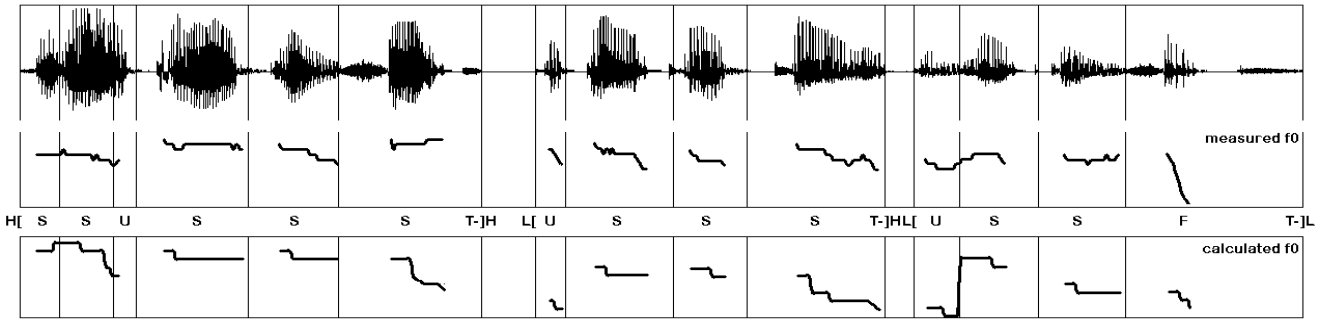

Fig.2 *He wore a pale blue shirt, a dark red tie and light green socks* – showing a. an example human waveform, b. the measured f0, c. generated text symbolic mark-up, and d. the calculated f0.

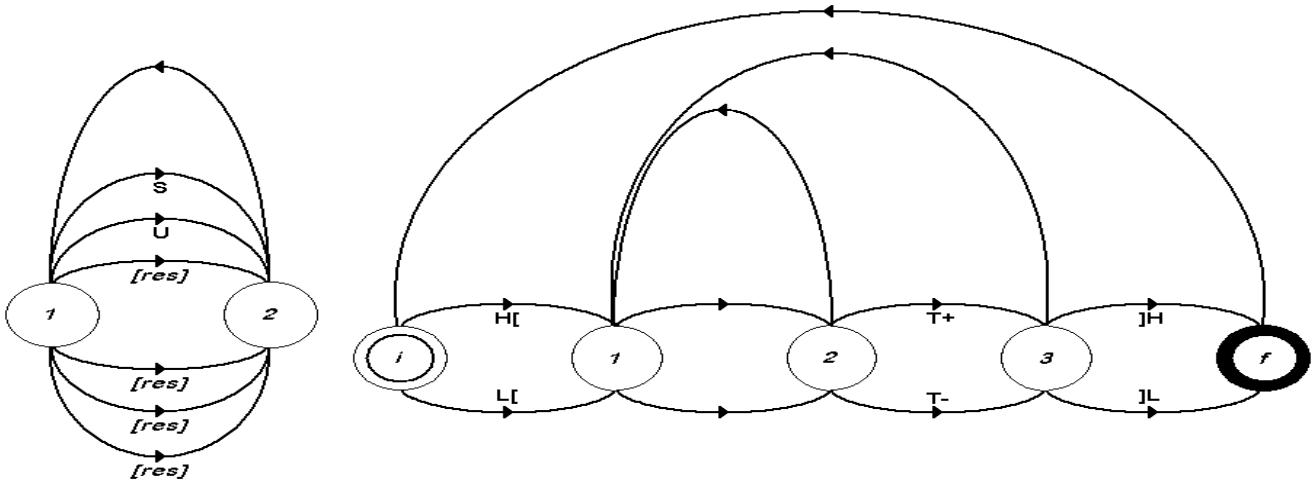

Fig.3 Finite state transition network showing the overall symbolic representational choices in our tts intonation model.