AN ADVANCED INTONATION MODEL FOR SYNTHESIS

Mark Tatham[3]*, Eric Lewis*[2] *and Katherine Morton*[1]
[1]University of Essex UK, [2]University of Bristol UK

ABSTRACT

The advanced intonation model for speech synthesis described here has a three level architecture. An initial abstract characterisation designed to represent intonation at the level of cognitive percept is rewritten to an intermediate representation which is speaker independent, yet which accurately reflects physical pitch contours. At this stage the contours lack the variability associated with natural speech. This representation is then further rewritten to provide an actual physical contour (now including variability and other 'natural' phenomena such as micro-intonation). One or two examples are given for stages one and two, and some indication of how we tackle stage three.

1. INTRODUCTION

Our synthesis intonation model has both physical basis and cognitive bases. We identified 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. An earlier paper [1] recognised a number of different types of physical process, pointing out that we could usefully distinguish between directly controlled processes not significantly constrained by processes intrinsic to the system, processes which manipulate or supervise [2] existing intrinsic processes to make them significant, and processes which are largely ignored in language. These concepts are in line with the principles established in the theory of ***Cognitive Phonetics*** (CP)[3].

CP puts forward two principles – conditions which must be met for a physical process to be usable:

- a sound or a prosodic effect must be able to be replicated so as to be perceived as the *same* sound or effect each time it is repeated – the ***principle of perceived identity***;
- 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 – the ***principle of perceived difference.***

These two CP principles are the basis of much phonological speech patterning – a cognitive symbolic representational system giving speakers and hearers shared understanding of which sounds are the same and which different, and hence which can and which cannot be used contrastively for representing morphemes in any particular language.

2. PHYSICAL INPUT

Within CP the progressive long-term lowering of sub-glottal air pressure while a sentence unfolds (declination) is treated as an intrinsic process, able to be systematically modified to produce long-term raising (inclination).

- **Long term** (longer than a word) intrinsic direction of *f0* change we associate with falling sub-glottal air pressure – *declination. Inclination* though is supervised [2] declination. In the underlying speech production model sub-glottal air pressure is progressively falling, unless it is actively manipulated to rise.
- **Mid-term** (often of word length) changes in *f0* direction are also important. Speakers can supervise changing sub-glottal air pressure to produce a mid-term 'push' in either direction, toward faster or slower vocal cord vibration. Push can be overlaid to produce a mid-term increase or decrease in downward or upward *trend* – we call this **turn-down** or **turn-up.**
- **Short term** (within a word) changes of *f0* direction arise, we assume, with local alterations of vocal cord tension, modulating the current inclination or declination.

3. COGNITIVE INPUT

Intonation is the cognitive *symbolic correlate* of *f0* change at the acoustic level**.** We assume there is *association* between cognitive and physical phenomena, and therefore the possibility of principled association between the corresponding cognitive and physical representations [4].

Because speakers and listeners are linguistically sensitive to a number of physical properties of *f0*, these must figure in our symbolic representation. Among the properties we have included are:

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

Speakers and listeners seem to have a baseline ex-

pectation for intonation – a normative representation which can be modified in *special* cases for adding emotional or intentional content to the message being conveyed [5] [6] – we call this norm the 'neutral intonation'. Similar categories, 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].

We favour the idea of a neutral contour *on a theoretical basis*, explicitly modelling the system as a two level process involving neutral intonation and then overlays for special effects.

As 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. Declination is a transition from ***H*** to ***L*** and a successful supervised reversal of the direction as a transition from ***L*** to ***H*** is inclination (after Pierrehumbert [7] and Silverman *et al.* [8]).

4. SYMBOLIC REPRESENTATION

The top domain of our symbolic representation is the ***sentence*** – though we do extend the model to paragraph intonation, not discussed here. We have already described [1] the need to represent sentence-wide *global* ***slope*** (a generic term) – inclination and declination, hence

L[……]H # – *inclination*

H[……]L # – *declination*

Each sentence has one or more ***intonational phrases***, and within these *l****ocal slope*** is represented as a 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 [9], 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.

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 …

One additional symbol is used – ***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 neutral intonation to produce special effects – in this case a local emphasis. Focus is assigned using a sentence parse. In this paper we do not discuss these overlay effects *per se*.

5. RELATING THE SYMBOLIC AND PHYSICAL REPRESENTATIONS

All high level synthesis systems reach the point in their procedures where the move has to be made from symbolic to physical representations. The idealised symbolic representation is characterised by lack of variability, contrasting sharply with the variation which is a key feature of the physical signal. The bulk of effort in the development of traditional phonetics and phonology has been in the direction of physical to symbolic – removal of variability – the task here is the reverse. We tackle the problem by progressing through levels of abstraction, in particular identifying an intermediate level. There are two good reasons for this:

a. it allows us to generate a notional *f0* curve without any particular frequency instantiation – this defers calculating actual physical values until we have an overall picture (for at least the sentence or paragraph domains) of *f0* trend;
b. it eases (but does not solve) the perennial problem [3] of relating abstract with concrete, cognitive with physical, etc.

We define an *f0* range for an individual voice. The overall *f0* available has values from 0 to 63, with the range for a voice falling within this. 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 and all ***S*** segments are notionally allocated a value associated with this baseline.

As we shall see, ***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 caused by 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.

Finally, the entire semi-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 (Figs. 1 and 2) 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 (not detailed here).

Let us work through an example (Fig. 2): *'He wore a pale blue shirt, a dark red tie and light green socks.'*

H[S | S | U | S | S | S T-]H
L[U | S | S | S T-]H
L[U | S | S | F T-]L #

1. Using our notional pitch range of 0-63 we begin by assigning, for this speaker, 32 throughout the first S; we assign 27 to the last *F* (an enhanced *S*). We regard these two stressed syllables as being 'keystone', anchor or pivot points for end-pointing frequency drop through a complete sentence.
2. Still within the # - # sentence domain, the value assigned to intermediate *S*s shows a regular drop from 32 to 27, in this instance.
3. The sentence *F* is assigned a value slightly higher than that generated by 2 – whilst not interfering with the overall drop pattern.
4. Syllables marked *U* are dropped further than their surrounding *S*s.
5. Boundaries – that is,]*X X*[– within a sentence establish phrase sub-domains – with the pitch at this point needing what we term 'reset'. Any preceding *S* has its pitch dropped.
6. The *S* following a reset point has its value raised, with subsequent *S*s adjusted to provide a smooth drop.
7. Turndown – *T-* – is implemented here by dropping the preceding *S* or *F*.

The result of applying these rules can be seen as the calculated *f0* curve in Fig. 2.

Note that the intrinsic declination baseline (#H[…]L#) is represented at the symbolic level. The *S* and *U* markers are placed about the declination baseline to bring them within the level of abstraction of our intermediate representation.

6. MICRO-INTONATION

Micro-intonation and special effects (see *Section 7*) are examples of adjunct modelling – models off to one side built for characterising non-core phenomena. Our system handles micro-intonation (local perturbations of the *f0* slope caused by coarticulatory or co-production aerodynamic effects) in either of two possible ways:

- The low-level word and syllable models assembled in the database used for concatenative synthesis [9] are held with their original *f0* preserved prior to normalisation and re-calculation. Pitch periods associated with micro-intonation are identified and blocked from participating in *f0* re-calculation when inserted into newly-generated sentences. The final perceptual effect is normally good, but fails in extreme use of the *f0* range. The approach is theoretically unsound, because it does not relate micro-intonation to the specific *f0* contour being generated for the current sentence.
- The *f0* trend within each database word or syllable model is used to calculate its relationship to attendant micro-intonation effects. This relationship is then used to re-calculate micro-intonation for any newly generated overall *f0* contour. Whilst computationally more complex than the alternative method, this approach is both theoretically more sound and perceptually more satisfactory.

7. SPECIAL EFFECTS

Special effects have been mentioned several times – this is a cover term for intonational effects going *beyond* descriptions of normal utterances to take in, for example, **pragmatically determined variations** [6]. Intonation is not the only parameter used in rendering such effects – the other prosodic phenomena of rhythm and stress are also involved. These effects are modelled as overlays on the neutral contours generated by the core model. It seems to us that this is a good approach to modelling intonational variants conveying emotional and intentional effects. In this paper we have not dealt with these, but the basic model does assume the general overlay concept. We have built in various hooks and other devices to ensure the *extensibility* of the model into those situations where the most basic neutral intonation is inappropriate.

8. CONCLUSION

Our intonation model is both cognitively and physically based, and is sufficiently generalised to assign intonation for many voices rather than just one single voice. The architecture calls for a symbolic representation of intonation and a representation of the physical *f0*. We have presented some heuristics for

deriving an *intermediate f0* from the symbolic intonation contour. The model is transparently extensible to phenomena beyond a neutral rendering of intonation, using the concept of overlays to incorporate pragmatically determined intentional and emotional effects.

7. REFERENCES

[1] Morton, K., Tatham, M. and Lewis, E. 1999. A New Intonation Model for Text-to-Speech Synthesis. In J. Ohala [ed.] *Proceedings of the International Congress of Phonetic Sciences*, San Francisco

[2] 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

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

[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] Tatham, M. and Lewis, E. 1999. Syllable reconstruction in concatenated waveform speech synthesis. In J. Ohala [ed.] *Proceedings of the International Congress of Phonetic Sciences*, San Francisco

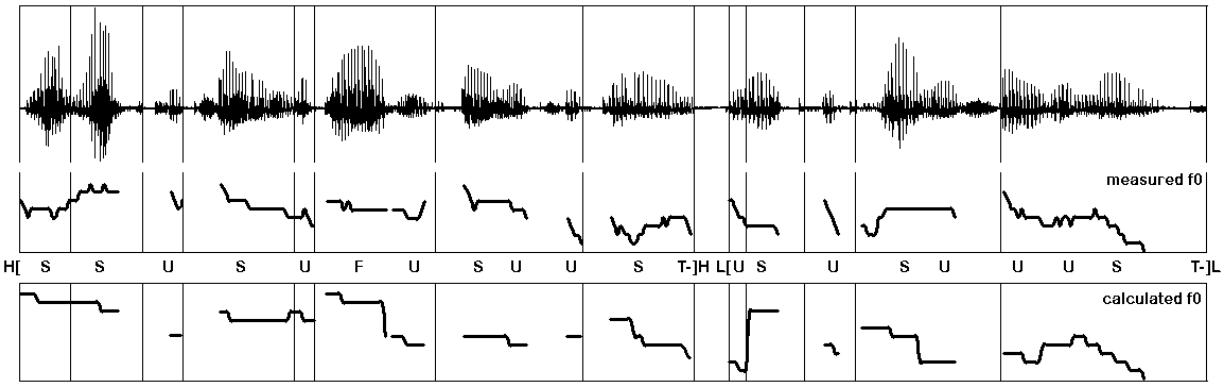

Fig. 1 *We have to chain the garden furniture down or it would vanish overnight* – showing *a*. an example human waveform, *b*. the measured *f0*, *c*. generatied text symbolic mark-up, and *d*. calculated intermediate f0.

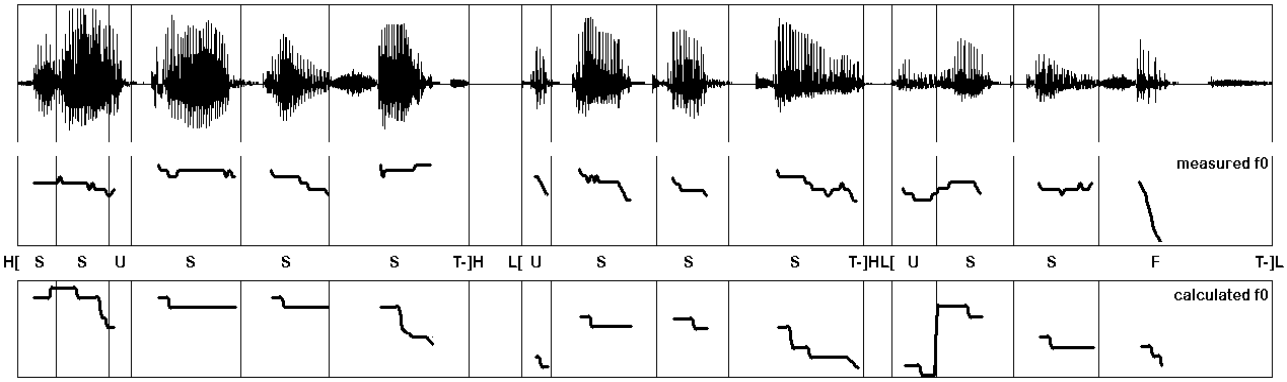

Fig. 2 *He wore a pale blue shirt, a dark red tie and light green socks* – showing *a*. an example human waveform, *b*. the measured *f0*, *c*. generated text symbolic mark-up, and *d*. calculated intermediate *f0*.