ASSIGNMENT OF INTONATION IN A HIGH-LEVEL SPEECH SYNTHESISER

Mark Tatham
Katherine Morton
Eric Lewis – University of Bristol

Reproduced from Proceedings of the Institute of Acoustics. St Albans: Institute of Acoustics. 255-262

INTRODUCTION

The text-to-speech **intonation model** we are developing derives from both linguistics and the acoustics and aerodynamics of speech production. Our initial premise is that in human speech production there are physical processes intrinsic to the speech mechanism, and that some of these processes are open to cognitive representation – as such they are able to enter into the domain of language processing.

The model defines three types of physical process:

1. **Incidental processes** which are intrinsic to the physical system and which exist apart from language; these are tolerated by language but are not directly involved in any encoding of linguistic content – e.g. breathing, *general* mechanical and aerodynamic inertia.
2. **Intrinsic processes** of the physical system, which can be *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 (these processes are the focus of **Cognitive Phonetic Theory** [2]). Supervised intrinsic processes contribute to the phonology of the language.
3. **Extrinsic processes** which can be changed at will and, at an extreme, 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 **Articulatory Phonetic Theory** [3]). Extrinsic processes contribute to the phonology of the language, and any accompanying intrinsic coarticulatory or coproduction processes are disregarded.

To clarify: the model distinguishes therefore between directly controlled processes which are 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).

Many or perhaps 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.

If these two criteria are met we have the basis of phonological speech patterning – a system enabling speakers and hearers to have a shared understanding of which sounds are the same and which are different

THE PHYSICAL BASIS OF THE INTONATION 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 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 in speech with a falling sub-glottal air pressure – that is declination. We regard inclination as successfully supervised declination. For us, sub-glottal air pressure is progressively falling, unless it is actively manipulated to rise.
- **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.
- We recognise also a category of **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**.

THE COGNITIVE BASIS OF THE INTONATION MODEL

Cognitive processing in language is usually modelled in symbolic terms, 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 the 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 the fundamental frequency of a speech wave, and it is these which 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 (and in other prosodic phenomena) 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 Siverman *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 the

prosodic part of a tts system. However we introduce the idea of neutrality *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. So, 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 try to 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 Siverman *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.

THE SYMBOLIC REPRESENTATION

The top level domain of the symbolic representation is the ***sentence***. Here we represent sentence-wide ***slope*** – 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***. In our model intonational phrases are defined by the syntax of the sentence. The sentence is parsed using a finite state grammar heavily dependent on syntactic category markers on words in a dictionary module in the tts system. We also take advantage of the distribution of punctuation marks in the input text [10] – though text authors are notoriously inconsistent in their use of punctuation. We have developed a set of heuristics which assign boundary markers for intonational phrases depending 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 in origin. This contrasts with the statistical approach adopted by some researchers [11].

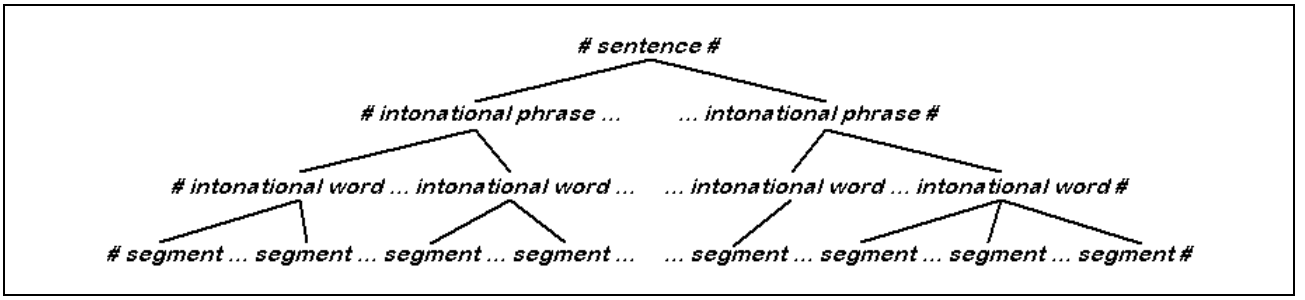

Fig.1 Diagram illustrating the hierarchical arrangement of intonation units within the model. The highest level is the sentence, and the lowest the syllable.

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.*

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. 2, 3 and 4) 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.'*

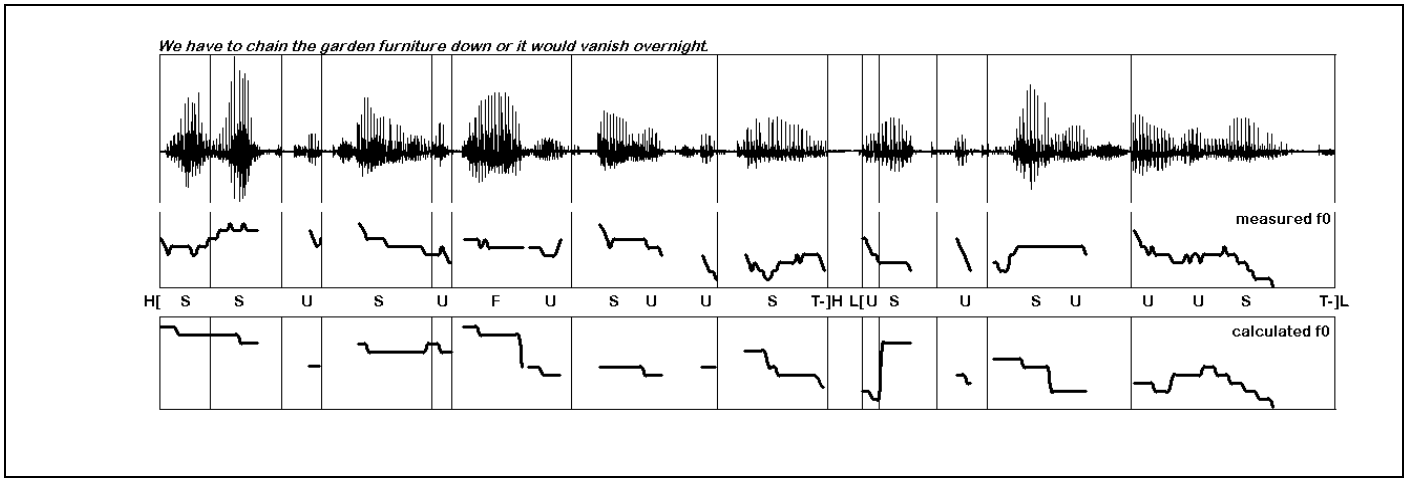

Fig.2 *We have to chain the garden furniture down or it would vanish overnight* – showing a. an example human waveform, b. the measured f0, c. generated text symbolic mark-up, and d. the calculated f0.

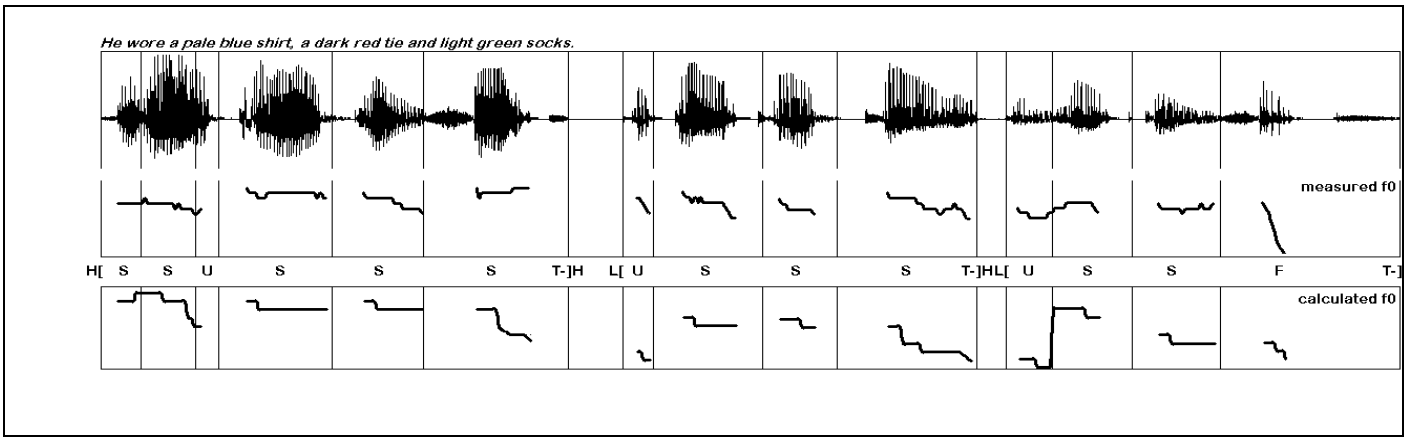

Fig.3 *He wore a pale blue shirt, a dark red tie and light green socks* – showing a. an example human waveform, b. the measured f0, c. generated text symbolic mark-up, and d. the calculated f0.

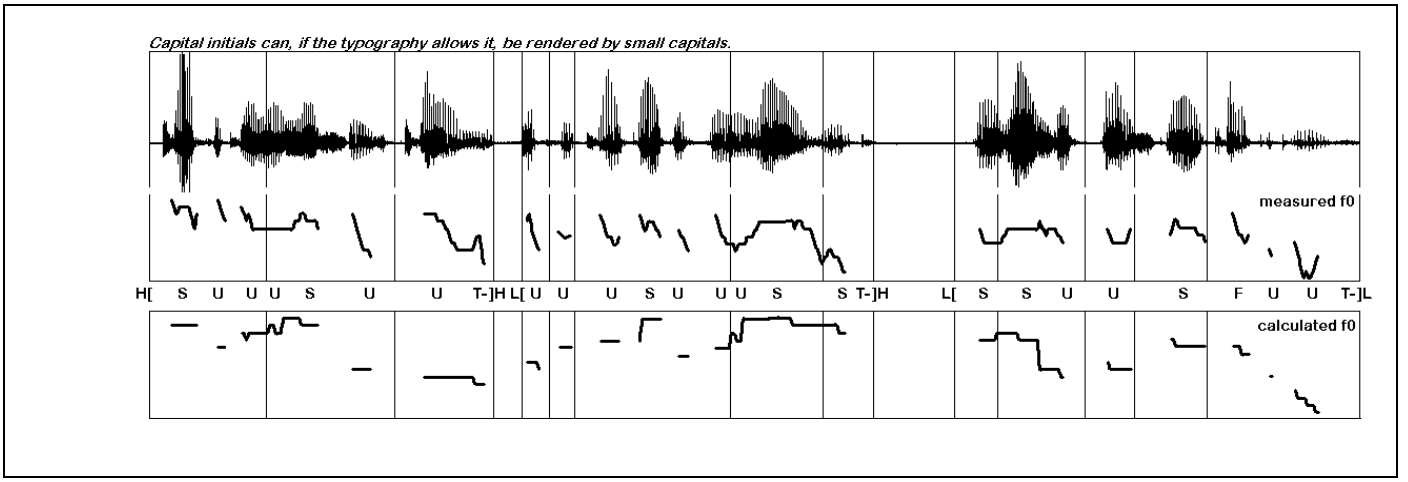

Fig.4 *Capital initials can, if the typography allows it, be rendered by small capitals* – showing a. an example human waveform, b. the measured f0, c. generated text symbolic mark-up, and d. the calculated f0.

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 – our subject could easily have spoken differently.
2. The **measured** *f0* from the waveform. Once again, this is not a canonical version – just one of a number of possibilities that our subject happened to use on this occasion.
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 but the way our system assigned a representation.
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 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.

- *Note:* In this presentation we do not discuss these overlay effects *per se* – but provision of pathways within the finite state transition network in Fig.5 is represented by the ***[res]*** (reserved) symbol.

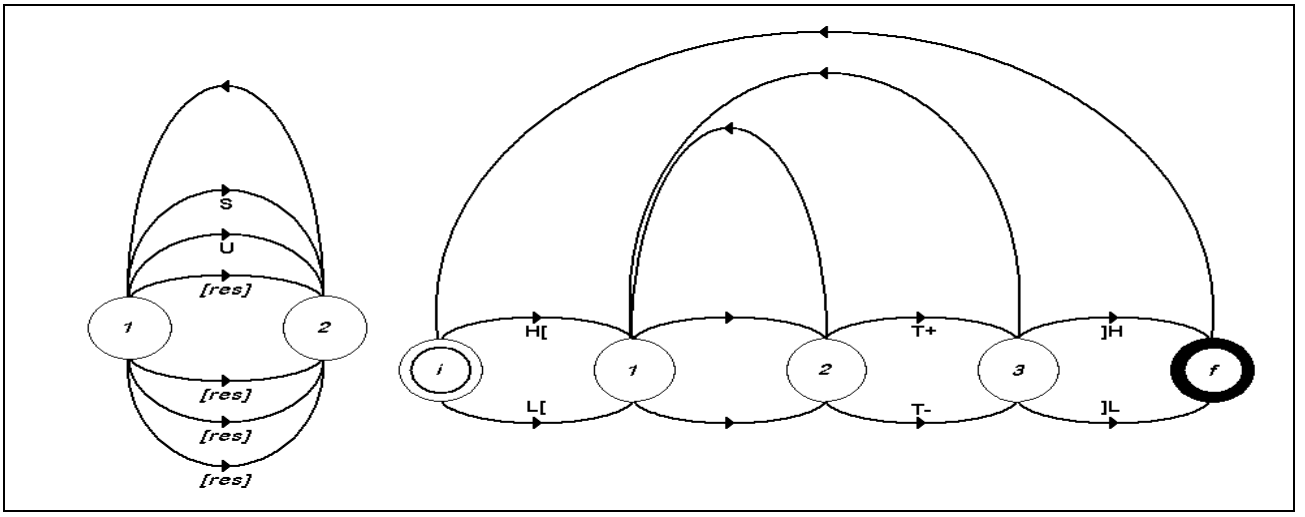

Fig.5 Finite state transition network showing the overall symbolic representational choices in our tts intonation model (see text).

Fig.5 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 (on 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 – 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. 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' on which we peg further symbols used in the representation of ***pragmatically determined overlays*** (not discussed in this presentation, but see Morton [5] and Morton and Tatham [6]).

FROM SYMBOLIC REPRESENTATION 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 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.

1. 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.

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.3-5 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 in Figs.3-5 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 several times 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** [5]. Intonation is not, of course, the only parameter used in the rendering of these effects – the other prosodic phenomena of rhythm and stress also play their roles. We have been modelling these effects as overlays on neutral contour generation as 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. In this paper we are not dealing with these effects and 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 the most basic neutral intonation is inappropriate.

CONCLUSION

In this paper we have presented the major properties of the our model of intonation for use in text-to-speech synthesis. The model has a number of important features reflecting 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

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

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

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

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

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

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

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

[8] K. Silverman, 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

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

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

[11] M.Q. Wang and J. Hirschberg (1991) Predicting intonational boundaries automatically from text: the ATIS domain. *Proceedings of the DARPA Speech and Natural Language Workshop*, pp. 378-383

[12] M. Tatham and E. Lewis (1998) Syllable recovery from polysyllabic words. In *Proceedings of Speech and Hearing 98*. St Albans: Institute of Acoustics – *this volume*

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

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

[15] R. Sproat (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