**Proceedings of the Institute of Acoustics**

# TOWARDS A PROCESS MODEL FOR INTONATION

A. Tams Electronic Systems Engineering, University of Essex, Colchester CO4 3SQ.
M. Tatham Linguistics, University of Essex, Colchester, CO4 3SQ.

## 1. INTRODUCTION

The inclusion of spoken language processing in multi-media leads to new applications for high quality speech synthesis, in particular hands free applications. These require highly natural sounding speech. Hence an appropriate model is that of reading aloud, which is easy to listen to with good intelligibility, for example a radio newscaster. This paper describes the synthesis of read aloud speaking styles with a novel process based intonational model, concentrating on the radio news broadcaster application. The emphasis on replicating the prosodics of a radio news broadcaster, since correct intonation and rhythm can lead to high acceptability of the synthetic speech.

Our definition of speaking style is based on Eskenazi [4]: a perceptual amalgam of the background, manner, and reaction to the intended audience. This overlaps, but does not include, emotional or attitudinal effects. Following Tams et al. [14] we have emphasised application for synthesis purposes. Two major perspectives have played a role in the definition of speaking styles. Phonetic and phonological techniques, studying the acoustic and linguistic characteristics of language are one approach. The second is the sociolinguistic and psychological perspective related to the use of language in a variety of contexts and situations. Eskenazi states that more attention should be paid to the sociolinguistic approach, calling for a data driven approach. The problem with the data driven approach is that the data is ambiguous, with no clear division between styles. Hence, the definition is extended to include a major role for the context, situation, audience, and aims of the communication - the speaker's environment.

Higuchi et al [6] used the Fujisaki model of intonation [5] to implement speaking styles for text-to-speech synthesis (hence TTS). They measured parameters of the model for four speaking styles, and derived rules for style conversion. This approach and many others were only partially successful, as a result of the use of average characteristics for a number of speakers.

The starting point for this approach is a suitable model of intonation. The most important requirement is that it can capture fine distinctions between styles. This paper attempts to address this problem. The variant of the Fujisaki model used captures f0 curves very accurately. We show that these can be matched to linguistic categories, and the model is extended for English. The next section describes the Fujisaki model. This draws on a specialised corpus for read aloud speaking styles desribed subsequently. The following section is the analysis of intonation, and the last section describes a model for TTS. A standard intonational phonological type model is introduced for the low level Fujisaki approach, and this is then extended in a novel way.

## 2. THE FUJISAKI MODEL

### 2.1 Overview of the Model

The basis for the intonation model is the source filter approach developed over two decades by Fujisaki for Japanese, of which Fujisaki et al [5] is a recent overview. Two sub systems are involved in the generation of the curve, with phrase and accent commands input to the mechanisms producing the phrase and accent components. The commands cannot be directly observed but can be inferred from the shape of their respective components.This model pays particular attention to the way the f0 contour is generated, treating the f0 contour as a linear superposition of the accent and phrase commands.

The functional model of Fujisaki is a control mechanism for f0 that generates an f0 contour from phrase commands and accent commands. The phrase command acts over the domain of the intonation phrase, shaped as an initial rise followed by a long fall to an asymptote line. This is generated by a phrase control mechanism, activated by a pulse command with varying magnitude. The accent command is a local peak on accented syllable, generated by the accent control mechanism. This is called by a binary step function, with duration and amplitude parameters.

The model is depicted in figure 1, and can be described mathematically, using a linear 2nd order equation for ln f0:

$$\ln f0(t) = \ln f\min + \sum_{i=1}^{I} ApiGpi(t-Toi) + \sum_{j=1}^{J} Aaj\{Gaj(t-T1j) - Gaj(t-T2j)\} \qquad (1)$$

$$Gpi(t) = \alpha_i^2 e^{-\alpha t} \qquad \text{for } t \geq 0 \qquad (2)$$

$$Gpi(t) = 0 \qquad \text{for } t < 0 \qquad (3)$$

$$Gaj(t) = Min\left[1-(1+\beta jt)e^{-\beta jt}, \gamma\right] \qquad \text{for } t \geq 0 \qquad (4)$$

$$Gaj(t) = 0 \qquad \text{for } t < 0 \qquad (5)$$

With the parameters:

| | |
|---|---|
| $f_{min}$ | asymptotic value of f0 in the absence of accent commands |
| I | number of phrase commands |
| J | number of accent commands |
| Api | magnitude of the ith phrase command |
| Toi | onset of the ith phrase command |
| T1j | onset of the jth accent command |
| T2j | offset of the jth accent command |
| $\alpha i$ | natural angular frquency of the phrase mechanism for the ith phrase command |

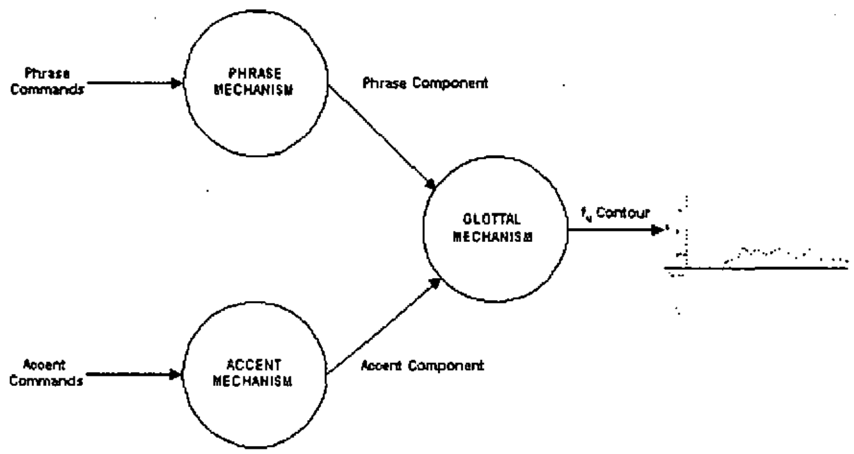


**Figure 1: The Fujisaki Model**

βj natural angular frequency of the accent mechanism for the jth accent command

γ ceiling of the accent component

## 2.2 Linguistic Representation

The model components are critically damped second order systems, with the two sets of parameters for the phrase and accent equations above. The model parameters, which include the angular frequency of the components above are generally regarded as constant for the utterance, though some variations of the model have used smaller domains.

Numerical optimisation of an analysis by synthesis procedure can be used to find the magnitude and timing of the underlying processes. In the Fujisaki model this is a hill climbing search of parameters guided by linguistic constraints, but in this work we have also combined it with statistical models (see below). These constraints are primary and necessary, because the search would produce an arbitrary number of accent and command phrases (over-fitting) to produce an optimal mathematical approximation to the contour otherwise.

Declination can be explained as a negative impulse to reset the phrase component. This model has been linked into physiological and physical mechanisms of the laryngeal system, based on f0 transition data in singing. The model has been applied to several languages including German [9] and preliminary studies for English [5]. The model has been used in this paper because it has physiological and physical justifications, is quantitative, synthesis is straightforward, and the mapping between the quasi-discrete inputs and continuous output of the accent and phrase mechanisms.

For Japanese, intonational phonological models have been developed, based on syntactic structure. This is not the case for English, and a solution to this was presented in Tams and Taham [15]. This is explained in more detail in the next two sections.

# 3. THE CORPUS

## 3.1 Overview

The RadioNews corpus described in this paper consists of British English radio news broadcasts and professionally read radio news data, designed to give wide coverage to the radio news broadcaster style.

This corpus was inspired by and based on the Boston University Radio News corpus, devised by Ostendorf et al [11]. In their corpus they have also included extensive phonetic alignments and ToBI labelling. This approach is not followed here, because with the growth in automatic labelling techniques, it is no longer a difficult and time consuming task to add a novel or additional annotation. Therefore, no standardised prosodic annotation is included as part of the corpus proper. This corpus contains approximately four and a half hours of speech from radio news broadcasts from three UK stations: BBC Radio 4 (R4), BBC Radio 1 (R1), and Classic FM (CFM). Between the stations there are marked differences in the form and contents of the broadcasts. It is divided into two sections of material, radio news and lab speech.

The radio news is primarily of nine data sets, recorded from the actual broadcasts (7-12 per data set). Eight of these are for an individual speaker (4 male, 4 female), with half from R4 and two each for CFM and R1. The data sets were designed to allow extensive coverage of a small number of speakers. The lab speech contains 22 recordings in 4 different read aloud styles (neutral, radio, advertisement, and bored), with examples for each station, by a professional speaker. The speaker has experience of radio news broadcasting and laryngograph data is available for these recordings. Additionally the speaker read a page from a novel, serving as a 'control' example of non news speech. The lab news is designed principally for speaking styles and variability research, and is approximately one and a half hours long.

The structure of the database was dictated by the speaking styles coverage criteria, practical considerations in recording the broadcasts and broadcasters, length of the recording session possible with the professional speaker, and anotation and analysis time constraints. In the corpus there is a move away from phonetically balanced to content balance. Compared to the Boston corpus, this corpus captures variation by careful selection of recordings for a single speaker. In terms of content, semantic labels are not included but the corpus has been designed to facilitate such studies by the careful selection of comparable broadcasts (in terms of contents, successive and parallel accounts). This enables the structure of the discourse (i.e. news stories) to be examined over time.

## 3.1 Annotation

All broadcasts are annotated with recording information, word boundaries, part-of-speech tags, and f0 data. Subsets of the data include additional syllableannotations will include syllable markings and hand marking of pitch periods, but so far this has only been completed for data set M1R4 (7 recorded broadcasts totalling 2 hours of speech).

The starting point of the word boundaries marking was an orthographic transcription. From the orthographic transcription a phonetic transcription was generated, using the Lexicon supplied with the Festival speech synthesis system [2]. This the is input to a phonetic alignment based on a dynamic time warping algorithm. This process is not error free, and has been hand corrected for data set M1R4. Part of speech tags are also computed by Festival, which implements a probabilistic tagger.

F0 data includes hand marked pitch periods. The most important requirement of this process is to mark the periodic sections of the waveform consistently, ensured by placing the pitch mark at a zero crossing. Since this point can be found automatically, this gives the process the necessary rigour. Recordings were first pre-processed by adjusting their bias to zero. Regions of voiced speech were selected, with boundary regions decisions influenced by phonological voicing representations. The most consistent speech feature, such as the start of a periodic 'hump' or a pitch excursion is then annotated. For all files, both lab and broadcasts, a second f0 representation was also computed using a pitch tracker.

## 4. ANALYSIS

### 4.1 Procedure

The analysis was used for adapting the Fujisaki model to English and data analysis. Adaptation is necessary because the model has difficulty handling all the accents of English (see below), with Japanese having fewer intonational contrasts.

The analysis used the M1R4 data set. This was partitioned into an analysis and test set at the broadcast level. Following his a data set of 200 utterances, representative of the radio broadcaster style, was selected from the analysis group of broadcasts . Half of the analysis set was selected on the basis of intonational behaviour and linguistic structure, the other half was randomly selected from the analysis group.

The For the analysis performed in this paper, all of the f0 contours are processed by five point median smoothing to remove segmental perturbations (the original contours are also retained). The alternative is to weight the fitting process to values of the f0 contour that have reliable pitch extraction, for example vowels, by use of an autocorrelation coefficient [9]. This approach was rejected because of the use of natural speech, as opposed to recordings with maerial selected to maxi,ise voiced speech. The analysis procedure consists of the following stages:

1. Processing the corpus annotations, extracting word information. This included utterance contents, boundaries of words within the utterance, orthographic, and part of speech information.

2. A linguistic and statistical analysis to generate candidates for phrase and accent commands. This perfoms a rudimentary parse, marking clause boundaries. This is based on the phrasing and acenntuation modules of Festival. additional analysis is performed using rule nbasedsyntactic constraints and the presence of pauses. Complex nominals are treated as a single unit.

3. The candidate accent and phrase comands are then input to an optimisation process.. Alternatively, canddate phrase and accent commands can be generated by hand using an adapted version of the SFS annotation tool and the rendering program.

Numerical optimisation is then performed for parameter values by two modules, render and search. Render is a trivial representation of equations (1) to (5), and the slow rise component described below. This takes as input files of accent and phrase commands, but no linguistic constraints are applied at this stage. This can be combined with a search module, subject to linguistic and well formedness constraints (each command is associated with a word, words cannot have more than accent). Alternative assignment of components can be chosen, for example the phrase break and accent assignment modules of Festival can be used.

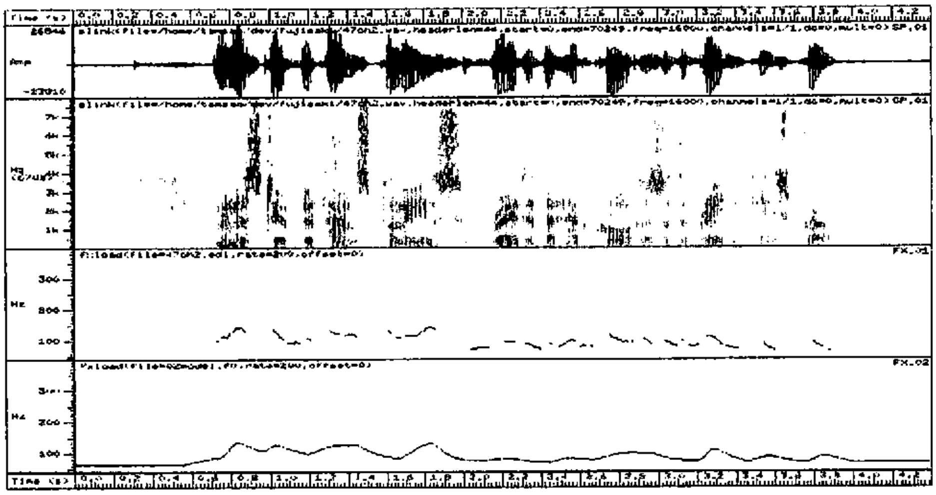

**Fig. 2 Rule Generated Contour**
The utterance is 'ministry of defence police have raided the headquarters of the greenpeace group', displayed using the SFS based tool. The parameters of the accent and phrase commands were obtained from a set of quantised values.

Figure 2 shows an example of an f0 contour synthesised using an implementation of the Fujisaki model for analysis purposes. This is not a numerically optimised curve, with accent and phrase commands selected according to linguistic constraints, and parameters quantified to discrete values.

Analysis shows that the magnitude of the phrase command is influenced by the length of the preceding phrase and syntactic cohesion of the boundary, with a maximum for utterance initial position. The location of the phrase command can be inferred from the portions of the f0 contour delimited by unaccented syllables (where f0 falls).

By comparison to the other speaking styles data in the corpus, different speaking styles show variation in accent realisation and phrase commands. The variation of model parameters for different speaking styles is also being investigated, though in most formulations of the model they are kept constant. However we have found that range is an important correlate of speaking styles, hence this assumption is not adopted in this work.

## 4.2 Extending the Model

The Fujisaki model does not have a clearly defined phonology for English, so an appropriate framework has been devised. Problems exist for long rising sections of the f0 contour [16], for example in interrogative statements, which are not very common in this corpus.

This is solved for German by Mixdorff [9] by introducing a slow rise component. This is like an accent component with onset T3 and offset T4, though with the time constant of the accent component, added as an extra term:

$$Gr(t) = 1-(1+\delta t)e^{-\delta t} \qquad \text{for } t \geq 0 \qquad (6)$$

$$Gr(t) = 0 \qquad \text{for } t = 0 \qquad (7)$$

This approach is adapted in this paper for English, but as a different gesture of the phrase command. This compensates the falling slope of the phrase command with an almost linear rise in ln f0. This change is to be consistent with the approach in the model for production described in the next section, and to solve theoretical problems with the phonology (noted by Ladd [7]).

The intonation phonology introduces boundary tones, similar to Mixdorff, but with a different phrasal structure. This draws on the higher level Spruce symbolic representation (Morton et al 1999). A sentence has a global slope set at L to H or H to L. This consists of one or more intonational phrases having local slopes at a lower level of abstraction, set at (L to H) or (H to L). Since this is a level that can be realised, the (L to H) is represented by modifying Gp with a Gr component (since the default is (H to L) by Gp). Hence Gr is part of the phrase mechanism, and is assigned from the sentence accent This is used to differentiate the L to H, and H to L global slope. A phrase contains intonational words which may carry a pitch accent, realised by component Ga.

The number of prosodic phrases in an utterance is denoted by the number of positive phrase commands. The onset of a prosodic phrase is marked by the adjustment of the declination line. Boundaries are marked by pauses, a pitch accent on the last syllable, or a boundary tone. The boundary can be shifted by accentuation requirements.This indicates interaction between phrasing and accebtuation processes, leading to a new approach in synthesis of intonation.

# 5. THE PROCESS MODEL

## 5.1 Rationale

Eskenazi [4] commented that it is difficult to model the variations in intonation for read aloud speaking styles because the intonational phonology is often the same. However the prosodic realisation is different, so a model must be able to explain this. Blaauw (1992) stated that speaking styles differences may be the result of interacting parameters during speech production and this hypothesis is the basis for the process model.

Two requirements for this approach can be identified: a need for indeterminacy (a choice can be made between two or more processes to satisfy a constraint) and concurrency (to model interacting processes). This solves the problem of the intonational phonology approach being underspecified for speaking styles, exploiting the concurrency implicit in the standard text-to-speech model and models of speech production. The assumption of this approach is that the interaction of parameters is significant for prosodic speaking styles models.

A requirement for indeterminacy has been identified in analysis by Wichman and Knowles [17]. suggest a simple mechanism for allowing choice in accent prediction by classifying syllables into those requiring an accent, those that cannot, and those that can choose to take a pitch accent.

The model being developed has a theoretical basis from process algebra, see Milner [8]. Concurrent systems are represented as a collection of independent sequential processes communicating with each other in order to exchange information. A model consists of several communicating processes and at a suitable level of detail each process can be thought as sequential. An event is an observable activity at some level of abstraction. Levels of activity (a functional description) in a system are the events, and which events are included depends on the level of abstraction.

## 5.2 Components

In moving to a concurrent model, the piepline architecture of most text to speech synthesis systems is discarded. This is necessary because the pipleine structure is not based on human speech production (apart from at a very high level of abstraction). The new framework draws on pyschological models of reading aloud, of which Raynor and Pollatsek [12] offers a comprehensive review of the reading components described in this section.. Processes of interest for an intonational model include visual extraction of the data from the text, lexical access, contextual processing, phrasing and accentuation. Each of these forms a component, of which a simplified description is given below. During reading each of these processes is performed concurrently.

Visual extraction of the data is based around eye movements. These can be classified into fixations and saccades. Saccades are a rapid eye movement from one location to another, a fixation is where the eye is stationary and takes information. The perceptual span describes the amount of information visually extracted in a fixation, normally measured in letters. This is a window of four characters to the left and fourteen to the right (for English readers). Fixations increase, varying with the beginiing of lines, longer words, frequency of words, function or content words, ends of sentences and clauses. In reading most words are fixated once for durations ranging from 50 to 250 ms. Short function words can be skipped and longer content words may be fixated more than once.

A lexical recognition process operates over the outputs of the visual extraction process, and during execution of the reading task is closely coupled with it. Within psychology the nature of this process, both in terms of strategy and mechanisms, is debated. In the literature, see [12], alternative routes include the dual route approach and the single route approach. In the dual route approach dicitionary based lookup operates on the outputs of the visual extraction process, with letter to sound rules performed in parallel. Recent work has suggested a single route of varied strategies, with a horse race to find the phonological match to the input text. hence, words and letters are processed concurrently. Mechanisms vary, with rules commonplace in dual route models and connectionist architecures for single route models. In this work a dual route is adopted to simplify the implementation, but in the general approach a single or dual route adopted. In fact it can be argued that there is a single route, with two competing strategies.

Leixcal access is defined as matching a word from the mental lexicon and making available properties of meaning, part of speech, the phonological representation, and orthography. The lexical process interacts with the context process. The context process is concerned with syntact processing. Researchers found that grammar structure increased lexical recall.Two further processing stages are performed, phrasing and accentuation. These interact, as described in the previous section (at phrase boundaries for example). This is the main advantage of the Fujisaki model, in that these can be formulated as separate processes. Each can be regarded as operating on its clock.. Realisation is equivalent to the glottal mechanism of Fujisaki using synchronisation and entrainment mechanisms to realise timing relations and constraints.

Figure 3 shows a simplified view of the process model for intonation. Not shown in the diagram is the connection between the context and phrasing diagram, omitted for clarity. In the visual extraction process the input is the text, represented in paragraph form. Output from the realisation process is the f0 contour. In comparison to a standard synthesis architecture the visual and lexical processes represent text processing, all other processes phonological level intonation processing except for realisation (phonetic level). Note that this structure shows the top level of the process organisation.

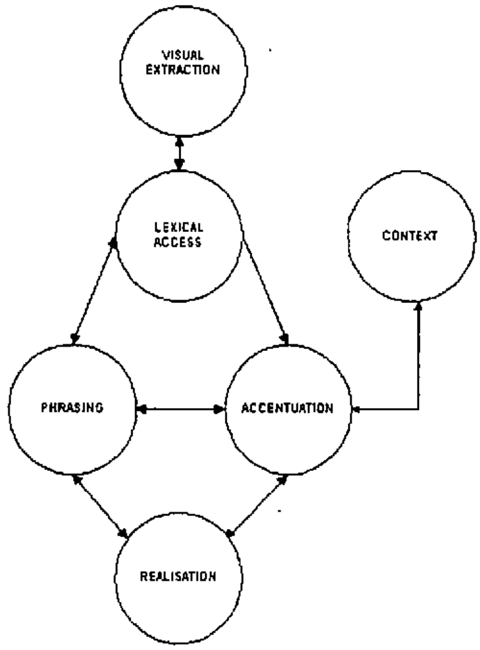


**Figure 3** **The Process Model**

---

One criticism of the Fujisaki model is that it is sometimes difficult to give a linguistic justification for the placement of accent and phrase commands, in particular the slow rise effect [16]. These can be viewed as a synchronisation restriction, where an accent or phrase command cannot be performed because of the execution of another generation process at a higher level of abstraction. In particular this is noted in the application of Gr, which serves as a modifier on Gp. This explains the variation in slow rise phrase command onsets and difficulties in aligning these with linguistic boundaries (though associated with the nuclear accent).

Festival and its functional architecture is used as a test bed for this new approach. This is made manageable by the adoption of a hybrid strategy. No attempt is being made to implement a wholly functional and concurrent TTS system. Rather the new approach is used for a model replacing prosodics modules, interfacing with a conventional architecture for further processing. In this sub system, there is no provision for streams of information and no large data transfers. Instead data transfer is by communication between processes. The current implementation uses a Java thread based implementation of the visual extraction, lexical access, and the Fujisaki model components. The lexical access simulates timing events, but is performed as a preliminary operation on the input. This is based on word class, frequency and the given/new distinction. It is envisaged that a more complex memory model will implemented in the future, based on Cahn [3].

The current implementation generates accentuation patterns, but timing information is generated from experimental results in [12]. This is problematic, since these are geneally expressed as mean values, hence a certain amount of tuning is required. To find parameters for the model is a bootstrapping problem, in particular the start sequences of the processing. In the literature the lookahead is regarded as negligible, compared to that normally assumed for synthesis (generally in the region of a phrase), at a couple of word fixations before reading starts. This limited lookahead is supported within the Fujisaki model, with phrases only requiring onset values. The methodology uses eye movement and lexical access pyschophysics as the input constraints, analysed f0 contours for the output constraints.

The model has serious limitations. Reading comprehension is ignored (though [5] suggests that eye movements are not directly related to the comprehension of information from the printed page). This model models the concurrency inherent in speech production, though it cannot be easily extended to a full synthesis system, since pipelining of complex data structures is not possible. Implementation is not straightforward, because of the thread based formalism and timing issues.

## 5. CONCLUSION

This paper falls into two parts: the first demonstrates the adequacy of the Fujisaki model (or approach) for British English intonation, the second shows how the Fujisaki model can be naturally placed in a process model. Although the treatment of intonation is derived from the model of Fujisaki, this paper is to the best of our knowledge unique in implementing it within an explicit computational implementation of the process model. This model offers a representation that can model interaction between parameters during the speech production process.

This research is concerned with modelling speaking styles. A corpus of appropriate speaking styles data, concentrating on radio news broadcast recordings, read speech, and laboratory recordings have been collected. This has been annotated and analysed with a revised Fujisaki model of intonation. This model uses two critically damped second order filters to generate f0 contours. The phrase component models long term effects such as declination (and its associated resets), the input parameter is a sequence of impulses. The accent component models pitch accents, and input to this component is a step function.

There is close agreement between the Fujisaki model and generated contours, suggesting some semblance of physiological and physical reality. It is not necessarily a universal model and additional laryngeal control is required for languages such as English. Further amendments are required, in particular the development of a phonological description to facilitate constraints on the phonetic parameters. A process model for synthesis is being developed, modelling components as concurrent processes. This approach is novel and has implications for the architecture of TTS systems.

## Acknowledgements

The laboratory recordings took place at BT Labs, thanks to the Speech Analysis and Synthesis team. Special thanks to Eric Lewis for use of the cue points program for extracting pitch marks from Cool Edit.

# REFERENCES

[1] E. Blaauw. Phonetic Differences Between Read and Spontaneous Speech. In *Proc. ICSLP92*, pp. 755-758, Banff, 1992.

[2] A. W. Black, P. Taylor and R. Caley. *The Festival Speech Synthesis System.* At http://cstr.ed.ac.uk/, 1998.

[3] J. Cahn. Generating pitch accent ditributions yjat show individual and stylistic differences. In *Third ESCA/COCSODA Workshop on Speech Synthesis*, Jenolan Caves, 1998.

[4] M. Eskenazi. Trends in Speaking Styles Research. In *Proc. Eurospeech93*, Volume 1, pp. 501-512, Berlin, September 1993.

[5] H. Fujisaki, S. Ohno and C. Wang A Command-Response Model for F0 Contour generation in Multilingual Speech Synthesis. In *Third ESCA/COCSODA Workshop on Speech Synthesis*, Jenolan Caves, 1998.

[6] N. Higuchi, T. Hirai, and Y. Sagisaka. Effects of Speaking Style on Parameters of Fundamental Frequency Contour. In *Progress in Speech Synthesis*, edited by J. P. H. van Santen, R. W. Sproat, J. P. Olive, and J. Hirschberg, pp. 417-428. Springer-Verlag, 1997

[7] D. R. Ladd. *Intonational phonology.* Cambridge University Press, 1996.

[8] R. Milner. *Communication and Concurrency*, Prentice Hall International, 1989.

[9] H. Mixdorff. *Modelling Patterns of German Model-based Quantitative Analysis and Synthesis of F0 contours*, unpublished PhD thesis, Technische Universitat Dresden, 1997.

[10] K. Morton, M. Tatham, and E. Lewis. A New Intonation Model for Text-to Speech Synthesis, in *Proc. of the International Congress of Phonetic Sciences*, San Francisco, 1999.

[11] M. Ostendorf, P. J. Price, and S. Shattuck-Hufnagel. The Boston University Radio News Corpus. In *University of Boston Technical Report ECS-95-001*, 1995.

[12] K. Rayner and A. Pollatsek. The Pyschology of Reading. Lawrence Erlbaum Associates, 1989.

[13] K. Ross, M. Ostendorf, and S. Shattuck-Hufnagel. Factors Affecting Pitch Accent Placement. *Proc. ICSLP92*, pp. 365-368. Banff, 1992.

[14] A. Tams, M. Tatham, and J. H. Page. Describing Speech Styles Using Prosody: A Pilot Study. In *Proc. Eurospeech'95*, Vol. 3, pp. 2081-2084, Madrid, September 1995.

[15] A. Tams and M. Tatham. Intonation for Synthesis of Speaking Styles. In *IEE Colloquium State of the Art in Speech Synthesis*, 00/058, pp. 6/1 – 6/11, April 2000.

[16] P. Taylor. Analysis and Synthesis of Intonation using the Tilt Model. In *Journal of the Acoustical Society of America*, vol 107, 3, pp. 1697-1714, 2000.

[17] A.Wichman and G. Knowles. How determinable are intonation units?. In *Proc. ICPhS 95*, Volume 2, pp. 223 – 225, Stockholm, 1995.