Intonation in a Text-to-Speech Conversion System

Alexander Ian Campbell Monaghan

Thesis submitted for the degree of Ph.D.

University of Edinburgh

1991

DECLARATION

This thesis was composed by me, and the work described herein is my own unless I have explicitly indicated otherwise.

ACKNOWLEDGEMENTS

It is impossible to work productively in a vacuum, and there are several people to whom I am grateful for supplying a less rarified working environment. Firstly, I would like to thank my supervisor Bob Ladd for his help, encouragement, humour and well-timed criticism: it has been a privilege and a pleasure to work with him.

Secondly, I have benefited from the support and co-operation of friends and colleagues at CSTR and in the Department of Linguistics at Edinburgh, particularly John Laver and Mervyn Jack who have given me the opportunity to pursue interesting avenues of research and the time to write that research up in its present form.

Thirdly, I have valued the reactions and comments of delegates at the various conferences and workshops where I have presented parts of this work: this exposure has helped to polish the finished product.

Finally, my greatest debt of gratitude is to my wife Alison who has done all the things I should have done while I was writing up this thesis: when Alison agreed to marry me I was a carefree researcher, and she has borne my transformation into a grumpy, bleary-eyed creature of the night without ever losing her good humour.

ABSTRACT

This thesis presents the development and implementation of a set of rules to generate intonational specifications for unrestricted text. The theoretical assumptions which motivate this work are outlined, and the performance of the rules is discussed with reference to various test corpora and formal evaluation experiments. The development of our rules is seen as a cycle involving the implementation of theoretical ideas about intonation in a text-to-speech conversion system, the testing of that implementation against some relevant body of data, and the refinement of the theory on the basis of the results.

The first chapter introduces the problem of intonation in text-to-speech conversion, discusses previous practical and theoretical approaches to the problem, and sets out the general approach which is followed in subsequent chapters. We restrict the scope of our rules to generating **acceptable neutral** intonation, an approximation to **broad focus** (Ladd 1980), and we present a rule-development strategy based on the idea of a **default specification** (which can be successively refined) and on the principle of making maximum use of all the information available from text.

The second chapter presents a framework for deriving an intonational specification in terms of **accents** and **boundaries** from a crude syntactic representation of any text sentence. This framework involves three stages: the division of text into intonational domains of various hierarchic levels; the assignment of accents to lexical items on the basis of stress information and grammatical class; and the modification of these accents and boundaries in accordance with phonological principles of prominence and rhythm.

Chapter 3 discusses the problem of evaluating synthetic intonation, introduces an original evaluation procedure, and presents two formal evaluations of the output of the rules described in Chapter 2. Further sections present our attempts to improve our treatment of the three major causes of errors in the evaluated output: prepositional phrases, non-words or **anomalies** (e.g. numbers, dates and abbreviations), and anaphora of various kinds.

The final chapter presents a summary of the main points of Chapters 1-3. We draw various conclusions regarding the nature of intonation, the development of text-to-speech conversion systems, and the generation of intonation in such systems.

TYPOGRAPHICAL CONVENTIONS

There are several typographical conventions which I have used throughout this thesis in an attempt to make my intentions clearer. Their intended interpretations are listed here.

Typeface	Interpretation
Italics	Single-word examples or single lexical items
Underlined italics	Multiple-word examples
boldface	First or new mentions of technical terms
Bold italics	Major concepts or principles
Initial Capitals	Subsequent mentions of technical terms
LARGE CAPITALS in the text	Local emphasis
CAPitalised SYLLables in exAMples	Locations of accents
?Example	Dubious examples
*Example	Ill-formed examples

Table of Contents

List of Tables

List of Figures

Chapter 1

Preliminaries

> as the quality of synthetic speech is steadily improving, listeners will make higher demands on the naturalness of synthetic intonation.
> Terken & Lemeer (1988:453)

1.1 Aims and Objectives

By the end of the twenty-first century, there will be computer text-to-speech (TTS) systems that can pass their part of the Turing test (Turing 1950): that is, they will be able to read text aloud in a manner which is not reliably distinguishable from the performance of a human reader. These systems will produce, in real time, an acoustic realisation of the text which sounds convincingly like the natural speech of a native speaker of the language of the text. This seems a reasonable prediction to make, in view of the current massive international research effort in speech and language technology. However, if we consider the capabilities which this task would require, it soon becomes obvious why we do not have such machines already.

We will assume for the present that the problems of reading in text (identifying the end of a sentence, dealing with non-words, etc.) and of driving a hardware synthesiser from a phonemic representation have already been solved (but see McAllister (1989) and Spiegel (1990)): the problem then becomes one of generating an appropriate phonemic representation from a list of orthographic words. The first step, and perhaps the most uncontroversial one, is to consult a lexicon to discover what information is available about each word in the list: the minimum information contained in such a lexicon would be the word's pronunciation and gross syntactic class (noun, verb, etc.); other information might indicate semantic features, finer syntactic classes, alternative syntactic classes, or even frequency of usage. In addition, there might be a morphological analysis, integrated with or separate from the lexicon.

Once all the information available from the lexicon has been extracted, the next step might be to parse the list of words using all the syntactic and semantic information just obtained. Such a parse would require a grammar of some type (transformational, unification-based, dependency-based, combinatory), and might produce a syntactic structure and a semantic representation. At this stage we have gleaned all the information available from a given text sentence. However, real text is not just a series of unrelated sentences: there are semantic and pragmatic relations between the sentences, and there is an organisation into topics and subtopics above the level of the sentence. Moreover, the interpretation of any sentence by a hearer is relative to the preceding context, linguistic and extra-linguistic, and this is reflected in its production by a human reader. Any machine satisfying the Turing test would therefore take account of these aspects of human communication, and would be able to relate input text to the discourse context and to the real-world situation at least to the satisfaction of a human listener.

The above requirements appear to constitute a need for such a machine to understand the text it is reading, and indeed machine understanding would solve a great many of the current problems for TTS systems: however, understanding is a notoriously difficult thing to define (Winograd & Flores 1986) and it is by no means certain that machines will ever have the sort of understanding which humans have. Fortunately, a case can be made for believing that it is not necessary for a TTS system to understand what it is saying in order to pass the Turing test. One of the most famous instances of a machine attempting the Turing test was Weizenbaum's ELIZA program (Weizenbaum 1966), a very simple and unpretentious program which nonetheless sustained a conversation with an unsuspecting human via teletype: we are not suggesting that the problem of text-to-speech conversion is as straightforward, or indeed that it should be approached in the same manner, but we would claim that it is not necessary to UNDERSTAND what one is saying in order to say it. A couple of examples should help to make the point clear.

> If the Poet Laureate reads his own work aloud, the audience is (usually) treated to an inspired rendition of the work. If he reads someone else's work aloud, particularly if he has not read the work before, the result is perhaps not as inspiring since the reader's knowledge of his material is not as good. If I read Edward Lear's work aloud, the result is likely to be reasonable, as I am familiar with the work and it is quite "readable": if I read the Poet Laureate's work aloud, the result may not be at all what he intended. All of these readings would be undeniably natural and should by definition pass the Turing test: the degree of understanding involved, however, and the listeners' reactions to the results, would vary greatly.
>
> Of the myriad academic journals currently published, only a tiny fraction will be comprehensible for any one person: yet most educated people, not to mention professional proofreaders, are capable of reading the majority of such text aloud and few academics can never have read an article which they did not understand.

The requirement of understanding the text, then, does not apply to human readers: so why should it apply to mechanical ones?

Knowledge of the real world to some extent, but more importantly knowledge of the language to be read, is in our view what allows humans to read text which they cannot claim to understand. The problems of text-to-speech conversion lie therefore in the field of linguistic competence rather than in the area of real understanding. For present purposes, however, linguistic competence must include the application of real-world knowledge to the construction of discourse-level linguistic structures, and also the application of inference to the resolution of linguistic ambiguity. It should therefore be clear why there are as yet no TTS systems which pass the Turing test.

Even this reduced amount of "understanding" which humans apply to the reading of a text is a very tall order for any automatic system. However, a text-to-speech system, as a system whose only function is to read out text, has one very significant advantage over a human reader: it is not expected to enter into a dialogue, but can reasonably restrict itself to monologues where the only relevant linguistic context is what it has already read out. The most one might expect to ask of such a machine interactively would be for it to stop, continue, repeat itself, or miss out passages: all these facilities are possible without adding any extra burden of understanding, although there are obviously intelligent strategies which could be incorporated to make, say, repetitions more comprehensible than the original production (slower, more articulated, etc.). A TTS system could certainly not be expected to respond to queries of the type "When did you say the weather would improve?" or even "How do you spell that?", let alone "Wasn't he the Managing Director?" or "Why do you say that?"

There is one other major point in a TTS system's favour: its output will be interpreted by humans. This may seem obvious — after all, there would be little point in a text-to-speech system interacting with a speech-to-text system — but consider the case of ELIZA. This very simple program, written in the early days of artificial intelligence (AI) research, was never designed to pass the Turing test: yet it succeeded remarkably well, precisely because the human involved automatically assumed that he was interacting with another human. This is an assumption which we humans automatically make, to the extent that even if we KNOW we are interacting with a machine we still attribute human knowledge, aims, attitudes, etc. to the machine (Murray et al. 1988; Murray & Arnott 1990). A human listener will therefore assume that the machine understands what it is saying and that it is adhering to Gricean maxims of co-operative interaction (Grice 1975), and will therefore make every effort to put a reasonable linguistic interpretation on its acoustic output. Moreover, humans are extremely good at normalising speech to counteract noisy environments, poor-quality transmissions, unfamiliar accents and physiological characteristics, and we are also very good at hearing what we expect to hear, both linguistically (Marslen-Wilson 1973, 1975; Tyler 1990) and phonetically (Couper-Kuhlen 1986:51ff.; Bard 1990). A TTS system is thus assured of a very sympathetic hearing, and can rely on the human listener making prodigious cognitive and physical efforts to interpret its output as linguistically and contextually appropriate.

The task of a TTS system therefore appears to involve applying linguistic and real-world knowledge to produce a rendition of the text which reflects such knowledge and which thus allows the listener to credit the system with the sort of understanding which normally accompanies such knowledge. The situation is, in fact, remarkably similar to that of the interstellar hitch-hiker (Adams 1979:25) who need only demonstrate his possession of a towel in order for potential benefactors to assume that he also possesses all the items generally considered more essential than towels for interstellar travel. How, then, does a system demonstrate this knowledge to a listener? There are basically two aspects of its acoustic output which the system can use for this purpose: word-level information and sentence-level information. The former will demonstrate that the system is familiar with the words of the language, and depends on lexical and morphological information which is largely independent of the context or meaning of the text: this aspect largely influences the segmental intelligibility of the output, as incorrect word pronunciations and lexical stress placements are highly distracting. The latter is more indicative of the system's awareness of the structure of the text and its relation to the context, and requires intelligent linguistic analysis: the results of this analysis are demonstrated not by the segmental quality of the output but by the appropriacy of the prosody.

Prosody is the listener's main index of a TTS system's apparent understanding: the speech of a child reading aloud may be correct at the word level without demonstrating any awareness of the structure and meaning of the text being read, but the correct assignment of prosody to the text is taken as a clear indication of understanding. It is generally agreed that prosody is also a major factor in determining the intelligibility of synthetic speech (Sorin et al. 1987:125; Barber et al. 1988:970; Terken & Lemeer 1988:453; Tatham 1990:235). Good prosody is thus crucial to the listener's impression of a TTS system, to the extent that poor prosody is perceived as worse than no prosody at all (van Bezooijen 1989b; Benoit 1990) since an indication of lack of understanding is more damning than no indication either way. It has been shown that within the prosodic phenomena of duration, F_0 and intensity by far the most perceptually salient is F_0 (Batliner & Nöth 1989:213; Thomassen 1979), and it is therefore F_0 on which most TTS systems, including the system discussed in this thesis, concentrate. Given these facts, it should be obvious that the main purpose of the linguistic analysis in a TTS system is to produce high-quality prosodic information, and that the main realisation of this information should be as F_0 specifications, i.e. in the form of an intonation contour. This is the view taken in the work presented here, and this is why the generation of appropriate intonation contours is seen as vital for high-quality TTS systems.

1.2 The CSTR TTS System

The TTS system in which the present work is set has been developed at CSTR over the past 6 years, and is the work of several researchers. The development of the system was funded by NEC from 1985 to 1988, and was continued as part of the Alvey Integrated Speech Technology Demonstrator between 1988 and 1991. The system can be divided into two main subsystems: a text-to-phoneme (TTP) system, which attacks the problem mentioned above of producing an annotated phoneme string from standard orthographic text, and a phoneme-to-speech (PTS) system which takes such a string and produces an acoustic realisation. (The term *phoneme* is used here to mean *abstract representation of speech sounds,* and should not be equated with any more closely-defined meaning.) These two systems are outlined briefly below. The Intonation Accent Placement module, and the interface between the syntactic analysis and this module which is described in Section 2.1, are entirely the work of the present author and constitute the major part of this thesis.

1.2.1 The Text-to-Phoneme System

The philosophy underlying this system is that high-quality TTS requires the use of large amounts of linguistic knowledge: the CSTR system therefore attempts to produce as full a linguistic analysis as possible from the input text. The system is modular, with each module containing a set of rules specific to a particular linguistic aspect of the input: as each module produces an analysis of one such aspect as its output, successive modules have increasing amounts of linguistic knowledge available as input. To facilitate rule development, the TTP system is implemented in PROLOG, a declarative logic-based programming language. The structure of the system is illustrated in Figure 1: a fuller description of the system design is given in McAllister & Shockey (1986). With the exception of the intonation module, the TTP system has changed little since 1988: it is therefore the September 1988 version which is outlined here.

The TTP system processes text one sentence at a time. Text in normal orthography is input to the system via a keyboard or is read from a file. The preprocessor, or Textual Anomaly Normalisation module, identifies any orthographic strings which do not conform to the system's limited notion of what constitutes a "word" (e.g. digit sequences or abbreviations) and converts them very simply into a form which can be processed by the other modules of the system.[1] The output of Textual Anomaly Normalisation is passed to the Word-Level Pronunciation Assignment rules, a group of modules whose task is to generate the pronunciations of individual words, and to the Syntactic Analysis module. The Syntactic Analysis module performs a crude analysis of the sentence structure, which is used both by the Word-Level Pronunciation modules and by the Intonation Accent Placement module. The Word-Level Pronunciation modules determine the morphological structure of words and assign a pronunciation to each morph, using a list of affixes and a dictionary of stems and exceptions: stems which are not found in the dictionary are assigned a pronunciation by a set of grapheme-to-phoneme conversion rules. In cases where a word's lexical stress pattern is not found in the dictionary, this is determined by rule in a Lexical Stress Assignment module. Any vowel quality changes associated with the lexical stress pattern and other phonological factors are effected by a Vowel Reduction module. The output of the Word-Level Pronunciation modules is processed by the Word Boundary Phonology rules, which perform phonological assimilations and deletions at word boundaries to simulate connected speech processes. Finally, the Intonation Accent Placement module identifies the types and locations of the major intonational events in the sentence. These processes produce an interpretation of the input text in the form of a richly-annotated phonemic string which is then passed to the PTS system to be converted into audible output.

[1] See Section 3.3 for more discussion of this module.

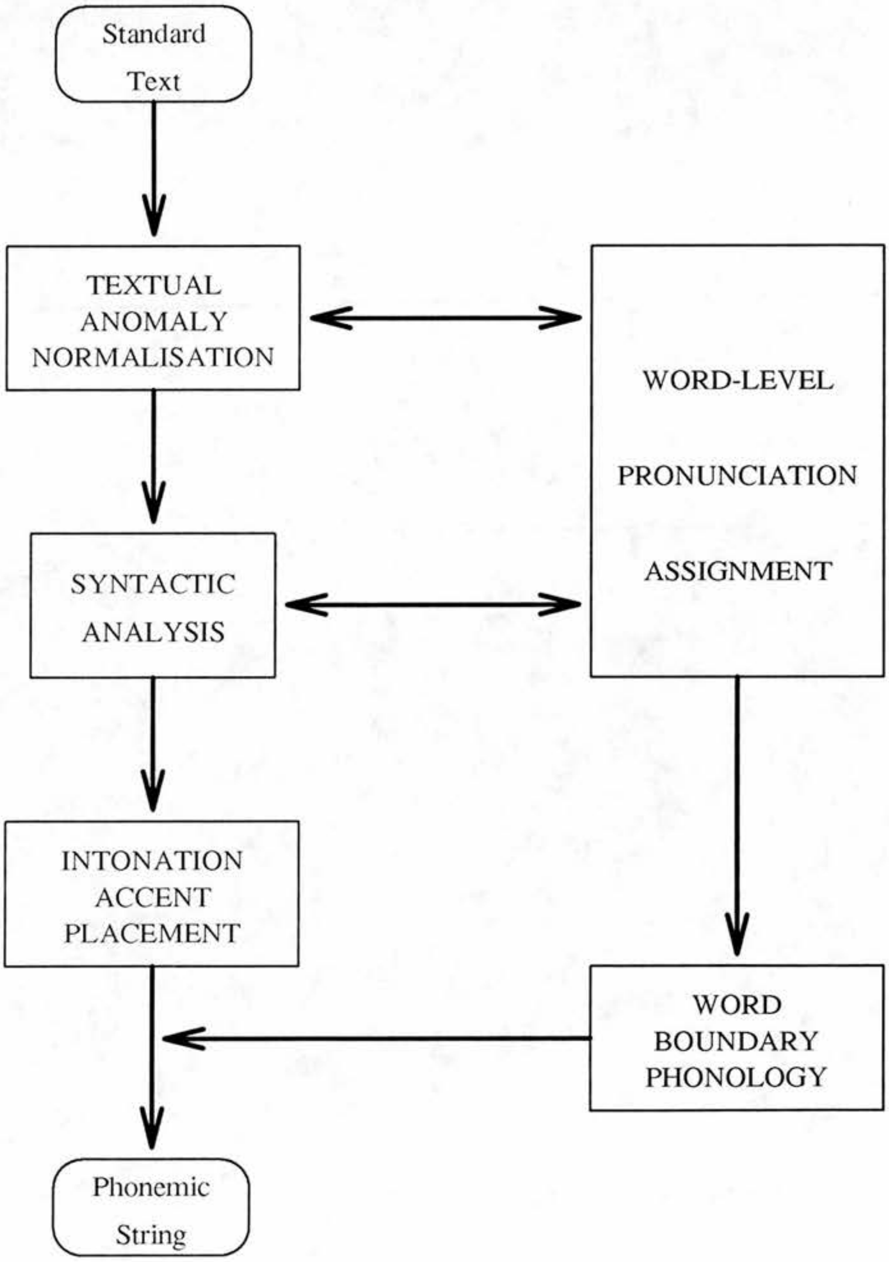
Standard Text
TEXTUAL ANOMALY NORMALISATION
SYNTACTIC ANALYSIS
INTONATION ACCENT PLACEMENT
Phonemic String
WORD-LEVEL PRONUNCIATION ASSIGNMENT
WORD BOUNDARY PHONOLOGY

Figure 1: The Structure of the CSTR Text-to-Phoneme System

1.2.2 The Phoneme-to-Speech System

In contrast to the TTP system, the PTS code has been developed largely since 1988: the major exception to this is the intonation model, which was implemented in 1987 but has undergone some revisions since then. Apart from control of F_0, which is discussed in detail below, the PTS system determines the allophonic and durational realisations of segmental phonemes. The system is based on an inventory of around 2,000 diphones, including some consonant clusters, taken from the stressed syllables of isolated nonsense words recorded from a male RP speaker. These diphones can be resynthesised using various versions of LPC and PSOLA synthesis techniques. Durations are assigned to diphones by Campbell's (1989) syllable-based duration rules, which take account of numerous prosodic features (stress, accent, syllable structure, etc.) in computing phoneme durations: the rules are derived by a neural net from a corpus of read text transcribed prosodically. No control of amplitude is currently possible, although preliminary ideas as to what amplitude effects are desirable are presented in Chapter 4. The PTS system is implemented in C, and produces acoustic output in close to real time. Further details of the system are given in Campbell et al. (1990).

1.3 The CSTR Intonation Model

The intonation model implemented in the CSTR system is similar to Pierrehumbert's (1980) target-and-transition approach. Intonation is modelled as a series of local events which specify pitch accents and intonational boundaries.

All intonational prominences are assumed to be pitch accents: there are two degrees of accent in our model, primary and secondary. We currently define primary accent as a pitch movement which is potentially nuclear, and secondary accent as any less prominent pitch movement. In certain cases, discussed in Chapter 2 below, a secondary accent may be realised as a tertiary accent, the difference phonologically being that a tertiary accent is associated with and subordinate to the following accent: tertiary accents are similar to the first half of a 't Hart & Collier (1975) "flat hat". Accents are associated with lexically-stressed syllables, such that every accent must fall on a stressed syllable but not every stressed syllable is accented: the question of the relation between stress and accent is a thorny one (Couper-Kuhlen 1986:19ff.; Selkirk 1984:252ff.; Bolinger 1972a:644), and will be returned to below.

Boundaries occur at the edges of intonational domains, and indeed boundaries and domains are to some extent mutually defining: the types and definitions of domains used in the TTS system are discussed in Chapter 2, but the existence of a hierarchy of prosodic domains similar to Ladd's (1988) is assumed in the CSTR intonation model. Domain boundaries may be realised in several ways: **boundary tones** of various kinds may be associated with them, in which case the boundary will be marked by a pitch target; pauses may be assigned at certain boundaries; the **register parameters** (see Section 1.3.2) may be changed, affecting subsequent pitch targets; or some combination of these effects may occur. All these possibilities occur in natural speech (Bruce et al. 1990:125; Couper-Kuhlen 1986:75; Crystal 1969), and impart varying degrees of prosodic prominence to the boundary.

To generate a contour from a sequence of specified pitch accents and boundaries requires two stages of processing: interpreting accents as phonological targets, and calculating absolute phonetic values for those targets. These targets are then linked by simple straight-line interpolation, with no smoothing or decay. The adequacy of such a simplistic approach to interpolation in perceptual terms has been amply demonstrated (de Pijper 1983; Willems et al. 1988), but it is envisaged that future development of the model will involve experimenting with smoothing and with the introduction of random (Kohler 1988) or natural (Monaghan et al. in preparation) pitch perturbations to enhance the perceived naturalness of the intonation contour. The modelling of segmental microprosody is widely acknowledged to affect the perceived naturalness of synthetic speech (Silverman 1987; Baart 1987; Sorin et al. 1987; Sato 1990), and work on this is also planned.

1.3.1 Target Assignment

The interpretation of accents and boundaries as targets involves reference to a **tune**, which defines the type of utterance (and thus the type of contour) to be produced. Tune choice depends on the speech act conveyed by the text, and is not currently implemented: instead, there is a default tune which produces satisfactory intonation contours for most declarative and WH-question utterances in English. It has been shown (Collier & Terken 1987:165ff.) that listeners do not find an invariant tune to be particularly unnatural or distracting.

The tune specifies the type of primary accent, the type of secondary accent, and the initial and final boundaries to be assigned: it is assumed that tertiary accents have an invariant interpretation. The default tune specifies a H*L[2] primary, a H* secondary, a mid initial boundary and a low final boundary: tertiaries are seen as H*H. The interpretations of these accents as sequences of targets are as follows:

> H* is interpreted as a mid target followed by a high target, i.e. a rise.
>
> H*L is interpreted as a mid target followed by a high target followed by a low target, i.e. a rise-fall.
>
> H*H is interpreted as a mid target followed by a high target followed by another high target, i.e. a rise followed by sustained high pitch.

The first two targets in an accent are placed one segment before and 60% through the accented vowel respectively: the third (or trailing) target of a primary accent is associated with the nucleus of the following syllable. The trailing high target of a tertiary accent is postponed until the next accent, thus linking the two accented items since tertiary accents encode close dependency on the following accented item.

Boundary tones are interpreted as a single target on the appropriate register line (high, low or mid): the case of rising final boundaries as distinct from high final boundaries is problematic, and we do not currently distinguish between these two boundary tones in the phonetic realisation although the phonological model allows for the two different categories.

[2] An asterisk indicates the "starred" or most salient tone, which is associated with the syllable nucleus.

An appropriate tune for polar ("yes/no") question intonation can be produced simply by changing the final boundary in our default tune to a High tone: Eady & Cooper (1986:413) found that polar questions differed from declaratives in their post-nuclear tail only, and our informal tests bear this out. However, we have as yet no reliable way of identifying polar questions automatically from text.

Cases where targets overlap, i.e. where there are too few segments between targets, can arise in this model, particularly domain-finally where a primary accent and a boundary tone may both require to be realised on a single syllable. Such cases will result in cross-overs in the F_0 contour, giving rise to two or more conflicting values at a particular point. This problem is currently resolved by taking the highest value for F_0 in all cases. A similar problem arises where the F_0 contour does not actually turn back on itself but where the targets are nevertheless so close together that their realisation creates an unnatural impression. Such cases occur on some utterance-final primary-accented syllables, and result in an unnaturally steep fall from the nucleus to the end of the utterance (unlike the shallow fall at the end of Figure 3): we currently spot such "concave" final contours, and remove the final Low target from the nuclear accent to give a straight fall from the nuclear peak to the end of the utterance.

1.3.2 The Phonetic Model

The interpretation of targets is based on a phonological **register** which defines high, mid and low lines. High and low are similar in some ways to top and bottom declination lines in other models, such as those described in Cooper & Sorensen (1981), Pierrehumbert (1981), Gårding & Bruce (1981) and Terken (1989a): mid is a neutral line to which the contour tends to return after pitch excursions. This register can approach the speaker baseline (a **downstep**) or move away from it (an **upstep**). In our model there are speaker-specific register parameters (minimum F_0 or baseline, default register width, initial height of register relative to minimum F_0) and speaker-independent parameters (excursion size relative to register width, step height relative to register height, current register height). These parameters are all inherently relative, with the possible exception of the speaker-specific parameters.

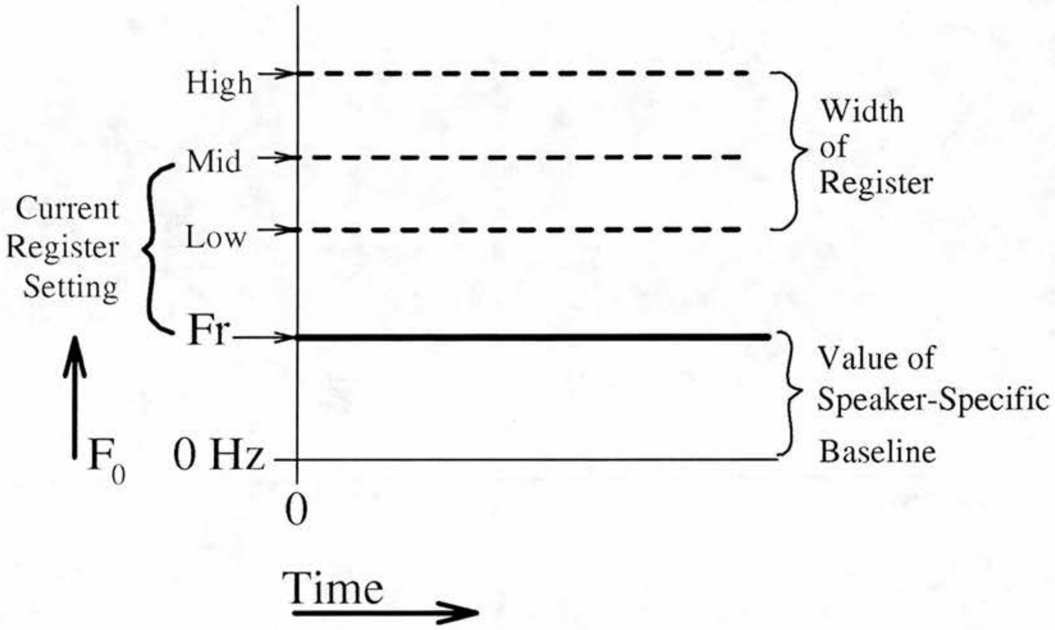

Figure 2: The Parameters of the Phonetic Model

The phonetic model was developed by D. R. Ladd, based on models by Fujisaki (e.g. Fujisaki & Nagashima 1969, Fujisaki & Sudo 1971, Fujisaki & Hirose 1983), Bruce (1982) and others: it is described at length in Ladd (1987) and Monaghan & Ladd (1990a, 1991). The model is illustrated graphically in Figure 2 and its mathematical workings are described in the following paragraphs.

For any speaker, a baseline value Fr is defined (in Hz). This value corresponds to the normal utterance-final frequency for declarative utterances in the speech of that speaker. A further speaker-specific parameter, N, defines the initial position of the space available for pitch movements, the **register**: Fr × N gives the value (in Hz) of the register midline at the beginning of most declarative utterances for that particular speaker. However, as stated above, the register is not fixed but may move up or down during an utterance: this occurs in steps at particular intonational boundaries, as discussed in Chapter 2.

Any given register setting f(N) is defined by the equation

$$f(N) = N \times d^i$$

where d is the step size (0.8 in our current model) and i is an integer expressing the number of steps the setting is away from the initial or default setting (for the default register setting N, $i = 0$). The value of the top of the register (i.e. High tone) at any given point is given by the equation $f(N) \times W$, where W is a speaker-dependent parameter determining the width of the register, and the bottom of the register (i.e. Low tone) is found by calculating $f(N)/W$. More generally, tonal configurations can be defined as sequences of values of T in an equation

$$f(T) = W^T$$

where $T = +1$ for High, -1 for Low, and 0 for the middle of the register. The actual F_0 values for the phonologically specified targets are then computed using the equation

$$F_0 = Fr \times f(N) \times f(T)$$

Once the absolute frequency values for the desired targets have been calculated and their absolute timing relations are known, the F_0 contour is constructed by joining successive targets with straight lines.

In the CSTR TTS system as currently implemented, the default values for all the parameters discussed above are stored in a PROLOG clause of the form:

```
ica_init([80,1.45,1.6,1,1,1,1]).
```

The first element of the list is the speaker minimum F_0, in this case 80Hz. The second element is the initial height of the register above the baseline, i.e. the height where the midline corresponds to the speaker's sentence-initial F_0: this is currently 1.45. The third element is the width of the register: this appears to be a linguistic variable as well as a speaker-specific one, and may be systematically varied during an utterance. The default value is 1.6. The fourth element is the size of excursion relative to the width of the register: excursions of size 1, as here, extend to the edge of the register. Again, this parameter may be varied for pragmatic reasons during an utterance. These default values have been established for the speaker whose diphones are used in the PTS system, using the procedures given in Monaghan & Ladd (1990a). The three remaining parameters in the list are used to control register height, as discussed in Chapter 2 below.

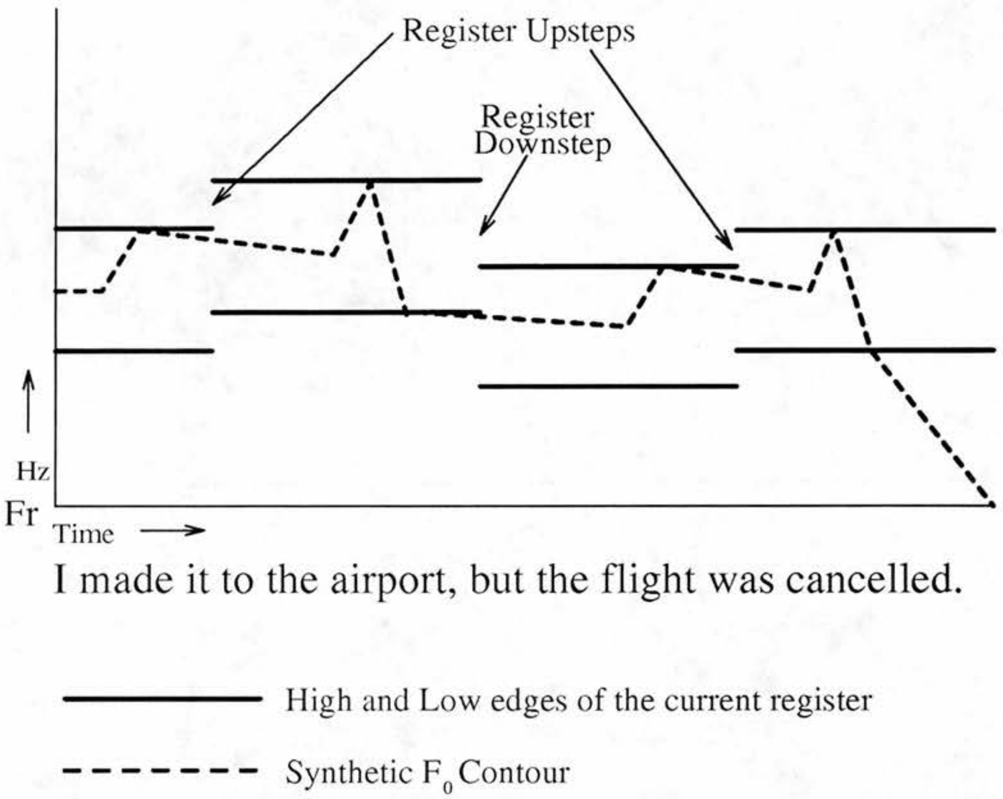
Register Upsteps
Register Downstep
Hz
Fr
Time
I made it to the airport, but the flight was cancelled.
High and Low edges of the current register
Synthetic F_0 Contour

Figure 3: Sample Intonation Contour produced by the CSTR Model

This contour shows secondary accents on *made* and *flight,* primary accents on *airport* and *cancelled* and a low final boundary: the two clauses are separated by a **register downstep,** and within each clause are two phrases separated by a **register upstep.**

The output of the phonetic model is illustrated in Figure 3, which shows a hypothetical F_0 contour for the sentence *I made it to the airport, but the flight was cancelled.* The solid lines represent the upper and lower (high and low) boundaries of the register, within which targets are scaled: the dashed line represents the F_0 contour, with targets at turning points and straight lines in between. The effects of register steps are illustrated, as is the fact that the contour may move outside the register during interpolations.

1.4 Literature Review

Previous published work both on intonation and on automatic text-to-speech synthesis is abundant, but work combining the two (i.e. concerned with generating synthetic intonation from text) is much scarcer and almost all relatively recent. The speech technology literature dealing with intonation synthesis[3] dates from the late seventies at the earliest, and for various reasons discussed below the linguistic literature before 1980 did not take any great interest in synthetic intonation. There is, however, a large amount of earlier material which is relevant to the study of automatic intonation generation, ranging from Dwight Bolinger's (1972a) classic article, "Accent is Predictable (if you're a mind-reader)", which as the title indicates takes a very pessimistic view of automatic accent-assignment, to Ignatius Mattingly's (1966) "Synthesis by Rule of Prosodic Features" which optimistically presents an embryo system for producing synthetic prosody from annotated text. This earlier material dates from at least as far back as Otto Jespersen's (1909) treatise on grammar. However, in the interests of brevity and to avoid needless repetition of the work of others, this chapter will concentrate on published work since 1970. For information on less recent work we have relied heavily on the various excellent bibliographies and reviews which are available and with which this chapter does not attempt to compete. (See e.g. Crystal 1966, 1969; Ladd 1979a.)

In an area of such great interdisciplinary interest as synthetic intonation currently enjoys, there is obviously a body of indirectly relevant work which is perhaps pertinent but which is also to all intents and purposes infinitely vast: electrical engineering, phonetics (acoustic, articulatory, descriptive, amongst others), computer science, artificial intelligence, cognitive psychology, and the philosophy of mind all fall into this category, and

[3] Intonation synthesis is distinguished here from F_0 synthesis. The latter is a prerequisite for the synthesis of voiced speech sounds, and has thus been possible and controllable since at least the beginning of this century (Klatt 1987:741): the former refers to the generation of automatically-determined intonation contours on the basis of some orthographic or linguistic representation.

these and many other fields have been either ignored or assumed in what follows. The task of comparing and distilling the wisdom of so many fields as it relates to intonation is one for which, unfortunately, we have neither the expertise nor the space required. Although we take others to task, here and elsewhere, for work which does not take sufficient account of the relevant literature, we are nonetheless forced to practise what we preach against and to ignore in large part the more peripheral work just mentioned.

The multi-faceted nature of intonation synthesis has led previous writers to "carve up" the literature along several dimensions. Baart (1987), who discusses only the more theoretic linguistic material, classifies this into four major schools of thought: syntactic-, semantic-pragmatic-, experimental- and focus-based views of intonation, each advocated by one or more prominent linguists in the second half of the twentieth century. Bing (1979a) sees the linguistic arguments as conforming to the classic "Levels vs. Configurations" debate begun by the American Structuralists and continued by Bolinger (1951), Lieberman (1965), Gunter (1972), Liberman (1975) and Ladd (1983b). Terken (1985) divides up the world according to rather different criteria, seeing the levels/configurations question and the syntax/semantics debate as largely resolved by the work of Ladd and Gussenhoven, but drawing a major distinction between the development and formalising of theories on the one hand and their experimental verification on the other.

Janet Pierrehumbert's (1980) thesis was perhaps the first successful attempt to review both theoretical and technical aspects of intonation synthesis. The importance of Pierrehumbert's work has been mentioned above in connection with the phonetic model assumed in this work: her contribution to making linguistically-interesting synthetic intonation a possibility has been considerable, and Pierrehumbert-style models are used by many of the workers whose research is discussed in this section. Her view of the literature is very similar to that of Liberman (1975), in that she divides previous work into the "Levels" and "Configurations" camps (1980:28), but she also draws a crucial distinction between linguistic intuition-based accounts and the more recent school of experimental verification (1980:54ff).

The simplest way to disentangle this confusing network of cross-cutting taxonomies, and also to our mind the best way to illuminate the different approaches and motivations at work in the area of synthetic intonation, is to divide the literature into that which is concerned more with fitting the data into a neat formalism, ignoring or "sanitising" anything which will not fit, and that which is concerned more with closely modelling or mimicking observed phenomena even at the expense of any wider pattern or insight. We shall refer to adherents of the former approach as **designers**, and proponents of the latter as **producers**. There are many dimensions along which these two classes differ, and equally many differences within each class (indeed, most of the distinctions used by other authors mentioned above will be referred to in the following discussion), but this simple dichotomy seems to be the most comprehensible and revealing for present purposes.

1.4.1 Designers

1.4.1.1 Common Ground

Designer approaches to language generally, and to intonation in particular, tend to concentrate on a small subset of the relevant data. This is justified on the grounds that language is composed of a large number of quasi-autonomous systems such as syntax, morphology, assimilation, syllable structure, etc. which each have their own rules and processes without any necessary regard for those of other systems, and that any one of these systems can therefore be investigated and formalised in isolation (Firth 1948; Chomsky 1980:246). This appears to be particularly true of intonation, which involves a multiplicity of levels with no clear mapping between them (although this view is currently being challenged by e.g. Steedman (1989) and Wheeler (1988)). A typical Designer account of intonation might concentrate on, say, the contours of tag questions or the relative peak heights of conjoined phrases, on the assumption that if a sufficient number of such areas are studied and formalised the entire jigsaw will eventually be pieced together and a coherent picture of language as a whole will emerge. Unfortunately, the number of conflicting views of any given area and the range of formalisms available at any one time work very successfully against such an outcome.

An additional problem with this approach is that the subsets of data chosen by Designers are often too small to reveal the underlying regularities which are being sought. The increasing use of data from more than one language (cf. Ladd 1983b; Gussenhoven 1985; Beckman & Pierrehumbert 1986) is an effort to avoid this problem, but it is still frequently the case that not enough data from any given language is examined. It is also frequently the case that the subsets of data to be studied are chosen on the basis of the assumptions underlying a particular formalism, and that as a result any regularities in smaller or larger sets which might lead to a revision or falsification of that formalism are ignored.[4]

Designer approaches tend to be founded in the linguistic or cognitive psychology traditions, and the assumptions of these traditions are therefore implicit in these approaches. It is thus assumed by many Designers that prosody is derived directly from syntax (e.g. Chomsky & Halle 1968; Selkirk 1984), that there is some interesting, elegant, abstract representation of intonation which is somehow mapped onto the uninteresting, chaotic, concrete phonetics of real utterances (e.g. Chafe 1974; Schmerling 1976; Bing 1979a), and that the intuitions of a trained linguist are much more likely to produce useful hypotheses than the data produced by naïve speakers. These assumptions and prejudices are based on the prevalent atmosphere and concerns in linguistics and other disciplines, such as the primacy in recent years of syntax over semantics, morphology, phonetics and other areas of language: they are no different in nature from the outdated view that English is a special case amongst world languages, or the trend for transformational models of language which was curtailed by Peters & Ritchie (1973).

1.4.1.2 Differences

Within Designer approaches there are two distinct major schools of thought regarding intonation. The first is based on the assumption mentioned above that intonation is directly derived from syntax, and the second claims that some notion of **focus** determines intonation. These two schools appear at first to be diametrically opposed in their

[4] A prime example of the problem of choosing a set of data on *a priori* grounds is the choice of isolated sentences as data for investigating the syntax-prosody correspondence (see page 25 below).

theoretical bases, but this is not in fact the case and indeed it is useful to combine elements of both in an account of intonational phenomena. Unfortunately, many of the proponents of these schools seem unable or unwilling to see the merits of combining the two approaches.

The distinction between syntax-based and focus-based theories of intonation has been drawn before (Schmerling 1976, Baart 1987), and the different theoretical arguments have been presented at length in numerous publications (Bresnan 1971, 1972; Lakoff 1972; Halliday 1967b; Bolinger 1958, 1986; Gussenhoven 1983a). However, discussions of both positions (e.g. Bierwisch 1968; Berman & Szamosi 1972) have tended to concentrate more on particular problematic examples than on the underlying hypotheses about language which are involved. Indeed, many recent authors, especially those for whom theoretical issues are of secondary importance[5], have followed Terken (1985) in assuming that work such as that by Ladd has resolved the theoretical distinction between syntax-based and focus-based approaches to intonation by explaining many of these problematic cases. This distinction has therefore rarely been presented as the fundamental difference in attitude and assumptions which we believe it to be. There is undoubtedly work on formal theories of intonation which does not fall unambiguously into one of these two schools, but that is generally because such work concerns itself with realising intonational specifications rather than deriving them: much of the work in the IPO paradigm ('t Hart & Collier 1975; de Pijper 1983; Nooteboom & Kruyt 1987) is of this nature. Nonetheless, distinctions such as those Baart (1987) draws between the work of Ladd, Fuchs and Gussenhoven (his "focus approach") and that of Bolinger, Schmerling and Keijsper (his "semantic-pragmatic approach") seem to us to be much less crucial than the distinction between all of these and the syntax-based approach of Chomsky, Bresnan and Selkirk which is given at least lip service by most of the European research in intonation generation, e.g. the Dutch national ASSP project (Quené & Kager 1989), the German national SYRUB project (Kugler-Kruse & Posmyk 1987), the Spanish national synthesis project (Rodriguez-Crespo & Escalada-Sardina 1990), and Olivetti's work on text-to-speech for Italian (Cericola et al. 1989) amongst others.

[5] See e.g. the work of Kager and Quené cited in Section 1.4.2 below.

This section presents some of the main issues in the syntax/focus debate, and attempts to explain why the debate arose and why it is no longer really relevant.

Syntax Determines Intonation

The case for determining intonation on the basis of syntax was perhaps first put convincingly by Halliday (1967b) in his grammatically-specified 5-tone system. However, the subsequent framework of Chomsky & Halle's (1968) *The Sound Pattern of English* (hereafter 'SPE') has had a much greater and enduring effect on linguistic and phonetic theories of intonation. As a syntactician, Chomsky was concerned to show the relation of syntax to phonology and phonetics and therefore took advantage of the frequent correspondence of a sentence's surface syntactic structure to its accentuation to claim that surface structure DETERMINES accentuation in these cases (Chomsky & Halle 1968:145). SPE's early disclaimer (p.23) that its rules "give accurate results only for simple constructions" is largely ignored in the remainder of the book, and even the acknowledgement (p.156) of "widely maintained but syntactically unmotivated" exceptions to their rules did not prevent Chomsky & Halle from expounding the theory that intonation is generally determined by syntactic structure. Bresnan (1971, 1972), Berman & Szamosi (1972) and others were quick to point out the many exceptions to this generalisation even in simple sentences, e.g. the ambiguity of

(1) I have plans to leave

which is resolved by the intonation. The usual explanation offered for these exceptions was that they were the result of the syntactic **transformational derivations** of the surface structures. Bresnan, for instance, claimed that in the above example the deletion transformation was preceded by a prosodic transformation which assigned accent to the deep syntactic structure to generate the appropriate intonational distinction (Bresnan 1972:326). Transformations were largely unconstrained at this time: transformational accounts of all types of linguistic phenomena were being proposed, and the psychological reality of syntactic transformations was widely accepted (Miller 1962; Clifton & Odom 1966), so it was only natural that prosodic transformations were simply added to the grammar where it was thought necessary. It apparently did not occur to the syntacticians until much later that these transformations were expressing semantic and pragmatic relations rather than syntactic ones: not until Bresnan (1982) proposed her Lexical Functional Grammar (LFG) as an alternative to transformations were such rules seen as reflecting choices ABOVE the syntactic level.

The collapse of unconstrained transformational approaches, triggered internally by Chomsky's (1970) reaction against the Generative Semantics movement (which advocated a wholly-transformational model of grammar) and externally by Peters & Ritchie's (1973) demonstration of the unacceptability of transformational accounts as models of human language on the grounds of their excessive power, led to a massive revision of syntactic theory but, crucially, not to any consequent revision of the syntax-prosody correspondence. SPE had made its mark in this area, and there was general acceptance that its basic principle of syntax determining prosody was correct and that the numerous exceptions would be handled by some complete account of syntax still to be developed. Theories such as LFG and Chomsky's various revised theories (van Riemsdijk & Williams 1986:170ff.) seemed to promise such an account, and so the syntax-prosody correspondence was assumed by writers such as Liberman (1975), Bing (1979a), Pierrehumbert (1980) and Selkirk (1984). However, in the light of work such as Bresnan's on LFG and with the appearance of other non-transformational theories of syntax (Gazdar et al. 1985; Steedman 1985), the importance of semantic and pragmatic factors began to receive acknowledgement from syntactic theorists and a model of grammar (Fig. 4) where semantics fed syntax and syntax fed prosody gained acceptance. This weakened the claim of the syntactic camp somewhat, since it could now be seen that semantics had a considerable influence on prosody (which was in fact what the transformational prosodic rules of the early seventies were attempting to capture), and so the central assumption regarding the syntax-prosody correspondence was revised to a claim that prosody could be PREDICTED from syntax.

The syntactic approach never attempted to account for all aspects of intonation: there is a long list of acknowledged exceptions, notably compound nouns and "contrastive stress". The former are cases where the Compound Stress Rule (CSR) proposed in SPE might be expected to apply, but unfortunately there are no reliable criteria for

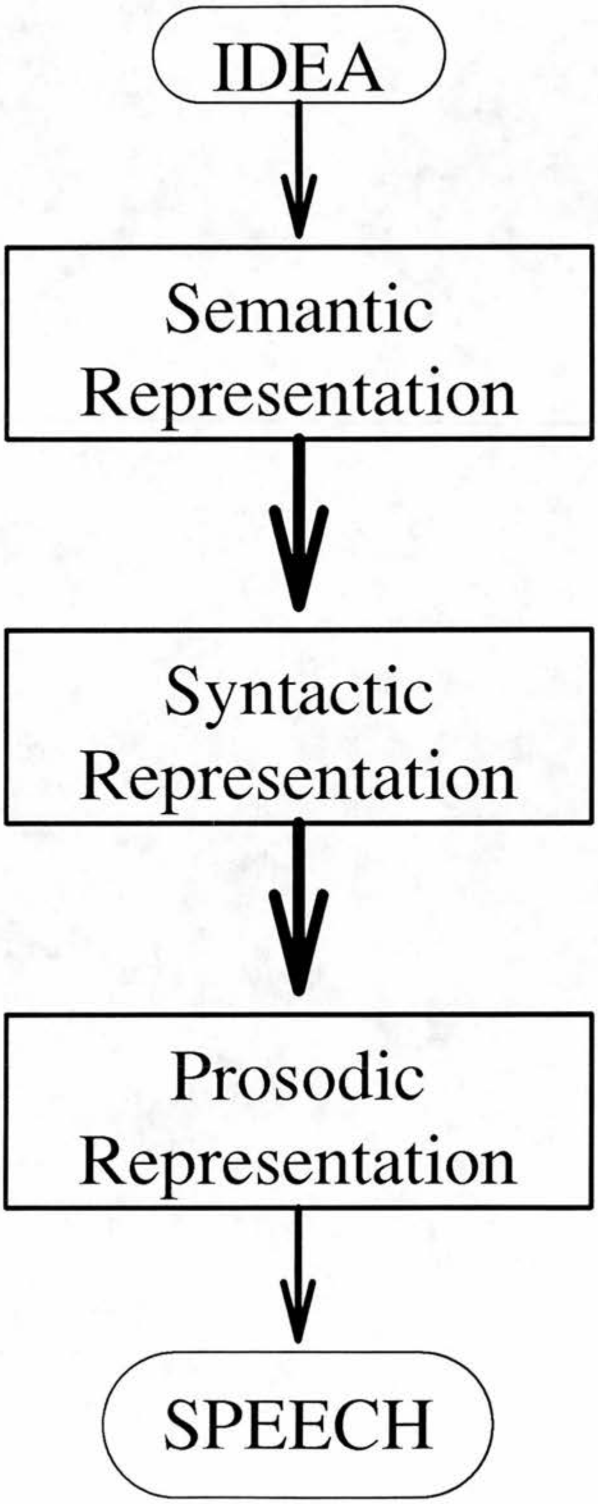

Figure 4: The Straight-Line Model of Prosody Generation

determining when compound stress is appropriate and when it is not (Selkirk 1984:243ff.; Sparck Jones 1985; Sproat & Liberman 1987). The latter includes cases where accent placement is heavily dependent on context, particularly those involving the deaccenting of anaphoric items. In fact, what the syntactic approach claims to predict is the location of the nucleus in cases of **normal stress**. Normal Stress is defined as the nucleus placement which naturally occurs when a sentence is produced out of context, such that there are no pragmatic factors involved. There is an implicit assumption that, just as speakers can assign lexical stress to isolated words based on lexical and morphological information, so they can assign the nucleus in an isolated sentence on the basis of syntactic structure. Normal Stress, then, represents a citation form which is claimed to be independent of contextual effects and therefore more basic than other accentuations: all realisations of sentences which do not conform to this norm are grouped together and seen as irrelevant to the grammatical description of prosody. Schmerling (1976:45ff.) summarises the arguments for this bipartite view, and cites numerous examples where there appears to be no Normal Stress version of a particular sentence such as (2a) or (2b).

(2a) Bill hit Mary and then she hit him
(2b) Bill hit Mary and then he hit himself

Such examples are generally explained by adherents of the syntactic approach by excluding them *post hoc* on the grounds that they have no "out of the blue" interpretation. This may seem reasonable on the face of it, but it fails to account for the difficulty of predicting the accentuation of Schmerling's (1976:41-2) *Truman died / Johnson died* examples: Schmerling's observation that these two simple sentences were uttered with very different accentuations (as a result of the relative predictability of the deaths) as the first utterances of their respective dialogues, i.e. as close to "out-of-the-blue" usage as is likely in natural speech, casts serious doubt on the validity of predicting contextless accent patterns and claiming that problematic cases are the result of unlikely contexts and can therefore safely be ignored. The real problem is that the number of sentences for which there is no obvious Normal Stress version is very large. As Bing (1983:143) points out, "no sentence can ever be completely divorced from a context. When a context is not given, a reader or listener supplies one." This made-up context is understandably more predictable and stable than contexts in real discourse, since the factors which determine it (world knowledge, Gricean principles) are less subject to change in the short term: however, this does not justify according it the special theoretical status which the notion of Normal Stress enjoys.

Another significant problem with the syntactic approach which is not acknowledged is the fact that the largest construct which it handles is the sentence. This is of course implicit in syntactic theory: syntactic constraints tend not to apply above the level of the sentence, and there is certainly no claim that syntax accounts for the ways sentences combine in texts or discourse. However, it is obviously difficult to motivate a claim that in

(3) Bill hit Mary. Then he hit himself.

there is some fundamentally different factor determining the accentuation compared with (2b) where the two sentences are combined into a single sentence. Given the possibilities for conjunction, parenthesis and subordination within a single sentence, it should be clear that any theory which attempts to predict the intonation of sentences while acknowledging its inability to deal with larger units is unlikely to meet with much success.

Speakers Determine Intonation

Amongst Designer approaches, the most well-known alternative to the syntax-based view of intonation is championed by Dwight Bolinger in numerous publications (e.g. Bolinger 1958, 1972a, 1983, 1987). The central tenet of this view is that intonation (amongst other things) is determined not by what syntactic structures the speaker uses but by what information (semantic, pragmatic, attitudinal) the speaker wishes to convey. This view pre-dates the syntactic approach by several years, but it has never been as widespread as the syntactic approach, mainly because until very recently there was not even a partial theory of the information conveyed by intonation whereas there have been highly-formalised theories of syntax since the early sixties.

Long before SPE appeared, Bolinger and others (e.g. Newman 1946, Kingdon 1958ab) were well aware of the limitations of the syntax-prosody correspondence and of the notion of Normal Stress: Bolinger pointed out in 1951 that a speaker's emotion was always a factor in prosodic realisations, stating that "All of the supposed intellectual meanings assigned to intonation carry some kind of emotional tone" (1951:204). Unlike the syntacticians, Bolinger draws examples from actual or plausible dialogue, and although these are often heavily dependent on context (e.g. Bolinger 1985:86ff., 1989:383ff.) they serve to illustrate the inadequacy of the syntactic approach when applied to units larger than isolated sentences. Bolinger's views on the relation of syntax to intonation are clearly stated:

> The distribution of sentence accents is not determined by sentence structure but by semantic and emotional highlighting. Syntax is relevant indirectly in that some structures are more likely to be highlighted than others. But a description along these lines can only be in statistical terms. Accents should not be mashed down to the level of stresses, which are lexical abstractions. . . . Whether one tries to set up prosodic rules for syntax or syntactic rules for prosody, the result is the same: two domains are confused which should be kept apart.
> Bolinger (1972a:644)

Although Bolinger's observations on the shortcomings of the syntactic approach to intonation were both accurate and irrefutable, the absence of any alternative formalism to relate prosody to other linguistic phenomena made Bolinger's position very unattractive to other linguists. As syntax became increasingly formalised and emerged as the fastest-moving strand of linguistic theory in the sixties, work on intonation turned to syntactic theory to provide underpinnings for accounts of intonational phenomena. It was not until multi-dimensional theories of phonology appeared (Liberman 1975; Goldsmith 1976) that there was any clear alternative to syntactic structure as a representation of suprasegmental phenomena. Metrical and autosegmental phonology provided structures which were specifically designed to represent non-linear relations between phonological constituents at various hierarchical levels, and the notions of relative (rather than absolute) strength or prominence, variable-length constituents, floating elements, the "Obligatory Contour Principle" (Leben 1973), rhythmic alternation and "Designated Terminal Element" (DTE) (Liberman & Prince 1977) were much more suited to handling suprasegmental phenomena such as stress and intonation, and much closer to traditional and intuitive ideas of nuclei and meter, than the syntax-based stress levels of SPE or the grammatical tones of Halliday. Accounts of intonation within a non-linear framework (van der Hulst & Smith 1982) soon outnumbered the transformation-based accounts, but such was the continuing influence of syntactic theory in the absence of any alternative level of representation which could relate prosody to other linguistic phenomena that even metrical phonologists (e.g. Liberman 1975, Ladd 1980) have tended to appeal to syntax for the input to their formalisms and to relate metrical constituents to syntactic constituents in the absence of a formal theory of semantics or indeed of any aspect of language above the sentence level.

The first successful attempt to formalise the relation between intonation and the types of semantic and pragmatic factors whose importance Bolinger had been pointing out for over thirty years came from the European school of intonation research, and specifically from Holland. Dutch work on intonation in a linguistic framework began in the fifties, and the large body of experimental phonetic and phonological work produced in Holland is discussed below: however, its contribution to the phonological theory of intonation was minimal until the work of Carlos Gussenhoven in the early eighties (Gussenhoven 1983ab, 1984, 1985). Gussenhoven proposed a "Sentence Accent Assignment Rule" (SAAR), which appealed directly to the notion of focus and to semantic constituents without any reference to syntactic structure and which was very successful in handling both simple and problematic cases in several Germanic languages. The SAAR, as expounded in Gussenhoven (1983a:391), is actually two sets of ordered rules, the first feeding the second:

Sentence Accent Assignment Rule (SAAR)
a = Argument. p = Predicate. c = Condition (i.e. not p or a).
x and y stand for any of a, p, c. Capitalisation shows [+focus].
(a) Domain Assignment: P(x)A → [p(x)a]
A(x)P → [a(x)p]
Y → [y]
(b) Accent Assignment: [ ] → [*]
In ap/pa, accent a.

> **argument** and **predicate** have the usual predicate-logic definitions: a **condition** is anything which is neither a predicate nor an argument.

The net result of the SAAR is thus an accent on every [+focus] argument and on

[+focus] predicates and conditions in cases where they form domains without an argument. Gussenhoven (1983a:379) defines focus as "a binary variable which obligatorily marks all or part of a sentence as [+focus], i.e. no sentence can be entirely [-focus]."

The publication of the SAAR was followed by a long debate in the literature between Gussenhoven and Bolinger over the usefulness of the SAAR in handling real data. Gussenhoven's assertion that accents were assigned by rule within a focus domain, while acknowledging the importance of semantic and pragmatic factors, contradicted the central claim of the Bolingerian "highlighting" approach that speakers place accents precisely on the words or syllables to which they wish to draw attention, regardless of any constituents semantic or otherwise. Bolinger (1985) presents numerous examples where the appropriacy of the SAAR is not obvious, but his main criticisms are that factors other than focus and constituency still have a rôle to play. His examples of "list intonation" (Bolinger 1985:85), for example, effectively make the point that "the mere fact of being last entitles the item in that position to receive the nucleus, and the initial position confers a similar distinction." Similarly, in defining what is [+focus], Bolinger rightly points out that the specificity and frequency of lexical items are both important factors in their accentuation or otherwise: "The less generic the term, the less likely it is to be left unaccented." (Bolinger 1985:95) However, much of Bolinger's paper is simply a polemic attack on a competing theory and an attempt to bolster up the idea of speakers determining intonation regardless of linguistic factors (p.102):

> It is the SPEAKER'S knowledge that primarily counts, not necessarily the knowledge shared with the hearer. It can hardly be otherwise, since the speaker is the person who applies the accent, and will do so according to his perceptions.

Bolinger must be well aware that the speaker has perceptions concerning the hearer's knowledge, and that as the hearer will interpret accents according to his own perceptions the speaker must tailor what he says to the very perceptions whose relevance Bolinger is questioning.

Gussenhoven's reply to Bolinger (Gussenhoven 1985) expands on his 1983 presentation, pointing out quite rightly that most of Bolinger's objections are answered in the original paper and are in any case tangential. There is obviously no reason why the

SAAR and degrees of speaker freedom are incompatible: all Gussenhoven is claiming is that once the speaker has made his choices the actual assignment of accents is governed by rules. This is really not so fundamentally different from Bolinger's system: in that system, once a word is highlighted there are rules (of lexical stress and syllable structure) which determine where the accent falls within the word; Gussenhoven is simply claiming that in fact similar rules determine, for domains of more than one word, which word will carry the accent. As Gussenhoven (1985:138) puts it, "My sole purpose in this reply has been to show that 'highlighting' is not applied to WORDS, but to semantic constituents that may have a wider scope than, and may in some cases not even coincide with the semantic content of, the word the accent is placed on, and that therefore the intervening level of structure cannot be dispensed with." In response to Bolinger's attack, Gussenhoven provides further data to support the SAAR in his examples from German (1985:131):

(3a) er hatte diese moRALvorstellungen internaliSIERT

(3b) er hatte die SELTsamsten moRALvorstellungen internalisiert

These are examples cited by Selkirk (1984:229) as problematic for the syntax-prosody mapping, and they are equally problematic for Bolinger's highlighting approach, since they involve the optional assignment of pre-nuclear accents to [-focus] items (in 3a, *Moralvorstellungen* is presumably contextually given). As Gussenhoven points out, the accent on *Moralvorstellungen* in 3a is purely a result of its pre-nuclear position in the utterance and a simple change of tense from past to present is sufficient to change that position and consequently remove the accent:

(3c) er internaliSIERT diese moralvorstellungen

It is difficult to see how an approach such as Bolinger's can explain why speakers should assign such different degrees of prominence to an item depending on its tense. However, although these examples serve as further evidence of the failings of Bolinger's approach, the onus is really on Gussenhoven to defend the SAAR rather than to launch a retaliatory attack on Bolinger's ideas since the inadequacy of the narrow highlighting approach was amply demonstrated in Gussenhoven (1983a) by his examples of "minimal-focus sentences":

> By minimal focus we mean any focus distribution that has less than the elements specified in the structural description of SAAR in its focus. An important subclass of minimal focus is polarity focus, discussed above. However, minimal focus may also arise when PART of an argument or predicate is [+focus]. It is minimal-focus sentences in general that make it clear that Bolinger's "highlighting" hypothesis is untenable. In such sentences, there is often so little in the way of words that is marked [+focus], that the resultant nucleus locations are scattered all over the place: the nucleus is desperately looking for semantically empty little words it can go to, and — not surprisingly — it is here that even closely related languages like Dutch and English part company.
> Gussenhoven (1983a:409)

This particular area of intonation has always been problematic for Bolinger, and indeed poses problems for Ladd (1980): nor are Gussenhoven's explanations entirely convincing, and many of his English examples seem to us to be rather unnatural (we are not in a position to judge the naturalness of the Dutch accentuations). However, it is clear that Bolinger's protestation that "the focusing of a preposition is like the focusing of any other word" (1985:85) is not only unconvincing in view of the well-known observations regarding the relative tendencies of function words and content words to be accented (Kingdon 1958b, Altenberg 1987) but, more revealingly, says nothing about which items are more likely to receive accent or why this is the case, let alone what determines the more unusual accentuations. Gussenhoven at least has a plausible answer for this in the SAAR and the claim that only in cases where the SAAR cannot apply, such as minimal-focus sentences, will other elements receive accent.

The debate continues in Gussenhoven et al. (1987) with a further exchange of views between the two main protagonists and an article by Keijsper which reviews the debate so far and attempts (unsuccessfully) to reconcile the two positions. Her failure in this is quite understandable, as the rival theories are both founded on practically untestable (and thus unfalsifiable) assumptions. Bolinger states his underlying creed, in a rather extreme form perhaps, as follows:

> Accentual choices are not made IN ORDER TO highlight, or to focus domains. Rather, the highlighting or the domains emerge from the speaker's affect, and are interpreted, informationally, by hearers.
> Bolinger (1987:143)

Regardless of how reasonable a view this seems, it is clearly impossible to falsify it without resorting to unnatural laboratory studies, at which point it could be claimed that the speaker's affect has been affected by this artificial environment. Bolinger's view thus remains intact until a more explanatory theory appears.

Gussenhoven's work is certainly a candidate for replacing Bolinger's view, and is arguably more explanatory in that it provides rules for phenomena which Bolinger is forced to ascribe to the vagaries of speakers' intentions. However, Gussenhoven too is quite explicit about his assumptions:

> the speaker is assumed to translate his communicative intentions into choices from a number of linguistic options, most importantly into a focus marking of the semantic constituents in his sentence (fragment). Sentence accent assignment rules translate these choices (again, mainly the focus marking) into sentence accents on particular words.
> Gussenhoven (1985:125)

Gussenhoven refers to his view as the "focus-to-accent" approach: it is, however, just as difficult to falsify as the highlighting approach. The crux of Gussenhoven's approach is that a given accentuation is predictable from a given focus structure: however, if one presents, as Bolinger does, a different accentuation from that predicted by the SAAR the obvious response is that it was derived from a different focus structure. Since there are very few constraints on focus structures and there is no clear way of establishing the focus structure of a particular utterance except from its accentuation, the argument rapidly becomes circular: if the accentuation is a natural one, there is almost bound to be a focus structure which could provide it; if it is not a natural one, there is no pressure on Gussenhoven either to predict it or to account for its unnaturalness. As Gussenhoven (1983a:396) himself points out, those who claim that the SAAR is inaccurate in its predictions of normal accentuations are falling "into the trap of taking the most likely Background ... and assuming that the subsequent reading, which is of course 'normal' in the light of our knowledge of the world, is also 'normal' in a linguistic sense." The focus distribution must be determined before the accuracy of the SAAR's predictions can be judged, and it is virtually impossible to show that for a given utterance a particular focus distribution is or is not the case.

The importance of Gussenhoven's work was that it demonstrated to intonational phonologists that there was a possible alternative to Chomskian syntax which could capture phonological intuitions and account for a great deal of problematic data and which they might use to feed their rules and representations. The notions of **focus domain** and **semantic constituent** allowed Bolinger's insights regarding the inappropriacy of syntax for predicting prosody to be incorporated into a formal theory of intonation within and above the sentence. Intonation research finally began to escape the restrictive assumption of a strong syntax-prosody correspondence, and work began on the higher-level factors (semantics, pragmatics, speaker intention) which play a vital rôle in determining intonation. Bolinger's conviction that the speaker has absolute control of intonation at all times, regardless of other factors, still remained intact: however, the work of Gussenhoven, Ladd and others was directed towards defining a formal system of linguistically-meaningful choices WITHIN WHICH speakers had total freedom to select the intonational specification which suited them.

The Best of Both Worlds

A compromise approach between the syntactic and the focus approaches has been tacitly followed by several Designers since Gussenhoven, although these have tended to be authors who are more concerned with intonation in its own right than with its relations to syntax or cognition. These authors have attempted to define their terms in such a way that they can be interpreted in either approach, starting with Ladd's equation of the syntax-based notion of **normal stress** with maximally **broad focus**: Ladd claims (1983a:157) that "with certain accent locations the focus is specified only very broadly — that is normal stress." This avoids problems such as sentences with no full-focus realisation (examples (2a&b) above) and explains the syntax-prosody correlation in most cases on the grounds that most isolated utterances, at least, have broad focus. Gussenhoven (1983a:387) gives a similar (in spirit) definition of Normal Stress, although the terminology is rather different: he defines Normal Stress as the accentuation "that results from the widest reasonable interpretation of the semantic material as the Variable with speech act V-addition." However, it is clear that this is not enough to explain intonational behaviour in cases of **narrow** focus: to do this it is necessary to appeal to a representation more abstract than surface syntax. Ladd takes the view that "focus is related to syntax in a fairly well-defined way," (1983b:166) and is therefore forced to construct a prosodic structure which closely resembles syntactic trees (Ladd 1986) in order to preserve this relation: Ladd's view therefore still corresponds to Figure 4 above. Gussenhoven, however, sees intonation as dependent on "the manipulation by speakers of certain semantic material with respect to a discourse background" (Gussenhoven 1983a:383), and although he recognises the fact that focus information can be reflected in syntax (1985:127ff.) he demonstrates convincingly (pp.129ff.) that this is not always the case.

Gussenhoven's data certainly do not fit a processing model such as Figure 4, and we would like to propose at this point that a model more like Figure 5 is closer to the truth. According to this model, prosody, specifically intonation, is driven by semantic information and syntax is also driven by the same semantic information. This parallel arrangement accounts for the large degree of positive correlation between syntax and prosody in many cases, particularly in isolated sentences, but also permits the less redundant cases where either the syntax or the prosody realises a particular aspect of the utterance but not both. Examples of the latter are, on the one hand, syntactic distinctions which are not reflected in prosody, and on the other the well-known instances of one-to-many syntax-to-prosody mappings: the former can be illustrated by examples such as

(4) I saw the man in the park with the telescope

which is syntactically at least 6 ways ambiguous but for which no intonational theory would want to produce 6 different prosodic structures; the latter include such instances as Schmerling's (1976:41-2) *Truman died / Johnson died* examples discussed above.

The fact that the syntax-prosody correspondence holds in many isolated sentences, i.e. that these sentences have some form of Normal Stress, must be explained in this model by the absence of significant differences between the semantic specification of the sentence and the syntactic one. This is of course always probable with an isolated sentence, since there are no other factors which affect its prosodic realisation. However, in most human utterances there are two other major types of information which may

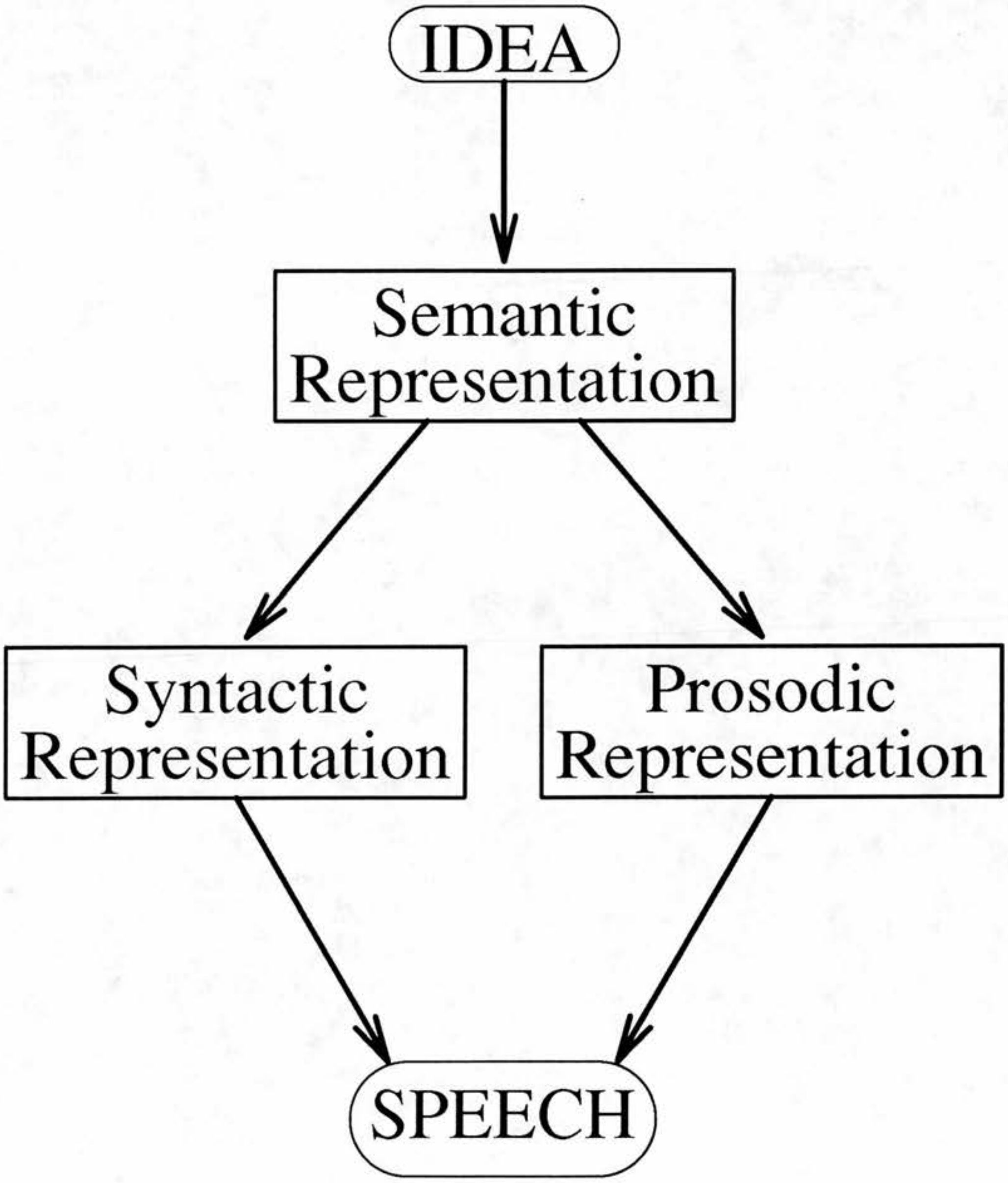

Figure 5: The Parallel Model of Prosody Generation

affect this realisation and which, in our view, account for most of the discrepancies between syntax and prosody and in doing so support the model in Figure 5. These are **intention** and **context**. Intention is entirely dependent on the speaker, and is thus very difficult to predict: it is described by Bolinger (1985:108) as "the speaker's sensations of importance," and indeed a more precise definition of such an explicitly subjective factor is as difficult to formulate as its effects are to predict. Bolinger's view of the primacy of intention is uncompromising:

> What counts is ... how the speaker feels about what he is saying. All these almost-true generalisations are only clues to that state of mind and can be overridden at any time, because of the subjective nature of accent.
> Bolinger (1987:142)

Nonetheless, it is possible to incorporate the effect of intention within a compromise solution. If we assume that the regularities described by the SAAR and by Normal Stress are in fact true generalisations about the rules and representations involved in generating intonation rather than just "pretty reliable guesses" (Bolinger 1985:108) which "are valid only statistically, in some broad and general way" (p.79), then we can see that intention must affect the input to rules such as the SAAR by determining the division of utterances into focus domains, for instance. The effect of intention is thus to specify the representation to which these more predictable rules apply, rather than (as Bolinger would claim) to replace or bypass such rules altogether. It is thus possible to say that both intention and more predictable processes such as the SAAR play a part in determining intonation.

The effect of **context** on intonation has been examined within the Designer approach by Bolinger (1972a), Ladd (1980), Prince (1981), Fuchs (1984) and others. The definition of context is usually avoided in these accounts: it is vaguely equated with shared knowledge, preceding discourse and general "givenness" (see Prince (1981) for a lengthy discussion). Despite this lack of a formal definition, there is widespread agreement (e.g. Daneš 1972:229; Horne 1987:51; Ladd 1984a:258; Levelt & Cutler 1983:215 and references therein) that context is a crucial factor in determining accent placement. Even hardened advocates of the syntax-based approach acknowledge the rôle of context in determining "non-normal" accentuations, as is clear from the definition of Normal Stress given above.

Fuchs (1984:135) ascribes most of the failings of the syntax-based approach to the effects of discourse context. Moreover, she criticises the concentration within that approach not just on isolated sentences but on the even narrower field of single-accent examples:

> Discussions of accent placement overwhelmingly are concerned with one-accent patterns — does 'the' accent go on this 'or' on that element? — while in spontaneous speech pluri-accent patterns abound.
> Fuchs (1984:136)

There has been very little investigation of the intonation of spontaneous speech within the Designer approach: with a few notable exceptions such as the large-scale data analyses of Crystal (1966) and Altenberg (1987), work on formal theories of intonation is based almost exclusively on artificial examples of isolated sentences or short dialogues. While it is indeed true that many aspects of intonation can usefully be studied on the basis of such examples, it is obvious that in such cases any effect of context will necessarily be minimal and that therefore a different class of examples must be studied in order to establish the effects of context on intonation. (See Monaghan (1990e) for details of how such a study might be carried out.) The necessity of studying linguistic structures larger than the sentence has been widely acknowledged in many areas of theoretical and computational linguistics: as G. Hirst (1981:2) points out, "many of the interesting problems of language do not occur in their full glorious complexity in a single sentence." Despite the continuing lack of a coherent theory describing how the various influences of semantics, word order, intention and context interact in determining the intonation of a particular utterance, we can hypothesise a hierarchy of effects on the basis of the literature discussed above. It is clear from Bolinger's examples, and indeed from common sense and a belief in speakers' free will, that the speaker's intentions are the overriding factor in determining the choices which a particular intonational specification realises: these are therefore at the top of the hierarchy. Both Fuchs and Gussenhoven clearly indicate that choices regarding focusing, including those determined by contextual information, precede and override the consideration of semantic structure: this is clearly the case in the traditional examples *John hit Mary and then he KICKED her* and *John hit Mary and then he kicked BILL*, where the difference is one of context (anaphoric reference) which overrides the semantic similarity of argument structure. It only remains to relegate low-level syntactic or phonological factors such as Bolinger's considerations of word order to the bottom of the hierarchy, and we have a clear precedence of factors from left to right thus:

Intention > Context > Semantics > Word Order

This is not a hierarchy which would be instantly accepted by many workers in the field of intonational theory, but nevertheless it is advanced here in all seriousness as a resolution of the syntax/focus debate. Indeed, we suspect that the main reason this hierarchy would not be more widely accepted at face value is that, as discussed above, most intonational theorists are firmly entrenched in one camp or the other and are more interested in pursuing the possibilities within that camp than they are in examining the views of the other camp or in reconciling the two views.

The most obvious question which might be posed by Designers regarding our proposed hierarchy of the determinants of intonation is: What has happened to syntax? The importance of syntax in such a hierarchy has been assumed by so many for so long that even the strongest proponents of the focus approach such as Ladd and Gussenhoven have not denied it a rôle in the determination of intonation. Despite this, we would contend that on the basis of the data presented in the literature, some of which is discussed here, there are no grounds for holding that syntactic structure has any influence on intonation. In fact, as is pointed out above, even those such as Bresnan who are most closely associated with the syntax-based approach to intonation have been forced to retreat to a position based on something much more akin to semantics than surface syntax (Bresnan 1972:326). Moreover, recent trends in syntactic theory such as the work of Steedman (1985, 1987) and Pollard (1988) stress the congruence of syntax and semantics to the extent that syntax above the level of word order effects loses its autonomy. Steedman in particular suggests that a view of syntax as little more than lexical categories and word order conventions is quite compatible with a semantic analysis which will drive prosodic rules such as Pierrehumbert's (Steedman 1989, 1990). The idea that syntax is at least to some extent a by-product of semantics and other factors has also been expressed by authors such as Schank (1975) and Joshi (1990). It would seem, then, that even syntacticians are wondering what has happened to syntax, and not just in the area of the syntax-prosody mapping. In the light of this widespread doubt, we feel perfectly justified in taking a stance which denies any significant rôle for syntax in the determination of prosody but which acknowledges the contribution of syntactic theory to the understanding of what DOES influence intonation.

1.4.1.3 European Work

With the notable exceptions of Gussenhoven and Halliday, most of the authors whose work has been central to the debate regarding the factors determining intonation have worked within the American school of linguistic research. There are, however, several other schools of intonation research, particularly in Europe. The traditional British school which produced Halliday has also spawned other notable contributions, and the Dutch tradition to which Gussenhoven's work is related has produced remarkable quantities of research on intonation. In addition to these schools, there are the French, Germanic and Praguian linguistic traditions, all of which have been represented in recent work on intonation synthesis. The major contributions of all these schools are presented in this section.

Traditional British research into the formal nature of prosody has been almost exclusively descriptive in character, again with the exception of Halliday: that is to say, British work has concentrated on describing the physical and perceptual nature of prosodic phenomena in their own terms without relating them to syntax, semantics or other aspects of language or cognition except in an informal or anecdotal manner. There have thus been several major works of descriptive phonetics such as Palmer's (1922) course of instruction in intonation, Kingdon's (1958ab) introductions to the linguistic systems of intonation and stress, and O'Connor & Arnold's (1961) practical prosodic transcription system, all of which share many characteristics typical of the British approach to intonation (such as a concentration on pitch movements rather than targets) but none of which provide any clear statement as to how the phenomena which they describe relate to the rest of the English language. Crystal's (1966) pioneering thesis was the first attempt within the British tradition to fit large amounts of phonetic data into a formal description of English prosody, and as such it is still the most complete and consistent phonetic account of this area. Crystal's work was not followed up to any great extent, and therefore remained largely unsurpassed during the seventies, but more recently British researchers have been attempting to apply the same descriptive approach to intonation synthesis. The work of Knowles and his colleagues at Lancaster (e.g. Knowles 1984; Knowles & Lawrence 1987) has attempted to derive prosodic rules for use in TTS systems from a corpus transcribed prosodically according to the O'Connor & Arnold system, and has revealed the difficulty of interpreting the transcription in phonological terms; research at UCL (e.g. Johnson & House 1986; House & Johnson 1986) is applying Halliday's (1967b) 5-tone system to intonation synthesis; and IBM(UK)'s research on TTS used a British nuclear tone model, based loosely on the systems developed by O'Connor & Arnold (1961) and Crystal (1966, 1969), to synthesise intonation contours (Williams & Alderson 1986). However, all these recent investigations have fallen foul of the problem that the British tradition is based on perceptual rather than instrumental or phonological analyses and that its descriptions and transcriptions are therefore highly subjective and not necessarily related to the acoustic parameters to be synthesised.

The Dutch perspective on intonation synthesis is also based on perceptual criteria. However, whereas the end of the sixties marked the beginning of a recession for British work on intonation, research on prosody was entering a boom period in Holland. In the mid sixties, the Instituut voor Perceptie Onderzoek (IPO) was set up in Eindhoven and its investigation of intonation synthesis for Dutch began. The basic principles of the IPO paradigm were stated in 't Hart & Cohen (1973) and 't Hart & Collier (1975), and a great deal of research has been carried out within this paradigm without the paradigm itself being significantly altered. IPO's research has been mainly directed towards relating listeners' perceptions of intonation to actual F_0 tracks via a series of intermediate levels: stylised F_0 contours, pitch movements, pitch contours and intonation patterns ('t Hart & Collier 1975:238). This approach is founded on the belief that relating F_0 curves and perception directly is not possible using current experimental techniques and that they should therefore be related indirectly, using **perceptual equivalence** between successive levels as a check on the validity of the relation.

Starting with a corpus of F_0 contours extracted from real speech, the first step in the IPO paradigm is to construct a **close-copy stylisation** of each contour: this is "a stylization which is perceptually indistinguishable from the original and satisfies this condition with the smallest possible number of straight lines." (de Pijper 1979:67) The second step is to derive a set of primitive pitch movements from the set of close-copies by applying further stylisation: these stylisations are presumably not "perceptually indistinguishable" from the originals, but are claimed to retain "perceptual equivalence" with them ('t Hart & Collier 1975:239). This reduces the intonation contours to sequences of elements chosen from a small inventory of pitch movements with invariant slope and duration ('t Hart & Cohen 1973:314ff.). Next, a grammar is constructed which is capable of automatically generating the corpus of perceptually equivalent stylisations from this inventory: this grammar can be expected to over-generate, but these unattested contours may well be present in a larger corpus. The IPO claim is that this grammar will reveal the constituents (accents, boundaries, interpolations, etc.) of the intonation contours which can be built from the basic pitch movements ('t Hart & Collier 1975:246), just as syntactic grammars take a set of word classes and reveal the phrases which make up a sentence. Finally, the set of contours generated by the grammar is presented to subjects who are asked to group them on grounds of perceived "similarity", in the hope that this will reveal the set of meaningful intonation patterns into which the contours should be classified.

The IPO paradigm has been applied to numerous languages, including English, and is in theory applicable to any corpus of F_0 traces. In addition, both the researchers and the experimental subjects appear to be capable of performing consistently in their judgements of stylisations. However, there are two major problems with this paradigm. Firstly, there is a profusion of perceptual criteria (indistinguishability, equivalence, similarity) none of which is clearly defined. Indistinguishability presumably relates to our perception of pitch rather than to our interpretation of intonation, and as such it is not clear what the function of close-copy contours is other than to demonstrate the limitations of human auditory perception: on the other hand, although the more stylised contours may fall into particular patterns their status is unclear as there are by definition perceptible differences between them and the natural contours whose significance is unknown. Secondly, despite attempts dating from de Pijper's (1979) efforts to develop procedures which will generate stylised contours automatically from natural speech input ('t Hart 1984; Willems et al. 1988; Terken 1989a), IPO still does not have any automatic stylisation procedure: this suggests that the human stylisers are applying more knowledge to the task than is reasonable, including knowledge of the structure and meaning of the language, and are therefore pre-judging what is perceptually important. Workers at IPO are aware of this problem themselves (Collier 1989:39; Terken personal communication), and in fact many have moved away from the use of stylisation to investigate the structure of intonation systems and have concentrated on the function of intonation in discourse: the use of stylisations in these investigations allows one to ignore variations in naturalness or emphasis of intonation and concentrate on the effect of context. In particular, the work of Nooteboom, Kruyt and Terken in various combinations has tested the relations between accent, focus and the **given/new** distinction (Prince 1981): Nooteboom, Kruyt & Terken (1981) introduces this line of investigation, states its use of and applicability to speech synthesis systems (p.9), and sets out the subjects for future investigation (pp.30-1) which include the pragmatic factors determining accent placement, the interpretation of accents and the contribution of accents to intelligibility and comprehension. The theses of Kruyt (1985) and Terken (1985) constitute perhaps the most significant Dutch contributions to the study of the functions of accent: the former concentrates on listeners' interpretations of accent, and the latter on the relation between accent and contextual givenness. Most of this work has been carried out in the theoretical framework of authors such as Ladd, Selkirk and Gussenhoven, with the assumptions that accent marks focus (Terken 1985:9) and that syntactic and prosodic domains are congruent (Nooteboom, Kruyt & Terken 1981:14): however, a more recent article (Nooteboom & Kruyt 1987) explicitly examines the theoretical claim that accent marks focussed (new or communicatively important) items in the light of the accentuation strategies of professional readers. Despite the authors' dissatisfaction with their experiments (p.1520ff.), their results are quite damning for both the assumption of one accent per focus domain and the claim that syntax determines accent placement within such a domain. Nooteboom & Kruyt (1987) took pairs of Dutch sentences where the second member of the pair contained both given and new material, and varied the accent patterns and constituent orders in that member so that the effects of linear order and newness could be examined. Subjects were asked to rate the appropriacy of the accent patterns on the traditional Dutch 1-10 scale (p.1516): their judgements clearly showed that the relation previously assumed by these authors between focus (indicated by accent) and newness does not hold, and moreover that the factors affecting this relation include the degree of specificity of the term to be accented, its position relative to other accents in the domain or utterance, and the relative newness of the other items in the utterance (pp.1520-1). With specific reference to assigning accents automatically, they found that accenting given information is much more acceptable than not accenting new information (p.1518) but that this only applies to accents which are not utterance-final (i.e. pre-nuclear accents): the placement of the nucleus was found to be crucial (p.1520). On the basis of these results, the authors conclude (p.1521):

> It seems very unlikely that we can teach machines to take such highly abstract semantic properties of text into account when generating accent patterns. Thus we can be fairly certain that speaking machines for some time to come, and perhaps for always, will go on producing inappropriate accent patterns. What we do not know is how often in reading out text this neglect of semantic aspects of accentuation will result in unacceptable accentuation or de-accentuation. It is imaginable that syntactic control of accent patterns in most cases leads to an acceptable and only rarely to an unacceptable result. This awaits further research.

This is very much the attitude taken by Dutch workers on intonation: the problems of realising a particular accent pattern, at least for the purposes of speech synthesis or experimental stimuli, are largely solved; the real problem is discovering not only what the factors are which play a major rôle in determining appropriate accent placement but also how these factors can be predicted and interpreted by an automatic TTS system. In the main, this seems to be an accurate assessment of the current situation.

Whereas the work of Dutch researchers both has direct relevance to English (because of the close relation between English and Dutch) and is frequently applied to English data, the applicability to English of work in other European traditions is generally much less obvious: moreover, much of this work is not accessible to the monolingual English community, being either unpublished or written in a language other than English. For example, recent work within the French Designer approach on intonation with relevance to synthetic speech is published mainly in French. As a result, this work has largely been ignored by researchers concerned solely with English. Nevertheless, the concerns of the researchers are very similar to those of workers on English and their ideas and results are consequently relevant to the present study.

Work on French prosody from the Designer perspective falls, like work on English, into the syntax-prosody camp and the non-syntactic camp. The former view is the standard one, with French syntacticians as convinced as any that prosody is directly related to syntax:

> certaines unités intonatives . . . devaient être incluses dans la base syntaxique si l'on voulait rendre compte du comportement de l'intonation dans les faits d'énonciation. On a déduit également des axiomes d'une théorie de l'intonation . . . que cette structure, qui se réalise toujours selon ses lois propres, s'associe cependant à la structure syntaxique.
> Di Cristo (1981:272)

Amongst those whose work is concerned at least partly with synthetic speech, the main proponent of this view is Philippe Martin (1981, 1982). Martin takes a very similar line to that followed by Selkirk and Bing above, claiming that prosody is predictable from syntax in the normal case although speakers can of course choose to diverge from this norm:

> la structure intonative fournit une esquisse de la syntaxe et constitue, d'une certaine façon une moyenne de préparer le travail de codage et de décodage de la structure syntaxique.
> Martin (1981:270)

This is the traditional view of the form and function of prosody in French. There is, however, considerable support for the view that syntax cannot provide sufficiently accurate prosodic predictions for high-quality TTS systems: the work of Liénard and Choppy (Liénard & Teil 1970; Choppy & Liénard 1977; Choppy 1979) is perhaps the most well-established voice in this camp, but it is by no means alone, and indeed seems to be gaining favour amongst recent authors (Sorin et al. 1987; Pasdeloup 1990ab; Guaitella & Santi 1990). The view taken in Choppy & Liénard (1977) clearly acknowledges both the relevance of factors other than syntax to the prosodic realisation of a text and the superiority of a treatment which can handle text regardless of its syntactic structure or even its grammaticality:

> Le programme de traitement prosodique ne comprend pas d'analyse syntaxique. Il ne s'agit pas ici de nier toute importance de la syntaxe, mais de dire que le rôle de la syntaxe n'est pas premier et unique. Les questions que nous posons sont (i) d'ordre théorique: aucune démonstration n'a été

> faite qu'un enfant de 3 ans dirigeait sa prosodie à partir d'une connaissance innée (ou acquise dans les premières années de la vie) de la syntaxe, (ii) d'ordre pratique: la prise en compte d'une analyse syntaxique dans un tel système de synthèse pose des problèmes importants relatifs à la rapidité du traitement et à l'encombrement de taille mémoire, et ne permettrait pas de traiter n'importe quelle phrase (phrases agrammaticales). Il semble que pour des phrases ambigues, la prise en compte de la syntaxe serait un élément déterminant pour la prosodie. Dans la plupart des cas, cette prise en compte de la syntaxe doit être assortie d'une analyse sémantique.
> Choppy & Liénard (1977:215)

This view is reflected in much of the current work on French prosody for TTS systems amongst those concerned with means as well as with ends. Valérie Pasdeloup's (1990a) thesis, for example, rejects the syntax-based accounts of French prosody proposed by such as Di Cristo & Rossi (1977) and Dell et al. (1984). Instead, she proposes principles based on physiological constraints and semantic constituents which constitute potential prosodic domains. In Pasdeloup's account, although the potential boundaries may be identified for any text sentence the assignment of domain boundaries in any actual utterance is also dependent on factors of speech rate and style: this gives her "the possibility to generate various acceptable prosodic structures for a given sentence." (Pasdeloup 1990b:193) The constraints on the length and composition of prosodic domains which Pasdeloup proposes are strikingly similar to those proposed by Gee & Grosjean (1983), but appear to have been formulated independently: her linguistic and psychological constraints have much in common with those we present for English in Chapter 2, and include such ideas as rhythmic alternation and subject-predicate boundaries.

Choppy and Liénard's work is also a large influence on the research of Santi and Guaitella at Aix. They have pursued the idea that punctuation is more than just a cue to syntactic boundaries, and indeed that in less stilted speaking styles there is little more than a coincidental correspondence between syntax and punctuation. Guaitella & Santi (1990:177) found that there are clear differences in the function of certain prosodic phenomena, e.g. silent pause, between read and spontaneous speech: in the former, there is an assumption that the speaker is marking a boundary (syntactic or otherwise); but "in spontaneous speech, the silent pause plays a rôle close to the one played by hesitation, that is to say not perceived as an intentional act of text structuration." They argue that a TTS system should be able to generate spontaneous-sounding output as well as read monologue (a view which is shared by authors such as Granström (1990) and Tatham (1990)), and that the conventional approach to punctuation and indeed to syntactic analysis in TTS systems is therefore inadequate. It is indeed difficult to see how the "one parse, one prosodic structure" philosophy of the syntacticians can account for such different realisations of the same text.

Perhaps even more inaccessible to English-speaking audiences are the ideas of the Prague School linguists. Nevertheless, this important European school of thought has published work on intonation such as that by Daneš (1960, 1972) and Firbas (1980) dealing with and written in English. The Prague School represents probably as abstract a view of intonation as can be found: the central notions are **functional sentence perspective**, which expresses the intended communicative effect of a particular sentence or utterance, and **communicative dynamism** (CD), which reflects the varying degrees of centrality of different items to the communicative effect of the sentence or utterance. Intonation is seen as dependent on the functional sentence perspective, in that it is determined by the speaker's intentions, but its effect on the CD of the various constituents of a sentence is dependent on syntactic and semantic considerations of **markedness** (Daneš 1972:226). In terms of the syntax-prosody debate, Daneš rejects the idea of syntactically-determined Normal Stress partly because not all sentences have a Normal Stress realisation (e.g. Czech negatives) and partly because in the Praguian view no utterance can be context-independent: "In other words: every utterance points to a 'consituation' (to use Mirowicz's term)." (Daneš 1972:221) He is highly critical of Chomskian attempts to fit intonation into a syntactic strait-jacket (p.230), and specifically states that the problems which examples such as the *plans to leave* ambiguity (example (1) above) pose for a syntactic account of intonation are entirely the result of the failure to take semantic and pragmatic factors into consideration:

> In other words: rather than saying that the intonation here works as a grammatical device (distinguishing, e.g. an object clause from an adverbial one, or determining the function of the conjunction), we should rather say that this is an accidental effect of two possible T-C [topic-comment] structures of the given utterance.
> Daneš (1972:229)

Where Praguian work concentrates on the cognitive and communicative aspects of intonation, the Germanic tradition is more concerned with what it terms "linguistic phonetics", i.e. the perceptually salient characteristics of F_0 contours. The Scandinavian research on intonation synthesis is almost all incorporated into the INFOVOX multilingual TTS system which is discussed below as part of the Producer approach: here we will concentrate on work on German language TTS, which falls more squarely within the Designer camp. With the exception of articles such as Richter (1984) and Batliner & Nöth (1989), most of the work on German TTS is based on the Kiel intonation model (Kohler 1986, 1988). This model is founded on observations of F_0 contours in isolated sentences, and assigns accent on the basis of syntactic information alone: however, the model is unusual in that "The positional variance of accenting peaks has been accorded great importance in the Kiel approach to intonation." (Gartenberg & Hertrich 1988:997) Indeed, the vast majority of Kiel's published work on intonation deals with the perceived differences of emphasis and meaning between different peak alignments (Gartenberg & Hertrich 1989; Hertrich & Gartenberg 1988, 1989; Kohler 1987): the results of this work demonstrate a clear correlation between peak position (early, medial or late) and listeners' perceptions of "the corresponding changes of meaning from 'established' to 'new' to 'emphatic'." (Kohler 1987:152) It appears that the **given/new** distinction is conveyed by emphasis, with **given** items accorded a lower degree of emphasis by early positioning of the peak, as Gartenberg & Hertrich (1988:999) found:

> All other things being equal, the later the peak the greater the degree of emphasis it conveys. This finding is particularly interesting when viewed as complementary to the frequently mentioned correlation between greater excursion size and the expression of emphasis.

These results are for German, and their applicability to English has not been tested: however, they could easily be incorporated into the CSTR TTS system given the appropriate discourse information to control the assignment of varying degrees of emphasis. The Kiel model seems to have gained in the area of peak alignment at the expense of prosodic structure: there is no hierarchy of domains even within a sentence, and downdrift across utterances is simulated by explicitly reducing the excursion size of successive peaks (Kohler 1990:191). Nevertheless, the model is sufficiently detailed and explicit to have been adopted as the prototype for INFOVOX's commercial German TTS system.

1.4.1.4 Implementations

Designer research is by definition less directed towards implementations than towards formalisms: there are, however, several implementations of this work either as research tools or as demonstrations of the adequacy of the formalism. The implementation of the Kiel model has already been mentioned, and this is the only working German TTS system based on a Designer formalism of which we are aware: implementations of Designer approaches to other languages are outlined here. All the TTS implementations discussed in this section are non-commercial in nature and of European origin: unfortunately, American efforts in speech synthesis seem to concentrate on commercial systems such as DECtalk, PROSE2000 and AT&T's development system, the detailed workings of all of which are closely-guarded secrets.

IPO's ideas on intonation synthesis have been implemented in working TTS systems for several languages including Dutch (Terken 1989a) and English (Willems et al. 1988). These two systems differ mainly in the phonetic details of the realisation of accents: the processes and principles on which they are based are identical. Terken (1989a) demonstrates that the IPO phonetic model for Dutch is comparable with natural intonation for isolated utterances but still not satisfactory across running text: there has been no evaluation of the appropriacy of automatic accent placement in the IPO TTS system. The situation is much the same for the British English phonetic model proposed in Willems et al. (1988), except that this model has yet to be evaluated on spontaneous utterances or running text.

The other Dutch implementation based firmly on Designer theory is that produced by Joan Baart at Leiden and described in his thesis (Baart 1987). This system assumes a full syntactic parse as input, plus Gussenhoven-style focus domains: it accents every [+focus] content word, and then deletes some accents on rhythmic grounds. Baart claims this system incorporates the conclusions of syntactic, focus-based and metrical approaches to intonation: however, it is restricted to processing isolated sentences and even then it encounters serious problems with the treatment of contextually **given** items (Baart 1987:161). As Baart points out, despite the linguistic knowledge which his system possesses, "This type of error is clearly hard to avoid, as long as machines do not really understand what a text is about." (Baart 1987:161)

Although there is no complete implementation of the traditional British model of intonation, in true Designer style there are formalisations of subsections of the model. Crystal's (1966) findings on tone-unit boundaries, for example, were formalised into a schema for automatic boundary assignment in Crystal (1975:16), and more recently Altenberg (1987:46ff.) re-examined Crystal's schema on the basis of data from the London-Lund Corpus (Quirk & Svartvik 1978) and produced a more explicit set of rules and a description of how they should be applied. Although these rules appear to give good coverage and reasonably accurate (c.75% correct boundary assignment) treatment of Altenberg's data, this is only a subjective and informal evaluation as the rules have not been implemented in an automatic system: moreover, the input which Altenberg (1987:120) assumes to drive his rules includes detailed syntactic and semantic analysis although he is aware that "It is uncertain to what extent an automatic parser can accomplish this". The problem of automatic tone-unit assignment in a practical TTS system thus remains unsolved.

Johnson & House (1986) present an implementation of the traditional British approach to intonation assignment WITHIN the tone-unit: given a division into such units, rules derived from a small corpus of read text assign one of four Crystal-style nuclear tones to each tone-unit on the basis of statistical tendencies. In addition, there is a large degree of permitted variation in the realisation of these tones so that a monotonous F_0 contour is avoided. This model is currently being revised and updated at UCL (House 1989), and forms the basis of research such as that reported in House & Youd (1990).

French language implementations of Designer research are more numerous, but again tend to handle small sets of data and very limited phenomena. They generally involve grafting a specialist module onto an existing TTS system such as those developed at ICP (Bailly 1986) and LIMSI (Teil 1975). Choppy's work on punctuation, for example, was implemented as part of the LIMSI speech output system, and more recently Aubergé's (1990) attempts at superimposing contours from different hierarchic levels (in a manner reminiscent of Fujisaki & Hirose's (1983) treatment of Japanese) was founded on the ICP TTS system. Probably the best example of a French language TTS system whose prosodic module is based on Designer ideas is the system developed at CNET (Sorin et al. 1987): this system takes the non-syntactic view, rejecting the syntax-prosody correspondence (partly, it must be said, on the grounds that no practical TTS parser is available) in favour of metrical and prosodic hierarchies determined from lexical information, positional factors and constituent length. As with the CSTR system, CNET's TTS strategies are aimed at getting the prosody right rather than producing full linguistic analyses: however, unlike the Producer approaches described below, this does not prevent them from making use of linguistic insights at all levels from the phonetic to the semantic.

1.4.2 Producers

1.4.2.1 Common Ground

The defining characteristics of Producer work on intonation synthesis tend, predictably, to be diametrically opposed to those of Designer research. Thus, where Designers are concerned to develop an elegant account of some carefully-chosen subset of intonational phenomena, Producers set about intonation as a whole and attack the problem from all sides simultaneously. The goal of the Producer approach is to model the entire range of intonational phenomena, using whatever methods are available. As an example of this approach applied to the highly problematic area of parsing natural language text, for instance, Producer research has recently shown (Garside, Leech & Sampson 1987) that the syntactic word class of the vast majority of words in a 6-million-word text corpus could be predicted from their orthography, specifically the last few letters of each word. Although the results of such research are impressive, it is tempting to compare such modelling of natural language by the use of data reduction and statistical analysis to the modelling of architectural monuments with matchsticks and glue: from a distance there may be a striking similarity between the model and the original, but such model-making does not tell us very much about the raw materials, function or underlying structure of the object we wish to model.

Devotees of the Producer approach tend to come from a background in engineering or computer science rather than the linguistic or cognitive traditions which underlie Designer research. This leads in many cases to a very naïve approach to speech and language. A common misconception illustrative of this naïveté is the assumption that a statistical average of several F_0 contours for a particular utterance type will produce a basic or archetypal contour for that utterance type. This assumption underlies much Producer work on creating natural F_0 databases (e.g. Aubergé 1990; Scheffers 1988; Mortamet et al. 1990), and yet the effects of such averaging could be predicted, in the light of linguistic knowledge about the effect of different speakers and lexical items on the F_0 contour, as a blurring or erasure of the most salient portions of the contour: in the case of Aubergé's work at least, this is exactly what happened. The inverse of this assumption is demonstrated in Traber (1990:143), where the naturalness of synthetic intonation is assessed on the basis of a statistical fit between natural and synthetic F_0 contours: again, there is no allowance made for the fact that some parts of the contour are more important than others for the perception of naturalness. Young & Fallside (1980) and Rodriguez-Crespo & Escalada-Sardina (1990) both take a revised version of this view, in that they extract average contours from natural speech for various syntactic phrase types or positions. However, although the syntax is controlled, the information content (lexical and propositional) is not and so even in informal assessments the shortcomings of this approach are soon evident: "The results were encouraging for utterances containing monosyllabic words only but were less so when words of more than one syllable were included." (Young & Fallside 1980:252) Another common assumption (e.g. King 1989:123; Yiourgalis & Kokkinakis 1990:412) is that large portions of F_0 contours can be predicted on the basis of the sentence-final punctuation: this may be partially true in some cases (see Section 2.1.4 below), but studies of such long standing as Uldall (1960, 1962) show that this is far from adequate. The inadequacy of a single contour for declarative or interrogative sentences in a multi-lingual TTS system is surely even more obvious, especially to non-anglophone researchers (Ladd 1983b), but nevertheless systems such as the Hungarian MULTIVOX (Olaszy et al. 1990:280) apply the same "sentence melody" to half a dozen very different languages.

Producer research on intonation differs markedly from the Designer approach in that it is largely corpus-based whereas the Designer approach puts little emphasis on corpora. As a consequence, Producer work has processed huge amounts of speech data, and despite its linguistic naïveté it has produced some very useful and interesting results. The work of Cooper and colleagues, for instance, has provided experimental evidence based on considerable amounts of data which indicates or confirms such notions as partial declination resetting (Cooper & Paccia-Cooper 1980:212), temporary register lowering during parentheticals (Cooper & Sorensen 1977:85), and the importance of boundary tones in characterising the difference between declarative and interrogative utterances (Eady & Cooper 1986:409). European Producers have also produced many interesting and useful results through the analysis of speech corpora: Quazza et al. (1989:508-9) found a 50% correlation between pauses and certain syntactic boundaries (i.e. half the occurrences of syntactic boundary X were marked by pause Y), indicating the optional nature of the syntax-prosody correspondence; Fallside & Young (1978) confirmed a finding of Olive & Nakatani (1974) that pitch was the most important suprasegmental parameter in synthetic speech quality; and despite his naïve approach, Scheffers (1988:981) contributed further evidence that "there is no one-to-one relation between measured F_0 and perceived intonation."

It is hardly surprising that Producers are often unaware of linguistic research on intonation, since most linguists are equally unaware of the interesting Producer work just cited. However, simply being aware of linguistic research in this area does not necessarily solve the problem. As their linguistic knowledge is limited, Producers are obliged to take Designer research at face value and to make of it what they can. Their lack of the knowledge necessary for a critical evaluation of the linguistic literature results in the acceptance by most Producers of whatever Designer ideas they encounter as immutable axioms of a precise science rather than the partial and transient accounts which they generally are. Consequently, much of the work in this approach is based on linguistic ideas which are significantly out of date or which only reflect one of a number of views within linguistics. Among instances of the problem of superseded linguistic theory forming the basis for Producers' researches are the reliance of Traber (1990:141) in building a neural network-based TTS system on syntactic accounts of German written in 1966 (the age-difference between the engineering theory and the linguistic theory involved being approximately 20 years); and the foundation of F_0 synthesis schemes on informal and outmoded accounts of suprasegmental phenomena, such as Witten's (1977:241) implementation of Abercrombie's (1967) "voice dynamics" and Young & Fallside's (1980) use of O'Connor & Arnold (1961) as the theoretical basis for their F_0 generation program. Young & Fallside (1980:242) also assume the existence of deep and surface syntactic representations, but many Producers go further and accept the existence of specific prosodic correlates for particular syntactic transformations: this is the case in the MITalk system, where O'Shaughnessy's F_0 algorithm specifically caters for "Boundaries created by a number of syntactic transformations" (Allen et al. 1987:107), and Cooper & Paccia-Cooper (1980:8ff.) likewise accept the Chomskian view that prosodic rules take account of the transformational history of a sentence. Broadly speaking, this is the position advocated by linguists such as Bresnan (1971, 1972) and Stockwell (1972:88), but linguistic ideas have progressed considerably since then: unfortunately, this progress seems to have passed many Producers by.

The most obvious example of an imbalance in the linguistic ideas which have influenced Producer work on synthesising intonation from text is the blind acceptance by the majority of these workers of the notion that prosody is determined solely and completely by syntax. This is stated bluntly as a universal truth at the head of many publications on Producer research: Barber et al. (1989:518), for instance, begin their account of intonation in an Italian TTS system with the statement that "Obviously, F_0 contour depends on the syntactic structure of the language." Such explicit statements of the syntactic view are common in the Producer literature (e.g. Russi 1990:117; Klatt 1987:774; Schnabel & Roth 1990:121), and its implicit acceptance is even more widespread (Carlson & Granström 1973:31; Rühl et al. 1984:243-4; Shi 1989:522; Cericola et al. 1989:386; Carlson et al. 1990:276; Sakai & Muraki 1990:329; Tatham 1990:239). Even when semantic and pragmatic information is available, Producers still rely on syntactic information to derive prosodic structure (e.g. Young & Fallside 1980; Sakai & Muraki 1990). There are of course exceptions (Kager & Quené 1987, 1989; Quazza et al. 1989; House & Youd 1990; Moulines et al. 1990), and there are those who derive prosody from syntax because it is available rather than through any belief in the processing model in Figure 4 (Kulas et al. 1986; Kugler-Kruse & Posmyk 1987;

Matsumoto & Yamaguchi 1990): the latter are at the core of the Producer approach, but the former are generally very close to the Designer frontier.

Whereas a Designer account of intonation might be judged on its power or coverage, or the types and number of primitives which it assumes, the universal criterion for assessing Producer models is the question of whether they "work", i.e. whether they can reliably produce convincing acoustic output in an acceptable amount of time. To make their mark on the scientific community, therefore, Producers are obliged to implement their ideas in a practical system. This situation produces relatively large numbers of implementations of all sizes and designs: whereas there are a handful or two of TTS systems which implement Designer ideas on intonation, there are literally hundreds of Producer implementations, all different. Paradoxically, given the fact that the main purpose of implementing one's ideas is to allow them to be assessed, very few of these systems have undergone any systematic or objective evaluation: the question of assessment is discussed at length in Chapter 3, but it is not an area in which the Producers have much more to show than the Designers despite their declared concern with actual performance. Nevertheless, Producer implementations of intonation generation from text exist for languages from Arabic (Ouado et al. 1987) to Bengali Datta et al. 1990) and there are even systems which handle five or six quite different languages (Olaszy 1989; Granström 1990) within the same architecture.

A final characteristic which, unfortunately, is shared by much Producer research on intonation synthesis is a lack of insight and flexibility in approach. Although Producers do not have the monopoly on blinkered views, they do tend more towards the "if it works, don't fiddle with it" school of thought. To return to the Lancaster analysis of word-endings in the LOB corpus again (Garside, Leech & Sampson 1987): this is a prime example of useful but blinkered Producer research. The fact that orthographic word endings are a good predictor of syntactic word class is an interesting finding, especially for those seeking strategies for word-class disambiguation in unrestricted text (Monaghan 1990b), but the REALLY interesting question is: What does this tell us about word-classes? The Producer approach is to restate the findings (i.e. "This tells us that word endings are a good predictor of syntactic word class!") and leave it at that: "Don't knock it, it worked" (Adams 1979:64). This is, however, missing an important fact which will be obvious to most linguists: word endings are highly correlated with morphological suffixes, and it is well known (Matthews 1974; Bauer 1983; Fudge 1984) that final suffixes determine word-class in the majority of cases. The logical next step, in this case to examine how much of the word-ending information is actually morphological (rather than orthographic) information and how much is something additional and perhaps not previously recognised, is thus rarely taken by the Producers. Such a position obviously stems from the aforementioned lack of awareness of linguistic and other literature, but it also indicates a further aspect of Producer philosophy: along with their concern to produce systems which work comes an indifference to the reasons WHY they work. Unfortunately for the Producers, this can also lead to an inability to see why things DON'T work: the case of the common-sense treatment of anomalies (see Section 3.3) is a very clear example of this.

There are many more instances of lack of insight amongst Producers: the work of Cooper and his colleagues whose useful contributions to the study of intonation are discussed above, for instance, is marred by such a lack. In Cooper & Paccia-Cooper (1980:45-6) it is pointed out that a certain durational effect occurs across a range of different syntactic constructions, and even that in all cases it distinguishes between two semantic interpretations: nevertheless, their obsession with syntactic explanations for prosodic phenomena leads the authors to postulate a syntactic meta-construction encompassing all the relevant constructions. The argument appears to be that since we know this has to be a syntactically-conditioned phenomenon its existence motivates a syntactic class. Eady & Cooper (1986) are prey to a similar degree of selective blindness: although their comparison of F_0 contours is much more sophisticated than those of Aubergé (1990) or Mortamet et al. (1990), being based on points of salience rather than the overall contour, their discovery (p.411) that sentence-initial focus in declaratives produces a flat, low contour over the rest of the sentence (whereas in all other combinations of focus and mood F_0 continues high for much longer) leads them to conclude simply (p.413) that "linguistic focus can be manifested in the acoustical attributes of the entire sentence." Fully five years after Pierrehumbert's (1980) account of F_0 contours in terms of tunes, then, Eady & Cooper (1986) fail to mention the possibility that this flatness is due to a lack of intonational events after the focal accent rather than to some holistic "sentence-initial-focus" contour.

Nebbia (1990) misses a very similar point, in spite of basing his system directly on Pierrehumbert's model: he includes an entire tune, much as in the traditional British approach (e.g. Halliday 1967b:55), simply to produce continuation rises at non-final declarative boundaries. This tune "minimally contrasts with the declaratives for the boundary tone choice" (Nebbia 1990:331) in that it has a high boundary as opposed to the normal declarative low boundary: however, such a boundary is certainly not obligatory in Italian non-terminal declaratives and indeed may be unnaturally high. It is a blind insistence on assigning some boundary tone, then, which results in this unnecessary and inappropriate third tune when in actual fact the mere absence of a boundary tone would preserve the declarative/interrogative symmetry of Nebbia's account and produce at least as appropriate a phonetic realisation of continuations. (See Section 2.1.3 below, where we take precisely this approach for English.)

Witten, in an implementation based on Halliday's 5-tone system, illustrates another aspect of this lack of insight amongst Producers. Despite having only a handful of different intonational possibilities for each "tone group", these possibilities are described in terms of no fewer than ten parameters (most of which are continuously variable) which specify such unintuitive quantities as "departure from linearity on each foot of pretonic" or "fraction along foot of the non-linearity position, for the tonic foot" (Witten 1977:255). As these parameters are manipulable at a lower level than the tone group if desired, Witten gains flexibility: but he pays for this in complexity and in lack of generality. Witten seeks to have the best of both worlds with a system which requires minimum input specifications (the 5-tone system) and yet is fully flexible in its output (the ten parameters), but he fails to see that a choice has to be made between these two: if the 5-tone system is not adequate, then some other input is required; if it is adequate, then there is no need for this multiplicity of parameters and certainly no way to control them. Witten, like many Producers, has produced a fine piece of craftsmanship without much thought for what it might be used for or how it might bc controlled.

1.4.2.2 Differences

For obvious reasons, there are few major theoretical differences between Producers, and certainly none which split this approach to the extent that the syntax-prosody correspondence divides the Designers. There are, however, many differences of assumptions and ambitions which delineate various sectors of Producer research. These differences tend to be found in the techniques on which Producer implementations are based and the data which they process, and so this section will mainly discuss characteristics of these implementations rather than explicit statements of theoretical positions.

Although the majority of Producers working on intonation synthesis come from a background of engineering or computer science, there are also numerous phoneticians and computational linguists. Among the most linguistics-based research subscribing to the Producer approach is the work of René Kager and Hugo Quené at Utrecht (Kager & Quené 1987, 1989; Quené & Kager 1989, 1990; Quené & Dirksen 1990). They come from backgrounds in phonetics and phonology, and have developed an algorithm for assigning prosody from text as part of the Dutch national speech technology programme ASSP. As stated above, Kager and Quené are on the edge of the Producer approach: their primary concern is to produce a working system, and to that end they have incorporated a great deal of *ad hoc* information and many heuristic methods into their system, but their interest in the underlying linguistic regularities which that system approximates comes a very close second. There are two principal steps in Kager and Quené's PROS algorithm: the first is to derive a **prosodic sentence structure** (PSS) from text, and the second is to assign accents to the domains in the PSS.

Quené & Kager (1990:2-4) clearly state that in their view the PSS should theoretically be derived directly from surface syntactic structure: they are thus firmly in the syntax-prosody camp. However, they are also well aware that there are two very good reasons why this is not possible in current TTS systems (p.4):

> Firstly, there is no algorithm for syntactic analysis (parser) available which performs satisfactorily for our purposes. Such a parser must be able to analyse any text, at a speed which exceeds the average speaking rate. ... Secondly, if such a parser did exist, it would run into great difficulties when analysing syntactically ambiguous sentences

They therefore do not attempt to produce such a syntactic parse, nor do they assume its existence:

> Instead, the PSS is derived directly from the orthographic input sentence, by rules which do NOT refer to a sentence's syntactic structure. Consequently, the resulting PSS can only approximate the theoretical prosodic structure, since not all relevant syntactic information is available for the prosodic analysis.
> Quené & Kager (1990:5)

Kager and Quené argue that, for Dutch at least, an adequate PSS can be constructed on the basis of orthographic cues (capitalisation, word endings, punctuation, etc.) and identification of impermissible sequences of word-classes such as **determiner+verb**. PROS has been carefully optimised for Dutch by the analysis of sizeable corpora, but these probabilistic refinements have been overlaid on a consistent, language-independent theoretical base: the PSS is based on Nespor & Vogel's (1982) Phi and Int domains. In PROS, a Phi is defined as "a lexical head (i.e. noun, verb, adverb or adjective), its left-hand specifiers, and all non-lexical words to the left up to the next lexical head." (Quen é & Kager 1989:214) Int domains generally correspond to the immediate daughters of S-nodes in surface syntax, but there are exceptions:

> Some constituents obligatorily form an Int, such as displaced syntactic constituents, parentheticals, and non-restrictive relative clauses. Complex NP's and subordinate clauses form Int's as much as possible, depending upon their length and upon the style of speech.
> (Quené & Kager 1989:214)

As with most TTS implementations, no further mention is made of variations in speaking style. Quené & Kager (1990:3) do, however, go into some detail on the matter of length constraints on domains[6]:

> Prosodic domains tend to be of equal length as much as possible, and their length increases in faster speech. To account for these effects, separate rules restructure the prosodic domains. An optional rule joins a Phi consisting of one lexical head with the Phi to its left under some syntactic conditions.

[6] See Section 2.8.3 for a discussion of the limitations of this approach compared with the one adopted in the CSTR system.

> Very short Int's can be eliminated by merging them with adjacent Int's, and very long Int's are broken down into shorter ones.

Once the PSS has been determined, this sequence of evenly-spaced boundaries is passed on to the accentuation rules. The Int boundaries are irrelevant at this stage: they will be uniformly interpreted as boundary tones accompanied by 250ms pauses in the final output, and no other use is made of them. Such an invariant interpretation is acknowledged (Quené & Kager 1989:215) to produce "disfluency in the speech output, which inhibits (rather than facilitates) its correct perception." This problem is to be investigated in ASSP project PROS2.

The function of Phi boundaries is much more complex. These boundaries demarcate the domains to which PROS's accentuation rules apply: they are given no direct phonetic realisation, but their location is crucial to the accent pattern assigned to the text. PROS's accentuation rules are based on Gussenhoven's SAAR (see page 27 above), but have been designed to compensate for the absence of the focus information which the SAAR requires: essentially, PROS assumes that all domains are [+focus] except in rare cases such as definite NPs and certain adjectives[7] (Kager & Quené 1989:105). Within each [+focus] Phi domain, PROS essentially follows a very simple accent-assignment strategy:

> Simple CW-FW algorithm
> Accentuate content words (CWs) and leave function words (FWs) unaccentuated. List FWs in a small lexicon (of several hundreds of forms).
> Kager & Quené (1989:103)

In Dutch, verbs apparently count as FWs most of the time, although there are specific environments (basically domain-finally (Kager & Quené 1989:106ff.)) where verbs receive accent. Predictably, this "CW-FW algorithm" assigns too many accents, and so PROS deletes some of them on rhythmic grounds: cases of three CWs in a row trigger deletion of the middle accent if the CWs conform to particular syntactic patterns such as **adverb+adjective+noun, quantifier+X+noun** (Quené & Kager 1989:216). This

[7] See Section 3.4 for a discussion of Kager and Quené's approach to anaphora in relation to our strategies.

strategy does not apply in many cases, and so PROS's output still contains too many accents, but this is a conscious decision. Kager and Quené are aware that erroneous accentuation is preferrable to erroneous DE-accentuation in most cases, and that as we pointed out above such errors are more tolerable in pre-nuclear position:

> It is known that perceptually, the final accent in an [Int] is the most prominent, or nuclear accent. For this reason, special care must be taken at the righthand periphery of sentences and [Int]'s.
> Kager & Quené (1987:245)

Perhaps the most well-known work on intonation synthesis within the Producer camp is the development of the INFOVOX multi-lingual TTS system for several European languages. INFOVOX is the commercial product of the RULSYS TTS development environment at Stockholm's Royal Institute of Technology (KTH): this system has benefited over the past 20 years from direct input by some of the most prominent Scandinavian phoneticians (Bruce et al. 1990; Carlson & Granström 1973; Gårding & Bruce 1981), as well as phoneticians working on other languages (Barber et al. 1988, 1989; Horne 1987; Kohler 1990). The name INFOVOX will be used here to refer to both the commercial and the development systems. Its prosodic modules are based on phonetic and phonological theory, and workers on INFOVOX, just like Kager and Quené, accept the view that syntax determines prosody in all the languages which they handle: however, unlike Kager and Quené, they do not attempt to supplement the meagre syntactic information automatically derivable from text with more heuristic-based techniques. INFOVOX currently handles languages including Danish (Granström et al. 1987), English (Bladon et al. 1987), French (Barber et al. 1988), German (Kohler 1990), Italian (Barber et al. 1989), Norwegian (Carlson et al. 1990) and Swedish (Carlson & Granström 1986). Although the details of prosodic realisations differ markedly between many of these languages, the general approach to prosody generation is intentionally language-independent. Figure 6 gives a schematic view of an INFOVOX TTS system: the implicit inclusion of prosody generation in the low-level phonetic rules is indicative of the Producer attitudes which this system embodies. The text is analysed syntactically, but this analysis is often on a very superficial level: "The minimum for a phonetic component is an input with function words and content words marked. The prosodic rules have to govern how these two groups of words should be associated" to form domains

(Carlson et al. 1990:276). In contrast to Kager & Quené's system, INFOVOX makes little or no attempt to refine its syntactic parse on the basis of prosodic knowledge: INFOVOX workers seem to assume that such non-syntactic information is irrelevant to the prosodic rules, and therefore this very crude syntactic analysis is passed on to the prosody module where the task of generating an intonation contour appears to involve a direct mapping from syntactic markers to phonetic targets: "In the intonation module, accentuation rules assign fixed F_0 values which are subsequently modulated by a microprosodic parameter and adjusted to a declination line." (Barber et al. 1988:971) Despite the fact that INFOVOX systems use a phonetic model of intonation which resembles that of Gårding & Bruce (1981), accentuation is wholly determined on the basis of syntactic factors and so no independent level of intonational phrases or accents is constructed. This model has been adopted largely unchallenged by many other TTS systems: similar architectures and attitudes can be found in the work of Gretter et al. (1990) on Italian, Bäckström et al. (1989) on Swedish and Bailly (1986) on French, amongst others, and the INFOVOX approach is very much taken as the archetype for a great deal of European Producer research on synthesising intonation which relies on minimal syntactic information only.

In addition to the amount of linguistic knowledge which different Producer implementations apply in deriving prosody from text, there is another dimension along which these systems vary: that of the system's ambitions. Roughly speaking, the ambitiousness of an implementation can be said to be inversely proportionate to the amount of linguistic input which that system assumes: thus, a very ambitious system might assume only unannotated text as input whereas a less ambitious system might start from a full syntactic parse or even from a prosodically-transcribed text. There is also the question of how reasonable these assumptions are: the assumption of a full syntactic parse is a common one but is also extremely unreasonable (see page 64 below), whereas some of the least ambitious systems assume a complete semantic and discourse-level representation which is not unreasonable for their particular application. This latter group of systems attempt to perform **synthesis from concept**, a variant on the text-to-speech conversion task in that a dialogue system is assumed whose output is not text but some conceptual representation which these systems attempt to convert into natural-sounding

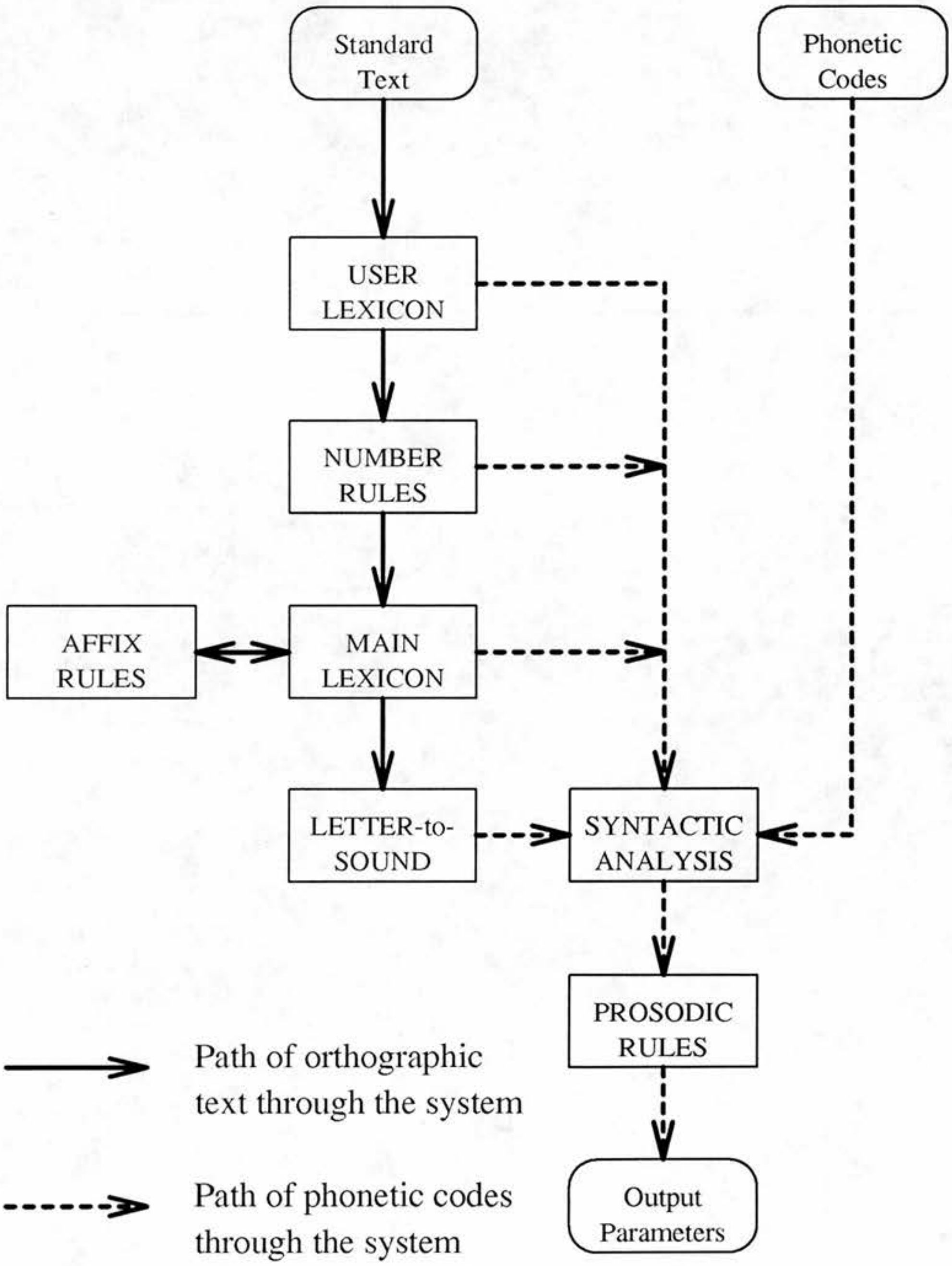

Figure 6: The Typical Structure of an INFOVOX Text-to-Speech System

speech. Early attempts at synthesis from concept such as Young & Fallside (1980) did not in fact see the problem as very different from TTS: they assumed that a dialogue system would output syntactic information similar to that presupposed by many TTS systems, except that a Chomskian "deep structure" would also be available. Even then, however, the assumption that surface syntax determined prosody was so strong that this "deep structure" was largely ignored:

> The deep structure only gives information as to sentence type (e.g. statement, question, etc.) and is not consulted for structural data.
> Young & Fallside (1980:242)

Youd & Fallside (1987, 1989) amend this approach to take more advantage of semantic information, but still rely heavily on surface syntax for accent and boundary assignment in their prosodic rules: the system still "takes a conceptual representation of the message, and generates a syntactically labelled surface structure" (Youd & Fallside 1989:514) which is only supplemented by focus information.

Although the work of Fallside and colleagues is criticised in more recent work on concept-based synthesis, many of their assumptions remain unchallenged. Sakai & Muraki (1990:329), in their presentation of a synthesis-from-concept system for Japanese, point out the inadequacy of syntactic analyses as conceptual representations, and indeed present a much more useful representation based on semantic relations such as "possession" and primitives such as objects and actions; but in actual fact their system appears to use little other than syntactic and morphological information. There is some mention (p.330) of marking the pragmatic function of verb phrases, but the examples which are discussed in detail (syntactic topicalisation, focus-marking morphemes, classes of conjunction) are all based on text rather than concept and indeed it appears that Sakai & Muraki (1990:331) generate text as an intermediate representation despite their arguments that text output is both unnecessary and unhelpful in dialogue systems.

Yamashita et al. (1990:241), again working on Japanese, similarly stress the advantages of producing speech output directly from a conceptual representation rather than from the prosodically more opaque medium of text. Their conceptual representation is rather less elegant than that in Sakai & Muraki (1990), involving the interpretation of certain modifiers and of verb cases rather than semantic relations. All the conceptual information refers to a set of rather arbitrary and application-specific "templates", but their encoding of discourse information such as intersentential relations is quite advanced:

> The concept descriptions are processed sentence by sentence. The conjunction templates describe the relation to the preceding sentence. This template requires a sentence as its argument, and generates the conjunction before the sentence.
> Yamashita et al. (1990:243)

Unfortunately, the legacy of Young & Fallside (1980) is still apparent in the control of prosody in Yamashita et al. (1990). In a Fujisaki-style model of Japanese prosody, which builds an intonation contour by superimposing parabolic contours at successive levels of utterance organisation, declination resets and fixed-length pauses are assigned directly in the conceptual representation (although some are optional depending on constituent length measured in morae) and absolute pitch values are imposed by explicit mathematical operators which are also part of the conceptual representation and which directly modify the accent component of the F_0 contour. This is much the same as Young & Fallside's (1980:243ff) direct manipulation of pitch and timing contours, so that despite the additional information available to Yamashita et al. (1990) there is little difference in the character of the eventual output.

Another variety of "unambitious" implementation, assuming a great deal of linguistic information in addition to standard text, can be found amongst Producers subscribing to the traditional British Designer account of prosody. The work of Witten (1977), on implementing prosody generation for a phoneme-to-speech system, is based largely on Halliday's (1967b) account of intonation and requires tones and boundaries to be marked in the input phoneme string (Witten 1977:242). Some of the problems with Witten's approach have been mentioned above, and a lack of ambition can justifiably be added to them. Witten's declared aim is to produce a flexible system, and that he does, but this "flexibility" seems to involve taking as clear and thorough an account of the units of English intonation as was available, casting in concrete such factors as peak alignment and slope (which are linguistically variable) (p.254), and allowing the user a degree of control over local pitch range, starting pitch and interpolation (p.255) which is simply not exercised by human speakers. Despite this unnecessary flexibility, Witten's system is unable to replicate Halliday's (1967b:16) tones 4 (rise-fall-rise) and 5 (fall-rise-fall).

More recently, IBM(UK)'s work on synthetic prosody tackled a similarly unambitious task. This work was based on the O'Connor & Arnold (1961, 1973) transcription system used by Young & Fallside (1980). The resultant implementation (Williams & Alderson 1986; Bell 1987; Alderson et al. 1988) requires prosodically-transcribed text as input, thus limiting itself to the task of providing a phonetic interpretation of the various O'Connor & Arnold diacritics and interpolating F_0 between them: the transcription indicates major and minor prosodic boundaries; nuclear and non-nuclear falls, rises, level and complex tones; minor prominences and changes in pitch range, leaving very little to be filled in. The aim was to derive such a transcription automatically from text by analysing a large (100,000 words) transcribed corpus and deducing statistical rules, but in sharp contrast to the transcription-to-speech problem this text-to-transcription task was clearly too ambitious. In any case, the work has now been abandoned.

A more common, and certainly more ambitious, task is to assign prosody from some form of syntactic analysis. There are two schools of thought amongst Producers regarding the use of syntax in assigning prosody for TTS: on the one hand, there are those who assume that some standard theoretical syntactic parse is sufficient to determine prosody; on the other hand, there are many who see syntactic analyses as a useful and accessible level of representation but who would like to supplement them with other analyses or representations. The former view is modelled on the "syntax = prosody" argument discussed above, and on the MITalk system which assumes this correspondence and explicitly states that other information is superfluous (Allen et al. 1987:40). Many systems have followed MITalk's lead, and it is common for Producer implementations either to perform a syntactic analysis in order to drive prosodic rules or to assume such an analysis and develop prosodic rules which make use of it. Shi's (1989) system for Mandarin Chinese, for example, derives syntactic information and passes it straight to the prosodic rules; Schnabel & Roth's (1990) German TTS system incorporates a syntactic parser which has been explicitly developed for "the insertion of syntactico-prosodic markers" (p.121); Cericola et al. (1989:388), in their TTS implementation for Italian, associate prosodic rules directly with grammatical rules, thus guaranteeing total congruence between their syntactic and prosodic analyses; Gretter et al. (1990:334) take a very similar approach, again for Italian; Tatham (1990:239) declares his intent to drive a Pierrehumbert-style intonation model from syntactic markers in a TTS system for British English; and there are many other systems for other languages which make the same assumption. Deriving prosody from syntax, even if one takes the view that there is a good correspondence between the two, is a reasonably ambitious task, and much interesting and inventive work has been done in these systems: however, the assumption that it is possible to provide prosodic rules with a reliable syntactic analysis of unannotated text seems to be highly dubious if not completely unreasonable. There have been many doubts expressed as to the possibility of parsing text deterministically (Grishman 1987:84ff; Monaghan 1990b; Matsumoto & Yamaguchi 1990:270; Moulines et al. 1990:312), and it is widely acknowledged that picking the "correct" parse from the output of a non-deterministic parser requires semantic, pragmatic and other information which is not generally available from text. This situation has produced the alternative approach where a full syntactic parse is neither assumed nor attempted and syntactic information is used in conjunction with other types of information to determine the prosodic realisation of text.

This alternative view has been taken in recent work by Kulas et al. (1986), Moulines et al. (1990), Russi (1990), Sorin et al. (1987), Fitzpatrick & Bachenko (1989), and Quazza et al. (1989) amongst others. Some of these authors have been principally concerned with producing a syntactic analysis of text, while others have attempted to use such an analysis to derive sentence prosody. Among the former, many have come to the conclusion that in a working TTS system a full syntactic analysis is impractical "wegen Rechenzeit und Speicherplatzbeschränkungen" (Kulas et al. 1986:199) and have therefore attempted to determine what syntactic information is actually required by TTS systems and how this can best be generated. Unfortunately, the results to date have not been very promising: both Fitzpatrick & Bachenko (1989) and Quazza et al. (1989) found that the amount of syntactic information required to assign appropriate prosody was considerable even for very restricted corpora (40 and 200 sentences respectively). However, Fitzpatrick & Bachenko (1989:193) suggest that "prosodic phrasing requires information from the syntax tree that sits rather low down on the tree." This suggestion also underlies the work of Willemse & Boves (1989, 1991) on syntactic processors for speech systems, in that their "Wild Card" parsing strategy does not attempt to assign structure above the phrase but concentrates instead on building smaller constituents, leaving their combination to other processes. (See Monaghan (1990b) for further discussion of the place of syntax in TTS systems.) Quazza et al. (1989) and Russi (1990), whilst acknowledging that "semantic and even discourse structure of a text must be taken into consideration" (Russi 1990:117) in order to produce high-quality prosodic output, maintain that as much syntactic information as possible should be extracted from text and to this end propose sophisticated parsing strategies, based in the former case on statistical tendencies in text corpora and in the latter on unification-based deterministic rules.

A good example of the use of non-syntactic information to determine prosody is provided by the TTS system for French developed at CNET (Sorin et al. 1987, Moulines et al. 1990). Here, the researchers are not primarily concerned with syntax and are prepared, in good Producer style, to make use of any information which will improve the quality of their system's output. Although their attitude to syntactic analysis is at times ambivalent (see below), workers at CNET have combined several types of information in their prosodic rules. The CNET system takes advantage of listeners' tolerance to insert more pauses than occur in natural speech, on the grounds that these will aid rather than hinder comprehension[8]: it therefore assigns a pause at most CW-FW boundaries, depending on the length of the domain thus formed and on the type of FW involved. The contour assigned to any domain is chosen from a set of 6 stylised contours extracted from a small corpus (Sorin et al. 1987:126). This approach thus combines the CW/FW distinction, domain length criteria and assumptions about the listener: however, Sorin et al. (1987:128) acknowledge "the limits of a 'syntax-independent' prosodic parser". Moulines et al. (1990) start from this position and attempt to add some syntactic analysis to the CNET system to reduce the errors highlighted in Sorin et al. (1987):

[8] However, Scharpff & van Heuven (1988) found that the location of pauses in synthetic speech is much more critical than CNET suppose, to the extent that infelicitous pause insertion can seriously hamper intelligibility.

> Even if one considers that the congruence between syntax and prosody is not complete, the lack of sufficient syntactical information leads in many cases to numerous prosodic segmentation errors which are unacceptable for every listener.
> Moulines et al. (1990:312)

Although workers at CNET see at least a partial syntactic analysis as essential, claiming that the identification of verbs and phrase boundaries is required for assigning intonation (Moulines et al. 1990:312), they also point out the problems of insoluble (and usually unimportant) syntactic ambiguities. They therefore reject standard syntactic parses in favour of rules specifically aimed at identifying prosodic boundaries:

> This parsing is based on the assumption that a prosodic boundary [in French] can be derived, in most cases, from the grammar category of the word and its 2 or 3 left and right neighbours. Globally very reliable (over 95% of the boundaries are correctly detected), this parsing module may fail in the (rare) cases of wrong grammar category assignment to a word (typically confusions between adjective or past participle) or when the input text is incorrectly spelled or punctuated.
> Moulines et al. (1990:313)

Despite the considerable success of this approach, there are still obvious shortcomings in the prosodic analysis: for instance, there is little indication of the type of boundary (and thus the type of prosodic treatment required) in many cases, and there is no indication of the overall prosodic structure of the sentence or text. The cases where the syntax-prosody correspondence is either ambiguous or non-existent, therefore, remain a serious problem which cannot be solved by current processing techniques:

> there is still room for progress in the specification of complete prosodic grammars able to provide natural, expressive reading of complete texts: to take into account the semantic and pragmatic aspects, the key problem remains the availability of efficient natural-language analysers.
> Moulines et al. (1990:317)

A further difference between Sorin et al. (1987) and Moulines et al. (1990) is the latter's use of a database of stored F_0 contours. CNET's successful work on concatenation synthesis has been extended to concatenating pieces of natural F_0 traces in order to produce high-quality synthetic F_0 contours. The basic principle is very similar to that underlying diphone concatenation (Campbell et al. 1990): if the steady states of intonation contours can be identified, and enough transitions between those states can be stored, then almost all the desired F_0 contours can be produced by locating the steady state positions in texts and then concatenating these transitions using the steady states as anchor points and transforming the transitions using waveform-manipulation techniques (Charpentier & Moulines 1989). The problem is obviously to identify the reliable steady states of F_0: CNET's system assumes (Moulines et al. 1990:313; Larreur et al. 1989:512) that word boundaries correspond to such steady states, and therefore assigns an F_0 pattern to each word depending on its length and prosodic characteristics:

> A melodic table provides frequency patterns adapted to all word lengths and all prosodic markers. ... Moreover, for each marker and for each word of a given length, several melodic contours are available; therefore, some melodic diversity may be introduced in synthesising longer texts.
> Larreur et al. (1989:512)

Given the amount of linguistic and phonetic data showing the independence of prosodic phenomena from constituents such as words (Liberman & Prince 1977; Goldsmith 1982; van der Hulst & Smith 1982; Ladd 1983b), it seems unlikely that the word is an appropriate unit to choose in concatenating F_0 contours. However, whether or not word boundaries constitute intonational steady states, they are not the only option: several other Producers, both at CNET and elsewhere, have attempted to apply concatenative techniques to F_0 synthesis using units other than the word. Traber (1990) presents a system for Swiss German which concatenates F_0 contours at the syllable level, and Emerard & Benoit (1988) suggest experiments with Markovian modelling to determine appropriate F_0 constituents (although this is currently limited to isolated monosyllabic words).

Although these Producer systems avoid the data-reduction error of Designer approaches to natural F_0 such as that of Aubergé (1990), and of course benefit potentially from the implicit naturalness of pre-recorded F_0 contours, it is not clear that the concatenation of partial F_0 traces will actually work. Concatenation may seem attractive for monosyllables or even short sentences with very restricted prosodic realisations, but for less restricted applications it may simply not be practical to record and store the number of natural F_0 contours required. The more that is known about the factors which determine intonation in natural speech, the greater the number of contours which will have to be stored in order to control these factors and thus produce appropriate prosody for new utterances. Moreover, given the amount of information required to choose the appropriate units to concatenate, there may be more sensible methods of generating a synthetic F_0 contour by supplying the same information to a phonological and/or phonetic model such as our own. The prerequisites for a concatenative system are (a) an inventory of all the necessary contour segments, multiplied by some factor to allow sufficient variation in the actual segment chosen; (b) a sufficiently detailed transcription to allow the appropriate units to be selected; and (c) a set of editing and smoothing techniques to ensure that the contour is properly aligned and does not contain perceptible discontinuities. If we assume that (c) is not a serious problem, the question is whether (a) can reasonably be produced and whether (b) could not drive a more efficient F_0-generation scheme. The first part of the question seems to require a certain amount of empirical research before it can be answered, but in view of the number of well-developed models of intonation currently available for application to synthesis the answer to the second part seems clear: if (b) could be produced, there would be more efficient and flexible ways of generating high-quality F_0 than concatenation. This is not surprising, since the basic argument for concatenation synthesis depends on a position of ignorance: concatenating pre-stored units allows us to take advantage of their inherent naturalness without understanding what the defining characteristics of that naturalness actually are, but the amount of knowledge which we now have about the characteristics of intonation contours renders this concatenative approach unnecessary.

The other technique common in Producer TTS systems which explicitly acknowledges a lack of understanding of the processes involved is the use of neural networks. These are applied to various stages of TTS conversion, including letter-to-sound transcription (Lucas & Damper 1990; Xiang & Bi 1990), syntactic analysis (Matsumoto & Yamaguchi 1990), and of course prosody. Although the CSTR TTS system employs neural nets to derive duration rules from transcribed corpora (Campbell 1987, 1989), they seem to us to be quite inappropriate for generating intonation contours: however, this view is not universal. Traber (1990) compares the performance of his F_0 concatenation system mentioned above with the output of a neural network trained on the concatenation data, and concludes that the latter is both more successful and more appropriate:

> So far, the resulting contours produced with the neural network are better than the ones produced with the patterns data base. Using a neural network for the generation of complete F_0 contours with high quality is feasible and may require much less human effort than other approaches.
> Traber (1990:141).

Many of the same arguments apply to the use of neural networks as to the concatenative approach to F_0, except that neural networks do not require the storage of large amounts of natural F_0 data. It remains true, however, that if the level of transcription required to train neural networks adequately can be produced it should be possible to take a more knowledge-based approach to synthesising intonation.

A final group of Producer F_0 generation schemes includes the work of Olaszy (Olaszy & Gordos 1987; Olaszy 1989; Olaszy et al. 1990) and Yiourgalis (Yiourgalis 1990; Yiourgalis & Kokkinakis 1990). Both these authors subscribe to a view of intonation generation in TTS as involving low-level phonetic control of F_0 in terms of absolute frequency and timing. The synthetic intonation described in Yiourgalis & Kokkinakis (1990) is crude in the extreme, but far from there being any acknowledgement of its crudeness, it is presented as capturing the important aspects of Greek prosody. The basis of their system for Greek, as described in Yiourgalis & Kokkinakis (1990:412ff.), is the adaption of a straight-line slope from 130Hz to 90Hz to the length of the utterance to be synthesised, with the addition of a slightly more detailed contour segment to cover the last 500ms of the utterance: the shape of this final segment depends solely upon the punctuation mark (one of apparently only four possibilities in Greek) which ends the corresponding text sentence, and is composed of from two to four straight-line sections of fixed slope and duration with no apparent regularities between the four possible patterns.

Olaszy and colleagues place a similar value on punctuation marks, but also impose microintonation (Olaszy & Gordos 1987:27) and local "unstressing" (their term for reduction of prominence):

> Unstressing is at least as important in speech as stressing. Unstressing rules decrease the fundamental frequency value by 3-6Hz inside the word. These

> rules are used for articles, prefixes, conjunctions, monosyllabic words in some languages, etc.
> Olaszy (1989:528)

This phonetic view of F_0 amongst Producers is not entirely surprising, since the O'Shaughnessy algorithm implemented in MITalk (Allen et al. 1987:100ff) took a similar approach. MITalk's 2-stage F_0 generation essentially involves calculating a declination line and perturbing that line on the basis of prominence information and segmental factors:

> The High Level System predicts a superposed F_0 contour by taking into consideration the sentence type, clause contour, phrase contour, and individual word contour. This contour is further amended in the Low Level System by considering the effects of individual segments.
> Allen et al. (1987:103)

This view of intonation as a phonetic phenomenon, and as composed of syntactically-determined contours, persists among Producers despite the fact that the various theories of intonational phonology discussed above have evolved since MITalk was conceived. Fortunately, as should be clear from the foregoing, there is a growing number of Producer implementations which take advantage of linguistic and cognitive theories of intonation and replace the phonetic and syntactic approach with equally practical and much more plausible models of prosody.

1.5 Tackling the Problem

The plethora of theories and techniques which have been applied to the problem of generating intonation from text may give the impression that every solution has been tried and that there is nothing new to be contributed or learnt in this area. This impression is totally erroneous: there is indeed a large and growing number of TTS systems which incorporate intonation rules, and most of these systems differ in their theoretical and methodological approaches, but two important facts have ensured that new avenues remain to be explored in the area of intonation synthesis. Firstly, as we have discussed above, there has been little exchange of ideas or results between adherents of different approaches to synthesising intonation: it is thus only recently that these different approaches have been combined in TTS work such as that of Kager & Quené and of researchers at CNET. Secondly, the field of intonational theory continues to expand and to generate original and computationally-tractable accounts of prosodic phenomena: there is therefore no lack of fresh theoretical material to be applied to synthesis systems.

It is the aim of the present work to apply the most promising of these theoretical and practical approaches to the problem of synthesising intonation from unannotated text, with the explicit assumptions that a coherent model is as desirable as a working system and that the development of a working system will provide an appropriate testing-ground for such a model. Before we can construct either system or model, however, we require a well-defined task for the system to perform and a coherent class of phenomena to be handled by the model.

1.5.1 Constraining the Problem

The aim of handling unrestricted, unannotated text is a very ambitious one for an automatic system, particularly since it is not clear that humans can perform this task consistently: as was pointed out above, there is a great deal of variation in human readings of unfamiliar text, and some of this variation leads to a loss of quality. It is clearly unreasonable to expect a machine to perform better than a human in this respect, and we are therefore reconciled to a less-than-perfect performance. However, there are further constraints which must be put on the task of generating intonation from text before it becomes reasonable for an automatic system to attempt. Since we have established that it is not currently possible for any automatic system to perform the syntactic, semantic and pragmatic analyses which are essential to humans' production of natural intonation from text, we must compensate for this lack of vital information by modifying our expectations of the system. We have done this by constraining the problem in three ways, all of which, in our opinion, are reasonable approximations of strategies adopted by human readers when faced with a similar lack of information. We have defined the phenomena which our system should handle as those of **acceptable, neutral, naïve intonation**: the most we expect from our system for any text is an intonation which conforms to this definition. The meaning of these restrictions is explained below.

Since there are many intonational variants which may be realised by different speakers producing the same utterance in a given context, it is not generally possible to identify a single "correct" intonation (Choppy 1979:186; Baart 1987:56): we therefore aim to produce one of the many **acceptable** variants rather than attempting to assess what the most "correct" intonation might be. **Acceptable** intonation must be plausible in context, but need not be the **most appropriate** intonation for a particular utterance. For instance, in the sentence:

(5) The Prime Minister was escorting Mrs. Churchill.

the **most appropriate** accentuation might place the nucleus on *Mrs.* (if the prime minister in question were Winston Churchill) but placing the nucleus on *Churchill* would still constitute an **acceptable** intonation. Any other nucleus placement, however, would be unacceptable unless a much wider context were provided: there is thus a strict limit on the number of **acceptable** nucleus placements in most cases.

Neutral intonation is similar in principle to **normal stress** or **broad focus**, in that it makes no special assumptions about the contextual or lexical meaning of the utterance (the basis of Normal Stress) and deliberately leaves the focus structure as ambiguous as possible (the definition of **broad focus**). The shortcomings of Normal Stress are counteracted by the acknowledgement that **neutral** intonation depends on factors other than syntax, and that when such factors are known the appropriate action will be taken to ensure an acceptable intonation. **Neutral** intonation is ideally congruent with Broad Focus, but given the difficulty of establishing focus structure from text there will inevitably be mismatches: however, as far as possible these mismatches should still result in an **acceptable** intonation. **Neutral** intonation is thus more flexible than Normal Stress, in that the neutral realisation of a sentence can vary with changes in context, but it is not as reliable as an intonation based on Broad Focus since it is assigned on the basis of less information. Moreover, **neutral** intonation also specifically excludes phenomena such as unusual emphasis, contrastive stress or stylistic effects: such phenomena are excluded from the model but, all other things being equal, it will produce an acceptable accentuation.

The specification of **naïve** intonation is perhaps more of a justification than a constraint. The basic notion that intonation can be assigned in the absence of understanding is clearly only valid if that intonation is not expected to demonstrate full understanding. As we discussed above, much human reading of text does not involve full understanding: however, there are cases where there is a clear difference between a **naïve** reading and an informed one. The choice of nucleus placement in example (5), for instance, could depend crucially on one's understanding of the referring expression *The Prime Minister* and the consequent interpretation of *Churchill* as coreferential or not: a **naïve** reading would assume no coreference in such cases. In cases where coreference is inevitable, however, the intonation should reflect this fact.[9] **Naïve** intonation, then, assumes no special or privileged knowledge which would not be available to every reader: the system is not the author of the text, and therefore no assumptions of understanding or generation by the machine should be made.

In sum, although the rules described in this thesis are intended to address the problem of generating intonation from unrestricted text, they are not designed to model spontaneous human monologue or dialogue. The expectations of the system are thereby

[9] See the discussion of anaphora in Chapter 3.

reduced, so that our system aims to produce **ACCEPTABLE NEUTRAL** intonation only. The ideal acceptable neutral intonation would approximate to that of a good newsreader who has not authored the text and who makes no assumptions of specialised knowledge on the part of the hearer: however, achievement of this standard is probably an unrealistic goal for an automatic system. A more plausible target might be to approximate an inexperienced newsreader on a regional network, rather than the top readers from national broadcasting.[10]

1.5.2 Defining the Problem

The problem which this thesis addresses can be stated as follows:

> To define a set of computationally-explicit rules which will allow a TTS system to derive a phonological specification of an acceptable neutral intonation contour: these rules should handle unrestricted text, without the need for any annotation or other human intervention beyond the typing of the text.

There are a few further requirements which are essential in producing an interesting and practical system. The most obvious of these is that the rules should be applicable in something approaching real time, i.e. they must be both efficient and deterministic. In addition, the system should not be cast in concrete: it should be flexible and extendable to allow for different applications and the incorporation of additional information. Not every application requires the same degree of formality of intonation, for instance, or the same physical characteristics in the synthetic voice. Moreover, it is foreseeable that the linguistic information which TTS systems currently lack will one day be available to automatic systems and it would therefore be unnecessarily short-sighted not to allow for its eventual incorporation. Finally, wherever possible the model should be compatible (or at least comparable) with intonational theory. This allows it to be applied in linguistic research such as the experimental work at Kiel based on synthetic speech stimuli (e.g.

[10] We would, however, hope to improve upon the intonation produced by most sports announcers.

Gartenberg & Hertrich 1988, 1989; Hertrich & Gartenberg 1988, 1989). Such a requirement also has the advantage that any implementation demonstrates the computational practicality (or otherwise) of the underlying intonational theory as a process model.

The combination of these requirements results in a demanding specification, despite the constraints on both system and model. We require a set of rules which will handle anything in a reasonably intelligent, if naïve, manner and which will produce phonetically interpretable output quickly and efficiently: yet these rules must also be maximally flexible and easily-extendable, which prohibits fine tuning and optimisation of any implementation. In order to achieve these goals, we need to apply sophisticated techniques and make use of development strategies which are unusual in TTS research.

1.5.3 Solving the Problem

We have adopted a target-and-transition model similar to Pierrehumbert's (1980) or Gårding & Bruce's (1981), which allows us to concentrate on a small number of points in the intonation contour rather than attempting to specify pitch at every millisecond. We have also assumed that the points at which targets require to be specified are exclusively pitch accents and prosodic boundaries, and the phonetic model described above allows us to specify these events in terms of a very small number of parameters. This allows us to perform all rule applications and computations at a purely symbolic level, manipulating atomic symbols rather than numerical values. Concentration on the abstract phonological specification of intonation contours makes it possible to avoid speaker-specific representations, and the adoption of intonational tunes allows us to write rules which manipulate accents and boundaries rather than targets or tones.

A major increase in speed of processing in our system results from the fact that the intonation rules handle only intonational events, with no reference to the segmental tier or to details of timing and alignment. The efficiency of symbolic processing in a language such as PROLOG, and its appropriateness for Natural Language Processing tasks, allows us to dispense with the complex time-related equations of most Producer implementations, thus increasing the efficiency of our system and retaining the elegance of a model based on a minimal set of entities (accents and boundaries). In addition, the processing of lists of symbols allows us to apply rules in a linear, left-to-right fashion without the problems of "lookahead" and lengthy search strategies which encumber systems such as those at IPO and Kiel which are obliged to compute pitch on the basis of overall utterance duration.

In addition to these computational techniques, there are two basic strategic principles which underlie our system design. The first principle is the common AI strategy of **default specification**, which allows us to underspecify intonation contours to as great an extent as possible. The second principle is a purely practical one: given the paucity of the prosodic information available from plain text, it seems obvious to us that any system aiming to derive high-quality prosody from such text must make maximal use of all available information from whatever source.

1.5.3.1 Default Specification

Given as large and complex a domain as the specification of text intonation, a strategy of broad generalisation from minimal information is indispensable for any automatic system. Moreover, such a strategy lends itself well to subsequent refinement and modification as more information or better generalisations become available. The use of default specification is a common method of expressing generalisations, particularly in AI programming. In our system it has been extensively applied to those areas where inadequate input is the rule and good heuristics the only solution, particularly in simulating the effects of information (such as semantics and pragmatics) which is crucial in determining the abstract phonological representation of any input text but which is generally not deducible from the input. A system of defaults, which can be refined and extended indefinitely, forms the basis of our rule-development strategy and gives our system maximal flexibility and generality: any exception to the rules can be catered for, assuming it can be specified precisely enough, and the system will always have the default specification on which to fall back.

Our system is also a tool for its own improvement, since rules can be revised on the basis of their current performance: wherever a regular error can be identified, a new exception clause can be added to the rules without changing their behaviour in other circumstances. The modularity of design and the ordering of the rules to give increasing generality of application are vital to this self-evaluation function.

1.5.3.2 Maximal Use of Available Information

There is very little linguistically "higher-level" information which can be reliably deduced from text by algorithmic analyses such as full parsers and semantic models. Extracting syntactic information from text is possible, as long as the text does not overstep the bounds of complexity or reasonableness which are imposed by grammar-writers (Berthelin et al. 1989; Willemse & Boves 1989, 1991): unfortunately, unrestricted text does overstep these bounds with monotonous regularity, which is hardly surprising since phenomena as common as chapter headings, punctuation marks and conjunctions generally fall well outside these definitions of reasonableness. Semantic analysis of text is also possible, but state-of-the-art systems such as SPICOS II (Niedermair et al. 1990) are limited to vocabularies of around a thousand items and to very restricted knowledge domains. Analysing running text at levels above that of semantics, such as pragmatic or discourse analyses, is currently impossible in any principled manner: this is due partly to the lack of any complete theory of linguistic factors at this level, and partly to the massive ambiguity of text compared with speech in this respect, since speech provides prosodic and other clues to emphasis and attitude which must be much more painfully teased out of plain text. In general, then, TTS systems cannot rely on any of the higher-level information which they so desperately need to produce appropriate prosodic output.

The sort of knowledge to which such systems must therefore resort is probabilistic, heuristic knowledge rather than hard-and-fast formal semantic or syntactic rules. Such knowledge can come from a variety of sources: there are statistical correspondences between prosody and lexical items such as the tendency of *even* to mark contrastiveness (Ladd 1983a); there are the syntax-prosody correspondences such as those noted by Halliday (1967ab); there is knowledge of the frequency of occurrence of words, collocations and entire phrases or syntactic structures; and there is knowledge of the semantic weight or import of particular items, to which Bolinger (1986) attributes much of the control of prosody. There is in fact no reason why any available knowledge, linguistic or otherwise, should not be brought into play: knowledge of spelling conventions, of other languages, of the writer's temperament, and of the historical and cultural setting of the text, for instance, could all conceivably contribute to the generation of appropriate prosodic characteristics. The point is that, in the absence of full understanding, assistance from any quarter should be gratefully accepted: statistical, intuitive, and unprincipled heuristics are the best cues available, and therefore they should be exploited. All these knowledge sources can be combined to produce an informed guess at the appropriate prosodic realisation of a text.

Based on this strategy, our rules attempt to make maximum use of all the information available from text: we have developed a set of heuristics which allows us to mimic the effects of semantic and focus structure on intonation. The system operates by using linguistic knowledge to specify the default case and then deducing exceptional cases from heuristics based on syntactic and lexical information. Despite the lack of reliable information regarding focus, semantic or even grammatical structure, an approximation to the hierarchical metrical structure which (according to most current linguistic theories) determines intonation can be constructed on the basis of a very basic syntactic analysis. This heuristic approach allows our system to produce an approximate prosodic structure from minimal linguistic information, which can be enhanced and corrected where more reliable prosodic cues are available.

Chapter 2

Accents and Boundaries

> accent assignment and phrase determination are the primary areas requiring improvement in order to further increase the naturalness of synthetic speech intonation
> Akers & Lennig (1985:2157)

Our work follows that of Ladd (1980, 1983ab), Pierrehumbert (1980, 1981), Bruce (1982) and others in assuming that at the phonological level an intonation contour may be specified exclusively in terms of two types of phonological event: pitch accents and prosodic boundaries. The task of our rules is thus to specify these events on the basis of textual cues, i.e. to bridge the gap between unrestricted input text and the phonological and phonetic models of intonation outlined above. The first step in this task involves assigning a phonological structure to the utterance, in terms of prosodic boundaries and the phonological domains which they demarcate. The second step is to assign accents to elements within these domains, based on their linguistic status and their rôle in a particular domain. There are, of course, interactions between the demarcation of domains and the assignment of accents, and we have developed various constraints or **well-formedness conditions** (WFCs) on these interactions to avoid conflicting specifications.

The first part of this chapter presents our strategy for splitting text sentences into phonological domains, and discusses in some detail the development of the various heuristics which accomplish this task. The second part illustrates the need for rhythm rules in intonation synthesis systems, and outlines the operation of our Rhythm Rule.

The third part discusses the implications of such a rule for lexical stress and the applicability of its output to controlling the synthesis of other segmental and suprasegmental phenomena such as vowel quality, duration and speech rate.[1]

[1] Parts of Sections 2.4, 2.7 and 2.8 below describe work which was presented for examination as an Honours dissertation (Monaghan 1987a). All these sections have been completely rewritten for the present thesis, and are included for the sake of clarity and completeness.

2.1 Phonological Domains

Several proposals have been made regarding the manner in which prosodic domains relate to other linguistic structure: Selkirk (1984:297), in a strongly syntax-based account of intonation, sees "the role of syntactic structure with respect to prominence patterns as one of demarcating the domains (continuous spans of the utterance) within which relations of relative prominence are defined"; Gussenhoven's SAAR takes semantic constituency to be the determining factor, as discussed above; Ladd (1980, 1986, 1988a) and Pierrehumbert (1980, Pierrehumbert & Beckman 1988, Hirschberg & Pierrehumbert 1986) combine syntactic and semantic constituency, and add contextual or pragmatic effects; and Gee & Grosjean (1983) claim a vital rôle for performance constraints such as phonological size and composition of domains. However, all these proposals agree that at the very least a full syntactic parse and grammatical analysis is required before prosodic domains can be assigned. Unfortunately, for reasons which are discussed at length in Monaghan (1990b, 1991b) but which are essentially due to the syntactic complexity and ambiguity of text, no parser exists which can consistently provide reliable syntactic analyses for TTS conversion other than for very limited domains. The type of detailed grammatical analysis presupposed by theoretical accounts of prosodic domain assignment is far beyond the capabilities of current syntactic analyses of unrestricted text, and we are therefore unable to implement any of these proposals in a working system.

The syntactic analysis performed in our TTS system is, like most current TTS parsers, limited to identifying major phrases such as verb phrases (VPs) and noun phrases (NPs), postulating clause boundaries, and disambiguating the word-class hypotheses generated by the morphological analysis: it produces a single very crude parse tree with a single word-class for each terminal node. Despite these limitations, in the absence of any other high-level (grammatical, semantic or pragmatic) linguistic information the syntactic analysis is the best indicator of prosodic structure which our system provides. The constraints of a real-time TTS system make the generation of a more detailed syntactic analysis impossible: these same constraints prohibit the construction of exhaustive metrical trees[2] or any complicated time-based lookahead functions, as these are computationally very expensive. Even the very limited amount of semantic analysis of which automatic systems are currently capable is ruled out on these same grounds. The obvious remaining option is to base the assignment of prosodic domains on our current crude syntactic analysis but in addition to use simple, syntax-based heuristics and semantic and phonological intuitions and generalisations to mimic the effect of the more complicated analyses which we presently lack. We have therefore developed a set of rules which take this crude syntactic analysis as input, supplement it with linguistic knowledge-based heuristics, and derive a structure in terms of prosodic domains. These domains then form the basis of our accent rules, which are discussed in Section 2.2.

Designing an interface between syntax and intonation is a complex task, for two main reasons. Firstly, syntactic and intonational rules do not always make reference to the same domains and constituents: a syntactic analysis would not normally identify list constructions, for instance, although these are very important to the intonational structure; and conversely, no intonation rules which we are aware of currently make any reference to the constituent "Adjective Phrase" in assigning domains or accents, although this is a common constituent in syntactic analyses. Secondly, as stated above, we know that most of the information which governs intonation is not available from even a very full and accurate syntactic analysis, but depends rather on semantics and pragmatics; the interface therefore needs to approximate such information on the basis of default specifications and a number of heuristic processes.

In a system designed to handle running text, it is essential to split the input into manageable chunks. The obvious first step in this direction is to process text one sentence at a time, and this we do: it is, however, still necessary to identify smaller units so that intra-sentence hierarchic relations such as those between clauses and phrases can be realised intonationally. These relations encode much of the communicative content

[2] Systems such as PROS2 (Dirksen & Quené 1991), which attempt to assign metrical structure to text, are not only expensive in terms of processing time but are also currently very unreliable.

of speech, and to obtain any degree of "naturalness" in synthetic F_0 they must be brought out in the intonation contour.

Since the information required to assign prosodic domains to text according to some principled and regular theory is simply not available to us, our syntax-intonation interface is forced to resort to unprincipled, irregular techniques in an attempt to compensate for this lack of information. This stage of our intonation generation scheme is thus a deliberate collection of generalisations, intuitions, and rules of thumb: to reflect this fact, and make its deliberateness explicit, we have given the name INTERFIX to the implementation of our efforts to span the chasm between syntax and intonational phonology. The two main functions of INTERFIX are: (a) to break the syntactic string into phonological domains within which the rules in the intonation module can apply regularly and independently of other domains; (b) to remove redundant information from the syntactic string and to add the semantic and pragmatic information necessary for the generation of appropriate intonation. This strategy places nearly all the burden of producing "intelligent" intonation on INTERFIX, so that its output should be sufficiently reliable and detailed for the accent modules which process this output to consist entirely of regular rules.

We shall here distinguish three stages in the evolution of the current INTERFIX program, from its first implementation to a version which produces a very satisfactory approximation of hierarchic prosodic structure. All three versions share the basic characteristics of INTERFIX, as does the most recent version described in Chapter 3. All versions, including the INTERFIX program currently implemented in our system, are insensitive to most syntactic information other than word classes and some major phrasal units. They therefore remove all other information in what is basically a structure-flattening process, leaving only the word-class information. Various other operations are performed before the information is discarded, but the basic principle is that everything other than domain boundaries and word-class information should be deleted before the syntactic analysis is passed on to the accent rules.

Table 1: Principles for Assigning Phonological Domains

1) Everything before the first NP constitutes a domain.
2) Everything from the first NP to the first VP constitutes a domain.
3) Everything from the first VP to the next **major phrase break** constitutes a domain.

2.1.1 INTERFIX 1.0

This was the prototype initial implementation to allow our accent rules to use the output of the CSTR TTS system's parser. The primary function of INTERFIX 1.0 was to split up the syntactic input into phonological domains. This was done according to the three principles set out in Table 1. The pre-NP domain is motivated by the fact that in English anything which proceeds the first NP is by definition marked for special treatment, since the unmarked clause structure in English has the subject NP in first position. The NP domain is the domain for which the accent assignment rules were originally designed (Monaghan 1987a), but is in any case motivated by the widely-acknowledged semantic subject-predicate boundary (Halliday 1967b; Burton-Roberts 1986ab; Ladd 1986; Gretter et al. 1990; Pasdeloup 1990b) which generally falls between the first NP and the rest of the clause. There appears to be no good reason for subdividing the VP domain on semantic or phonological grounds, so this gives up to three domains per major phonological phrase - a pre-NP domain, an NP domain and a VP domain. There was, however, no clear definition of **major phrase break** in INTERFIX 1.0: the default definition was the end of the input. This meant that any input text sentence was split into a maximum of three domains, the last of which could contain several full clauses.

The pre-NP domain was treated specially by the accent assignment rules of INTERFIX 1.0, as certain items appeared to require accents in these domains but not in others. In particular, adverbs and conjunctions such as *indeed* and *although*, which convey inter-sentential relations, produced adverse reactions from listeners if they occurred unaccented in a pre-NP domain but were quite acceptable unaccented in other domains. Examples such as those in (6) illustrate this alternation.

(6a) beCAUSE of this disaGREEment, HESeltine left the CABinet

(6b) HESeltine left the CABinet because of this disaGREEment

(6c) beFORE the match had even STARted, ALex broke his LEG

(6d) ALex broke his LEG before the match had even STARted

(6e) toDAY we are FASTing

(6f) we are FASTing today

INTERFIX 1.0 therefore marked pre-NP domains for special treatment, so that the accent-assignment rules could assign accents to such items in these domains only.

With the exception of the phenomenon exemplified in (6), all domains were treated identically by INTERFIX 1.0. The items in each domain were uniformly subject to the accent rules described below, and all domain boundaries were realised in exactly the same way. The standard treatment of domain boundaries was to insert a register downstep and not to insert any boundary tones: the only boundary tone assigned by INTERFIX 1.0 was a final Low boundary tone at the end of the input sentence. This treatment produced acceptable results for short, simple sentences, but more complex structures involving embedding or major prosodic breaks could not be handled appropriately by this version.

The fact that our rules assigned items before the first NP in a sentence to a separate domain caused problems in the numerous cases where there were no such items. Although these cases were by definition in the majority, being the unmarked cases, our rules were ordered to search for the marked case first: this is normal practice in implementing such rules, as the definition of the unmarked case often subsumes that of the marked case. Because of this ordering, however, and because in addition our rules could not take into account any items which they had already processed, every case of unmarked syntax (i.e. with an NP in sentence-initial position) led to the assignment of an empty pre-NP domain. As a result of our uniform treatment of domain boundaries, these empty domains produced spurious and undesirable sentence-initial downsteps. The simple remedy for this problem in INTERFIX 1.0 was to remove any such downsteps from the beginning of the prosodic representation, and this worked quite well as a temporary solution: however, the problem became considerably more serious in later versions of INTERFIX.

2.1.2 INTERFIX 1.1

The failings of Version 1.0 were numerous. Some are mentioned above, such as the limit of three domains per input sentence and the failure to assign boundary tones: others included the lack of reference to commas or other text punctuation, and the lack of any attempt to simulate hierarchic phonological structure. Version 1.1 was a first attempt to remedy some of these failings, based on the loose correspondence between clause boundaries in the syntactic analysis and boundaries between phonological domains.

INTERFIX 1.1 still assigned phonological domains according to the principles in Table 1 above. The one crucial difference was that a **major phrase break** was defined as a full syntactic clause boundary, so that at each such boundary the process of domain assignment recommenced. The use of this new definition of major phrase breaks had two important consequences: first, the maximum number of domains assigned per sentence was increased from 3 (in Version 1.0) to infinity by introducing recursion at clause boundaries; second, the resultant domains were generally smaller and also much more uniform in length, since there was no longer the problem of assigning several clauses to the third and final domain.

Both these effects appeared to take us in the right direction, i.e. towards a more naturalistic phonological structure with domains which were more consistent both in size and in constituency. However, they were not without their disadvantages. Many more empty pre-NP domains were assigned by this method, and default downstepping at domain boundaries resulted in extremely steep downtrends in many cases.

The assignment of empty domains was a problem even in INTERFIX 1.0, but as it was restricted to occurring sentence-initially in that version any downsteps which were inserted unnecessarily could easily be deleted. In Version 1.1, however, superfluous downsteps were assigned at most sentence-internal clause boundaries and their removal would therefore have entailed considerable extra processing of the entire input sentence.

The problem of very steep downtrends, resulting from our default strategy of downstepping at domain boundaries in Version 1.1, was further complicated by these superfluous downsteps. Although this problem was not unexpected, in the absence of a clear alternative the existing default treatment was allowed to stand. The consequences were, however, more serious than we had anticipated, in that the cumulative downsteps in any input sentence of two or more clauses soon resulted in unacceptably steeply falling contours. Indeed, in multi-clause sentences the shrinking of the register caused by this rapid descent towards the speaker-baseline resulted in accents being realised by pitch-excursions of a size more appropriate to microprosodic phenomena. Figure 10 shows an F_0 trace produced by INTERFIX 1.1, which illustrates this problem. The fact that Version 1.1 was actually much worse in this respect than Version 1.0 led us to develop a completely different approach to the treatment of domain boundaries.

2.1.3 TGs

In view of the failure of our previous strategy to mimic the phonological structure of natural intonation, we re-assessed the problem and attempted to formulate a new strategy which would solve the problems of the old one without putting unrealistic demands on the syntactic analysis. We knew that the principled division of the syntactic analysis into appropriate phonological domains is essential to the perceived quality of synthetic speech, but that the mismatches between syntactic structure and phonological structure are many and varied such that only a full semantic and pragmatic analysis would allow all these mismatches to be resolved, and it is not currently possible to generate any such analysis automatically. Indeed, not only are the details of the mapping between syntax and phonology unclear but the existence of any consistent correlation between syntactic and phonological structure is still the subject of considerable theoretical debate. In accordance with our strategy of default rules and limited exception clauses we therefore sought to assign a phonological structure to text sentences which is fairly easily derivable from a crude syntactic analysis, can be assigned left-to-right, and yet provides a useful hierarchy of phonological domains.

We decided on a three-tier hierarchic model of phonological domains or tone-groups

(TGs), with an input text sentence as the largest domain (tg(2) in our formalism), a major syntactic constituent (e.g. NP, VP) as the smallest domain (tg(0)), and a full clause as the intermediate domain (tg(1)). This system was seen as a minimum taxonomy, as the domain of a tg(2) was already given, the domain of the accent assignment rules corresponded to a tg(0), and at least one intervening level was required to model hierarchic effects on register. We implemented this model in INTERFIX 2.0, with the expectation that it would constitute an improvement on Version 1.0 but that there would still be significant problems in our approximation of prosodic structure.

As well as the register effects associated with boundaries in Versions 1.0 and 1.1, the three-tier model of TGs provided different tonal realisations for the three levels of boundary. Boundaries of tg(0)s were realised by the register-step alone, while tg(2) boundaries were assigned a boundary tone of the type specified in the current tune together with a short pause (with a durational value of one syllable). The realisation of tg(1) boundaries was variable, to allow for the difference between restrictive and non-restrictive relative clauses: we decided that tg(1) boundaries which were marked by punctuation should be assigned a boundary tone and pause in addition to the effects on register presented below, whereas those which were not so marked should be realised by the appropriate register effects only. Such a strategy approximates the findings of Stockwell (1972:90), Bing (1979a:151-2) and Bruce et al. (1990:128) that realisations of intonational boundaries are variable and rarely correspond to a real pause. Despite the problems with the reliability of punctuation in text which are discussed below, this treatment appeared to produce highly satisfactory realisations and has therefore been retained unchanged in the current version of INTERFIX.

The fact that the majority of TG boundaries have no boundary tone associated with them under this scheme means that F_0 is usually interpolated across these boundaries. Moreover, the fact that most TGs begin with a Mid target and end in a nuclear fall results in the majority of cases in a rising interpolation from Low to Mid across boundaries which are not assigned a boundary tone. This interpolation produces the impression of a "continuation rise" at most non-final boundaries in the acoustic output, although in actual fact no such rise is assigned by our rules. We attribute the perception of utterance-internal "continuation rises" in natural speech to this same phenomenon of interpolation across a boundary, rather than to a category of boundary tone or tail as claimed by several authors (e.g. Crystal 1969; Bing 1979a; Cooper & Sorensen 1981; Pierrehumbert 1981), and as a consequence we do not model such rises explicitly but instead allow them to arise through the underspecification of most of the synthetic F_0 contour. This approach not only produces the appropriate impression of incomplete or unfinished contours (the speaker has a following target in mind), but also accounts for the observed variability in F_0 at boundaries supposedly marked by "continuation rises" (Lieberman 1967:53; Bing 1979a:81) since the precise characteristics of the rise depend not on the boundary with which it is associated but instead on the type and position of the accents on either side.

2.1.4 Punctuation and Boundaries

The inclusion in the intonation rules of reference to sentence punctuation has always been desirable, since punctuation is one of the few indications of the intended prosodic realisation of a text. Although we did not wish our domain-assignment rules to depend upon punctuation-marks to the extent of the rules proposed in Yiourgalis & Kokkinakis (1990) or Choppy (1979), it was in keeping with our general policy of making maximum use of all the information available from text to incorporate rules for the interpretation of sentence punctuation. It was hoped that punctuation would provide indications of both hierarchic structure and domain boundaries. Unfortunately, it was not clear how reliable or consistent punctuation would be as an indication of phonological structure, and before the development of INTERFIX there was no obvious location for such rules.

Version 2.0 includes rules for the interpretation of commas, which both override and reinforce the domain identification rules discussed above. Rules for identifying lists and parenthetical phrases on the basis of punctuation assign boundary tones which override the TG structure, and as described above commas are also employed to determine whether boundary tones are assigned at tg(1) boundaries. The identification of lists and parentheticals in text is very important if a naturalistic phonological structure is to be produced. In particular, vital register effects can only be modelled if such constructions are identified. The heuristics employed by Version 2.0 are relatively crude, but their accuracy is very high: of 34 instances of lists and parentheticals in a small test corpus designed to pose problems for Version 1.1, all but one were correctly identified by Version 2.0.

A list is defined as ***two or more syntactic elements of the same category at the same level separated by commas and followed immediately by a conjunction with an optional comma before it.*** This definition allows our rules to identify lists of any number and type of element from words to clauses, and also makes allowance for variations in punctuation style. We can thus interpret all the commas and conjunctions in (7) as separating elements of lists.

(7a) sausages, egg, beans and chips

(7b) I came, I saw, and I conquered.

(7c) He ran through the door, across the hall, up the stairs and into the attic.

(7d) My supervisor eats, drinks and breathes intonation!

A parenthetical is similarly defined as ***any syntactic element flanked by commas at the same syntactic level.*** This definition assumes that the TTS system's parser will handle commas: the parsing strategy in the CSTR TTS system simply treats any comma as a sister of the constituent on its immediate left. This means that a parenthetical can be any constituent with a comma at each side. The end of the input sentence may replace the righthand comma, but the beginning of a sentence cannot serve as the lefthand boundary of a parenthetical. This may be because a sentence-initial parenthetical by definition constitutes a marked, pre-NP domain and is therefore a special case, or it may be the pressure to accent the earliest item in an utterance which forces a different treatment of sentence-initial parentheticals: whatever the reason, both the accent patterns and the realisations of the final boundaries of the parenthetical phrases in (8) and (9) differ in natural speech according to their position in the sentence, with the (a) versions conforming to our treatment of parentheticals but the (b) versions behaving more like our pre-NP domains.

(8a) John's brother, you know, reads four novels a day.

(8b) You know, John's brother reads four novels a day.

(9a) Tomorrow, William, we must feed the pine marten.

(9b) William, tomorrow we must feed the pine marten.

The parenthetical rules are ordered after the list-identifying rules, since the former are less specific and would therefore apply erroneously to input which should be treated by the latter. This ordering prevents the examples in (7) from being labelled as parentheticals, but still allows the correct labelling of true parentheticals such as those in (10).

(10a) I was given sausages, which I can't stand, and chips, which I adore.

(10b) I came and, incidentally, I conquered.

(10c) John, the useless lump, doesn't read books.

(10d) Fred, John's brother, has read every book by Kurt Vonnegut twice.

These simple definitions account for the majority of such constructions in the sentences we have processed. Their implementation and interpretation are discussed in detail below: however, they obviously depend for their input on two rather unreliable sources of information. Firstly, the use of syntax-based definitions, here as elsewhere, presupposes a syntactic analysis which can consistently provide the appropriate information. Although our punctuation rules do not make excessive demands on the syntactic analysis, there are still cases where the syntax is too ambiguous to be resolved by current parsers. This is particularly true in the case of constructions involving conjunctions (Grishman 1987:84ff.): since these are crucial to our identification of list constructions, the problems which syntactic analysers experience with examples such as those in (11) will affect the application of our rules. There is no way to ensure that a parser will assign list elements to the same level of structure (unless the syntactic rules are specifically designed to do this, which would defeat the object of using syntactic rules to determine constituency).

(11a) fish and chips, bacon and eggs and prunes and custard

(11b) Tom won the doll, the goldfish and the panda, and Dick shot the teddy.

Secondly, our rules assume a level of consistency in the use of punctuation which is impossible to guarantee in unrestricted text. The optional nature of the final comma in a list is a concession to the variation in the use of punctuation, both by a single author and between different authors, but we cannot make sufficient allowance for the range of options which authors make use of in punctuating anything but the most formal and legalistic texts. There is nothing which obliges writers to demarcate all parentheticals with commas or other punctuation, and indeed English is much more flexible in this respect than a language such as German where the use of commas is almost grammaticalised. We must accept, therefore, that rules based on regularities of punctuation can only be approximate and will be more or less prone to errors depending on the style and provenance of the text to which they are applied.

2.1.5 INTERFIX 2.0

INTERFIX 2.0 was the first version to incorporate the TGs introduced above. Their implementation and its consequences are discussed in this section, as is the implementation of our rules to interpret sentence-internal commas. Version 2.0 also incorporated rules to assign boundary tones to domain boundaries, according to the principles presented above.

2.1.5.1 TG Implementation

The implementation of our three-tiered model of TGs consists of two distinct stages. First, the various domains must be recognised and marked by INTERFIX: second, these marks must be interpreted by the phonetic model. The first stage refers exclusively to the output of the syntax module, and assigns domain boundaries left-to-right by the rules in Table 2 which encode an interpretation of the principles given in Table 1 above. The output of this stage is a completely flat sequence of domains containing boundary

Table 2: Rules for Assigning TG Boundaries

1) Assign a tg(1) boundary at the start of a syntactic clause.

2) Assign a tg(0) boundary at the start of the first NP in a clause.

3) Assign a tg(0) boundary at the start of the first VP in a clause.

symbols but with no explicit hierarchic structure. No tg(2) boundaries are assigned, as the entire input is by default treated as a single tg(2) in Version 2.0. The interpretation of TG boundaries by the phonetic model required an expansion of the register parameter in the F_0 equations given in Section 1.3.2. The single phrasal parameter Fp which Ladd's (1987) original equations used was expanded to give a tg(0) parameter, a tg(1) parameter, and a tg(2) parameter, all of which are equal to 1 by default. At any tg(X) boundary, then, the default treatment is to downstep the tg(X) parameter by a factor of 0.8 and reset any tg(Y) parameters (where $Y < X$) to 1. The current register setting $f(N)$ is then calculated as the product of the default setting N and these three parameters. Thus, successive tg(0) boundaries will create a decline in F_0 which is arrested when the tg(0) phrasal parameter is restored to its initial value of 1 at tg(1) boundaries: similarly, successive tg(1) boundaries will progressively lower F_0 until a tg(2) boundary resets the tg(0) and tg(1) phrasal parameters to their default values. This produces a natural-sounding contour in which both hierarchical and local relations can be expressed in a regular manner. Figure 7 gives a schematic illustration of the path of typical register shifts in this scheme, and the contours in Figures 9-11 show their effect on the course of F_0 in the acoustic output.

Since the default is to downstep at every boundary, an approximation to declination models (Thorsen 1985; Kugler-Kruse & Posmyk 1987) will generally be produced - however, there is nothing to prevent our model upstepping at certain boundaries and this allows us to produce more natural-sounding output than a declination model. A further notable advantage of our model over the declination approach is that we avoid any reference to absolute timing in modelling pitch relations: there is no requirement to calculate absolute or relative timings for the control of register, since all events are strictly local and are not determined by factors such as overall utterance length or

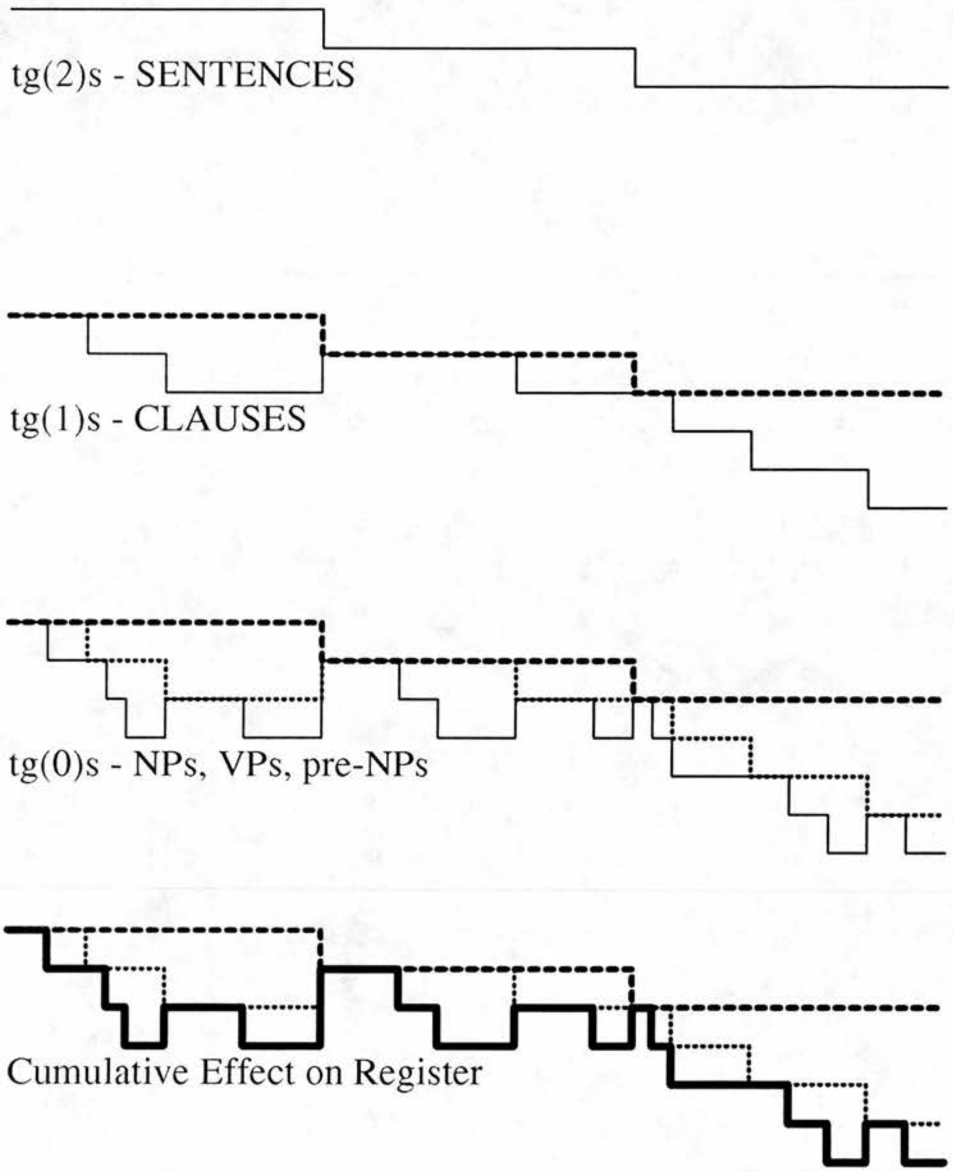

Figure 7: The Effect of Default Downstepping on the Register in our Three-Tier Model of Prosodic Domains

absolute degrees of slope which play a major role in many TTS systems (e.g. Young & Fallside 1980; Gårding & Bruce 1981; Sorin et al. 1987; Isard & Pearson 1988; Willems et al. 1988). This makes our model more flexible and thus more widely applicable, and at the same time avoids considerable computational expense and complexity.

Both theory and implementation allow for downstepping and resetting of the tg(2) parameter, to reflect the hierarchic relations between sentences. It is intended that this should correspond to the organisation of text into paragraphs, and that in each paragraph the first sentence should be *upstepped* and the last sentence should be *downstepped.* This is in line with other authors' results for paragraph-level register effects on intonation (Lehiste 1975; Silverman 1987). Rules interpreting paragraph breaks in this way have been implemented, but the information to drive the rules is not currently extracted from text. Because of the fact that the rules only require the specification of paragraph boundaries, sensitivity to white space in text is all that they demand of the text pre-processing module in any TTS system: however, to date the CSTR TTS system as a whole processes each text sentence in complete isolation, being quite unaware of any suprasentential phenomena, and consequently does not pass these breaks on to the intonation rules. Further levels of prosodic structure, such as the organisation of a text into topics or the relations between more than one text, could be simply and elegantly incorporated into our rules merely by defining the boundaries of such units and adding a further parameter to the register equation for each successive level to be included.

2.1.6 Implementation of Punctuation Rules

Once list commas and parenthetical commas have been identified and marked by INTERFIX, their interpretation by the phonetic model involves manipulating the same TG-dependent register parameters introduced above. The three values which define the register setting at the point where the comma is encountered are stored on a stack: in the case of parentheticals, they are retrieved at the end of the parenthetical; in the case of lists, they are retrieved at the end of each element of the list. This storage and retrieval of the register parameters results in a temporary suspension of the register: register changes within parentheticals or lists are free to occur independently of the surrounding context, but at the closure of the suspension the superordinate register setting is resumed where it left off. The use of a stack allows such constructions to be embedded, such that a parenthetical may contain a list one element of which contains another parenthetical. In this case, the outer parenthetical triggers the storage of the register parameters which are retrieved at the end of that parenthetical: next, the list construction will cause register parameters to be stored and retrieved at the beginning of each new item, and for the item containing a parenthetical there will be a further suspension where the current parameter values are stored for the duration of the parenthetical and then retrieved for the remainder of the list item. At the end of the list, the parameters stored at its beginning are retrieved, leaving only the parameters for the outer parenthetical on the stack: at the end of that parenthetical, the stack is cleared and the register is reset using the retrieved values to exactly what it was before the entire embedded construction occurred. Such complicated embedded constructions are a relatively common phenomenon in formal and technical text.

The result of our treatment of lists and parentheticals is that all elements of the same list begin at the same register setting and any parenthetical is downstepped relative to the surrounding material: this is in agreement with suggestions in the literature (Crystal 1969:273; Stockwell 1972:107; Choppy 1979:188; Cooper & Sorensen 1981:85; Knowles forthcoming). Figure 8 gives a schematic illustration of the register shifts associated with lists and parentheticals by our rules.

Commas which do not mark parenthetical or list constructions generally fall into one of two categories. The first category comprises commas which occur at domain boundaries and distinguish, for instance,

(12) However, you managed to do it

from

(13) However you managed to do it, . . .

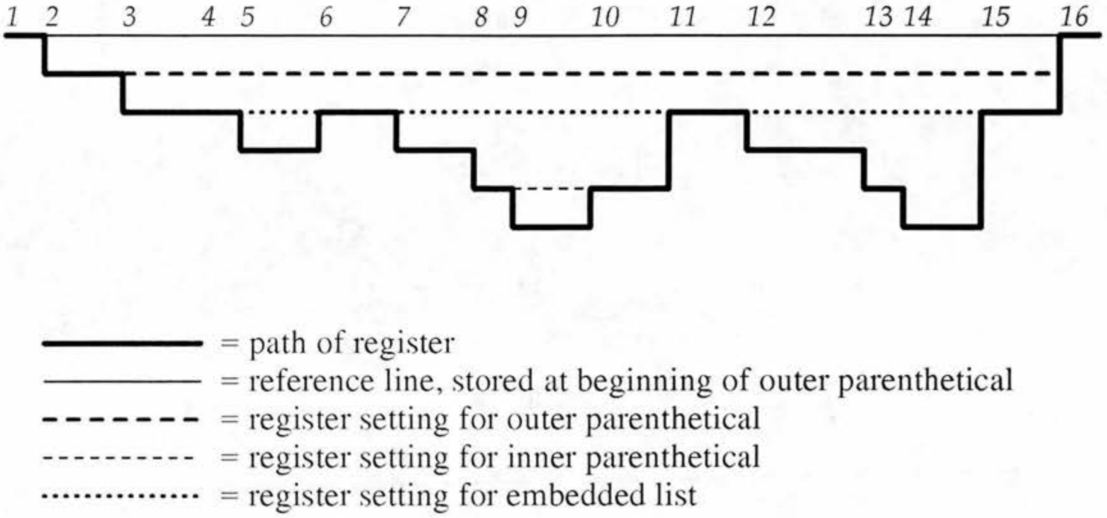
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
= path of register
= reference line, stored at beginning of outer parenthetical
= register setting for outer parenthetical
= register setting for inner parenthetical
= register setting for embedded list

Figure 8: The Effect of Lists and Parentheticals on the Register

This example illustrates the treatment of a sentence which contains a parenthetical within a list within a parenthetical. The register setting at point [1] is that of the matrix sentence: this is stored at point [2], where the outer parenthetical begins. The parenthetical downsteps the register immediately. At point [3], there is a tg(0) boundary within the parenthetical, e.g. between NP and VP, which causes a downstep. Point [4] is the beginning of a list within the parenthetical: the register is not immediately downstepped, but the register value is stored. There is a tg(0) boundary within the first element of the list at point [5]. The second element of the list begins at point [6], causing the register to be reset to the value stored at [4], and there are two tg(0) boundaries within this second element at points [7] and [8]. Point [9] is the beginning of a second parenthetical within the second element of the list, and this causes an immediate downstep: point [10] is the end of this inner parenthetical, so the register returns to the level stored at point [9]. A third element of the list starts at point [11], causing a reset to the value at point [4], and this third element contains three tg(0) boundaries at points [12], [13] and [14]. The whole list ends at point [15], causing a return to the level of the outer parenthetical (downstepped at point [3]), and finally the parenthetical ends at point [16] returning the register to the setting stored at point [2]. This is an unusually complicated case, but its purpose is to illustrate the potential of our model.

These commas are treated as indicating a stronger degree of boundary, and are therefore translated into boundary tones and pauses by INTERFIX 2.0, giving a boundary strength similar to that of a tg(1) boundary but without the associated effects on register and thus with no change in the perceived prosodic structure. This treatment adds appropriate emphasis to these boundaries but preserves the original hierarchical structure. The second category contains all remaining commas, and includes those in strings of adjectives (e.g. *The big, bad wolf*) and between prepositional and other minor phrases: these do not generally convey any vital phonological information, and are simply deleted.

Commas do not constitute boundaries in their own right (although they may coincide with boundaries, as stated) and therefore do not affect the operation of the Rhythm Rule. They are present in the input to the Rhythm Rule, but they are simply ignored and passed to the phonetic rules.

2.1.7 Comparisons

The three contours which follow illustrate the F_0 output of the three versions of INTERFIX discussed above. They show the treatment of the input sentence *Yesterday he finally delivered the manuscript, which was due last month, although the introduction and the index were still missing* by each version. The syntactic analysis did not vary between versions, so all differences are the result of modifications to the intonation module.

Although there is little variation in the first clause, the contours produced for the second and third clauses differ markedly from version to version. This is because almost all the advances presented above are designed to improve our treatment of multi-clause input: the treatment of a one-clause sentence would be much the same whichever version of INTERFIX were applied. The important differences in the output of successive versions only become apparent when more complicated input is to be handled. For example, the hierarchic phonological relations between the three clauses are simulated by the output of Version 2.0 in a manner of which neither of the earlier versions was capable. Similarly, the effect of improved domain assignment rules in later versions on accent placement are shown in the contours below. The limit of three domains per sentence in Version 1.0 results in an extremely long final domain in Figure

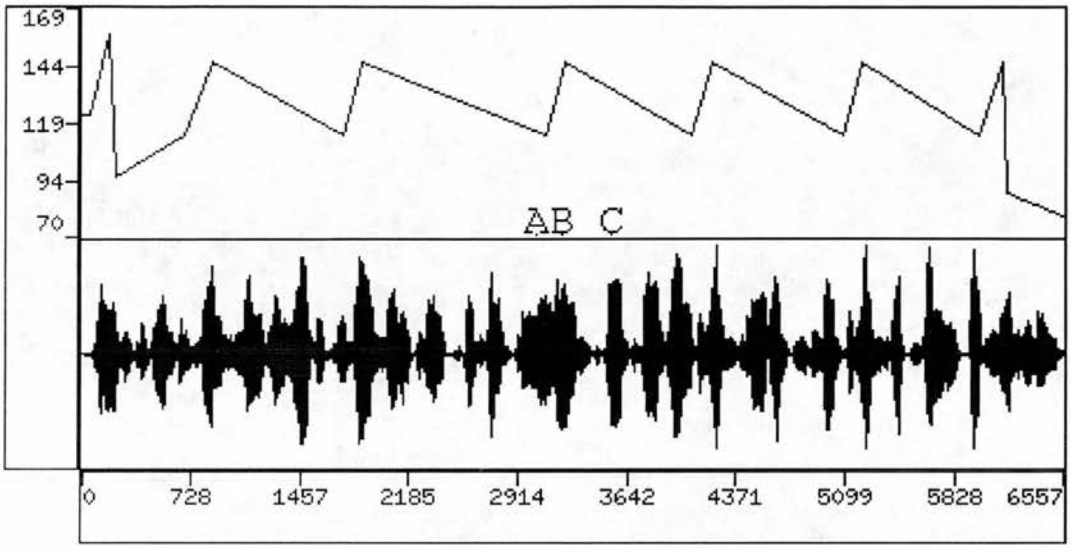

Figure 9: A Sample of the Output of INTERFIX 1.0

The letters A, B and C indicate the positions of the words *due, last* and *month* respectively, in this figure and in Figures 10 and 11.

9, and consequently in a serious accent-placement error: the placement of an accent on word B but none on A or C in the sequence *due last month.* Both Version 1.1 and Version 2.0, with their improved domain assignment rules, avoid this error and produce the highly acceptable accentuation *DUE last MONTH* as a result: despite the rapidly decaying contour produced by Version 1.1, the accents assigned to A and C can still be seen in Figure 10, illustrating the improvement in the operation of the accent rules which was brought about by a better definition of prosodic domains in later versions of INTERFIX.

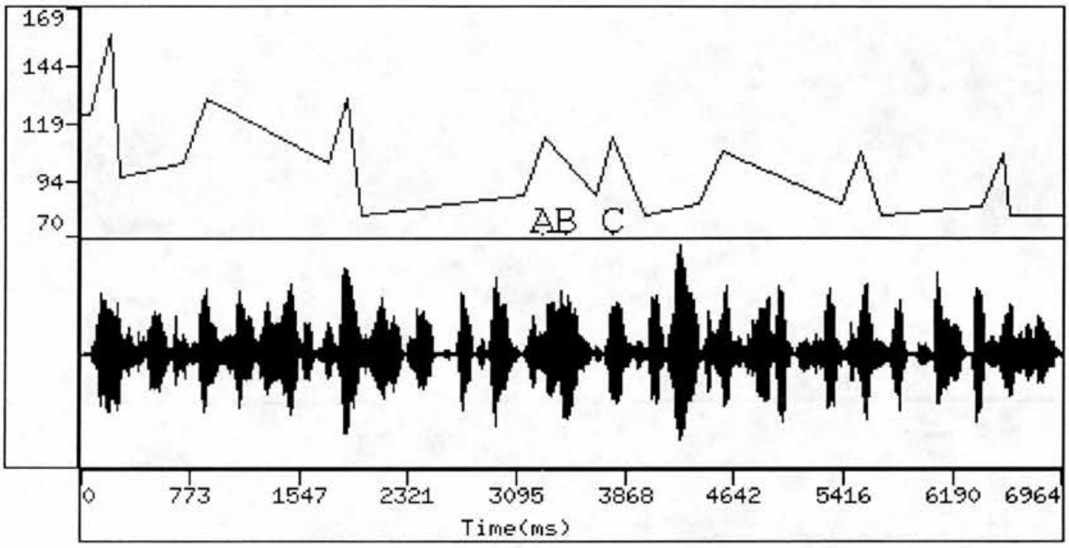

Figure 10: A Sample of the Output of INTERFIX 1.1

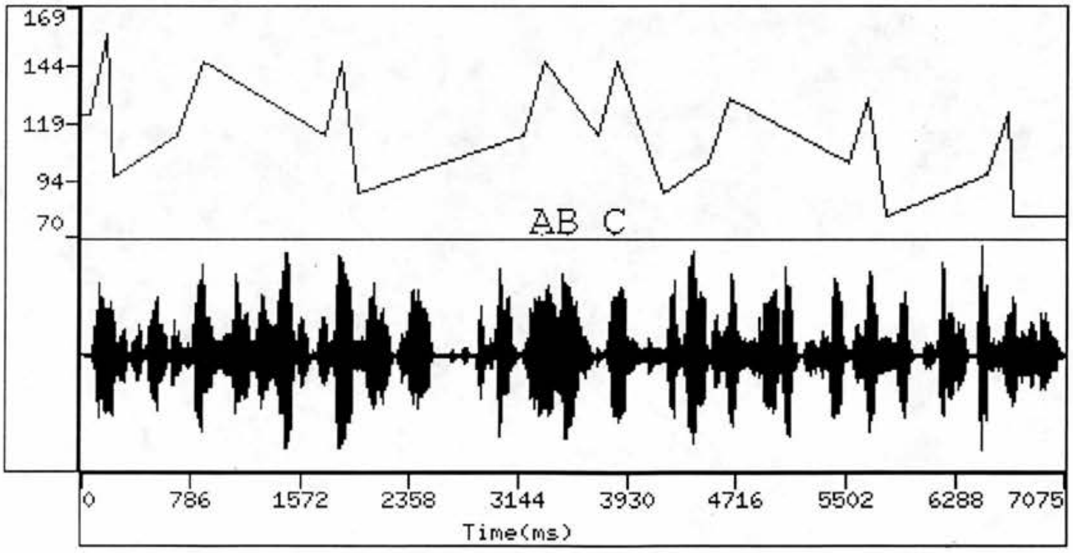

Figure 11: A Sample of the Output of INTERFIX 2.0

2.2 Accent-Assignment Rules

As we pointed out above, automatic semantic and pragmatic analyses of text are not currently available: the intonation component of our text-to-speech system therefore has only lexical and syntactic information to work on. The lexical input supplies the locations of lexically-stressed syllables, and this information determines the placement of pitch targets within words. The syntax provides word-class information, which drives the accent-assignment rules. The accent-assignment rules are essentially a small set of default rules which process each tg(0) domain in isolation and within that domain assign primary ('1'), secondary ('2'), or no ('-') accent to lexical items based on the word-class of those items. We currently define primary accent as a degree of accent which is potentially nuclear, and secondary accent as any less prominent accent. Primary accent is assigned to some content words (CWs), such as nouns and proper nouns: secondary accent is assigned to other CWs, such as main verbs, adjectives, and adverbs. No accent is assigned to other items, which are considered to be function words (FWs). The CW/FW split in accentuation strategies was mentioned above (p.58) in describing the intonation rules developed at Utrecht (Quené & Kager 1989, 1990) and CNET (Sorin et al. 1987), and is a common basis for accent rules in TTS systems: the assignment of different **degrees of accent** is, however, much less usual, and may even be unique to our rules. The direct determination of relative peak height from word-class information, as in the O'Shaughnessy algorithm (Allen et al. 1987:101ff.), is quite different from the degrees of accent which we assign for two reasons: firstly, there is no difference in peak height associated with different degrees of accent in our model; and secondly, as is shown in Section 2.3, the degrees of accent which we assign are phonological rather than phonetic in nature and may be changed or even deleted by phonological rules. The observation which underlies our division of CWs into two classes is that in any domain the nucleus falls not on the rightmost accent, but on the rightmost POTENTIALLY NUCLEAR accent: for example, in the everyday situation of a child experimenting with the use of saucepans as military helmets, *head* and *stuck* are both equally new but the former receives the nucleus in both (14a) and (14b) regardless of word-order.

(14a) DON'T do that: you'll get your HEAD stuck!

(14b) DON'T do that: it'll get stuck on your HEAD!

The division of CWs into, roughly, arguments (potentially nuclear) and predicates (not potentially nuclear) allows us to distinguish between items such as *head* and *stuck* whilst allowing us the option, in (14b), of assigning a non-nuclear accent to *stuck*: this seems to produce the appropriate results in most cases, although there are obviously exceptions to such a simple strategy, some of which are discussed below.

These rules work reasonably well for simple sentences such as

(15) The aardvark apologised to the Alsatian.

However, treating NPs with multiple adjectives or noun-noun compounds using these rules leads to unacceptably over-accented output: a typical NP from an academic journal,

(16) various linguistic factors of sentence stress assignment and detection in spoken and written Russian texts

would be assigned six primary and five secondary accents by these rules. This is clearly undesirable overkill, to say the least.

Some criteria for removing (or at least reducing) certain of these accents in complicated noun phrases (hereafter 'big NPs') such as (16) are therefore required, and the obvious candidates are semantic knowledge and information on grammatical function. Unfortunately, preliminary attempts to incorporate some lexical semantic knowledge did not reveal any interesting regularities: accent assignment heuristics incorporating various semantic distinctions (e.g. material/nonmaterial) did not perform much better than chance, and more sophisticated semantic processing is beyond the current scope of text-to-speech systems both in terms of processing constraints and because there is no well-developed theory of semantic representations for natural language. The use of detailed grammatical information to identify heads and modifiers in big NPs also proved to be impractical because of the inefficiency of highly-detailed parsers and the performance demands for real-time text-to-speech conversion: however, this might be an option for future systems.

In the absence of cues from syntax and semantics, we have developed a set of heuristics based on the principle of rhythmic alternation, according to which accentual and other prominences are relational in nature and depend for their realisation on a metrical structure of alternating "weak" and "strong" elements (Liberman & Prince 1977; Ladd 1980): we have also incorporated other phonological insights wherever possible. These heuristics operate on the **over-accented representation** produced by the default accent-assignment rules above, and are designed to produce maximally acceptable, neutral default contours for the majority of English sentences: as stated in Chapter 1, we make no claims to model natural human speech and we do not attempt to handle contrastive or emphatic contours.

Table 3: Domain-General Rhythm Rule

(Over-Accented Representation)

1. Delete all accents to the right of the rightmost primary.
2. Reduce all primaries except the rightmost to secondaries.

(Fully-Accented Representation)

3. Delete every ODD-NUMBERED secondary LEFTWARDS from the primary.
4. Apply Well-Formedness Conditions (WFCs).

2.3 Domain-General Rhythm Rule

Our Rhythm Rule was originally developed to handle big NPs (Monaghan 1987a), but the improvements in our system's automatic identification of phonological domains discussed above have allowed its generalisation to all domain types. The Rhythm Rule currently handles one tg(0) domain at a time, treating each such domain entirely independently of all other domains. The rule takes the **over-accented representation** discussed above and produces a **fully-accented representation**, through the application of phonological generalisations, and finally a **rhythmic representation** according to the principle of rhythmic alternation. This procedure provides a useful approximation to metrical foot structure without resorting to lengthy time-related equations or constructing metrical trees. The basic principles of the Rhythm Rule were originally worked out in collaboration with Dr. D. R. Ladd, but the present author takes full responsibility for their implementation and refinement. The different stages of the rule are set out in Table 3. Clause (1) of the Rhythm Rule deletes all accent markers after the rightmost primary, which is the default nucleus of the domain: all accent markers to its right are therefore post-nuclear, and post-nuclear tails are by default accentless. There are two reasons for deciding on the rightmost primary as the default nucleus. Firstly, placing the nucleus as late as possible allows us to produce acceptable contours for more sentences: this is in line with Newman's (1946:176) observation that "the last heavy stress in an intonational unit takes the nuclear heavy stress." Crystal's (1969:224) data bears this out, as do Berman & Szamosi's (1972) examples, and our own informal experiments (Monaghan 1987a) indicate that in assigning accents for synthetic speech output a postponing of the nucleus is in most cases preferable to a long tail. Secondly, in English and many other languages new or important information (and therefore the nucleus) tends to come towards the end of an utterance (Halliday 1967b:22; Brown & Yule 1983:126ff., 156; Sperber & Wilson 1986:216). This generalisation is similar to Bolinger's (1985) "climactic" ordering of information and the Praguians' CD (Daneš 1972).

Domains which contain no primary accents, such as intransitive VPs or copular phrases, would be left accentless by clause (1). We have found that such phrases should retain their accents in the default case: these domains are therefore identified in advance and treated specially. Effectively, the last accent (if any) in such a domain is temporarily promoted to a primary and then demoted again after the Rhythm Rule has been applied.

Once the nucleus has been identified, clause (2) reduces all other primaries to secondaries. We currently allow only one primary accent per tg(0) domain, since we have found that this produces more acceptable output in most cases. The output of the first two clauses is termed the **fully-accented representation** because we believe that under normal circumstances any desired contour can be constructed from some subset of the accents available in this representation. As is stated above, this does not include unusual emphasis or stylistic variation. Subsequent clauses are therefore optional, depending on speech rate and style: this optionality is discussed in Section 2.8.

The third clause of the Rhythm Rule implements a basic rhythmic alternation in the string of accents, deleting roughly half of the secondaries. The basic principle of deleting alternate accents produces highly acceptable output in most cases. A detailed illustration of our treatment of an example which we feel provides a more familiar parallel to Thompson's (1980) classic example of rhythmic deletion

(17) SAN francisco GOLden gate BRIDGE

is given in the next section.

The need for clause (4) arises because of constraints on higher-level domains, e.g. clause-level or sentence-level constraints. We have developed a set of Well-Formedness Conditions (WFCs) to implement such effects, and these are discussed in detail in Section 2.5. As an example, although the leftmost secondary in a DOMAIN can be deleted by clause (3) of the Rhythm Rule, the leftmost secondary in any SENTENCE must be preserved. Thus, even if the domain *has green feet*[3] comes out of the Rhythm Rule with a single (primary) accent on *feet,* the accent originally assigned to *green* will have to be restored in a sentence such as

(18) He has green feet.

because the first **accentable item** in any utterance must retain its accent.

[3] We assume that all forms of verbs such as *have, be, do* behave intonationally as auxiliary verbs (unaccented) in the default case.

Input Text:					
	b	b	c	radio	news
Accent Assignment:					
(Over-Accented Representation)	1	1	1	2	1
Clause 1: Delete all accents to the right of the rightmost primary					
	1	1	1	2	1
Clause 2: Reduce all primaries except the rightmost to secondaries					
(Fully-Accented Representation)	2	2	2	2	1
Clause 3: Delete every ODD-NUMBERED secondary LEFTWARDS from the primary					
	2	d	2	d	1
Clause 4: Apply WFCs.					
(Rhythmic Representation)	2	d	2	d	1
	B	b	C	radio	NEWS

Figure 12: An Example of the Operation of the Rhythm Rule

2.4 An Example

To illustrate precisely how our Rhythm Rule works, the stages involved in generating rhythmic output are stepped through here. Figure 12 shows the representations derived at each stage: capitalisation indicates syllables which have a pitch accent associated with them.

The input to clause (1) of the rule is the over-accented representation [1, 1, 1, 2, 1]. (If *radio* were identified as a noun or a proper noun rather than an adjective, the input would consist entirely of primaries: the final output, however, would be unchanged.) As the rightmost accent is a primary, clause (1) performs no deletions and passes the list of accents on unchanged.

Clause (2) has rather more to do in this case, reducing all the primaries except that on *news* (already identified as the nucleus) to secondaries. If *radio* had been parsed as a noun or proper noun, its accent too would have been reduced at this stage. The fully-accented representation, [2, 2, 2, 2, 1], is passed on to clause (3).

Rhythmic alternation is introduced by the third clause of our Rhythm Rule, which finds the nucleus (rightmost accent) and works back from it, deleting alternate secondaries starting with the first. The output of clause (3), in this case the list [2, d, 2, d, 1] (where 'd' indicates a deleted accent which may or may not be reinstated by a later WFC), is passed to the final clause.

In this example there are no WFCs to apply and so the accent string is output unchanged. It only remains to map the final rhythmic representation onto the stressed syllables of the utterance to produce the desired result:

(19) B b C radio NEWS

Some further examples of the output of the representations produced by our Rhythm Rule, which were judged highly acceptable in the evaluation reported in Section 3.1.2, are given here. '|' indicates a domain boundary, with material on each side treated independently by the Rhythm Rule.

(20) uNIted nations CONference on trade and deVELopment

(21) WEST lothian DIstrict general HOspital

(22) QUAsi-autonomous NATional government organisATion

(23) and that's the NEWS at FIVE minutes past ONE

(24) DEists | beLIEVED in god as a creAtor but reJECTed reveLATion

(25) it was a conTINuing belief in PROvidence | which suSTAINED
voltaire's DEism

There are of course many exceptions to the Rhythm Rule as stated here. Some of the more regular ones are described in Appendix C, and have been incorporated into our current rules but will not be discussed here: others are not so easily predicted, and may never fall within the scope of automatic systems.

2.5 Well-Formedness Conditions

Given the reliance of our Rhythm Rule on particular domains, it was inevitable that as the definitions of domains evolved there would be consequences for the appropriateness of the accentuations which the rule produced. In several cases, work on INTERFIX has also led to improvement and refinement of the Rhythm Rule, either by revealing more effective strategies or by resolving a particular issue one way or the other. These interactions have been formalised as phonological constraints or **well-formedness conditions** (WFCs) which apply to certain domains and override their default treatment by the Rhythm Rule. The defining characteristic of our WFCs is that they do not apply to tg(0) domains in isolation, unlike the accent-assignment rules and the Rhythm Rule. The current version of INTERFIX includes a small number of domain-specific and domain-general WFCs which have been determined by analysing the regular errors made in the treatment of different domains. The development of the major WFCs currently implemented is described in this section.

Our original Rhythm Rule was formulated for isolated NPs, as stated above, and contained a rule preventing the deletion of the initial accent in an NP. The version extended to VPs and full sentences preserved this rule, as it appeared to apply equally well to pre-NP domains and reasonably well to most VPs, particularly long and complicated VPs. It was clear from informal assessment of INTERFIX Version 1.1, however, that the preservation of domain-initial accents in series of short domains produced unacceptably over-accented output. We were faced with a dilemma: a rule which worked very well on particular domains appeared to be inappropriate for full sentences. Should it be included in a domain-general rhythm component, and if not, did we need domain-specific rhythm rules after all?

The original implementation of our Rhythm Rule also expected a domain to contain a primary accent, an understandable expectation in a rule developed specifically to handle big NPs. Allowances were made for domains with no primary accent, as described above, but these were initially assumed to be very infrequent. This assumption was soon found to be incorrect, and indeed domains were often found to contain no accents at all, e.g. NPs consisting solely of a pronoun. However, these domains were still relatively rare until the implementation of INTERFIX 1.1 produced smaller and more numerous domains. The problem of empty domains assigned by INTERFIX, particularly in pre-NP position, was discussed above: in addition, the Rhythm Rule's performance was significantly impaired by the frequency of non-empty domains with no accent.

A third problem arose precisely from our treatment of domains with no primary accent but with one or more secondaries. The strategy of treating the rightmost secondary as a primary for the purposes of the Rhythm Rule worked well in most cases, but the fact that this accent was nonetheless realised phonetically as a secondary meant that in cases where such a domain occurred sentence-finally the final accent in the utterance was non-nuclear: the absence of a nuclear fall led to an impression of continuation or suspension rather than finality at the end of such utterances, as we would expect from our view of continuation rises and related phenomena presented in Section 2.1.3 above. However, the treatment of such domains in positions other than sentence-finally was highly satisfactory, lending further support to our treatment of utterance-medial continuation rises as the absence of boundary tones. Again, we faced a dilemma: was our basic rule flawed, or was there some other explanation?

The first and third problems seemed to indicate some sort of "utterance effect", i.e. a constraint on utterance-initial or utterance-final domains which did not apply to other domains. In the former case, we had a rule ("Preserve the first accent") which applied very well to NPs when they constituted a whole utterance in themselves but not when they formed only part of an utterance. In the latter case, we had a strategy which was appropriate for domains which did not contain the nucleus of the entire utterance but which was quite inappropriate for the final accent in an utterance. This situation led us to formulate **well-formedness conditions** on the specification of an utterance or tg(2) domain, which could rectify the failings of the Rhythm Rule at the tg(0) level. The sub-rule of preserving a domain-initial accent was deleted from the Rhythm Rule, and our treatment of domains containing secondary (but no primary) accents was continued unaltered, but the output of both these processes was made subject to the following pair of tg(2)-specific WFCs:

1. The first accentable element in a tg(2) must be accented.

2. The final accent in a tg(2) must be a primary accent, i.e. it must constitute the nucleus of the tg(2).

The first of these tg(2)-specific WFCs expresses the insight that the accenting of the first accentable element in a domain is specific to tg(2) domains and is not a domain-general constraint. Indeed, if it is applied domain-generally it will produce unnaturally over-accented representations. Implementing this WFC required removing the relevant accent-preserving clause from the Rhythm Rule, as mentioned, and checking the final phonological string to see that it conformed to the new restrictions. Since unaccented accentable items are marked as such by the 'd' symbol, checking this string merely requires a left-to-right pass as far as the first accentable element: if this is already accented, no action is required; if not, it is assigned a **tertiary** accent. Originally, this WFC replaced a 'd' with a '2', but this appeared to give too much prominence to the accented item in many cases and so we decided to assign a tertiary accent, corresponding to the first half of an IPO "flat hat" contour ('t Hart & Cohen 1973), when this WFC applied. Tertiaries indicate close dependency on the following accent, and since a reinstated tg(2)-initial accent must generally be assigned to either a verb with a following argument (in a VP domain), some modifier of a following argument (in an NP domain), or an adverb or subordinating conjunction (in a pre-NP domain) this treatment is generally appropriate.

This WFC, together with a minor alteration in the relevant accent-assignment rule, provides an elegant treatment of those items which are assigned accents in neutral utterances ***if and only if*** they are the first accentable element in a tg(2). This is characteristic of items such as subordinating conjunctions and certain adverbs and prepositions, which are often accented utterance-initially but never utterance-medially: examples of this behaviour are given in Section 2.1.1 above. (The only exception to this seems to be when such items occur after a colon: although this also depends on their semantic function, we are aware that this might indicate the need for a tg(2) boundary at some colons.) These items are therefore assigned a 'd' *ab initio* by the accent-assignment rules, which is promoted to a '3' in precisely those cases where it constitutes the utterance-initial accentable item. This treatment replaces the strategy adopted by INTERFIX 1.0 of assigning secondaries to such items in pre-NP domains. The output of this WFC produces the accentuations illustrated in example (6) above.

The second tg(2)-specific WFC checks that the last accent in any tg(2) is a primary accent, and if this is not the case it promotes that accent to a primary in order to comply with the restriction. This strategy assumes that such accents should be preserved in domains which contain no primary accent, as mentioned above, and even made nuclear in certain cases, and this seems to be borne out by the resultant intonation contours. The WFC replaces our previous assumption that the last accent in any domain must be nuclear, and successfully resolves the problems of over-accented output caused by that assumption.

The remaining problem mentioned above, that of empty domains and non-empty domains with no accents, was also handled by a WFC: this time, the condition was applied to all domains, from the smallest to the largest, and was stated as follows:

> 3. Any TG must contain at least one accent. Domains which could constitute TGs but which do not contain any accents (e.g. a subject NP consisting only of an unaccented pronoun) are subsumed by a neighbouring TG of the same level. Such domains may be accentless only if no accents were ever assigned to them: they may not become accentless through the operation of rhythmic or similar accent deletion rules.

This domain-general WFC captures the effects of two related constraints on the interaction of semantics and phonology. Firstly, if a potential domain contains no accentable (i.e. communicatively important and/or unpredictable) items then it cannot function as a domain. This seems intuitively correct, since the relations between any such semantically empty domain and its fellows would be vague in the extreme: if a domain is not informative enough to merit any accent-assignment, its claim to the status of a domain and its effect on other domains must surely be minimal. Moreover, since no pitch targets would be placed in such a domain there would actually be no phonetic indication of its independent status, making the distinction between an accentless domain and the pre-head or tail of some accented domain inaudible in the output: this ambiguity would clearly be theoretically undesirable, and could serve no purpose in the phonetic realisation. Secondly, if a potential domain contains any accentable items then at least one of

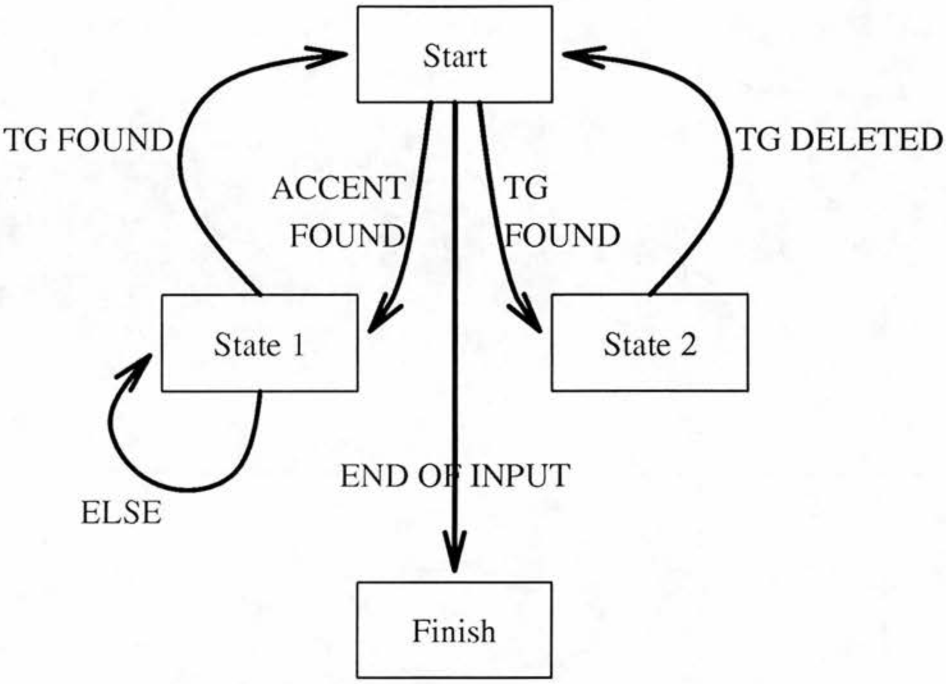

Figure 13: Deleting Superfluous TG Boundaries

them must retain its accent: this is a little less intuitive, but amounts to saying that no domain which has at any stage been assigned accents (and thus judged to be semantically important) can later be subsumed into another domain. Such a policy of preserving accents and domains where possible appears to produce more natural-sounding results than a less conservative policy, although the latter might be preferable in synthesising faster or more casual speech styles (see below).

In addition, if this WFC is implemented between domain assignment and the phonetic model, all superfluous downsteps and empty domains are removed, thus solving the problem of overly-steep downtrends and avoiding unnecessary processing. The way this is currently achieved is by a left-to-right pass through the output of the accent assignment and Rhythm Rules, ensuring that this conforms to the WFC's requirement of at least one accent per TG by simply deleting any TG boundary which is not separated from the preceding TG boundary by an accent. This is done as shown in Figure 13

Because of the order in which TG boundaries are assigned, i.e. the highest level of boundary first in a nesting arrangement, this WFC will always preserve the highest boundary level. Thus, if a tg(1) contains any accents the relevant tg(1) boundary will be preserved: only if a tg(1) contains no accents in any of its tg(0)s will the entire tg(1) be subsumed. The number of tg(0)s per tg(1) in the output of this WFC varies between 1 and 3, since TGs which come out of the Rhythm Rule with no accents are subsumed by the following TG. Only if the entire utterance is accentless will an unaccented TG be permitted by the WFC.

Even the few WFCs which we have been able to isolate so far have significantly improved the output of our accent assignment module by reducing the mismatches between an optimum phonological structure analysis and the input to that module. We expect that the interaction of our three-tier TG framework with constraints on accent assignment will shed light on several of the remaining problems in the output of our intonation system, and thus motivate further WFCs.

2.6 Incorporating Stress-Shift

Stress-shift is the phenomenon whereby in certain environments accent may fall on a syllable with secondary lexical stress in preference to the primary-stressed syllable. Thus, speakers produce

(26) SAN francisco INTernational AIRport

although when pronounced in isolation *international* is accented on the third syllable, thus:

(27) interNAtional

Despite much discussion of stress-shift in the linguistic literature (Liberman & Prince 1977, Selkirk 1984:273ff.), no other TTS system of which we are aware has any strategy for handling this phenomenon. However, stress-shift can be incorporated very easily into our Rhythm Rule. To produce (26), no modification is required to the rule stated above: we need only allow the accent-assignment rules to assign accents to secondary- as well as primary-stressed syllables. This will result in TWO accents being associated with *international*: one for the main stress and one for the secondary stress. Both accents are of the appropriate degree (primary or secondary) for the item's syntactic category, in this case secondary accents for an adjective. Given this, producing (26) is no different from producing (17) or (19), since our Rhythm Rule refers only to sequences of accents and takes no account of word boundaries.[4] Our implementation of stress-shift is much simpler than that proposed by Selkirk (1984:273): her proposal is the position of the lexical stress within the word be changed and then that the accent

[4] Butterfield & Cutler (1990) conclude that the generation of intonation in natural human speech makes no reference to word boundaries.

be re-associated with the new position; our view is that both positions are always available, and the normal operation of the Rhythm Rule simply determines where the accent falls. Selkirk's approach also forces her to put an explicit constraint on how far a stress may be moved (p.280), whereas the simpler approach taken here involves no movement whatsoever. However, our approach assumes that the presence of secondary lexical stress is synonymous with the possibility of stress-shift, and this is clearly not the usual interpretation: traditionally, secondary stress has been more closely linked with vowel quality and syllable structure than with accent (Jespersen 1909; Chomsky & Halle 1968; Liberman & Prince 1977). Yet the simplicity and elegance of an account of stress-shift which needs nothing more than a firm link between lexical stress and sentence accent leads us to present an alternative definition of secondary stress which has much in common with Bolinger's (1986) system of stress and accent. The notion of **accentability**, mentioned above in connection with WFCs, is the crucial idea: lexical stress IS accentability in our view.

In our model, primary stress and secondary stress are equivalent in all respects except one, which is that the nucleus must be associated with a PRIMARY stress. This identification of lexical stress with accentability places two empirical constraints on the occurrence of secondary stress. Firstly, all and only those syllables which participate in stress-shift must be assigned lexical stress: some empirical investigation is required to identify all such syllables. Secondly, the rightmost lexical stress in any word must be the PRIMARY stress: rightward stress-shift within a word is not available in English. (Both these constraints apply only to neutral, unmarked intonation: we are well aware that a speaker may assign stress and accent freely for special communicative purpose.) Items such as *celebrate*, whose primary stress falls on the first syllable, are consequently not marked for secondary stress: the difference in vowel quality between the final syllables in *celebrate* and *celibate* therefore has to be made on some other basis, and heuristics based on syllable or morphological structure would seem to be the most obvious candidates. In a practical system, however, any problems which our treatment of secondary stress might pose for vowel reduction rules are far outweighed by the information on syllable types which is available in our system after the operation of the Rhythm Rule. This information can be used to drive segmental phonological and phonetic rules, as well as controlling other aspects of prosody such as duration and intensity.

Table 4: Syllable Types

WORD		SYLLABLE		
Stress	Accent	Stress	Accent	TYPE
yes	yes	yes	yes	1
yes	yes	yes	no	2
yes	yes	no	no	3
yes	no	yes	no	4
yes	no	no	no	5
no	no	no	no	6

2.7 Syllable Types

The output of the Rhythm Rule provides subsequent processes with six distinct syllable types, shown in Table 4. The distinctions are based on relative prominence information, most of which is not available before the operation of the Rhythm Rule, and they attempt to combine phonological, acoustic and pragmatic effects in a very simple manner. This six-way distinction provides a useful approximation of metrical foot structure without expending any effort on constructing a metrical hierarchy: the distinctions are all drawn in the course of producing a rhythmic accent alternation. In our model, stress-markers which are not associated with an accent in the over-accented representation are deleted, whereas stresses whose associated accents are deleted by the Rhythm Rule are retained to preserve the distinction between lexically unaccentable and (rhythmically or pragmatically) deaccented items. Such a detailed system of syllable types can be used by duration-assignment routines (Campbell 1989, 1990; Bruce et al. 1990) and allophonic variation heuristics including vowel-reduction rules, and may well be more useful than traditional stress information: these distinctions are similar to those suggested in Thompson (1980:135) as being adequate for duration assignment. In turn, durational and segmental effects will contribute to the perceived rhythm of the acoustic output.

In current text-to-speech systems, such as DECtalk, INFOVOX and CSTR's TTS system, vowel reduction rules are generally applied well before any accent rules and so the prosodic structure of an utterance cannot influence vowel quality. However, there are good grounds for suggesting that a significant part of vowel reduction should be performed at some stage of the intonation assignment: for example, it is not clear until after the operation of the Rhythm Rule which words should be prominent in the output. Although word-internal reductions based on morphology or syllable structure could well be effected before intonation, there are several areas of vowel reduction which might be better handled after the accent rules. For example, the vowels in semantically empty or redundant items (verbs such as *go* and prepositions such as *to,* etc.) could safely be reduced after the operation of the accent rules if they were not accented, whereas if they were reduced before the assignment of intonation any non-neutral (pragmatic or contrastive) accent assigned to such items would require the vowels to be reconstructed: depending on the detailed working of the system, the recovery of such deleted information might simply not be possible. Given the syllable typology above, more general types of reduction could also be effected to model particular speakers or styles: all unstressed vowels, or even all unaccented vowels, could be reduced in synthesising faster or more casual speech styles; unstressed vowels in unaccented words could be elided; and so on. We therefore feel that the output of our Rhythm Rule provides both a useful set of distinctions to drive rules governing phonetic and phonological reductions and a strong argument for a review of TTS architectures to allow for the influence of prosodic structures on segmental phonetic realisations.

2.8 Speech Rate

It should be made clear again at this point that we regard the Rhythm Rule as an optional process. Although our system currently applies the rule by default, the synthesis of different speech styles and rates would make it desirable to reduce the number of accents even further in some cases and not to reduce them at all in others. As was stated above, the **fully-accented representation** is assumed to contain all the information necessary to synthesise any intonation contour occurring in fluent speech: in some accents of English, especially at slower speech rates (personal communication, anonymous reviewer), all the accents in this representation may be realised as pitch prominences. We have found to date that the output of the Rhythm Rule as it stands is generally more acceptable, particularly for running text: however, the options which the rule provides can produce prosodic specifications appropriate to a range of speaking styles and rates from the laboriously stilted to the extremely reduced. This section sets out the possibilities which our system offers for variation in speech rate.

2.8.1 Background

Until very recently, in most text-to-speech systems prosody was seen as the icing on the cake rather than a vital ingredient, and this was justified by references to the relative intelligibility of flat, uninflected synthesis as compared with synthetic prosody (e.g. van Bezooijen 1989a; Benoit 1990). However, in developmental TTS systems (e.g. Quené & Kager 1989; Hirschberg 1990ab) prosody is now attracting much more attention as its importance to high-quality synthetic output is recognised. Nevertheless, despite the research effort devoted to intonation and duration, speech rate still receives little attention. This may be due to the fact that theories of speech rate on which to base such work are practically non-existent (Couper-Kuhlen 1986:185), or it may be because the information which determines speech rate, such as speaker attitude and discourse context (Couper-Kuhlen 1986:173; Klatt 1987:760) is not available to current TTS systems.

It should be obvious that automatic speech output systems cannot produce synthetic speech output without at the very least some implicit notion of speech rate. For example, a TTS system implementing Klatt's (1979) duration rules implicitly assumes the speech rate for which Klatt's inherent and minimum phoneme durations are appropriate. Similarly, a system trained on a particular corpus of analysed speech will produce duration rules which assume a speech rate related to the rate at which the corpus materials were originally produced, whether or not this is explicit in the rules. Some deterministic procedure has to specify the absolute duration of speech segments, otherwise there can be no resultant waveform. By definition, then, every working TTS system has a default speech rate setting. However, although some TTS systems can control speech rate explicitly, e.g. AT&T's NewSpeak (Hirschberg 1990ab) which manipulates durations automatically based on syntactic and pragmatic information, this capability is still very much a rarity.

There are essentially two ways in which speech rate can be manipulated. The most obvious is to change the phonetic durations of speech segments across some domain, such as an utterance or a paragraph. This involves applying rules which directly alter the absolute durations of speech segments, and is the approach taken by most TTS systems which control speech rate at all. The alternative, rather less obvious approach to controlling speech rate is to modify the phonological representation BEFORE any phonetic duration rules apply. The types of modification which might be considered include changes in boundary location at various levels (e.g. syllable, word, phrase, and utterance boundaries) and changes in the prominence of particular elements as the result of stress or accent assignments. These will obviously have effects on the absolute durations of speech segments, given the nature of duration rules, but they will also affect the **perceived speech rate**. Perceived Speech Rate differs from Absolute Speech Rate in that it is determined by the hearer's perceptions rather than by the absolute physical characteristics of the speech waveform.

We take the view here that perceived rate is similar to perceived pitch or phoneme in that speakers' expectations influence their perceptions. It has been shown (e.g. Pierrehumbert 1979) that speakers judge pitch with reference to the phonological structure of the utterance as well as to the absolute frequencies involved: similarly, the phenomenon of vowel intrinsic pitch demonstrates the ability of listeners to normalise for such factors. The well-known tendency of listeners to hear what they expect regardless of the phonemes which were actually produced (examples such as *We can recognise speech* versus *We can wreck a nice beach* are well known) is a further instance of linguistic factors (in this case pragmatics) overriding acoustic information, and Summerfield & Haggard (1972) and Klatt & Cooper (1975) demonstrate that listeners can apply the technique of normalisation to judgements of speech rate. More recently, Gussenhoven & Rietveld (1987) showed that these judgements are also sensitive to prosodic structure: they found (p.283) that different prosodic structures could cause differences in perceived rate of around three per cent, and concluded that such differences "may be explained by the implied presence or absence of intonation phrase boundaries." (p.273) Several other authors (e.g. Schmerling 1976:89ff.; Nespor & Vogel 1982:254) have also pointed out the dependency between speech rate and prosodic structure. The placement of prosodic prominences and boundaries in the CSTR TTS system is largely controlled by the Rhythm Rule, and this would therefore appear to be the appropriate place for changes in Perceived Speech Rate to be implemented.

2.8.2 Controlling Absolute Speech Rate

The duration rules implemented in the CSTR TTS system were developed by Nick Campbell and are based on a syllable-level model of durations (Campbell 1989, 1990). The rules were derived by running neural nets over a prosodically-transcribed corpus of English read speech, and are therefore sensitive to prosodic markings such as accents and boundaries. However, this sensitivity is not a peculiarity of our system: other corpus-based approaches to synthesising duration also take account of prosodic structure (e.g. Larreur et al. 1989), and non-corpus-based rules such as Klatt's (1979) and Witten's (1977) are sensitive to suprasegmental features. The CSTR rules have an explicit rate parameter which allows the base syllable duration to be altered linearly, but we are well aware that this is not a true reflection of rate differences in human speech. Unfortunately, no good metric of natural rate variation is available (Campbell 1987), and we have therefore opted for this crude approximation.

The prosodic specification of an utterance drives the duration rules, by specifying accents and boundaries, but it does not supply the value for the rate parameter. Rather, the default rate value has been chosen to match the default accentuation strategy. Thus, duration is directly dependent on prosodic information in the default case. For a particular utterance, text or application, however, the rate parameter can be changed to mimic a different style of delivery, giving a dimension of control over Absolute Speech Rate which is fully independent of the prosodic specification.

The prosodic specification provides our syllable-based duration rules with the six distinct syllable types shown in Table 4 above. Additional information on prosodic boundaries, syntactic classes and syllable structure is also available to the duration rules. On this basis, a highly natural durational specification can be produced which listeners find hard to distinguish from human durational patterns (Monaghan 1991a).

2.8.3 Controlling Perceived Speech Rate

Listeners' perceptions of rate are at least partly dependent on their expectations of the relation between prosodic specification and Absolute Speech Rate. It follows, then, that modifications to prosodic specification (more accents, fewer boundaries, etc.) will alter the Perceived Speech Rate of an utterance. The two main aspects of prosodic structure in the CSTR TTS system, domains and accents, have been presented above: the former are determined largely on the basis of the syntactic analysis, and the latter are determined by the Rhythm Rule. Their respective influences on Perceived Speech Rate are discussed in the remainder of this section.

2.8.3.1 Domains

Even TTS systems with sophisticated prosodic rules such as PROS (Quené & Kager 1989) generally make the same type of implicit and inflexible assumptions about domain length as Klatt's rules make about segmental durations. In PROS and many other systems, there are absolute constraints on the size of prosodic domains: the assumption is that such constraints depend on the number of elements within each domain on the segmental level (words, syllables or segments), and that restructuring is required when there is a large variation between the number of such elements within a particular domain of type X and the number of such elements within other domains of the same type. This seems unnecessarily simplistic and indeed inappropriate in many cases: firstly, since the perception of domain size is determined by prosodic rather than segmental characteristics, it seems obvious that any constraints should be framed in prosodic terms; secondly, if any flexibility in speech rate is to be allowed, these absolute constraints must be made relative and sensitive to some rate parameter. Our system of domains takes account of both these points, and consequently produces more appropriate and flexible output.

There are three levels of domain in our model, as described above: tg(0), or minor phrase; tg(1), or major phrase; and tg(2), or utterance. The default definitions of these domains are that a tg(0) corresponds to a subject NP or a VP, a tg(1) to a full clause, and a tg(2) to a text sentence. However, these domains can be easily redefined to produce different prosodic structures for the same input text: redefining tg(0) to be a subordinate clause and tg(1) to be a main clause would give a much sparser structure, corresponding to a faster speech rate; redefining tg(0) to be one major lexical item or one accent, tg(1) to be an NP or VP, and tg(2) to be a clause would give a denser structure of the sort associated with a slower speech rate where each sentence effectively constituted a paragraph. Such redefinitions would thus produce quite different domain specifications, generating an impression of more or less careful, reduced or information-rich speech and thus changing the listener's perception of the speech rate. This is only the tip of the iceberg, however: the resultant domains form the input to the Rhythm Rule, and the accentuation which this rule assigns is an even more flexible source of modifications to the Perceived Speech Rate.

2.8.3.2 Rhythm Rule

As stated above, the Rhythm Rule processes one tg(0) domain at a time, regardless of how such a domain is defined, and applies its rules purely on the basis of the accents which have been assigned to that domain (i.e. regardless of syntactic or semantic information). The various stages of our Rhythm Rule produce successively more reduced prosody. The first stage, which produces the **fully-accented** representation, is currently obligatory

Input Text:						
	The	man	with	the	red	tie
Accent Assignment:						
(Over-Accented Representation)	-	1	-	-	2	1
Clause 1: Delete all accents to the right of the rightmost primary.						
	-	1	-	-	2	1
Clause 2: Reduce all primaries except the rightmost to secondaries.						
(Fully-Accented Representation)	-	2	-	-	2	1
Subsequent clauses are OPTIONAL, depending on speech style						
Clause 3: Delete every odd-numbered secondary leftwards from the primary.						
	-	2	-	-	d	1
Clause 4: Apply WFCs.						
(Rhythmic Representation)	-	2	-	-	d	1
	the	MAN	with	the	red	TIE

Figure 14: Incremental Operation of the Rhythm Rule

and provides the basis from which various degrees of reduction can be produced. All subsequent stages are optional: the choice of which additional stages to apply is dependent upon the speech rate and style to be synthesised.

To illustrate precisely how our Rhythm Rule can produce different degrees of prosodic reduction, its application to

(28) The man with the red tie

is shown in Figure 14. All intonational diacritics are associated with lexical stresses: primary and secondary accent are represented by 1 and 2 respectively; deleted accent is represented by d; hyphens represent function words which have not been accented. Capitalisation indicates accented syllables.

The input to the Rhythm Rule is the Over-Accented Representation produced by the CW/FW accent-assignment heuristic. The first stage produces what we term the Fully-

Accented Representation, from which various degrees of reduction can be produced. This stage is obligatory for fluent speech[5]: all subsequent stages are fully optional. There are two processes in the first stage of the rule: the determination of the nucleus, and the reduction of non-nuclear primary accents. These are largely interdependent in a tg(0) domain, since such a domain is defined as having only one primary accent. The example in Figure 14 illustrates an important fact about the Rhythm Rule, which is that some processes may apply without changing the representation: the rightmost accent is already a primary accent in this case, so Clause (1) effects no change. There is therefore a degree of ambiguity in the output. Clause (2), however, makes it clear that both *man* and *tie* belong to the same tg(0) domain by demoting the primary accent which *man* would otherwise retain: thus the granularity of domains, one of the cues to Perceived Speech Rate, is made explicit at this stage.

If a more reduced form of the prosody is required, the second stage of the rule can now be applied. Rhythmic alternation is introduced by Clause (3), which finds the nucleus (rightmost accent) and works back from it, deleting alternate secondary accents starting with the first. The output of Clause (3), in this case the sequence [- 2 - - d 1], is now considerably reduced from the input to the rule. This sequence is passed to the final clause. In some cases there are still prosodic representations at this stage which require modification by the Well-Formedness Conditions (WFCs) discussed above: however, in this example none of these WFCs applies as this is already a highly-natural accentuation for this phrase. It only remains to map the final rhythmic representation onto the stressed syllables of the utterance to produce the desired result:

(29) the MAN with the red TIE

[5] Even this stage may be omitted if a non-fluent, word-by-word delivery is desired, as might be appropriate in a proof-reading system, for instance. However, the same effect might be better achieved by reducing the size of tg(0) domains, as mentioned above.

2.8.4 Discussion

The CSTR TTS system is able to manipulate speech rate in two distinct ways. Firstly, by modifying a parameter at the level of the phonetic duration rules, it can exercise control over the absolute lengths of speech segments in the domain of those rules. Secondly, by changing the prosodic specification of accents and boundaries, using the Rhythm Rule as discussed above, it can alter the input to the phonetic duration rules: this will affect the Absolute Speech Rate, since the phonetic rules are sensitive to differences in prosodic specifications, but more importantly it will also condition listeners' expectations and thus affect the Perceived Speech Rate. Although the effects of these two manipulations must obviously interact, their control is completely independent. The question of mutual constraints, i.e. the appropriateness of certain prosodic specifications for certain Absolute Speech Rates, is still to be investigated.

More problematic are the questions of predicting speech rate and choosing how to manipulate it. The former question is part of the larger problem of understanding text: to predict changes in speech rate during the realisation of a particular text, we must first disentangle the pragmatic and attitudinal information which is to be conveyed, and this is far beyond the capabilities of any current TTS system. The latter question lends itself more easily to investigation, as it addresses the problem of how best to realise rate changes once they have been predicted: does the phonetic parameter control rate changes within texts, or only between texts, or only between speakers; do certain text styles require different prosodic specifications, or is the prosodic structure entirely at the whim of the speaker? Our intuitions suggest that the phonetic parameter will be best suited to modelling inter-speaker differences, and that text-internal changes will be most effectively modelled by prosodic restructuring on the basis of pragmatic information, but the results of empirical investigations are clearly required before this question can be resolved.

2.9 Summary

2.9.1 Domains

The principled division of the syntactic analysis into appropriate phonological domains is essential to the accurate operation of both accent assignment rules and phonological structure-building heuristics. The former determine the relative prominence of lexical items and of syllables within those items, and are designed to apply to regular domains. The latter determine the perceived hierarchic prosodic relations (subordination, conjunction, etc.[6]) between domains. Together these two rule sets define the abstract F_0 contour in all its important aspects.

The identification and classification of phonological domains based on syntactic structure allows our system to handle long stretches of text in a highly naturalistic manner without recourse to hierarchic structure-building analyses and using only three levels of domain. Our flat left-to-right approach is fast and efficient, yet it appears to capture most of the information which more complicated analyses could provide. The current INTERFIX program will process any text sentence into a structured series of phonological domains which conforms to naturalistic constraints and which reflects a large proportion of the semantic and pragmatic organisation deducible from text. Our emphasis on left-to-right symbolic processing and our rejection of the more traditional time-dependent approaches to phrasal F_0 effects (so-called "declination") is supported by the quality and flexibility of our system's output: each domain is processed in isolation by the lower-level rules, and the resultant strings are concatenated before being fed into the phonetic model, which allows register steps and boundary tones to be easily inserted between domains rather than attempting to map them onto the entire string. We therefore believe that it is well worth pursuing this more abstract, phonological approach.

There are still many exceptions to our phonological default treatment which INTERFIX cannot recognise or treat appropriately, and most of these will probably never be

[6] These relations should not be confused with syntactic or semantic uses of these terms.

handled by such a program. However, the system of domains presented above provides a consistent and flexible framework for handling several phenomena. Examples of problems which we feel the framework provided by INTERFIX will help to solve are the tg(2) boundaries at colons and similar punctuation marks which we mentioned above, the problem of commas which mark a double boundary (e.g. *Bob, who won, incidentally, . . .* where the last comma ends both parentheticals and must therefore be interpreted as such), and the accenting of stranded or transposed elements such as negative markers, sentence-final prepositions and auxiliaries: some of these are investigated further in subsequent chapters. More ambitious areas for future work might include investigating the interaction of semantic and grammatical information (e.g. the relation between animacy and grammatical function discussed by Faber (1987)) in determining intonation, and the as yet untouched area of suprasentential relations and their effects on F_0.

2.9.2 Accents

In the absence of adequate automatic semantic and pragmatic analyses, there is a need for heuristic rhythm rules in text-to-speech synthesis if we are to produce acceptable intonation contours. Our Rhythm Rule presented above appears to satisfy most of the requirements for such a rule: it is simple to implement, allows the incorporation of lexical, syntactic and phonological information, and (with appropriate treatment of a small number of regular exceptions) produces acceptable output for most unmarked sentences.

The interpretation of secondary stress required by our Rhythm Rule may seem to pose problems for vowel-reduction routines in speech-output systems: however, the benefits of the syllable-type distinctions available in the output of the Rhythm Rule for vowel quality, duration and other rules more than compensate for this. We argue above that at least some vowel-reduction rules should be applied AFTER the Rhythm Rule, and for those aspects of vowel reduction which are usually related to lexical stress information we suggest two alternative metrics: morphological structure and syllable structure.

Fine durational, segmental and rhythmic distinctions CAN be drawn without recourse to time-functions or metrical structure. The simple rules and heuristics discussed above are all used in the CSTR TTS system to do precisely that. Although these procedures are still relatively crude, we believe that they can be refined considerably by further experimentation and that they could form the basis for a very sensitive prosodic component to be used in future text-to-speech research.

2.9.3 The Intonation Model

The implications of the work presented in this chapter for models of intonation in TTS systems are considerable. The model which our rules imply agrees in many respects with the findings of theoretical and experimental work on intonation which is not generally taken into account in TTS implementations. There are three main points which merit discussion at this stage, and which (when added to our assumptions of a target-and-transition model and of a specification solely in terms of accents and boundaries) define quite closely the theoretical view of intonation which underlies our rules.

Firstly, in accordance with the view which we stated in Chapter 1, we believe that domain assignment is determined by the interaction of semantic constituency and focus structure. This view is very similar to that expressed in Gussenhoven's SAAR discussed above (p.27), and indeed in our opinion the SAAR remains the best characterisation of the basic factors underlying domain assignment in natural speech. However, neither semantic constituency nor focus structure is available to the intonation rules of a TTS system and the standard approach to generating intonation in such systems is therefore to ignore such factors and concentrate on more readily-available information such as syntax. We have taken a rather different approach in developing the rules described above: although we have perforce relied on lexical and syntactic information alone in assigning domains, there are two crucial assumptions underlying our rules which link them closely to the theoretical position expressed in the SAAR. The first of these is the assumption that, in the majority of cases, there is a close relation between major syntactic constituents and the semantic constituents to which Gussenhoven's rule refers: NPs usually correspond to Gussenhoven's **arguments**, intransitive VPs correspond to his **predicates**, and so forth. The second assumption which we have made is that the input to our rules is entirely [+focus] in Gussenhoven's terms, i.e. that no element or constituent is defocussed, such that a Broad Focus treatment is appropriate in all cases. Both these assumptions are, of course, only justified for some proportion of cases: there are bound to be exceptions, and the frequency of such exceptions will depend on the style, subject matter and complexity of the particular textual materials involved. Consequently, although these assumptions are crucial to our approach in the default case, any information which reliably indicates exceptions to our assumptions should override the default treatment. Several such exceptional cases, and our attempts to provide an appropriate treatment of them, are discussed in the next chapter.

Secondly, our Rhythm Rule clearly lays a great deal of importance upon the location of the nucleus in any tg(0) domain. The various clauses of the rule essentially identify the nuclear accent and then adapt the accent pattern of the entire domain to suit the location of the nucleus. The prime importance of the nucleus is not a new concept: the traditional British approach to intonation, with notions such as Halliday's (1967b:22ff.) "tonicity", has always given pride of place to the nuclear accent; Bing (1979a:25) ascribes most of the communicative effect of intonation to the shape and location of the nucleus; Selkirk (1981:386) sees the location of the nuclear accent as determining Normal Stress; and indeed the concentration of adherents of the syntactic approach on the nucleus alone, criticised by Fuchs (1984:136), is demonstrated by their use of capitalisation to distinguish a single point of prominence per sentence.

The primacy of the nucleus is therefore widely acknowledged, and we subscribe whole-heartedly to this view whilst refining it in certain minor respects. We assign a nucleus to each tg(0) domain, rather than to a clause or a sentence, and we allow other phonological conventions such as boundary tones and the tendency towards rightmost prominence to determine which of these tg(0) nuclei will function as the nucleus of the entire utterance. We also avoid Fuchs' quite justified criticisms by assigning non-nuclear or **head** accents where possible. This approach seems to us to combine the best of all previous authors' observations.

Finally, given the prime importance of the nucleus for both the naturalness and the interpretation of intonation contours, the locations of head accents must by definition be less crucial. Although various suggestions have been made in the linguistic literature as to how the placement of non-nuclear accents is determined (such as Bolinger's (1989:238ff.) claim that such accents simply reflect a degree of speaker interest slightly below that conveyed by the nucleus, or Ladd's (1980) brief observations on the relation between head accents and rhythmic structure), no clear theoretical explanation has been expounded. We take the view that, at least as far as listeners' perceptions of intonation are concerned, head accents are to a large extent optional and may be assigned relatively freely. The two main limitations on this freedom appear to be that head accents should fall on the lexically-stressed syllables of content words (rather than in some less neutral position) and that they should conform to some rhythmic pattern of alternations (avoiding "stress clashes" (Liberman & Prince 1977) and long stretches of unaccented speech). The failure of head accents to respect any other limitations, such as the **given/new** distinction or syntactic head-modifier relations, has been amply demonstrated (Schmerling 1976; Fuchs 1984; Boisson 1985; Terken & Nooteboom 1987; Nooteboom & Kruyt 1987). We have therefore based our rules on a model of intonation where, in full-focus domains, accent placement is determined by the position of the nucleus and by principles of rhythmic alternation only: the question of what determines the nucleus location is, however, far from trivial and is returned to below.

2.9.4 Conclusions

The rules presented in this chapter constitute a coherent, principled and flexible scheme for assigning prosodic specifications to unrestricted text based on minimal syntactic and lexical information. Our approach has been to incorporate as much linguistic knowledge as possible into the rules and to make use of heuristic generalisations in order to compensate for the paucity of the information currently deducible from text in an automatic system. This work has led us to two main conclusions.

Firstly, it is possible to produce a much richer and more reliable prosodic specification than most TTS systems attempt, despite the lamentable quality of automatic text analysis at the syntactic level and above. This bears out our view that ***linguistic competence or knowledge of the language***, rather than true understanding of the text, suffices to predict appropriate prosody in most cases: this is precisely the type of information which we believe many human speakers make use of when reading aloud from unfamiliar material, and this is what we have attempted to build into our rules.

Secondly, sparse prosodic specifications such as ours (in terms solely of accents and boundaries) nevertheless supply a great deal of information regarding the possible prosodic realisations of a particular sentence. This information can be used by phonetic realisation rules not just for F_0 but also for duration, intensity, vowel quality and other prosodic parameters. Moreover, this information can be controlled in such a way as to influence speakers' perceptions of the output on a stylistic level: absolute and impressionistic control of casualness, rate and emphasis are all available by varying the granularity of prosodic domains and the operation of our Rhythm Rule. This degree of control far surpasses that available in any other TTS system of which we are aware.

Chapter 3

Errors and Solutions

> What we do not know is how often in reading out text this neglect of semantic aspects of accentuation will result in unacceptable accentuation or de-accentuation.
> Nooteboom & Kruyt (1987:1521)

The rules described in the preceding chapter were developed on the basis of linguistic intuitions, informal tests, and comparisons with small corpora. We were well aware that they were inadequate in many cases, but we were equally surprised at the number of cases where they performed very well. The areas in which these rules were lacking were largely predictable: our impression was that errors occurred most frequently in complex syntactic constructions and in sentences with rich or unpredictable pragmatic contexts. On the basis of these subjective impressions, the next step would have been to examine these problematic areas in more depth, perhaps with reference to a small corpus of error-prone sentences. However, having constructed a robust and coherent set of rules, we decided that subjective judgements had served their purpose and it was now time for a formal evaluation of our prosodic output. It was hoped that this would provide us with both a clear indication of the general quality of our output and some pointers to particular shortcomings.

The present chapter presents this evaluation, the errors which it revealed, and the rules which were subsequently developed to provide some of the necessary solutions. The first section describes the methodology behind the evaluation experiment, the experiment itself, and its results; subsequent sections describe our approach to handling specific phenomena which proved to be problematic for our rules.

3.1 Evaluation

When the evaluation described in this section was planned, very little formal evaluation of synthetic intonation had been undertaken anywhere and even less had been documented: our methodology was therefore of necessity innovative in nature rather than being based on an established tradition of intonation evaluation. At the time of writing, it is still the case that there are no accepted standard methods or criteria for the assessment of synthetic prosody (Pols 1990:297; Terken & Collier 1990:205), although there have been many more published studies, some of which are discussed in Section 3.6. Our methodology was consequently developed from first principles specifically to meet our own short-term requirements, but we feel that it is both a useful means of comparison across TTS systems and a widely-applicable method of identifying levels of performance and problematic areas for a particular system.

In the absence of an established evaluation methodology for synthetic intonation, there were three major questions which required to be answered in order to design an appropriate procedure. Firstly, there was the question of what to evaluate. Previous evaluations of synthetic speech output (e.g. Luce et al. (1983), Pisoni (1987) and Pisoni et al. (1987)) paid scant attention to prosody, concentrating on segmental intelligibility as measured by rhyme tests such as the Modified Rhyme Test (MRT) (House et al. 1965) and the Diagnostic Rhyme Test (DRT) (Voiers 1983). Those researchers who did explicitly assess the quality of their synthetic prosody tended to limit such assessment to direct comparisons between natural and synthetic F_0 contours, either on a perceptual level (e.g. 't Hart (1979) and de Pijper (1983)) or on the basis of some purely physical distance metric (e.g. Choppy & Liénard (1977) and Wothke (1990)). However, neither of these methods was appropriate for our purposes, as we were concerned not with the closeness of the contours which our system produced to those of some pre-recorded utterance but rather with the appropriateness of the phonological specification (in terms of accents and boundaries) for generating an acceptable intonation contour. Moreover, we were concerned to measure our system not against random isolated utterances, as was the norm in previous studies, but against naturally-occurring unrestricted running text.

We therefore decided to evaluate the **acceptability** of our phonological specifications ***in a real context***, by eliciting native-speaker judgements of our system's output for multi-paragraph texts.

The second major question concerned the purpose of the evaluation: what did we hope to learn as a result of evaluating our output? Two distinct motivations for evaluating the performance of a TTS system can be distinguished in the literature. The first, a desire to discover the types of error to which a particular system is prone, requires **diagnostic** procedures: the second, a need to quantify the absolute or relative performance of a system (on some fixed scale, or as compared with the output of other systems), is **documentary** in nature. Diagnostic evaluation is a development tool, a means of identifying areas where a particular system needs improvement or where it is performing well enough to be left alone: tests such as the DRT are specifically designed to address these requirements at the level of segmental quality. Documentary evaluation aims rather at verifying a system's progress, and at comparing the achievements of system X with those of system Y or with the output of an earlier version of system X: techniques such as the use of "semantically unpredictable sentences" (Grice 1989; Hazan & Grice 1989) have recently been developed to assess segmental quality in this fashion. Documentary evaluation, by definition, is intended to prove a point, and therefore careful attention to the experimental methodology and to the analysis of the results is essential if the exercise is to serve the intended documentary purpose. Diagnostic evaluation, on the other hand, is answerable only to the evaluator: in many areas, especially those where there are clearly major deficiencies in a system's performance, the information which is required to continue the progress and development of the system can be obtained from relatively informal evaluation studies.

As stated above, the motivation for evaluating our automatic intonation was two-fold: we hoped to gain both a documentary indication of the general level of acceptability and some diagnostic pointers to areas of particular difficulty. Our need for a formal, rigorous evaluation was thus equalled by our desire for information to direct further development, and it was therefore decided that an objective, quantitative assessment methodology was required but that it should be one which would allow us to analyse particular errors as well as providing an overall measure of the system's performance. In order to conduct a scientifically rigorous evaluation without expending a disproportionate amount of effort, we decided to limit the experiment to a relatively small (1,000 words) corpus of running text: this was intended to provide a sufficiently large sample of the system's output for documentary purposes while still allowing scope for extensive manual analysis of errors.

Naturally-occurring unrestricted running text was chosen for the evaluation materials in order to present our system with a realistic task and also to reveal the largest number of diagnostic errors. As House (1987:134) points out, running text is the ultimate test of synthetic intonation: "For the synthesis of isolated sentences, patterns may be readily specified which are plausible", but larger stretches of coherent text present a much greater challenge.

Finally, the question of how to evaluate the system's output required to be addressed. There were two main factors which determined the answer to this question: the quality of the acoustic output from the CSTR TTS system, and the intermediate representations which the intonation rules produce. From a purely intuitive point of view, intonation — including at least F_0, but probably also some aspects of segmental timing and the distribution and duration of pauses — would appear to be intrinsically a part of acoustic output and therefore to be best evaluated on that basis. However, if we attempt to evaluate intonation in acoustic output, we immediately encounter serious problems of interpretation unless we can reliably control all the relevant segmental and supra-segmental factors in the acoustic realisation: objective judgements of prosody in poor-quality speech are notoriously difficult to make (Young & Fallside 1980; Terken & Lemeer 1988). The quality of segmental synthesis available at CSTR when this evaluation was undertaken was very poor, being limited to formant synthesis based on the Holmes RP allophones (Holmes et al. 1964): any assessment of automatic intonation in such output, even by trained listeners, was prone to overriding interference from the segmental quality. Moreover, the CSTR TTS system at that time exercised very little control over duration and none at all over amplitude, and our own experience in working with the system was that these other prosodic factors could completely mask the perceived course of F_0 and create quite spurious impressions of the underlying prosodic specification. It would therefore have been a very time-consuming and rather uncertain task either to abstract the intonation judgements away from the impression of overall synthesis quality or, alternatively, to produce sufficiently controlled acoustic stimuli based on a more high-quality synthesis method such as simple digital analysis/resynthesis[1]. In order to avoid these problems and any uncertainties which they might introduce in our results, we decided to evaluate the phonological prosodic specification at a **symbolic** level of representation rather than at the acoustic level: this allowed us to concentrate directly on the output of the intonation rules with no fear of interference from problems related to the acoustic realisation, but left us with the problem of deciding which symbolic representation should be evaluated.

3.1.1 Evaluating Symbolic Output

In our phonological model, and indeed in most current phonological models of synthetic intonation, three fairly distinct aspects of intonation can be identified as potential candidates for evaluation in symbolic output. These are:

1. Accent Placement: the location of major pitch prominences ("sentence accents", "sentence stresses"), signalling contrast, focus, backgrounding of **given** information, etc.

2. Domain Demarcation: the placement of boundary tones and register shifts which signal semantic and discourse-level organisation. This includes such effects as greater pitch range at the beginning of paragraphs, the proper location of pauses to distinguish possible readings of sentences, the resolution of attachment ambiguities, etc.

3. Tune Choice: the selection of rise, fall, fall-rise, etc. as the realisation of the accents referred to in point [1], together with the choice of pitch movements at boundaries (point [2]). This choice is most directly related to sentence-type or

[1] See Section 3.6 for some further discussion of the problems which attend the combination of natural segmentals with synthetic prosody.

> speech acts — question, statement, command, etc. — although there are attitudinal effects of tune choice that are more difficult to define with any degree of precision.

Of the various levels of representation available in the output of our system, the **rhythmic representation** produced by the Rhythm Rule was chosen as the most appropriate for evaluation, on the grounds that it is the most unambiguous and easily-interpreted representation. This representation gives degrees of accent within each tg(0) domain, thus providing much more information than a simple accented/unaccented contrast. In addition, the interaction in our model between domain boundaries and degrees of accent means that the acceptability of the latter depends to a large extent on the assignment of the former: judgements of accent-placement decisions therefore implicitly judge the domain-assignment rules. The rhythmic representation thus allowed us to combine aspects [1] and [2] above. The question of Tune Choice, although represented symbolically in the phonetic model at the level of high and low tones, was decided to be both peripheral to the phonological representation and insufficiently variable in that the system made use of only the default tune (for declaratives and WH-questions) when in fully-automatic mode.[2]

Evaluation of a fourth area, namely the accuracy of the phonetic model by which the phonological intentions under points [1]-[3] are realised, must necessarily take account of acoustic output. This includes, but is not limited to, problems as diverse as: the slope of rapid pitch changes; the alignment of F_0 turning points relative to segmentals; the overall range within which pitch movements take place, and the rate and direction of such changes within that overall range; and modifications in the durations of segments associated with major intonational prominences or prosodic boundaries. However, although our system makes use of the particular phonetic model described in Chapter 1 it was not our intention to evaluate this model: the present evaluation was concerned exclusively with the phonological representations which our rules produced.

[2] In actual fact, there were no instances of sentences requiring the choice of any different tune in the evaluation materials.

Plainly, any evaluation of synthetic intonation at a symbolic level is not going to be as straightforward as a symbolic assessment of processes such as word-level grapheme-to-phoneme transcription or lexical stress assignment rules. We have stated elsewhere (Monaghan & Ladd 1990b:306) that the most important prerequisite for the evaluation of symbolic output is the existence of categorically correct answers (e.g. orthographic *guilt* transcribed as [g i l t] rather than [g w i l t], or *reform* stressed on the second syllable rather than on the first): the main difficulty in evaluating symbolic representations of intonation is their failure to satisfy this prerequisite. This failure comes about in three interrelated ways:

1. There is more than one correct answer. For any given word string, there are likely to be several symbolically distinct intonation patterns that could all occur in natural speech and would all convey the same syntactic structure, discourse situation, semantic meaning, etc. (Choppy 1979:186; Brazil 1984:46).

2. There is no categorical division between correct and incorrect intonation patterns. A great deal has been written about the systematic exploitation by speakers of gradient variability in intonational meaning for particular communicative effects (Bolinger 1972a, 1986; Ladd 1980; Uldall 1960, 1964). In addition, as was pointed out in Chapter 1, human beings will go to considerable lengths in order to construct a context which allows any given intonation to be judged appropriate.

3. There is no agreed symbolic representation of intonational phonology, which is perhaps not surprising given the plethora of different views and models of intonation discussed in Chapter 1. However, the bulk of this disagreement was hardly touched on above, centring as it does on the question of what constitutes intonation and primarily on the various different theoretical and descriptive resolutions of the relationship between categorical and gradient phenomena.[3] This creates difficulties in obtaining consistent and objective interpretations of intonational symbols and structures: analogous disagreements over the symbolic representation of segmental phonology are almost non-existent.

[3] For fuller discussion, see Ladd (1980), Bolinger (1983, 1985), and Brazil et al. (1980).

These problems should not be exaggerated, however. In the first place, points [1] and [2] pose no particular difficulties for us: as stated in Chapter 1, we are not aiming to produce the one-and-only "correct" representation, but merely one of the many acceptable ones; moreover, if humans are accustomed to presuming intonation contours to be appropriate until proved otherwise (point [2]), we deem it perfectly reasonable for our rules to take advantage of this fact. However, there are obviously varying degrees of appropriateness, particularly given an extensive context: it was therefore decided that a scale of **acceptability** would be more appropriate than the usual binary (correct/incorrect) choice. Point [3] unfortunately rules out the use of large numbers of naïve subjects in symbolic evaluations of intonation: however, it does not in any way preclude evaluation by trained phoneticians or phonologists provided that the intended interpretation of the symbols assessed is made very clear to the judges.

Accent placement seems particularly amenable to evaluation at the symbolic level by trained judges, for two reasons. First, accent placement can be indicated simply and for the most part unambiguously, without reference to acoustic properties. For example, the difference between the two versions of speaker B's response in the following two dialogues can be marked by superscript digits to show the location of primary and secondary accents.

2 1

(30a) A: I haven't seen you around much lately. B: I've been to Austria.

1

(30b) A: Fancy a holiday in Austria this summer? B: I've been to Austria.

Second, the correctness or appropriateness of accent placement is, compared with other aspects of intonation, quite determinate. It is indisputable that swapping the accent placement patterns in the two dialogues just given would render speaker B's responses decidedly inappropriate and perhaps even uninterpretable. This is in marked contrast to the difficulty of assessing the appropriateness of various more "phonetic" aspects of intonation such as F_0 peak alignment or rate of F_0 change.

In our opinion, then, the only important modifications to typical evaluation procedures that were needed in assessing the output of our accent assignment rules were: (1) the use of expert judges to assess the symbolic output, and (2) the replacement of the inappropriate "correct/incorrect" dichotomy with an acceptability scale. These modifications were introduced in our evaluation experiment to assess the performance of our system's accent placement rules, which is described in the rest of this section.

3.1.2 Evaluating Accent Placement

To produce a corpus of synthetic accent patterns for the evaluation of accent placement, we applied the rules described in the previous chapter to hand-generated, crude phrase-structure syntax analyses of four selected texts of approximately 250 words each. The syntactic analyses were produced by hand to avoid any confusion between errors in the intonation rules and errors in the CSTR TTS system's syntax module, and we ensured that these analyses provided no more detailed information than the rough word-classes and major phrase boundaries which our automatic parser generated. The four texts whose treatment we evaluated were transcripts of radio broadcasts (two news broadcasts and two Open University broadcasts), part of the Spoken English Corpus (Williams & Alderson 1986:6-7) kindly made available to us by Dr. Briony Williams, a CSTR colleague working on a speech recognition project.

The resulting accent patterns were rated for appropriateness by three expert judges: the present author (AM) together with Dr. D. R. Ladd (DRL) and Dr. Williams (BJW). Dr. Ladd was well-acquainted with the phonetic and phonological models involved, having been instrumental in the development of both: Dr. Williams was included in the study as a check on the possible biases of the other two judges, as she was for several years associated with the development of intonation in TTS systems at IBM(UK) and her ideas on synthesising — and even transcribing — intonation diverge from theirs in numerous respects.

3.1.2.1 Procedure

It was necessary to select a unit of text within which to evaluate accent patterns. It is not sensible to evaluate each word to see if it has an appropriate level of accent, nor to evaluate each accent to see if it is placed on an appropriate word: there are two main reasons for this. First, accent is inherently relational: the appropriate level of accent on a given word must be defined relative to that on some other word or words, and cannot be judged in isolation. This characteristic of accent is expressed in our system by various accent adjustment rules such as the Rhythm Rule, where an accent is adjusted relative to some neighbouring accent(s). The second reason why we cannot evaluate on a word-by-word or accent-by-accent basis is that the operation of the accent adjustment rules will propagate errors throughout a tg(0) domain. Thus, in an NP such as *the black and white cow* the principle of alternation embodied in our Rhythm Rule will give us *the BLACK and white COW* but the addition of one word can turn this highly acceptable accentuation into a most unnatural one such as *the black and WHITE striped COW*. Given all these considerations, we took as the basis of our evaluations ***the overall appropriateness of the accent pattern within each tg(0)***. Each tg(0) unit was assigned one of four ratings:

4: entirely acceptable; possible as a realisation in natural speech

3: quite acceptable, but with certain unnatural features

2: only marginally acceptable, with seriously unnatural features

1: entirely unacceptable; likely to lead to misinterpretation or uninterpretability.

Under this procedure, any major error that is propagated throughout a tg(0) would receive only a single 1 rating rather than a whole string of 1 ratings for each individual inappropriately placed accent; conversely, a major error that, because of interaction with another major error, FAILED to be propagated throughout a tg(0) would be judged just as badly as one that was so propagated. Similarly, a tg(0) domain containing a string of appropriately accented words would receive only a single 4 rating, not one for each word or accent in the domain. We felt that this procedure would not give undue weight either to the system's failures or to its successes. The accent patterns and tg(0) boundaries are shown in Appendix A: the accent-placement rules are summarised in Appendix C.

It should be noted that although each tg(0) was assessed for appropriateness in its own right, that appropriateness was judged on the basis of the full textual context of the particular tg(0). This meant that, rather than adjudging any conceivable accent pattern as acceptable, the judges insisted on an appropriate accentuation for the content in context. This is the most stringent criterion for running text, since it demands pragmatic and semantic appropriateness in addition to syntactic and lexical criteria.

The intention in using a scale without a middle point was to force the judges to choose between a broadly positive rating (3 or 4) and a broadly negative one (2 or 1). This understanding of the rating scale was made clear to the judges in advance.

Before rating the four texts to be evaluated, the three judges first worked through a fifth text together, in order to be sure that they had a common understanding of the assessment task and were in rough agreement about the use of the rating scale. As a further precaution, in order to check that the rules were performing according to our intentions and that there were no errors in either the program or the syntactic parses, all five texts were assigned accents and boundaries manually by the present author in a simulation of the automatic rules: an accent-by-accent correlation of better than 99% was achieved between this manual treatment and the output of the program. The automatic accentuations of the four texts that form the basis of the actual evaluation were then rated without consultation among the judges: some discussion of inter-judge agreement in the results is presented below.

3.1.2.2 Results

On the basis of the judges' ratings, each tg(0) in the four texts was assigned a score from 1 (the worst possible rating) to 10 (the best). This was done by summing the 3 scores assigned by the 3 judges and subtracting 2 from the total. (Since the individual judges gave ratings from 1 to 4, the raw summed scores range from 3 to 12; by subtracting 2 we arrive at the more manageable 1-10 scale.) Figure 15 gives a histogram of the distribution of tg(0) scores on this scale.

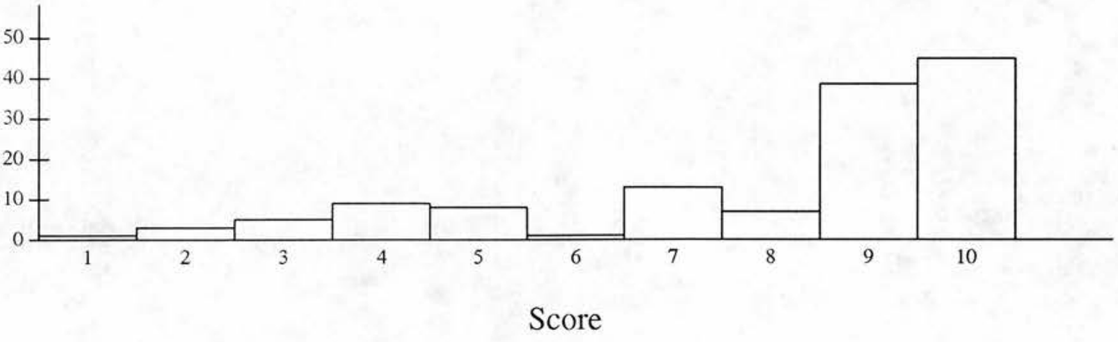

Figure 15: Distribution of Scores

The marked positive skewing in Figure 15 shows that by and large the accent patterns were evaluated favourably, with very few ratings in the lower half of the scale. However, it is difficult to draw more concrete conclusions without setting a cutoff level on the 10-point scale, and because of the aggregate nature of the scores any such point will be arbitrary to some extent. Units with scores of 9 or 10 can definitely be counted a successes for the system, since they can only result from uniformly positive ratings (3-4-4 or 4-4-4), but this does not allow for those units which were assigned 4-3-3 or 3-3-3, both uniformly positive. However, if we set the cutoff any lower, we might well include units that received at least one negative rating. Pooling the ratings in these cases does obscure the difference between units that had only positive ratings (3 or 4) and those that had mixed ratings — e.g. scale point 7 could indicate judgements of 3-3-3 or 2-3-4 or even 1-4-4. Of the total of 114 tg(0) units in the 4 texts evaluated, 67 (or 59%) had scores 9 or 10. Closer examination of the raw scores assigned by individual judges revealed that a further 11 units which scored 7 or 8 in total had been assigned the uniformly positive scores 4-3-3 and 3-3-3, and if these 11 instances are included this gives 78 (or 68%) positive results. By contrast, only 4 of the total of 114 units had scores 1 or 2 on the 10-point scale (indicating uniformly negative ratings).

3.1.2.3 Inter-Judge Agreement

Pooling the ratings gives us a graphic impression of the system's performance which is easily understood, as presented in Figure 15. However, pooling is justified only if the judges are in substantial agreement. If there were many disagreements among the judges, there would be many tg(0) units with intermediate scores; moreover, these intermediate scores would be ambiguous, indicating either units on which the judges had agreed on an intermediate score, or units on which the ratings had been very different. The fact that there are not many intermediate scores in the results shown in Figure 15 suggests a fairly substantial level of agreement, but it is nonetheless appropriate to take a closer look at the question of inter-judge agreement.

In order to determine whether it was appropriate simply to pool the three judges' ratings, two approaches to assessing inter-judge agreement were applied. The first was to compare the three judges with respect to their use of the four rating categories. The second was to look at the agreement for each tg(0) unit, computing the number of such units on which there was complete agreement, partial agreement, etc. The results of these two measures are reported in the next two paragraphs.

The judges' use of the four rating categories is shown graphically in Figure 16. All three judges agree in having a positively-skewed distribution: however, the skewing of AM's judgements is more strongly positive than DRL's, while BJW has a bimodal distribution with more 1-ratings than 2-ratings. A more meaningful or detailed statistical comparison of these distributions would have been difficult to extract: a second measure of agreement was therefore investigated, namely comparing ratings for particular tg(0)s.

In assessing the agreement on ratings for the individual tg(0) units, four degrees of inter-judge agreement were defined. These were:

a) Complete agreement: The three judges' ratings were identical.

b) Partial agreement: The three judges' ratings were all either broadly positive (3 or 4) or broadly negative (2 or 1).

c) Partial disagreement: None of the judges considered the tg(0) either entirely acceptable (4) or entirely unacceptable (1), but two rated it positive (3) and one negative (2), or vice-versa.

d) Complete disagreement: Any other combination of three ratings (e.g. 4-4-1, 4-1-1, 3-2-1).

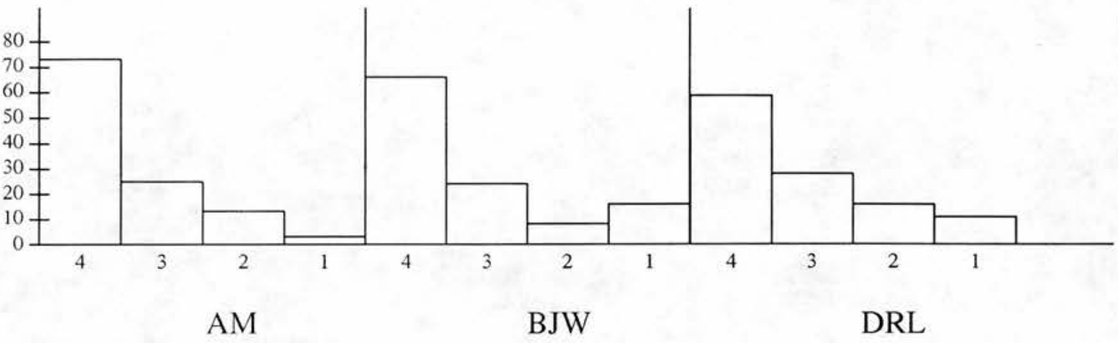

Figure 16: Distribution of Judges' Ratings

Of the 114 tg(0) units, there was complete agreement on 54, or just under half, and partial agreement as just defined on a further 33. This means that for 87 units (or 76%) the judges agreed on a classification broadly positive vs. broadly negative. There were only 5 cases of partial disagreement, but 22 of complete disagreement. The number of cases of complete disagreement is interesting, and is illustrative of the special difficulties presented by synthetic intonation in the evaluation of TTS systems. Of the 22 cases, there were 9 in which the ratings were divided between entirely unacceptable (1) and broadly positive (3 or 4) — i.e. combinations of ratings like 4-1-1, 3-3-1, etc. Many of these seemed to involve errors of accent placement that newsreaders themselves commonly make, such as

(3) the NEPHew of miss world ORGaniser julia MORley

(4) the aSSEMbly has been effectively HIjacked by the unionist PARties

(5) RAIN in some southern AReas will clear aWAY

These cases clearly indicate a large degree of variation in the tolerances and expectations of individual judges. This is a serious problem for the precise interpretation of the results produced by evaluation studies, and the proportion of cases affected by such variation is likely to increase with increases in the number of judges or with the use of naïve subjects.

3.1.2.4 Analysis of Errors

The 36 (32%) units which did not receive broadly positive results, i.e. those which were scored 1 or 2 by at least one judge, were analysed to see if any pattern of errors could be discovered. We anticipated that some of the errors would be rectifiable on the basis of the limited syntactic and lexical information available to the intonation module at run time, but that the majority of errors would be the result of semantic and pragmatic phenomena whose proper treatment requires levels of analysis not currently available to automatic systems.

A thorough analysis of the problematic tone groups revealed a small number of recurrent causes for most of the errors, and a small number of errors whose cause or solution remains obscure. Of the recurrent causes, more were syntactic or lexical than we had anticipated but some were the result of higher-level (semantic and pragmatic) processes. From this analysis, it appears that many of these problems may admit of a relatively simple solution. Unfortunately, the anticipated errors resulting from the absence of semantic and pragmatic information were nonetheless in the majority.

The errors split loosely into two groups, "tractable" and "intractable". The tractable errors, accounting for ten of the problematic units, included one instance of a colon used to punctuate a clause break: we had not yet incorporated the treatment of colons into our domain rules, but this was a problem of which we were already aware (see Section 2.5 above). Other errors in this group involved failure to treat long prepositional phrases (PPs) as separate domains and the lack of rules for treating number strings (126, 23587, etc.) specially: both these problems are addressed below.

Twenty-six of the unsatisfactory units involved errors due primarily to semantic or pragmatic content. There were several cases of semantically empty (redundant or predictable) lexical items (verbs such as *say* and *do*; nouns such as *committee* and *meeting*), which were accented by our rules but should have been deaccented; two cases of contrastive intonation ('not an AUdi but a BRITish car'), which can be successfully assigned only on the basis of a true understanding of the text; and many instances of anaphora, such as

(31) here he reJECTed the leibnizian VIEW

(32) EDward CROzier, a former PERSonal assistant to the MORleys

(33) further RAIN is likely toMOrrow

In examples (31-33) the view, the Morleys and the rain respectively were all **given**, i.e. they had all been mentioned previously in the text, and should therefore not have received accents: unfortunately, there is currently no reliable way of identifying such items automatically. There were also some cases where the interdependence of such semantic and pragmatic factors was too complex for even a careful analysis to reveal which factor caused the particular problem. Such errors will clearly persist in TTS systems for the foreseeable future, i.e. until automatic understanding of text by machines becomes a reality: however, some short- and medium-term solutions to these types of error are presented in the following sections.

3.2 Prepositional Phrases

It appeared that several of the tg(0)s which were assigned unsatisfactory accentuation in the evaluation experiment would have been assigned more acceptable accent patterns if the final PP in the sentence had been treated as a separate domain. This is particularly clear where the tone group in question was unusually long. Two examples of sentence-final tg(0)s from this evaluation data (Appendix A) illustrate the need for PPs to form separate domains in these cases:

(34) has praised Mrs Thatcher for standing firm at the Anglo-Irish summit.

(35) has welcomed a report by an Australian Royal Commission on the effects of Britain's atomic bomb testing programme in the Australian desert in the fifties and early sixties.

The intonation rules were accordingly revised to incorporate an improved treatment of PPs, and it was hoped that this would improve the accent patterns assigned to all of the problematic tone groups where the sentence-final PP was the major cause of unnaturalness. Additional routines to identify sentence-final PPs and treat them as separate tg(0)s were implemented: a sentence-final PP was defined for this purpose as the material following and including the rightmost preposition in a sentence, except where there is an intonational boundary between that preposition and the end of the sentence. The evaluation data were then processed anew with all sentence-final PPs being treated as separate domains.

In the majority of cases, the changes which this treatment produced in the output improved its subjective acceptability: indeed, there were some very large improvements in accent placement. However, two unforeseen problems also emerged. The first of these appears to have a simple solution, and although the second is a more serious problem its very existence in many ways validates our approach to the assigning intonation from text.

3.2.1 Two Types of PP

There appear to be two distinct types of sentence-final PP in the data: the first consists of a preposition and some material which that preposition modifies, usually an NP argument; the second consists of a "stranded" preposition, one with no following argument, and it is this type which causes the first problem. If such a preposition forms an intonational domain on its own, that domain will receive no accent (prepositions are by default not accentable) and will thus be subsumed into a neighbouring domain: no change to the accent string will be effected. In order to assign appropriate accent strings to sentences which end in a preposition, such prepositions must be made **accentable**: this would allow the nucleus to fall on the preposition, as is the case in examples such as

(36) Why don't you kiss and make up?

(37) He just bundled me up and hurled me over!

(38) We rang your house, but you weren't in.

but it would also force the content word preceding the preposition to receive an accent, as it would then be the last content word in the penultimate domain.

The alternative is to mark the preposition as accentable but not assign it to a separate domain, and this allows us to account for a further set of data. The contrast between

(39) Before he left, he locked me up.

and

(40) Before he left, he locked the cat up.

is reasonably common, and seems to depend on the presence or absence of a lexical NP between the verb and the preposition/particle: thus, (41) patterns with (39).

(41) . . . he locked up.

If we assign a primary accent to prepositions in these examples we will always place the nucleus on the preposition, but if we assign a secondary accent to sentence-final prepositions the Rhythm Rule and the various WFCs in the accent placement rules will give us the correct nucleus placement in all these cases. Our rules have therefore been modified to do precisely this, and these modifications have been integrated into the current version of the CSTR TTS system.

The second, and much less tractable, problem was identified when it was noted that the revised treatment of PPs actually caused a significant decrease in acceptability for two tone groups whose treatment had been judged highly acceptable in the evaluation experiment. On closer examination, it was clear that the original assignment of such felicitous accentuations to these tone groups was purely fortuitous, and had resulted from a chance correspondence between the contextual or pragmatic status of certain lexical items and their treatment by the Rhythm Rule. This is of course the sort of circumstance which our heuristic approach is specifically designed to take advantage of, and consequently there was no reason why we would have drawn attention to these cases in our original analysis of the evaluation results. However, although the revised treatment of PPs has destroyed this chance correspondence, the fact that the only two cases of loss of performance are ones where a clear contextual reason (in one case an anaphor, in the other a semantically empty item) can be given is gratifying for two reasons. Firstly, it suggests that our new domain-assignment routines come closer to modelling the linguistic realities, since the element of chance is reduced: there have been no inexplicable deteriorations in performance as a result of these revisions. Secondly, it shows that the inclusion of information on semantic and contextual deaccenting, which we knew was already required to solve previously noted problems, would be sufficient to cope with these additional problematic cases.

3.2.2 Evaluation

As indicated above, our revised treatment of sentence-final PPs was tested by rerunning the intonation module over the four evaluation texts. The resultant changes in accentuation were examined by the present author, who found that all the problematic sentence-final PPs from the original evaluation now received a much improved accentuation In addition, the accentuations of all other sentence-final PPs were at least as good and in some cases better than previously (with the exception of the two cases noted above). However, since our new strategy had been devised specifically to handle these particular problematic cases, we felt that this was perhaps not a fair test of the applicability of this strategy to unrestricted text. We therefore ran our revised intonation rules over a fresh corpus of unrestricted running text to see how this treatment would fare in a more open test.

We chose a corpus of academic abstracts which had been collected in related work on developing a text-to-speech system for a specific application involving reading out such abstracts. The abstracts were all from published papers authored by members of CSTR, and all dealt with various aspects of speech technology, but they were not restricted in any other relevant way. The corpus comprised 1,200 words of text made up of 58 sentences. Both the version of our rules evaluated above and the revised version incorporating our rules for PPs were applied to a syntactic analysis of this corpus, which was hand-generated by the present author on the same basis as the syntactic analysis of the four texts discussed in Section 3.1 above. The resulting accent placements were compared, and the points where they differed were examined. With almost no exceptions, the output of our revised rules was judged on the basis of this informal examination to be preferable to the output of the previous version. A more formal evaluation was considered both premature and unnecessary at this stage.

3.3 Anomalies

In the evaluation experiment presented in Section 3.1, several of the errors were attributable to the inappropriate treatment of number strings by the intonation module: there appears to be a fairly rigid set of rules for the accentuation of these constructions in neutral speech, and our system simply did not include the appropriate rules. Fortunately, number strings form part of a group of constructions whose semantic structure can be deduced with some certainty and which are clearly marked in text. Real text contains a high proportion of character strings which are not normal words, and which have therefore been regarded as a problem for text-to-speech systems and are usually either ignored or converted into words by such systems (Booth 1987; Barber et al. 1988:967; Carlson et al. 1990:272; Schnabel & Roth 1990:121; Wothke 1990:221). These are all constructions containing characters other than lower-case letters, referred to as **anomalies** because of their failure to correspond to the lower-case alphabetic "norm" for most text. Such forms include dates (1/2/34, 1986, '87), number strings (123, 12.34, 12,345, 123456), times (12:34, 12.34pm), and various types of abbreviation (KGB, UNESCO, Ph.D): in the more sophisticated of current text-to-speech systems these constructions are generally identified by a preprocessor module which attempts to determine the nature of any anomaly. If this particular subset of anomalies were marked for special treatment by the intonation rules, the amount of information deducible about their structure and function might well be such that, far from presenting a problem for accent-placement rules, they could consistently be assigned a highly natural-sounding accentuation. To that end, a set of **accent grammars** describing the appropriate intonational treatment of all the classes of anomaly represented in our data were developed and tested.

There are several distinct classes of anomaly which differ from each other in their intonational behaviour. The present description only addresses five common types: years, times, dates, number strings and abbreviations. These include all the types which occurred in the evaluation data above. In the context of our intonation rules, the behaviour of each type can be described by answering two questions: what is the relation of this construction to a prosodic domain, and which items within it should receive accents if an optimal accentuation is to be produced? The meaning of these questions requires some explanation.

The accents which should be assigned to a construction are determined by the semantic and pragmatic functions of its constituents, as well as the predicted effect of rhythmic factors on those accents. In our model, accents are assigned to almost all content words and approximately half of these accents are then deleted by the Rhythm Rule. If a particular domain does not behave in accordance with our Rhythm Rule, the assignment of accents may need to be modified accordingly: there may be reasonable pragmatic or other grounds for this, as will be seen below.

A minimal prosodic domain is defined as the domain of operation of the Rhythm Rule. Within such a domain, the Rhythm Rule is sensitive to the differences in intonational behaviour between predicates (verbs and adjectives) and arguments (nouns and proper nouns) and allows effects such as stress-shift to be modelled. Stress shift or other rhythmic effects across domain boundaries are not permitted, since accents in one domain cannot influence accents in another. If the intonation of a particular constituent depends on a neighbouring constituent, the two constituents should therefore share a domain: if their behaviour is independent of each other, they should be in separate domains. As an example, the difference between the realisations of *FIFteen* in *there are FIFteen MEN in a RUGby team* and *the NUMber fifTEEN is an INteger* results from the prosodic domains involved. In the former, *fifteen* and *men* are in the same domain and so the accent on *men* shifts that on *fifteen*: in the latter, *fifteen* is immediately followed by a domain boundary and so there is no influence from subsequent accents. These two factors — domain boundaries and conformity to the Rhythm Rule — were used to determine the rules required in the accent grammars for the five types of anomaly discussed below.

3.3.1 Years

Years (written as four digits, or as two digits preceded by an apostrophe) are always accented on the first and last lexically-stressed syllables when pronounced in isolation:

only contrastive usage licenses accents on other syllables. Thus, (42) and (43) are the only acceptable non-contrastive accent assignments for these utterances.

(42) NINEteen eighty-NINE

(43) TEN sixty-SEven

However, following material can affect this accentuation by causing the deletion of the second accent, thus:

(44) the NINEteen eighty-nine FESTival

(45) the NINEteen eighty-NINE edinburgh FESTival

The contrast between (44) and (45) is a result of the principle of rhythmic alternation which prohibits accents on adjacent items in the same domain, and therefore years do not necessarily constitute domains in themselves. However, it does not appear to be possible to delete or even to shift the first accent on a year constituent in non-contrastive usage:

(46) $*$the FAmous nineteen eighty-NINE edinburgh FESTival

(47) $*$the FAmous nineTEEN eighty-NINE edinburgh FESTival

This suggests that such a constituent must start a domain, but that it may combine with following material (up to the next prosodic boundary). Since rhythmic deletion can apply to these domains, they must be processed by the Rhythm Rule: marking the unaccented items in the year constituent as pragmatically deaccented (which, arguably, is what they are) will allow the Rhythm Rule to apply correctly to mimic the observed behaviour of these domains. The details of the implementation of these ideas are discussed in Section 3.3.6.

3.3.2 Times

Time constructions (e.g. 22:10 and 5:55, pronounced as *twenty-two ten* and *five fifty-five* respectively) appear to behave in exactly the same manner as years:

(48) it's TEN forty-THREE

(49) the TWELVE thirty-two exPRESS

(50) *the FAmous four FIFty from PADDington

They can therefore be handled by the same rules, although different pronunciations (such as *oh five fifty-five*, or *five to six*) or the addition of specifiers such as *a.m.* and *p.m.* may require special treatment.

3.3.3 Dates

Constructions giving days, months and years can be pronounced in one of two ways: *1/2/34*, for instance, may be expanded to (51) or (52).

(51) the FIRST of the SEcond thirty-FOUR

(52) the FIRST of FEBruary nineteen thirty-FOUR

A mixture of these two realisations is also possible. All these forms appear to share the same accent pattern, which simplifies things greatly and avoids the need to choose one or the other. However, the accent pattern for the year as part of a date is not the same as that for a year alone: although *nineteen* also seems quite acceptable with an accent in many cases, this can lead to unnaturally over-accented intonation as in (53).

(53) *the THIRD of MAY NINEteen TWELVE

We therefore decided to assign accents to the first and last accentable items and to the month in these constructions: this seemed a reasonable compromise between avoiding over-accented output such as (53) and the risk of accenting too few items in examples such as (54).

(54) ?the TWENty-seventh of sepTEMber seventeen seventy-SEven

Dates appear to constitute domains in themselves, in that their accentuation is not affected by preceding or following constituents. Examples such as (55) are simply ungrammatical, as dates cannot normally function as premodifiers in English, and (56) shows that preceding accents need not affect these constructions.

(55) ∗The fourteenth of july seventeen eighty-nine events

(56) on the MORning of the FOURTH of juLY seventeen seventy-SIX

Dates of the form _12/9_, meaning the twelfth of September, behave in the same way as those specifying a particular year except that the accent which would have been assigned to the year is not assigned in these cases. (57) and (58) illustrate the behaviour of such forms as self-contained tg(0) domains, unaffected by the proximity of accents in neighbouring domains: (59) shows the impossibility of using these forms as premodifiers in English.

(57) the TWELFTH of the NINTH is a TUESday

(58) he's igNORing the FOURTH of juLY

(59) ∗The twenty-fifth of December celebrations.

3.3.4 Number Strings

For present purposes, a number string is any string of digits (interrupted only by commas and decimal points in appropriate places) which occurs in text and does not function as a date or similar construction. Number strings are expanded as sequences of cardinal numbers and the decimal point where appropriate.

In general, expanded number strings consist of items which can receive accents and items which cannot. The former include the "units" zero to nine and the "tens" ten to ninety; the latter include the words *hundred, thousand, million*, and so on. The existence of these two distinct classes of items is evident from examples such as those in (60):

(60a) FOUR thousand SIX hundred and EIGHT

(60b) THREE million SEven hundred thousand POUNDS

There is a clear tendency for accents to fall on "units" and "tens", and not on other items. Nor is this simply one possible strategy, as the examples in (61) show:

(61a) ∗FOUR THOUsand six HUNdred and EIGHT

(61b) ∗THREE MILLion seven HUNdred thousand POUNDS

It is clearly desirable, then, to distinguish between these two classes of items in assigning accents to number strings.[4] Accent patterns were previously generated by assigning accents to all the accentable items and then applying the Rhythm Rule from right to left to delete every second accent. This results in accentuations such as the following:

(62a) SIXteen thousand SEven hundred and twenty-EIGHT

(62b) TWO million seven hundred and SIXty-four thousand and TWO

(62c) TWO hundred and sixty-ONE million sixTEEN hundred and fifty-NINE

It should be obvious that (62c), and even (62b), would not be judged entirely natural: however, if instead of ignoring the unaccentable items we allow them to delete the accent of the item which precedes them, we produce the following which are rather more acceptable:

(63) SIXteen thousand seven hundred and twenty-EIGHT

(64) TWO million seven hundred and SIXty-four thousand and TWO

(65) TWO hundred and SIXty-one million SIXteen hundred and fifty-NINE

(66) NINE hundred and FIFty-seven thousand two hundred and forty-ONE

The solution would appear to be a compromise, allowing the accent to be deleted only if there is more than one accentable item before the next unaccentable one: our implementation approximates this compromise, and the evaluation discussed below shows just how effective a compromise it is.

There are two further problems in deciding which items in a number string should receive accents. The first seems to depend on whether the construction is functioning as a modifier, and the second involves the decimal point.

There appears to be a regular exception to the unaccentability of words such as *hundred, thousand*, and so on: in cases where these words come at the end of a domain

[4] There is some variation between speakers as to both what accent patterns they produce for number strings and what patterns they find acceptable: however, the results in Section 3.3.7 below show the general acceptability of our strategy.

(generally, the end of a noun phrase), they can and must be accented. The reason for this behaviour is not clear — it may be that numbers in this position are functioning differently (e.g. not modifying a following noun), or it may be simply the result of the phonological pressure to place accents at the right edges of domains - but the behaviour itself is clear enough:

(67) PROject TWO THOUsand
(68) the POpulation of SCOTland is about FIVE MILLion
(69) PICK a NUMber between TEN and three HUNdred
(70) she was aWARDed SIX hundred THOUsand

This behaviour can be modelled quite easily in our system by assigning such words to a special syntactic class and checking whether the final item in a domain belongs to this class. This is clearly not a particularly theoretically principled solution, but until the reasons for the observed behaviour have been determined there is some justification for the view that any solution that works is as good as any other.

In number strings incorporating a decimal point, the word *point* never receives an accent in non-contrastive usage: however, it does have the effect of splitting the construction into two sections which behave very differently. The section before the point behaves as though the section after the point were not there, and the section following the point behaves unlike a number string. Thus, we find accents as we would expect before the point but something rather different after it:

(71) TWO hundred and seventy-SIX
(72) TWO hundred and seventy-SIX point FIVE three eight one NINE
(73) TWO hundred and seventy-SIX point FOUR TWO

A strategy of assigning accents to the first and last decimal places appears to produce acceptable accentuations for up to five decimal places. Moreover, the simplicity of this strategy has much to recommend it in an automatic system. Although larger numbers of decimal places than this may sound somewhat unnatural if no accents are interposed between the first and the last place, the rarity of such strings in text allows this problem to be disregarded at least for the present: the accentuation in (72) and (73) is not difficult to achieve within the framework of our current rules, as is described in the section on implementation below.

Something of the relation between number strings and domains should be clear from the above examples, and from more famous examples such as (74). The influence of subsequent material on the accent patterns of number strings in such examples indicates that such constructions do not necessarily constitute a domain in themselves:

(74) FIFteen MEN

(75) there were TWENty-five PEOple on the BUS

(76) the BISHop orDAINED THIRty-seven PRIESTS today

(77) JOHN'S friend PAUL had a BUDget of NINEty-five thousand POUNDS

Examples (76) and (77) indicate that number strings must start a domain, although in informal tests some listeners judged these (and (67) and (68) above) to be unnaturally over-accented. The assignment of a domain boundary at the start of number strings is consistent with the behaviour of other anomalies, and again the evaluation of our implementation presented below shows that a domain boundary in this position is appropriate in most cases.

3.3.5 Abbreviations

The term *abbreviations* covers a multitude of sins: almost all textual anomalies could reasonably be described as abbreviations. Its usage here is much more restricted, in that it encompasses only those alphabetic anomalies which are not acronyms. By this definition, *NEC, FRCP, B.Sc, Ph.D* and *RSSPCC* all qualify as abbreviations but *DEC, FRIBA, CoHSE, RS232* and *3M* do not. Those which do qualify do not appear to differ significantly in their intonational behaviour, despite variations in their orthographic forms:

(78) EN ee SEE (NEC)[5]

(79) PEE aitch DEE (Ph.D)

(80) BEE ess SEE (B.Sc)

(81) TEE gee double-you YOU (TGWU)

(82) ARE ess ess pee see SEE (RSSPCC)

As with the strings of individually-pronounced digits after a decimal point, these strings of individually-pronounced letters appear to require accents on the first and last items only. The same argument also applies regarding the possible unnaturalness of this treatment for very long abbreviations: a corpus of 2,500 random abbreviations, which we collected as part of the development of this particular accent grammar, contained only one seven-letter and but a handful of six-letter exemplars. Our "first and last" treatment agrees with a suggestion by Kingdon (1958a:188), who also points out on the same page that the accentuation of abbreviations bears no relation to their meaning or to the structure of the full orthographic form.

The accent on the final element of an abbreviation can be deleted as a result of subsequent accents as in (83), and this even occurs in very long abbreviations as in (84):

(83) this is the BEE bee see NEWS at nine o'CLOCK

(84) i went to the ARE ess ess pee see see OFFices today

However, the accent on the first element does not seem to be affected by preceding accents:

(85) the SECond EE ee see SUMMit

(86) doctor OWen's ESS dee PEE

It therefore seems likely that a treatment whereby an abbreviation starts a new domain but need not finish it will yield appropriate accent patterns for these cases.

[5] For the purpose of marking accent placement, abbreviations in these and subsequent examples are written out with each character expanded to a pseudo-word such as *ay, bee, see* (for A, B and C). Capitalisation indicates accent.

3.3.6 Implementation

All our rules for the intonational treatment of anomalies will require modifications to the text preprocessor module and to certain intervening modules, as well as to the intonation rules, before they can be implemented in the full CSTR TTS system. Although the preprocessor already classifies and parses anomalies in a sufficiently detailed manner, its rules have not yet been modified to produce output in the required format. Similarly, the other modules of the system must be modified to process or ignore this new format as appropriate. This work has yet to be undertaken. Despite this, we foresaw no great difficulties in implementing all the above proposals within the framework of our intonation module. Such a local implementation would allow us to test our anomaly-handling rules by providing hand-generated input, although the rules would not apply to text input since the rest of the TTS system would not produce the input specified in the rules. The implementation of all these rules, as extensions of our intonation rules specifically designed to handle such constructions, has therefore been accomplished and is discussed in this section.

Implementing our anomaly rules involved modifications to INTERFIX, to the Rhythm Rule and to our accent assignment rules. INTERFIX was revised to recognise anomaly phrases marked by hand and to treat them appropriately. It assigns date phrases to a self-contained tg(0) domain, and interprets the start of any other anomaly phrase as a tg(0) boundary so that all such phrases begin a tg(0). All anomalies except dates are then passed to the accent-assignment rules and the Rhythm Rule: dates are not subject to the Rhythm Rule, since their accentuation is not influenced by any other items in the same domain.

Years, times and abbreviations are all treated exactly the same by the accent-assignment rules: their final lexical stress is assigned a primary accent, their initial lexical stress is assigned a secondary accent, and all other elements are assigned a 'd'. As in all cases, primary and secondary lexical stress are treated equally as sites for accent assignment. Any difference in the number of words or stresses in such items has no bearing on our rules.

For our purposes, number phrases are composed of two sorts of item: **digits** (e.g.

four, seven, eleven, nineteen, twenty), which are assigned secondary accent, and **orders** (e.g. *hundred, thousand, point*), which are deaccented. Once these accent markers have been assigned, a set of post-processing rules applies to the accent string. These rules reduce some of the secondary accents left-to-right, according to the following procedure:

1. They keep the first accent which they encounter, but they then set a flag which causes them to delete all accents until they encounter an **order** item.

2. When they encounter an **order** item, they unset the flag.

Since there are rarely (if ever) more than two **digit** stresses between **order** items, this procedure never deaccents too many items and in practice approximates very closely the suggestion above of only deleting an accent between **order** items if there is another accent to its left.

In addition, these post-processing rules transform the last accent marker, whatever it may be, into a primary to ensure that any number which ends in an **order** item will have primary accent on that item at this stage. The result of this post-processing is thus to preserve secondary accents on any digit immediately to the right of an **order** item and assign a primary accent to the rightmost item in a number phrase whatever its class, but to delete any other accents.

As an example of how this process works, we will work through the various stages involved in the treatment of a tg(0) domain such as

(87) 15,000 men

when everything in the domain except the noun *men* is flagged as a number phrase by INTERFIX. The first stage is for the accent assignment rules to assign secondary accents to the two lexical stresses in the digit *fifteen*, a deaccented marker to the order item *thousand* and a primary accent to the noun *men*. This produces an accent string such as [2 2 d 1]. Next, the number phrase is subject to the post-processing rules, which take the partial accent string [2 2 d] (corresponding to just the number phrase, rather than the whole domain) and preserve the first accent but delete the other: they also assign a primary to the last item, to produce the string [2 d 1]. This is now reunited with the standard primary accent on *men*, giving us [2 d 1 1], and finally this string is subject to the Rhythm Rule which reduces and then deletes the primary on *thousand* to give [2 d d 1] as the final accent pattern. This produces a very acceptable accentuation:

(88) FIFteen thousand MEN

Date phrases, once they are recognised by INTERFIX, undergo a much simpler process. Their first and last lexical stresses are assigned secondary and primary accent respectively, and the month (marked as a distinct word-class) is assigned an additional secondary accent: all other stresses are treated as deaccented. Since all calendar months in English have one lexical stress only, there are no problems of stress clash and so no need for any rhythm rule to apply.

3.3.7 Evaluation

Having implemented our rules for accent-placement in anomalies, it was necessary to test them both to ensure that they had been correctly implemented and to assess their accuracy. An evaluation experiment was therefore carried out along similar principles to those presented in Section 3.1 above. We decided to evaluate the anomaly rules at the symbolic level, providing hand-generated syntactic analyses to INTERFIX and assessing the output of the accent-placement rules.

By applying a version of the intonation module which incorporated all the rules described in the previous section to hand-generated crude phrase-structure analyses, a corpus of synthetic accent patterns representative of the accent placements assigned to anomalies by our rules. The syntactic structures were assigned independently by Dr. Colin Matheson, a CSTR syntactician. They included the information required by our anomaly rules, but were otherwise no more detailed than the analyses routinely generated by our automatic syntactic parser. The data to be evaluated consisted of 108 different anomalies drawn at random from electronic mail messages and from Milligan (1972). The 108 anomalies were divided between 93 isolated phrases or sentences: some of these were simply one single anomaly, others were sentence fragments containing an anomaly, and the remainder were full text sentences containing up to three anomalies each in various positions. This varied data was chosen in order to assess the performance of our rules in all the possible circumstances to be found in unrestricted text.

The resulting accent patterns were rated for appropriateness by three expert judges: the present author (AM) together with Dr. G. Lindsey (GL) and Dr. Williams (BJW). Neither of the latter two judges was familiar with the rules to be evaluated, and both differ considerably from each other and from the present author in their views on intonation.

3.3.7.1 Procedure

As in the experiment described in Section 3.1, the judges evaluated the output of our anomaly rules on the basis of ***the overall appropriateness of the accent pattern within each tg(0)*** and each tg(0) unit was therefore assigned one of four ratings:

4: entirely acceptable; possible as a realisation in natural speech

3: quite acceptable, but with certain unnatural features

2: only marginally acceptable, with seriously unnatural features

1: entirely unacceptable; likely to lead to misinterpretation or uninterpretability.

None of the judges had any difficulty in using this scale or in assigning a rating to each tg(0). In addition, the judges were asked to indicate the source of any unnaturalness or unacceptability where this could be readily determined.

3.3.7.2 Results

On the basis of the judges' ratings, each anomaly in the data was assigned a score from 1 (the worst possible rating) to 10 (the best). This was done by summing the 3 scores assigned by the 3 judges and subtracting 2: this converted the raw summed scores (which ranged from 3 to 12) into a more intuitive 1-10 scale. The tg(0)s which did not contain an anomaly were ignored in compiling the results, and in any cases of a tg(0) containing more than one anomaly the score for that tg(0) was assigned to both anomalies, unless

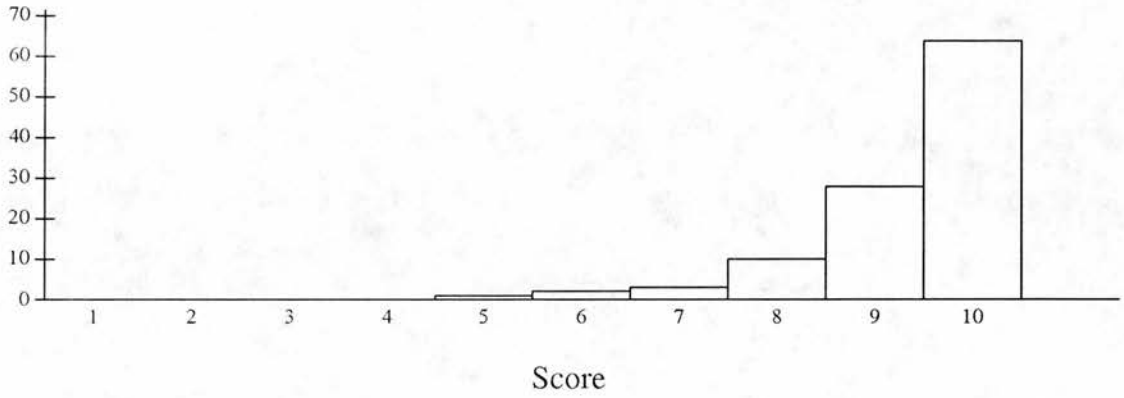

Figure 17: Distribution of Scores

one of the anomalies was indicated as the sole source of unnaturalness or unacceptability: in such cases the judge's rating was assigned to the problematic anomaly and any other anomalies in the same tg(0) were assigned a rating of 4. A histogram of the results is given in Figure 17.

The results can immediately be seen to be very good, and significantly better than those in Figure 15 above. None of the anomalies scored lower than 5, and scores of 5 and 6 were each only assigned to one and two of the 108 anomalies respectively. 102 anomalies (94.4%) were given uniformly positive scores (4s and 3s only), and no anomaly was assigned a rating of 1 on our 4-point scale by any judge. These are clearly very high ratings, and they testify to our success in taking what were previously problematic phenomena and making them a strength of our system.

3.3.7.3 Inter-Judge Agreement

To justify our pooling of the three judges' ratings, we applied the same two methods as in Section 3.1 above for assessing agreement between judges. Figure 18 gives a graphical representation of the individual judges' use of the rating scale: although BJW's ratings are noticeably less positive than those of AM and GL, it is clear that the distribution of ratings by all three judges have a very positive skew. The difference, in fact, is almost

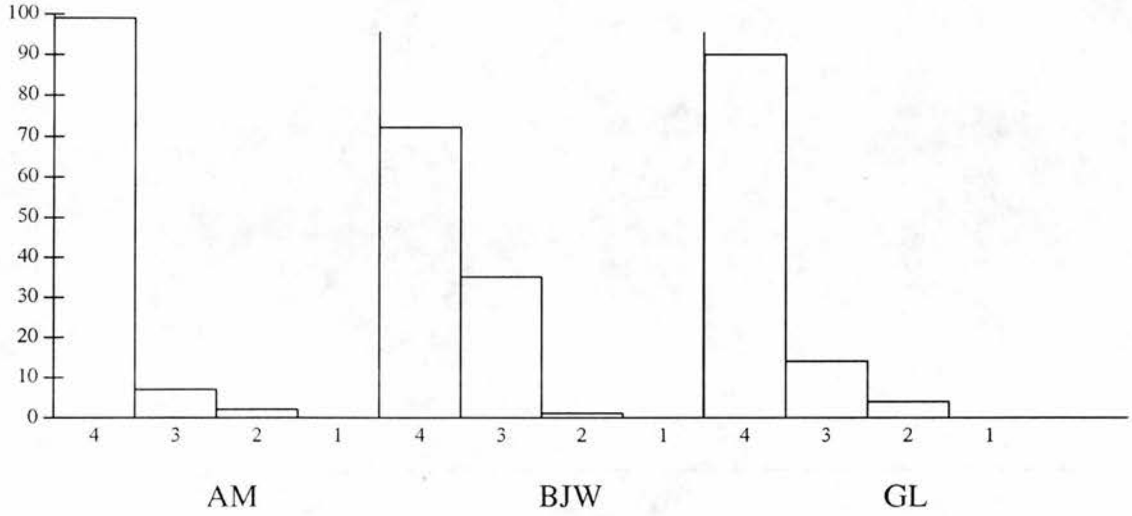

Figure 18: Distribution of Judges' Ratings

all in the relative proportions of 3s and 4s which were assigned by each judge, and since these are both positive ratings this variation is not very important.

For the second measure of agreement, comparing ratings for particular anomalies, we used the same four degrees of inter-judge agreement as above. These were:

a) Complete agreement: The three judges' ratings were identical.

b) Partial agreement: The three judges' ratings were all either broadly positive (3 or 4) or broadly negative (2 or 1).

c) Partial disagreement: None of the judges considered the tg(0) either entirely acceptable (4) or entirely unacceptable (1), but two rated it positive (3) and one negative (2), or vice-versa.

d) Complete disagreement: Any other combination of three ratings (e.g. 4-4-1, 4-1-1, 3-2-1).

Of the 108 anomalies, there was complete agreement on 66 (61%) and partial agreement on a further 36 (33%). For 94% of the anomalies, therefore, the three judges agreed on a classification broadly positive vs. broadly negative. There were only 2 cases of partial disagreement, and 4 cases of complete disagreement, despite the fact that the judges only used three of the four scale points.

Both our measures of inter-judge agreement show a very high correlation between the ratings assigned by different judges. This fact supports the pooling of the scores as above, and emphasises the very high level of acceptability of the accentuations which our revised rules assign to anomalies.

3.3.7.4 Analysis of Errors

There are very few cases (less than 6%) where a negative rating was assigned by any of the three judges, and none where all the judges were agreed in assigning broadly negative ratings. Every one of the six anomalies which was not assigned a broadly positive rating by all the judges is therefore a case of partial or complete disagreement between the judges as defined above.

Two of these errors are also the only two cases of time phrases modified by specifiers such as *am, pm* and *GMT*. These two cases exemplify the problem anticipated above, that our rules do not handle modified times appropriately: all the judges were agreed that the accentuations assigned to (89) and (90) were not optimal, and this is clearly an area where there is room for improvement in our rules.

(89) 17:14 BST

(90) 3.30am.

Three more errors all occurred in the same sentence, and were all given a rating of 2 by GL but 4 by the other two judges: there is no obvious explanation for this, and it must for the moment simply be attributed to individual preferences or differences of interpretation. The one remaining error was again judged less than optimal by all the judges, and appears to have been caused by a wrong decision regarding the structure of a compound which was propagated throughout the domain by our Rhythm Rule: this type of error will occasionally arise with such a rule, but its frequency is so low as to be negligible.

Various problems which we had anticipated, such as our lack of specific rules for telephone numbers and post codes, did not seem to arise: the judges found the treatment of these anomalies by the number or abbreviation rules quite acceptable. However, we feel that it would be advisable in a working system to develop further rules for such readily-identifiable constructions as well as improving our treatment of specified time phrases.

3.3.8 Discussion

There are several assumptions and assertions made in the above section, some of which deserve more discussion than the present chapter allows. The following few paragraphs attempt to point out areas for further investigation and to elaborate briefly on some of the less complex issues raised.

The most obvious question arising from the preceding description of our intonational treatment of anomalies in text is whether such a treatment is appropriate: are the questions of accents and domains really the ones which need to be answered, or should we be looking at the precise function (grammatical, semantic or pragmatic) of anomalies in text and assigning intonation from that? There are two levels of answer to this question. From the standpoint of someone trying to build a machine to do the impossible, i.e. assign natural intonation from unrestricted text, there is a compelling case for maintaining that any approach which works is a good approach to take. The results of our evaluation certainly indicate that a very large degree of improvement can be made to the treatment of these previously problematic phenomena by incorporating the rules developed using our heuristic approach: a major cause of errors in our assignment of intonation to running text has been successfully eradicated by this approach. The counter-argument, of course, is that there is little point in trying to make an impossible task easier and that what we ought to be doing is investigating the higher-level factors governing this task and trying thereby to bring it into the realm of the possible: from the standpoint of a theoretical linguist, this is probably the only justifiable course of action. To the extent that it is possible to do both, this would seem to be the best short- and long-term solution but the issue remains unresolved.

Two further points warrant some discussion. The first of these is the issue of extendability to contrastive usage and other explicit exceptions. The rules of our intonation module presented in Chapter 2 explicitly exclude contrastive and emphatic usages, but they are designed to be flexible enough so that when information on such usages is available it can be easily incorporated. In some ways, particularly in the case of the exceptional behaviour of domain-final *hundred, thousand*, etc., the treatment of anomalies already seems to incorporate the use of such higher-level information and it might be expected that the mechanisms for recognising and treating anomalies would be capable of extracting other information as well. It must be emphasised, however, that we believe some element of understanding to be essential to the treatment of contrast and non-neutral emphasis and that the mechanisms suggested for handling anomalies do not incorporate any such element. It is only by virtue of the exceptional textual characteristics of these anomalies that any higher-level information is deducible from them: even the identification of dates or years written out in full, as in *nineteen eighty-nine*, presents major problems on which the present observations have no bearing. Indeed, there is an empirical question to be answered regarding the factors which determine whether, for instance, a date is written in full or as digits: it may well be that alternative textual forms correlate with different higher-level specifications and consequently have distinct intonational characteristics. No work on such factors has been carried out to our knowledge.

Secondly, in view of the preceding point, it should be determined whether there are any other classes of items which are readily identifiable from text and whose appropriate interpretation is relatively clear. Two classes of item come to mind in this context: proper names, generally identifiable by their initial capital letter, and punctuation. The latter has already been investigated above and the results have been partially incorporated into the CSTR TTS system's intonation routines, and the former is high on the list of areas requiring investigation in TTS systems generally although the problems involved in interpreting proper names have been described at length by philosophers and linguistics alike, among them Strawson (1959), Levi (1978) and Sproat & Liberman (1986, 1987). There may also be other classes of items amenable to the above approach which will be revealed by corpus analysis in the future: on the assumption that there are other such classes, we have attempted to ensure that our model will be readily adaptable in order to take advantage of them.

3.4 Anaphora

It was clear from the results of the evaluation experiment in Section 3.1 above that the single biggest problem with the output of our accent-placement rules was their failure to take account of pragmatic deaccenting, either of anaphora or of semantically redundant lexical items. As has been frequently pointed out in work on synthetic intonation (e.g. Quené & Kager 1989, Hirschberg 1990), the information which governs such deaccenting (pragmatics, semantics, discourse structure, etc.) is not currently available to automatic systems: it was therefore necessary to investigate alternative approaches in order to produce practical heuristic methods for tackling these problems which would minimise the occurrence of such errors in the short (and probably medium) term. This section presents a number of such heuristics which have been developed on the basis of the data from the two test corpora mentioned above. Some of these heuristics are dependent upon the particular semantic domain (in a system giving railway timetable information, for example, items such as *train* and *platform* might be redundant), and some are domain-independent (e.g. deictic modifiers should always have the same effect, as should contrastive stress).

3.4.1 Deaccenting

Deaccenting is a phenomenon which has been addressed at length in the linguistic literature (Ladd 1980, 1984a; Bing 1983; Selkirk 1984; Fuchs 1984; Baart 1987). The consensus view, among these and other authors who see deaccenting as a linguistic choice, is that it is a local phenomenon which involves a decision to remove accent from a particular item with the observed effect that that accent "moves" to another item. The processes which determine which item receives this displaced accent are still the subject of considerable disagreement: Ladd's notion of **default accent** determined by a local change in metrical structure is favoured by many, but Gussenhoven's "Polarity Focus Rule" (1984:49ff.) suggests a mechanism whereby certain grammatical classes attract such accents and Bolinger (1989:224ff.) defends at some length the claim that deaccenting, like accenting, is entirely controlled by the speaker's communicative intent.

We take the position here that the site where the displaced accent falls is determined by rule, but we hold that deaccenting is a very local phenomenon and that the rules which determine where accents fall in these cases are the same sorts of rule as those which resolve stress clashes (Liberman & Prince 1977:309ff.). Once the speaker has decided to deaccent a particular item, for whatever reason, the task of assigning accents around that item is performed by rhythm rules and other general phonological principles. It is therefore possible to treat phenomena such as anaphoric deaccenting and contrastive deaccenting as very local, without any reference to global metrical or syntactic structures, and allow the repercussions of a Bolingerian unilateral decision to deaccent a particular item to be handled by rules such as the Rhythm Rule which exist for quite independent reasons.

The most common cause of deaccenting in text, as was demonstrated by our evaluation experiment in Section 3.1 above, is the occurrence of anaphora. G. Hirst (1981:7) defines an anaphor as "a reference whose antecedent is a concept or entity EVOKED implicitly or explicitly by the preceding text or situation." This is a rather broader definition than we require, since we are only concerned with anaphora which constitute exceptions to our standard intonation rules. For present purposes, anaphora are defined as items whose prosodic status is affected by the fact that they refer to items which are contextually **given**. The problem is essentially that many (but not all) anaphoric items require to be deaccented: if they are erroneously assigned an accent, the detrimental effects on intelligibility and perceived naturalness are severe. The process of deaccenting is completely straightforward in our system: in fact, many anaphoric items are automatically deaccented by Clause (3) of our Rhythm Rule (see Figure 12 and page 108 above), but those which are not remain a serious problem. For instance, when the 1,200-word corpus of technical abstracts mentioned above was processed by the intonation module more than 20 of the 58 sentences produced serious errors attributable to the inappropriate treatment of anaphora. The formulation of strategies which will minimise the number of such errors is therefore a high priority, and the remainder of this section presents three types of strategy which we have developed to meet this need. All these strategies have been implemented in the intonation module of the CSTR TTS system.

3.4.2 Lexis

The problem of content words which are semantically redundant or empty accounts for many of the unacceptable accentuations in our evaluation data (Appendix A). Items such as *walk, give, operation, area, sense,* and *example* frequently do not constitute candidates for accent despite their word class, as they do not contribute any significant information. There are numerous instances of such items in the data, and many of them could be marked in the lexicon as semantically redundant: the problem is that such items are not always redundant, and other items such as *organiser, world* and *view* are occasionally redundant but usually quite meaningful. The question of which items are to be marked as redundant is therefore not an empirical one, but a fair number of clear cases can probably be distinguished.

The lexical information available to text-to-speech systems is basically whatever is contained in on-line lexica. This can often be supplemented by morphological, syntactic or other analysis, but the lexis-based heuristics presented here rely on the information which they require being present in some form in the appropriate lexicon. Lexical heuristics are effective for treating subsets of lexical items, often as small as a single item, which do not conform to the general behaviour of some larger set, such as the general rule that content words can be accented and function words cannot, and the most obvious way of dealing with these lexical items is to mark them as exceptions in the lexicon. Such items fall loosely into two classes for the purposes of text-to-speech applications: domain-specific items and domain-general items.

In restricted domains, or in a particular text or discourse, certain lexical items may be relatively redundant or uninformative: these may be **explicitly** anaphoric, referring back to some entity or concept already mentioned explicitly, or they may be **implicitly** anaphoric, referring to something which has not been mentioned but which is taken for granted in the context. Both these types of anaphora correspond to Prince's (1981) category of "evoked entities". The phrase <u>*lexical item*</u> is an example of the former type in this paragraph: a mention of <u>*doctoral dissertations*</u>, although not explicitly mentioned previously, would constitute an example of the second type in the present context. Both phrases, if not highlighted, would ideally be given reduced prosodic prominence. In the 1200-word corpus mentioned above, which was restricted to speech technology abstracts, the items *speech* and *system* accounted for almost all the errors attributable to anaphora in the intonational output, and only one of the 86 tokens of these items required to be accented. It would therefore seem sensible to mark the lexical entries for these items so that they are not assigned accentual prominence, and accept the very low (less than 1.2%) error rate. There is, however, a problem in establishing which items in any given domain, or worse still in any paragraph or discourse, should be marked as redundant. In many cases this could be done manually on the basis of known errors (as above), statistics (word frequency, etc.) or experts' intuitions, and could even be updated interactively: eventually, such information might be determined automatically, but at the moment it must be entered by hand.

It does appear from our data that it is the lexical items rather than the concepts or entities which they represent which are deaccented: occurrences of *spoken language*, for example, are not deaccented in the same way that the lexical item *speech* is. Brown (1983:75), in a study of accent placement in dialogue, presents similar results: contrary to her expectations, she found that speakers' accentuation of particular phrases or lexical items depended not on the anaphoric status of the entity referred to but rather on whether the phrase or item had been uttered previously. The situation is, however, not as simple as it appears: in a larger study of descriptive text, Bell (1987) found that in general the "newness" of a particular lexical item or phrase was not a reliable indicator of its accentuation. Nevertheless, it seems from our results that there are some reliable indications of deaccenting to be deduced from lexical information.

There are certain lexical items which require to be treated as anaphora but whose behaviour is relatively independent of the particular semantic or discourse domain in which they occur. These are items which convey very little new information even on first usage, and whose main function is as the default specification from a set of items which can fill a particular slot. Examples include items such as *Street* in the set { *Road, Place, Street, Avenue, Drive, Lane, ...* }, *cake* in the set { *tart, pie, flan, cake, mousse, turnover, ...* } (British English), and *Land* in the set { *Bay, Point, Land, Hill, Cave, River, ...* }: notice the default nucleus placements (indicated by capitalisation) in the (a) examples as compared with the other examples in (91-93). These items are most

(91) a. OXford street b. oxford ROAD
c. oxford CIRcus d. oxford LANE

(92) a. APple cake b. apple PIE
c. apple FOOL d. apple SORbet

(93) a. SNEEZE land / BOGgle cake / GRZANSK street
b. sneeze ISland / boggle STEW / grzansk BOUlevard
c. sneeze COVE / boggle JAM / grzansk TERRace

easily handled by marking the individual lexical entries for obligatory deaccenting and accepting the occasional error as above.

Larger classes such as demonstrative pronouns { *this, those, ...* } or redundant nouns { *thing, person, place, time, idea, ...* } are also domain-general in their refusal to conform to the default behaviour for content or function words. The former may be accented when they stand for a full NP (but see below for more discussion), and the latter are generally unaccented regardless of any particular local context. Both can be handled by marking them in the lexicon, e.g. by simply including demonstrative pronouns in the set of content words or by adding an explicit argument to the appropriate lexical entries or grammatical classes to indicate prosodic markedness.

3.4.3 Syntax

Syntactic heuristics depend on the rough correspondence of anaphoric deaccenting with particular syntactic constructions. As was pointed out in Chapter 1, this correspondence can only be approximate since it is not syntax which determines the prosodic behaviour of anaphoric items. Nonetheless, it is generally claimed (Chafe 1976:25; Brazil 1984:47; Selkirk 1984:198ff.; Allen et al. 1987:107; Baart 1987:53) that such syntactic constructions as clefting, topicalisation, definite NPs and certain types of predicate have well-defined prosodic characteristics.[6] Unfortunately, at least in our ex-

[6] For a review of this position, see Delin (1989) Chapter 2.

perience, this is not the case. For instance, contrary to common assumptions (Allerton & Cruttenden 1979; Schmerling 1976:34; Kager & Quené 1989:105), definite NPs do not appear to have a greater tendency to be deaccented in our 1200-word corpus, and studies by Kruyt (1985) and Terken (1985) found similar results with very little difference in the likelihood of accentuation between old and new information. Moreover, it is well known (e.g. Bolinger 1972a, Ladd 1980) that **given** items may be accented anyway at the speaker's whim, particularly if they are complex and if they occur in pre-nuclear position, and that even "brand-new" items in Prince's (1981) terms need not be accented (Terken 1980:49; Brown 1983:74ff.; Fuchs 1984:144). However, it would seem that NPs introduced by deictic expressions (such as *the latter approach . . . , this month . . . , those red boots . . .*) do tend to have deaccented (anaphoric) heads and that modified expressions in general are more likely to contain anaphoric items: in the corpus of abstracts mentioned above, all the deictic forms modifying nouns had the effect of deaccenting their head noun. Although the use of definite modifiers is no indication of the focal status of an NP, then, deictic forms appear to be reliable indicators of deaccented heads. We have therefore attempted to identify such forms and incorporate this behaviour into our system. There will of course be exceptions to such a heuristic: examples such as *this evening* or *It's that man again* will obviously pose problems, but the generalisation is nonetheless a useful one and the exceptions to the rule will be fewer than the errors which it avoids. Work is continuing on identifying further syntactic indications of prosodic markedness which will improve our treatment of anaphora.

3.4.4 Pragmatics

As reliable pragmatic information is not currently available to automatic systems, and will not be for some time, these heuristics are intended to mimic the effect of pragmatic factors on accent assignment. Our current heuristics are based on the semantic/pragmatic content of lexical items: once the output of these heuristics has been analysed to discover which particular phenomena are still problematic, these persistent errors will be examined to see if more sophisticated strategies can be formulated.

The basis for our present strategy is the observation that nouns modified by *other,* *alternative, first, second, last,* and similar modifiers can be reliably treated as pragmatically deaccented: the semantics of these items is such that they are almost always used in contexts where the head of the NP which they modify is predictable. Such modifiers force these nouns to be interpreted as **given**, and therefore deaccenting is probable (although not always possible and certainly not obligatory, as stated above). These modifiers can be marked in the lexicon as indicators of anaphora, and the items which they modify can then be deaccented by rule. There remains the problem of identifying the head of the phrase, but in most cases this is trivial and such items need merely set a flag which deaccents the next head: the implementation of this is discussed below.

More ambitious pragmatics-based heuristics might include routines to determine whether a repeated item should be deaccented or to spot parallel structures (contrastive stress, conjunction, etc.) so that they can be treated appropriately: these are much more difficult problems to resolve, but they are also correspondingly less frequent! To date, attempts to recognise anaphora in text on general pragmatic grounds of **givenness** have met with very little success: Bell (1987) reports that a strategy of deaccenting repeated lexical items was completely unsuccessful in the IBM(UK) TTS system, as mentioned above, and more linguistically-oriented work such as that by Burton-Roberts (1986ab) and the systems discussed in Carter (1987) requires impractical amounts of linguistic analysis before it can provide reliable indications for deaccenting. It appears that the heuristic approach advocated here will continue to be the only viable method of modelling pragmatic factors for synthetic intonation for the foreseeable future.

3.4.5 Implementation

All the heuristics mentioned above are currently implemented in the CSTR TTS system. Some have involved modifications to the lexicon, and others have required special rules to be written. The rules are incorporated in our accent-placement module, and operate on the output of INTERFIX such that they see only one tg(0) domain at a time. This section outlines the behaviour of these heuristics and the output which they produce.

All the examples of domain-independent redundant lexical items given in Section 3.4.2 have been flagged in our main lexicon, and domain-specific redundant items have been treated in the same way in domain-specific lexica (of which we currently have two; one for speech technology, and one for pathology): this produces lexical entries explicitly flagged with an argument to indicate prosodic markedness. Content words flagged in this way are automatically deaccented by our accent-placement rules, and function words which are flagged for markedness will conversely be assigned an accent. In both cases, however, these accent-assignments are subject to modification by the Rhythm Rule and the WFCs described in Chapter 2. This treatment generally results in appropriate intonational behaviour for these items. Demonstrative pronouns such as *this* and *these* are marked for accentuation only if they function as the head of a domain, i.e. if there are no other accents in that domain: this heuristic interacts with the treatment of demonstrative determiners, as follows.

Both determiners and pronouns are given the same word class if they are demonstrative: *syn-det([1])*, where the *[1]* argument indicates their marked status. This reduces needless word-class ambiguity (Monaghan 1990b:110-111), and allows the behaviour of these items to be determined by their prosodic context. Any *syn-det([1])* item is initially assigned a ‘d’, the lowest degree of prosodic prominence associated with a **content word**, and also sets a flag which deaccents a following noun (if any). This flag only applies to the current prosodic domain. Thus, in domains such as

(94) This specimen

where *syn-det([1])* modifies a noun the noun will be deaccented; in domains such as

(95) These red ones

where the head of the phrase is a pronoun and there is no noun to be deaccented, the *syn-det([1])* will have no effect on following accents; in domains such as

(96) This

where *syn-det([1])* functions as a pronoun it will attract the nuclear accent for that domain. There are still problems with this strategy, of course, but on the whole it appears to produce the desired results in the small corpora we have investigated.

The pragmatic import of items such as *other, alternative, latter,* etc. has been implemented by marking these items in the lexicon as domain-general indicators of anaphoric deaccenting, and by providing special rules for their treatment in the accent-assignment module. When this module encounters such an item, it assigns it the highest degree of prosodic prominence, i.e. primary accent, and sets the same flag as above: the result of this is that the item is almost guaranteed to become the prosodic head of its domain, since other items bearing primary accent (i.e. nouns) to its right will be deaccented and the rightmost primary in any domain is the prosodic head by default as described in Chapter 2. Again, this seems to be the behaviour which we require in the output of our intonation rules in most cases: however, there are some minor problems with the present treatment. In particular, there are cases where these items are followed by content words which they perhaps do not modify but which are nonetheless in the same domain. In a tg(0) domain such as *other members of the same organisation*, for instance, the preferred accentuation would place the nucleus on *organisation* as in (97): unfortunately, this accentuation is not possible in the current version of our rules, since the special interpretation of *other* will cause all nouns in the same domain to be deaccented, producing the much less acceptable accentuation in (98).

(97) OTHer members of the SAME organiSAtion
(98) OTHer members of the same organisation

This problem is currently unavoidable, however, since we cannot identify the precise scope of such modifiers any more closely than by assuming that it includes the remainder of the current domain. The assignment of a secondary accent to *other* instead of a primary would allow us to produce accentuations such as (99), since following secondaries would take precedence: however, there would then be the possibility that in examples such as (100) the accent assigned to *other* would be deleted by the Rhythm Rule, producing highly unnatural output.

(99) OTHer clowns dressed in PINK
(100) other PINK and green STRIPED clowns

With our present Rhythm Rule and without introducing new categories of accent, there appears to be no better solution than the one which is presently implemented: such errors as it produces must therefore be tolerated for the time being.

The small number of other TTS systems, such as PROS ((Kager & Quené 1989) and NewSpeak (Hirschberg 1990ab), which incorporate strategies to handle anaphora appear to have a much less flexible approach to these phenomena. For instance, PROS recognises anaphora of two types — definite NPs and certain adjectival modifiers similar to those just discussed — but it only has one strategy for handling both cases: namely the strategy of deaccenting the entire domain, which consistently produces accentuations such as that in example (98). Moreover, the inappropriateness of deaccenting all definite NPs was discussed above. Our rules for handling anaphora therefore appear to be much more flexible and successful than those applied in other developmental TTS systems.

3.4.6 Summary

The three classes of heuristic presented here are intended to handle as many anaphoric phenomena as possible without requiring a TTS system to perform any unreasonably complicated linguistic analysis. Many of the strategies suggested are relatively simplistic and will require refinement before they can be optimally incorporated in a working system: however, all the general principles are applicable to any text-to-speech system and most are not domain-specific. Our heuristics are based on careful examination of a relatively small corpus, together with subjective impressions of other corpora and linguistic intuitions. They reliably produce surprisingly acceptable accentuations for a wide range of problematic phenomena: nevertheless, they are obviously a poor substitute for accurate semantic and pragmatic analysis, and should be seen as stopgap measures in the absence of these analyses rather than an alternative method of deriving such information. However, although the availability of reliable higher-level linguistic analyses would remove the need for such rules, we anticipate that the general principle of prosodic markedness will still be appropriate in more linguistically-motivated treatments of anaphora.

All the heuristics outlined in this section are currently made use of by the CSTR TTS system. The output of these strategies appears at this stage to be highly satisfactory and widely applicable, and it is hoped that it will provide a reliable basis for assigning prosody to truly unrestricted text. Preliminary evaluation by naïve users in a real application domain (Monaghan 1991b) indicates that systems incorporating such strategies can produce intelligible and acceptable output for unrestricted running text.

The principles on which individual heuristic techniques are based may be few and far between, but this is not true of our overall approach: we believe that the sorts of strategy adopted here are also adopted by human readers in the absence of full understanding, and that the rôle which a reader's informal ideas about prosody and probability play in his/her realisation of a complex text is similar to that of the various rules discussed above. We have avoided the complicated mathematics of probability and the advanced parsing algorithms currently popular (e.g. Fitzpatrick & Bachenko 1989; Marcus & Hindle 1990), partly because these are not our principal interest but more importantly because we do not believe that high-powered maths is the correct solution to the TTS problem: it is our expectation that the most interesting and convincing TTS systems will increasingly be those which incorporate linguistic knowledge of all kinds, algorithmic and heuristic, to compensate for machines' lack of understanding. That said, we explicitly state that any reliable information, whatever its source, on the probable prosodic realisation of text should be exploited unscrupulously: all's fair in speech synthesis!

3.5 Discussion

The evaluation experiment discussed in the first section of this chapter introduced and validated a new procedure for the assessment of synthetic intonation. Assessment at a symbolic level was shown to be practicable for the more phonological models of intonation developed in recent work on TTS systems, and the results of such an assessment were interpreted for both diagnostic and documentary purposes. Subsequent sections presented our attempts to resolve the major errors revealed by the results of our evaluation experiment. The strategies adopted to handle the phenomena which cause these errors are all in accordance with our stated purpose (Section 1.5.1) of producing **acceptable neutral** intonation: however, the cases which they handle are clearly exceptions to our assumptions (Section 2.9.3) that all input text is [+focus] and that major syntactic phrases correspond to Gussenhoven-style semantic constituents. The following paragraphs discuss the implications of these exceptions for our model of intonation set out in Chapter 2.

Although the rules presented in this chapter still aim to generate a phonological specification of **neutral** intonation for any input text, many of them (particularly those which handle anaphora) are specifically designed to identify cases where such a neutral specification does not correspond to a **broad focus** reading. These cases were shown to be problematic for the rules described in Chapter 2, and therefore required to be treated differently: heuristics to identify such cases on the basis of lexical and syntactic information were developed accordingly. However, these heuristics still conform to our strategy of defining the intonation contour solely in terms of accents and boundaries which can be interpreted locally: we have found no reason as yet to attempt to build detailed metrical trees or to include other types of event in our phonological specifications.

Our treatment of anaphora is based on deaccenting probable anaphoric content words. In contrast with metrical approaches to deaccenting (e.g. Liberman & Prince 1977; Bing 1979a; Ladd 1980, 1984), which see the process of deaccenting as a reversal of the labelling on sister nodes of a binary tree dominating an arbitrary amount of structure, our deaccenting rules demote an accented item on a purely local basis without forcing any corresponding promotion of some other item. This treatment allows us to apply our rules in a left-to-right manner without concerning ourselves with non-local repercussions. The question of where to assign any **default accent** (Ladd 1980:81ff.) consequently does not arise in our system: nucleus placement in domains containing deaccented items is determined by the same principles of rightmost prominence and degrees of accentability as apply in full-focus domains, and the location of head accents depends on the accentability hierarchy and the application of the Rhythm Rule. Ladd (1980:54ff.) criticises a similar account of deaccenting by Vanderslice & Ladefoged (1972), where the phenomenon is treated simply as the removal of a feature [+accent] from a particular item: Ladd (1980:84ff.) demonstrates that different degrees of accent are required to account for both Broad Focus and Narrow Focus accent placement, and we have incorporated such differences of accentability in our model; however, the claim (Ladd 1980:56) that an overall metrical structure is required to model deaccenting phenomena is never really justified in relation to degrees of accentability. Our rules appear to produce acceptable output, handling both leftward shift and rightward shift of the nucleus as well as the deaccenting of non-nuclear accents, without any reference to metrical structure.

It is true, of course, that in the case of our rules handling modifiers such as *other, latter, different* and so forth (Section 3.4.4) the current implementation does assign a different degree of accent to these items, and this could be seen as promotion along the lines of a Default Accent theory. However, this promotion is merely to prevent deletion by our Rhythm Rule of the accents assigned to such modifiers and indeed their promotion has its associated problems as discussed in Section 3.4.5. It is our view that accents such as these which are assigned on the basis of pragmatic information should not be susceptible to deletion by the Rhythm Rule: however, we currently have no way of protecting these accents other than by assigning to them the maximum degree of prominence available in our model.

Domains which contain anaphorically deaccented items are in our view at least partly [-focus], since we equate deaccenting with the marking of [-focus] items. In place of our earlier assumption that all input was [+focus], we now insist that every domain must contain some [+focus] constituent (marked by accent)[7]. This rules out domains which are entirely [-focus]: such domains would be assigned no accents in our model, and are assumed to be absorbed into neighbouring domains as pre-heads or tails in the manner described in Section 2.5. Within any domain, then, there may be [-focus] elements in addition to the obligatory [+focus] constituent: the breadth of focus is indicated by accent placement, particularly by the location of the nucleus (as would be expected, given the primacy of the nucleus discussed in Section 2.9.3). Ambiguous focus structures will arise, especially when deaccenting applies to head accents which are susceptible to deletion by the Rhythm Rule, but such ambiguity is common in natural speech (Brown 1983) and is allowed for in theoretical accounts of intonation (e.g. Ladd 1980:74ff.; Gussenhoven 1983b:168ff.).

The rules which were introduced above to improve our treatment of anomalies and of PPs appear to constitute exceptions to the generalisation that the semantic constituents relevant to prosody correspond to major syntactic phrases. Many of the anomalies treated by these rules, such as number strings and abbreviations, do not have to constitute a complete major syntactic phrase, and our rules for sentence-final PPs imply a distinction between these constituents and all other Gussenhoven-style **conditions** in that the latter do not trigger domain boundaries. We have no theoretical explanation for these exceptions, and indeed our present rules do not handle rare cases such as (35) above (p.151) entirely appropriately. It was stated in Chapter 2 that our assumptions of some limited congruence between syntax and semantics were intended to be overruled in cases where information to the contrary was available, and anomalies and sentence-final PPs seem to be such cases: however, the only justification which we can offer for treating these particular cases as exceptions to our earlier assumptions is that, judging from the various evaluations discussed both above and below, such a treatment improves our intonational specifications. New or improved treatments may emerge from further work

[7] The question of whether focus is realised by accent is a complicated one (see Ladd (1980, 1983a) for a lengthy discussion), but the tendency for accent to be associated with focussed constituents is widely attested (e.g. Terken 1985; Eady & Cooper 1986; Nooteboom & Kruyt 1987).

and may lead to a clearer understanding of the linguistic or other reasons behind the intonational behaviour of the phenomena discussed in this chapter, but for the present our treatment of these phenomena can only be judged by its results.

Although we now have many different syntactic and semantic constructions assigned to tg(0) domains by our rules, we have not altered our uniform treatment of the boundaries between these domains (see Section 2.1.3 above). Ladd (1988) has shown that in experimental conditions speakers are able to modify the realisations of prosodic boundaries between domains at the same level of structure on the basis of the semantic relations between those domains: however, his data (pp.538-541) also show that speakers do not always make use of this ability, and we have not as yet found any reason to exercise similar control over the realisations of individual boundaries in the output of our system. It would appear that, in synthetic speech at least, listeners are not disturbed by uniform realisations of boundaries between domains at a particular structural level regardless of semantic or other relations between those domains. We suggest that in such cases the relations which might have been realised by variations in boundary strength will be conveyed by other factors such as syntactic form, linguistic context and world knowledge: the trade-off between prosodic and other aspects of speech in conveying information is returned to in Chapter 4.

Our rules for handling the various exceptional cases discussed in this chapter could conceivably have introduced a large number of possible choices between multiple prosodic treatments of lexical and syntactic items. However, this was emphatically not the case: not only do our rules continue to restrict prosodic realisations to the assignment of accents and boundaries, as discussed above, but they also restrict the possible treatments of any item to a choice of two — "the default treatment" or "not the default treatment". For any lexical item there is a default accent assignment — primary, secondary, 'd' (in certain cases discussed in Sections 2.6 and 3.3.4) or nothing — and the only other possibility is to deaccent that item (assigning it a 'd'):[8] such deaccenting indicates [-focus] portions of a domain, as discussed above. Similarly, at any boundary

[8] The case of sentence-final prepositions, discussed in Section 3.2 above, is an exception: these items are assigned an accent, although the default treatment of prepositions is to leave

between lexical items there is a default boundary assignment — tg(2) boundary, tg(1) boundary, tg(0) boundary or no boundary — which is currently determined from the syntactic analysis: the only alternative to this is to fail to assign a boundary where the default treatment would have assigned one, and this alternative is currently applied in accentless domains which are assumed to be completely [-focus].

This system of binary choices between a default specification and some non-default specification seems to correspond to the **marked/unmarked** dichotomy invoked in many areas of linguistic theory (Chomsky & Halle 1968:402ff.; Fodor et al. 1974:497; Gazdar et al. 1985:29ff.), and this is how authors such as Ladd (1980, 1983a), Gussenhoven (1983a) and Selkirk (1984) have interpreted the choice between deaccenting and not deaccenting: however, we reserve the use of markedness for another purpose. In the present case, we seem to be concerned with the indication of focus rather than any other factor: the breadth of focus which listeners perceive is determined by the placement of accents and boundaries in the output of our rules. The presence of an accent indicates focus, as does the presence of a boundary between major focussed constituents: in both cases, the default specification corresponds to a maximally Broad Focus while the choice of a more exceptional specification corresponds to some degree of Narrow Focus (with the possibility of ambiguity as noted above). Our rules therefore produce something other than Broad Focus in the cases of anaphora discussed in Section 3.4, and a clarification of the relation between Broad Focus and our declared aim of **neutral** intonation is consequently appropriate at this point.

In Section 1.3 we defined **neutral** intonation as being as close an approximation to Broad Focus as possible given the restricted information available to a TTS system. Broad Focus is defined (Ladd 1980:74) as the accent placement which "leaves the focus broad or unspecified": however, many of the cases discussed in Section 3.4 are precisely those where a Broad Focus realisation is not acceptable and where this fact is predictable despite the limitations on the information available to TTS systems. We disagree with

them unaccented. However, we see this process not as an exceptional treatment of perfectly ordinary prepositions but rather as the identification and appropriate (default) treatment of a class of verbal particles.

Ladd's claim (1980:98) that deaccenting and focus are compatible: in our view, focus structure determines deaccenting in that nuclear accents may not be placed on [-focus] items.[9] However, we do not agree with Gussenhoven's (1983a) "one focus, one accent" approach. We see focus structure, together with the relative accentability of focussed constituents, as determining nucleus placement with Normal Stress corresponding to maximally Broad Focus: this is largely in agreement with Ladd (1980:73ff.). Where we differ from Ladd's approach is in our view of the relation between Broad Focus and markedness: Ladd (1980:76) equates these two notions, whilst pointing out that many sentences do not have a Broad Focus realisation, and therefore concludes that these sentences cannot have an unmarked realisation either; we agree entirely that sentences such as Ladd's examples (repeated here as (101) and (102)) involving reflexives and the word *even* have no full-focus realisation, but we see their focus structures as being completely determined by semantic and contextual factors and consequently as being **unmarked**.

(101) even a TWO-year-old could do that

(102) john was killed by himSELF

Examples such as (103) and (104), however, reflect intonational choices by the speaker which explicitly contradict the pragmatic content of *himself*, and it is these cases for which we reserve the term **marked**.

(103) john was KILLED by himself

(104) john WAS killed by himself

Although it is true, as Bolinger (1951), Schmerling (1976) and Fuchs (1984) amongst others have pointed out, that the notion of Normal Stress as a syntactically determined property of isolated sentences is not tenable, since it is demonstrably untrue that "every

[9] The optional nature of head accents has been discussed above.

sentence has a 'normal' pronunciation" (Schmerling 1976:49), we are of the opinion that ***every sentence in a particular context*** DOES have a normal or **unmarked** accentuation, and that it is this which is seen in (102) as opposed to (103) or (104). Ladd's examples of sentences which do not admit of a full-focus interpretation are therefore not cases of **marked** focus in our view, since any other focus structure would be much less normal: the difference between the accent placement in (102) and those in (103) and (104) is to our mind clearly one of **markedness**. A similar case can be seen in Ladd's (1980:75) example (repeated here as (105)) in defence of Normal Stress. Speaker B's reply in (105) clearly has Narrow Focus on *wonderful*, but it is in no way unusual or **marked** in the context: on the other hand, the reply in (106), despite being perfectly compatible with Broad Focus, gives a distinctly unusual emphasis to *man* and is therefore **marked** in our terms. It is the context, then, rather than any particular focus structure, which determines what the **unmarked** realisation of a particular sentence will be: this is similar to the Praguian notion of neutrality relative to a particular linguistic context (Daneš (1972:223ff.).

(105) A: What kind of a man is John?
B: oh, he's a WONderful man
(106) A: What kind of a man is John?
B: oh, he's a WONderful MAN

Our concept of **neutral** intonation is intended to cover these unmarked realisations, i.e. accent placements and boundary locations which can be predicted not from the syntax but from the **pragmatic context**, and which do not convey unpredictable contrast or emphasis. Thus, the cases of Narrow Focus produced by our rules in Section 3.4 are nevertheless examples of Neutral intonation since they are the least marked realisations ***in the context***. There are, of course, many examples of Neutral intonation which our rules cannot handle — the precise behaviour of *even* is an obvious example — but the most common causes of errors in our system's output appear to be handled satisfactorily by the rules described above.

3.6 Final Evaluation

The major modifications to our rules which we have discussed in this chapter appeared to improve our intonational output in several respects and to allow us to provide an appropriate prosodic treatment of various classes of phenomena which were previously problematic for our system and for TTS systems generally. However, despite the encouraging results of the various small-scale evaluation experiments described in Chapter 3, we felt that a more global assessment of our revised system was required to validate our approach to intonation synthesis and to demonstrate the high level of performance of our rules as a whole. The final section of this chapter presents such an evaluation and discusses the advantages and disadvantages of the experimental methodology chosen.

The primary aim of the evaluation experiment described in this section was to determine the perceived naturalness of the accents and boundaries which our rules produced. We decided that this final evaluation would be based on naïve listeners' judgements of the acoustic output of the system, despite the difficulties discussed in Section 3.1: an evaluation on this basis was seen as both complementary to the symbolic-level evaluations in preceding sections and necessary to demonstrate the capacity of our rules to generate high-quality synthetic speech. It was therefore necessary to compensate as far as possible for the phonetic and acoustic shortcomings of this output by using an experimental control of resynthesised rather than raw natural speech. This was intended to allow as fair an assessment as possible of the phonological specification produced by our rules independently of the phonetic quality of the synthesis, as well as giving some indication of the degree to which the degradation entailed by current signal-processing techniques mars the perceived naturalness of our output.

The problem of defining "naturalness" was avoided by the reasonable assumption that human listeners know what natural intonation sounds like and are therefore capable of judging synthetic stimuli according to that criterion. This is the main advantage of evaluating acoustic rather than symbolic output: we may be quite confident as to the validity of the listeners' judgements. Unfortunately, the use of acoustic output effectively precludes the assessment of running text, since there is no reliable way for listeners to judge overall coherence: listeners do not have the possibility of looking back at the previous sentence or paragraph, as the judges in Section 3.1 did; nor is it clear that lengthy passages of text can be reliably compared in any useful way. We were therefore obliged to base this evaluation on isolated sentences: however, this was not seen as a serious problem since our rules had already been evaluated on running text with good results in Section 3.1.

3.6.1 Method

The evaluation experiment essentially involved pair-wise comparisons of synthetic stimuli, with the subjects being required to state a preference for one stimulus over the other in each pair. As mentioned above, it is assumed that subjects' preferences will correlate with increased perceived naturalness, thus providing a reliable index of the relative quality of the stimuli. As far as possible, all factors other than the intonation contour were carefully controlled in the stimuli, so that "perceived naturalness" corresponds in effect to "appropriateness of intonation".

In order to obtain responses from a wide spectrum of listeners, the experimental task was performed by 165 normal-hearing subjects of both sexes: the subjects' ages ranged between 18 and 63 years, and they were drawn from a variety of backgrounds although the majority were university students. Subjects were remunerated for their participation. The experiment was administered by Hazel Sydeserff, and was performed directly after the experiment reported in Sydeserff et al. (1991) in which the same subjects responded to ninety multiple-choice questions presented in a mixture of synthetic and natural speech.

3.6.1.1 Stimuli

20 sentences were selected from the 200 sentences of the ATR phonetically-balanced sentence set developed at CSTR for the collection of speech databases. The principal criterion for selecting these sentences was the avoidance of long voiceless stretches which would make F_0 difficult to track both for speech analysis software and for human subjects. Each of the 20 selected sentences was synthesised in three different versions.

All three versions shared the same durational and segmental characteristics: these were derived from the transcribed natural speech in the ATR database, the values for the former being copied directly and the latter being replaced with the appropriate diphone units for the phonemic transcription. The diphones used were recorded from isolated nonsense words spoken by the same speaker whose readings of the ATR sentences we used: we hoped that this would help to minimise mismatches between segmental and prosodic characteristics in the stimuli. In addition, all three versions were synthesised using the PSOLA waveform-concatenation synthesis technique (Moulines et al. 1990) as implemented at CSTR by Paul Taylor. Only the F_0 contour differed between the three versions, and the different contours were generated as follows:

Version A: The syntactic analysis which forms part of the ATR database was used to drive the intonation assignment rules, and the resulting abstract intonational specification was aligned with the segmental tier using our standard alignment rules. This version therefore applied both the phonological and the phonetic models from our intonation model in deriving F_0, based on an independently-produced syntactic analysis.

Version B: The natural speech version of each sentence was transcribed, on the basis of auditory judgements and inspection of the output of an automatic pitch-tracker, to produce an abstract intonational specification in terms of the accents and boundaries of our model: this specification was then realised by the phonetic model and imposed on the segmental tier as in Version A. This version therefore only applied the phonetic model from our intonation rules.

Version C: The F_0 contour was extracted from the natural speech realisation of the sentence using an automatic pitch-tracker (Phillips 1985)[10], and was reimposed

[10] Evaluation of this pitch-tracker has been carried out at CSTR, using a corpus of 50 sentences from each of two speakers (one male and one female), gives a figure of under 2% for halving and doubling errors combined when compared with the output of a laryngograph. This figure compares favourably with the performance of other pitch-trackers on the same data.

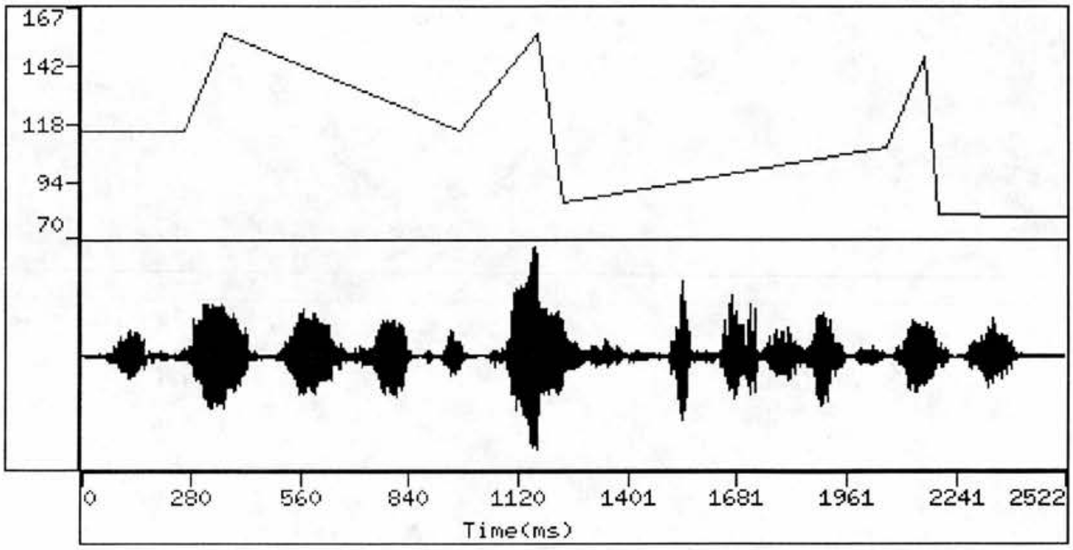

Figure 19: The Version A Contour Assigned to ATR Sentence 027:
i'll DRAFT those new proPOsals before the next MEEting

on the diphones of the segmental tier. This version involved neither the phonetic nor the phonological component of our intonation-generation software.

Version A, effectively the fully-automatic intonation contour, was the version whose quality was to be evaluated: the rules used to generate this version are given in Appendix C. Version C, the fully-natural version, was intended as a control against which to measure the performance of Version A: the reason for using re-synthesised F_0 and diphone concatenation was to keep the effects of signal processing and segmental quality constant for both versions. Version B was included to allow us to separate the influence of the phonetic model from the appropriateness of the phonological specification: Version B is as close an approximation to the original natural intonation specification at a phonological level as any TTS system could reasonably be expected to produce, and therefore a comparison of this with Version C was expected to reveal the extent to which inadequacies in the phonetic model impair the quality of our synthetic intonation.

The texts, transcriptions and syntactic analyses of the 20 chosen sentences are given in Appendix B: sample F_0 contours are shown in Figures 19-21.

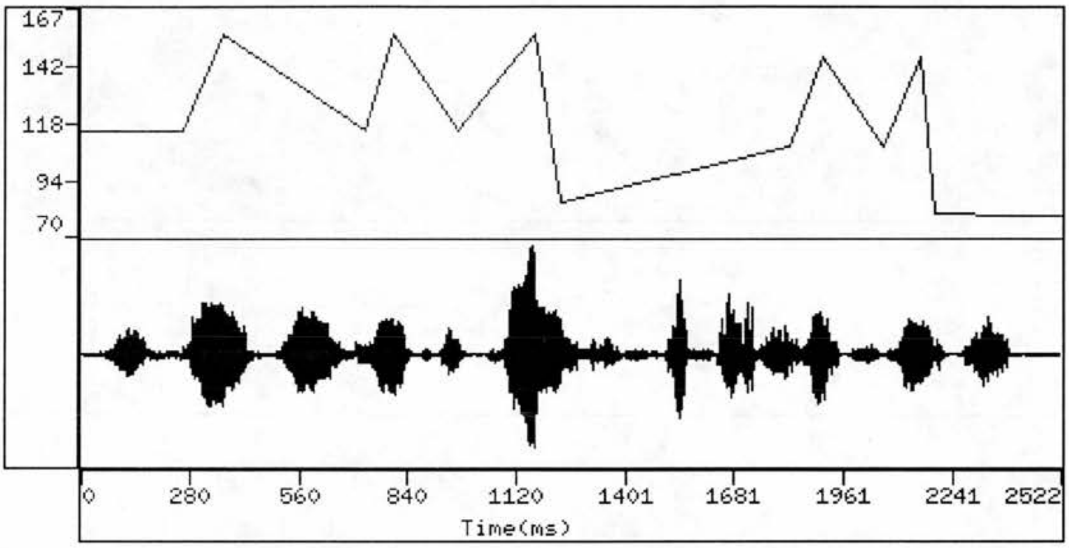

Figure 20: The Version B Contour Assigned to ATR Sentence 027:
i'll DRAFT those NEW proPOsals before the NEXT MEEting

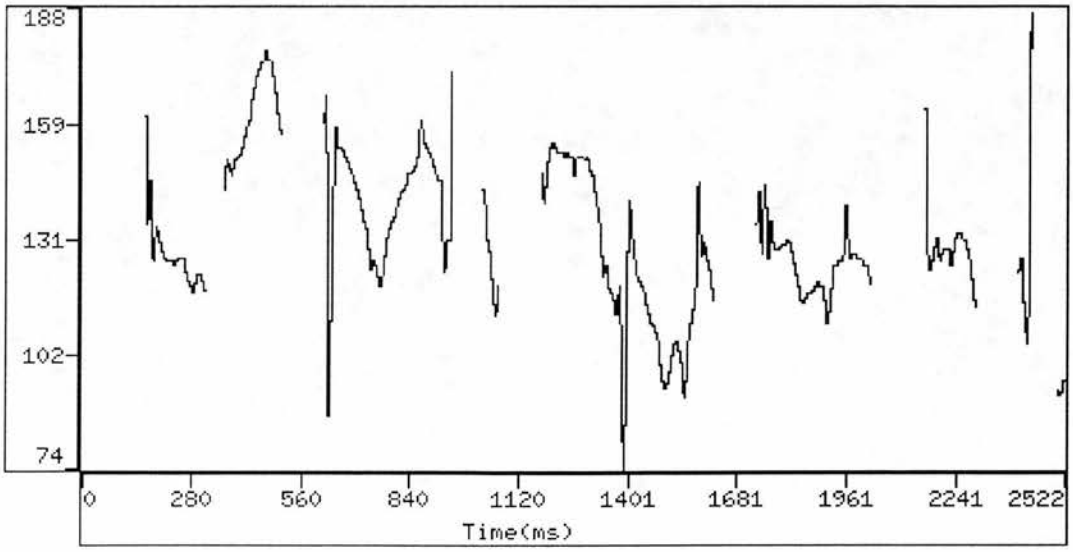

Figure 21: The Version C Contour Assigned to ATR Sentence 027

3.6.1.2 Presentation

Subjects were presented with pairs of stimuli which differed, if at all, only in their F_0 contours: a single ATR sentence served as the base for both members of each pair, providing the segmental and durational characteristics. Each of the 165 subjects was presented with a total of 50 stimulus pairs, which were made up from the three versions as follows: 10 A-C pairs, 10 C-A pairs, 10 B-C pairs, 10 C-B pairs, 3 A-A pairs, 3 B-B pairs and 4 C-C pairs. Each of the 20 ATR sentences formed the basis of 2 pairs (one A-C or C-A pair, and one B-C or C-B pair) and half the sentences also provided an X-X control pair (A-A, B-B or C-C).

For each pair, the subjects were asked to tick one of two boxes to indicate their preference. The instructions were printed on the answer papers, and the exact wording of the instructions which the subjects received was as follows:

> *In this section you will hear several pairs of sentences. In each pair, one sentence is natural human speech that has been distorted by a computer and the other has been generated by a computer trying to imitate the same talker. You are required to indicate which sentence in each pair is the natural version by ticking the appropriate box in the answer book.*
>
> *Ticking the left-hand box means:*
>
> *'First sentence of the pair is spoken by the human being'.*
>
> *Ticking the right-hand box means:*
>
> *'Second sentence of the pair is spoken by the human being'.*
>
> *REMEMBER:*
>
> 1. *You must judge and give an answer for each sentence pair.*
> 2. *There will be a short pause after each sentence pair, and you will be warned that the next sentence pair is coming up by means of a short tone.*
> 3. *After presentation of 10 sentence pairs, you will hear two short tones. This helps you to orientate yourself in the test.*
> 4. *After presentation of 30 sentence pairs you will be instructed to turn the page.*

The following reminder was printed at the top of the second page:

> *Tick the box which corresponds to the sentence which you consider to have been spoken by the human being.*

The stimuli in each pair were separated by a 2-second pause, and each pair was separated by a half-second tone at 500Hz and a gap of 2 seconds. The duration of individual stimuli varied between 2.1 and 4.3 seconds. Pairs were presented in groups of 10, with each group separated by a 2-tone sequence of 300ms duration and a 4-second pause. The order of presentation was randomised, with a constraint that the same ATR sentence did not form the basis of consecutive pairs.

The ten identical pairs were included as a check on subjects' preference for either the first or the second stimulus regardless of their relative perceived naturalness, since in the "forced-choice" design which we employed in this experiment the subjects have no "don't know" option.

The details of the presentation were determined in a small pilot experiment, and were designed to minimise subjects' boredom while allowing them ample time to make the required judgements. This pilot experiment found that subjects took very little time to make their judgements, but that larger numbers of judgements (80 in the pilot experiment) rapidly produced boredom and loss of concentration in many subjects. As a result of this finding, many of the control pairs were removed from the design of the main experiment. No subjects complained of boredom or insufficient time in the main experiment.

3.6.2 Results

For each condition (A-C, C-A, B-C, C-B, X-X) there were 165 × 10 = 1650 responses. Responses were transcribed by hand from the answer sheets, and a check on 10% of the answer papers revealed a negligible (less than 0.5%) error rate in transcribing the responses. The raw correct response scores for each condition are given in Table 5.

Table 5: Raw Scores and Percentages for the 5 Main Conditions

A-C	correct =	1045/1650	(63.3333%)
C-A	correct =	931/1650	(56.4242%)
B-C	correct =	1106/1650	(67.0303%)
C-B	correct =	828/1650	(50.1818%)
X-X	correct[11]=	865/1650	(52.4242%)

It is clear from the results in Table 5 that subjects found great difficulty in discriminating between natural speech and any of the automatic prosodic stimuli. Although all the results for the comparison of different stimuli (with the exception of C-B comparisons) are statistically significant ($p < 0.01$), which tells us that the stimuli are perceptibly different, the highest score which subjects achieved is only 17% better than chance (A-C comparisons, Table 5).

If the responses to the control pairs are broken down further, it can be seen that there is a much more skewed distribution of responses to the fully-automatic A-A pairs than to either the B-B pairs or the C-C pairs as shown in Table 6.

Table 6: Breakdown of Responses for Control Pairs

A-A	answered "1" =	278/495	(56.1616%)
B-B	answered "1" =	254/495	(51.3131%)
C-C	answered "1" =	333/660	(50.4545%)

All the scores for the control cases are not significantly different from chance, with the exception of the A-A comparisons where this significance is due to the subjects' judgements for one particular stimulus pair. The experiment can therefore be said to have been well-designed and sufficiently balanced: although there is a marked order effect in the scores for the test cases, this should simply cancel out since we have equal numbers of presentations in each order. Table 7 gives the scores for A-C/C-A comparisons and B-C/C-B comparisons with the scores for both presentation orders combined to cancel out the order effect.

[11] For the control identical pairs, "correct" corresponds to a response indicating the first member of the pair.

Table 7: Scores for Test Conditions Disregarding Order

A-C/C-A	correct =	1976/3300	(59.8788%)
B-C/C-B	correct =	1934/3300	(58.6061%)

The scores in Table 7 clearly show the difficulty which subjects experienced in deciding which stimulus was more "natural": subjects produced correct judgements in less than 60% of cases for both fully-automatic and semi-automatic stimuli compared with natural speech. We interpret these results to mean that the quality of our synthetic intonation is not reliably distinguishable from that of natural intonation in the great majority of cases, since subjects generally performed little better than chance in this discrimination experiment.

In order to make a more fine-grained evaluation of these results, we analysed the significance of the scores for each of the 40 test pairs using a Chi-Square significance test ($p < 0.01$). We assumed a probability of correct judgement of 50% in the null case, since none of the control pairs[12] showed a significant deviation from this. The results of this analysis are given in Table 8.

Table 8: Significance of Scores by Sentence

Case	% significantly correct	% significantly incorrect	% non-significant
A-C	60	0	40
C-A	70	20	10
B-C	70	0	30
C-B	30	20	50

The best scores, in the B-C case, show that subjects are only confident of the difference between the two stimuli in 70% of cases. By contrast, in the C-B case subjects are only identifying the correct stimulus reliably in 30% of cases, and they are also judging the automatic version as significantly more "natural" in 20% of cases. The overall conclusion to be drawn from the results in Table 8 is that subjects are not sure in most cases which stimulus is the human version: moreover, when the subjects

[12] With the exception of the one pair noted above.

are significantly confident about their preferences they sometimes prefer the automatic version to the natural one.

These results provide both a clear indication of the high degree of perceived naturalness achieved by the output of our intonation rules and a benchmark standard against which other systems can be measured. The viability of our experimental design is now amply demonstrated, as is the high quality of the synthetic output which we evaluated.

3.6.3 Discussion

3.6.3.1 Findings

The raw results in Table 5 clearly show that while subjects prefer Version C over either of the synthetic contours, which is what one would expect, the differences in scoring are nonetheless quite small: 67% correct judgements in the most extreme case, and barely over 50% in the least differentiated condition. These appear intuitively to be very satisfactory results for our intonation rules. Further analysis shows that in many cases subjects were unable to judge which was the more "natural" stimulus and in some cases they perceived the automatic intonation to be more "natural" than the human version. These are very encouraging results for an automatic system, but their interpretation as absolute indicators of any particular level of naturalness is quite problematic. Subjects' preferences as elicited in our experiment give no clear indication of absolute levels of preference. There are, however, various relative performance levels which can be deduced from the results of this experiment.

Firstly, it is clear that there are shortcomings in our current phonetic model. Table 5 shows a very large difference in naturalness between the two stimuli in the B-C cases, where the major empirical difference lies in our phonetic model. The fact that the output of our phonetic model is not entirely confusable with natural speech is not surprising, since there are several features of natural intonation which it simply does not handle. The two most obvious such omissions are microintonation and smoothing. Microintonation is widely acknowledged (Silverman 1984, 1987; Baart 1987; Gillott et al. 1990) to contribute significantly to the perceived naturalness of speech. Our phonetic model currently makes no allowance for this phenomenon, although we do have plans to incorporate segmental factors into the CSTR TTS system by extracting microprosodic perturbations of F_0 from natural speech (Monaghan et al. forthcoming). Similarly, our model currently constructs an F_0 contour from straight-line segments only with no smoothing of angles or transitions, which is quite different from what happens in natural speech where transitions are constrained by the physical properties of the larynx. We intend to incorporate some n-point smoothing, together with some slight random perturbation of the track of F_0, in an attempt to produce less precise and mechanical output, but this work has yet to be carried out.

Secondly, there is a considerable effect of presentation order in the results in Table 5: for all the test pairs, there is a strong tendency for subjects to prefer the second stimulus. There are several possible reasons for this, ranging from the question of what subjects are actually judging to how a subject's hand moves over the answer sheet, but no evidence from the present experiment to support a particular hypothesis. We will therefore refrain from further speculation. As pointed out above, the balanced design of our experiment allows us to ignore this skew in interpreting the results.

Thirdly, when presentation order is ignored there is very little difference in subjects' preferences between Version A and Version B when compared with Version C. There was no direct comparison between Version A and Version B, and it is therefore impossible to tell what differences subjects would have perceived in such a comparison. It is also possible that chance interactions between the prosodic and segmental characteristics of the various versions may have affected the perceived difference: a direct comparison might therefore have revealed a larger difference than the indirect comparisons which we performed, and the present results are consequently less conclusive than they appear. The control of prosodic and segmental interactions is a problem in any assessment of synthetic intonation, and is discussed at length below. However, it was certainly not our impression, on listening to the stimuli in advance, that there was any conspicuous effect of this kind. It has also been suggested (Terken personal communication) that there may have been an effect of excessive stimulus length which made it difficult for subjects to judge the stimuli with any confidence: however, examination of the data shows a correlation between stimulus length and subjects' judgements of only 0.277 (t=1.22 with

18 degrees of freedom, i.e. less than 90% confidence). We are therefore confident that the intonational specifications produced by our phonological model are of a similar standard to the best transcription which an automatic system could be expected to produce. This is of course less surprising in an experiment based on isolated sentences than it would be for running text, but it is a good result nonetheless. Although it is possible that subjects' judgements of the Version B stimuli were affected by some prosodic inconsistencies (see below), this is just as likely to have marred the perception of the Version A stimuli: it seems reasonable to conclude, therefore, that our automatic phonological specifications of intonation provide an adequate basis for quite natural-sounding synthetic output.

Given the absence of previous similar evaluations of our own or other systems, it is of course difficult to put the results of the present experiment in a wider perspective. The most closely comparable evaluations are probably those carried out by van Bezooijen (1989a) and van Bezooijen & Pols (1989) to assess the output of Quené & Kager's (1989) PROS algorithm, and the less ambitious evaluation of prosody in French TTS systems which is reported in Benoit (1990). However, the present experiment differs from these in several important respects. Van Bezooijen's experiments make no use of fully-automatic prosody, all versions being hand-edited to some extent, and rely heavily on the interpretation of a pre-defined 10-point scale on which subjects are asked to score stimuli: we have attempted to avoid both these factors by avoiding manual intervention in all versions and by giving subjects a simpler choice. With respect to Benoit's experiments, an important difference between his work and our own is that Benoit tested intelligibility rather than naturalness, and this together with his choice of monotone F_0 as a reference or control condition makes it difficult to compare his results with ours: we would not be surprised if speech on a monotone were comparable with our prosody or with natural speech in terms of intelligibility, particularly when assessed simply on a sentence- or word-count basis, but it is certainly not equally natural! We feel that our approach has significant advantages over these previous experiments, and has set a new standard for the evaluation of synthetic intonation in acoustic output.

3.6.3.2 Problems

Not surprisingly, as the present evaluation is the first example of an original experimental approach rather than the latest in a long line of similar investigations, there are certain problems with both the materials and the design. Although none of these is felt to be a major failing, they will be discussed here in order both to clarify the significance of the present results and to indicate problems which subsequent investigations may be able to avoid.

Regarding the materials used in this experiment, there are two minor problems which merit discussion. Firstly, in the preparation of the stimuli which we used it was necessary (because of the limitations of the signal-processing techniques available) to prepare the Version C stimuli in a slightly different manner from the other stimuli: in Version C, the natural F_0 contour imposed on the segmental information was extracted from the original natural utterance by an automatic pitch-marking algorithm, whereas in the other two versions the F_0 contours were generated automatically and aligned with diphones whose pitch-marks had been checked manually. This resulted in a slightly different synthesis quality for Version C stimuli, which may or may not have assisted listeners in identifying the "natural" member of the stimulus pairs. This problem is of course solvable, but not immediately, and may have unfairly affected the relative performance of the more automatic intonation versions.

Secondly, given the well-attested interaction between pitch-prominence and other forms of prosodic prominence such as duration, amplitude and vowel quality (Bolinger 1958; Vanderslice & Ladefoged 1972; Vaissière 1983), there is a possibility that listeners' perceptions of the naturalness of stimuli was affected by the coherence (or lack of coherence) of several prosodic factors rather than simply the quality of the intonation contour. In the case of the Version B and Version C stimuli particularly, lack of coherence could have resulted from differences in amplitude in the original natural utterance which we were not able to reproduce in the PSOLA diphone synthesis: it is known (e.g. Vaissière 1983; Batliner & Nöth 1989) that speakers differ in the degree to which they employ variations in amplitude in addition to F_0 cues to convey prominence, so that although the F_0 contours were carefully controlled the overall impression of prominence may not have been as natural as in the original. However, these differences are likely to have been negligible compared with the possible impression of prosodic incoherence in Version A: here, both the vowel qualities and the segmental durations of the stimuli were derived from the natural accentuation whereas in many cases the automatic accentuation was quite different. The close interaction of F_0 and duration is generally accepted, and our own work on synthesis has demonstrated that durational cues can even override F_0 cues, particularly at boundaries, so that inappropriate combinations of accentuation and duration may well have affected listeners' perceptions of the Version A stimuli. Evaluation of the CSTR duration model (Monaghan 1991a) has shown that its output is of very high quality, and it might be preferable in future evaluations to use synthetic rather than natural durations in combination with synthetic F_0 to avoid the problem of prosodic mismatches. The rôle of vowel quality in the perception of prosody is less well understood (Koopmans-van Beinum & van Bergem 1989; van Bergem & Koopmans-van Beinum 1989), but again it is possible that the interactions between vowel quality and accentuation are such that inappropriate amounts of vowel reduction in Version A interfered with listeners' perceptions of naturalness. It is of course possible to alter the diphone specifications in the relevant cases: however, all these modifications are taking us further and further from a direct and controlled comparison of intonation contours. As is so often the case with language, there is no easy way to control all the relevant factors.

A central assumption of the experimental design is that naïve subjects are capable of judging the (relative) naturalness of synthetic stimuli. This assumption has been made by several previous evaluation studies (e.g. van Bezooijen 1989a; Scheffers 1988; Terken 1989a), and indeed van Bezooijen and others (Nooteboom & Kruyt 1987; Willems et al. 1988) have made the stronger assumption that listeners can judge the ABSOLUTE naturalness of such stimuli. The results of the present experiment appear to support the weaker assumption at least, in that there are clear differences in listeners' judgements and these differences conform to what we would have predicted, namely that the more "natural" contours are preferred to the more "synthetic" contours. Nevertheless, by avoiding the problem of defining rigid criteria of naturalness we have created a different problem: how do we know that what listeners are judging is naturalness rather than anything else? The answer is: we don't, but this is not in fact a serious problem in our view. Unless the main object of synthetic intonation is to play a part in passing the Turing test, there is no reason to believe that naturalness is actually the appropriate assessment criterion: indeed, some authors have argued (Hirschberg 1990a; Blauert & Schaffert 1985) that people do not want synthetic speech which is indistinguishable from human speech, and portrayals of fictional intelligent computers in radio and television broadcasts appear to confirm this view. What is definitely important for a practical system is whether it conforms to human preferences, and it seems likely that if some factor other than perceived naturalness has influenced listeners' judgements in the present experiment then that factor is precisely these preferences. We maintain, therefore, that even if we cannot be sure that speakers are assessing the undefined "naturalness" of the stimuli we can rely on them applying a criterion or criteria of equal relevance and importance to the assessment of synthetic speech.

3.6.3.3 Future Work

The present experiment is easily replicable for other TTS systems, and puts the onus on such systems to measure themselves against the performance of our intonation rules. Appendix B gives most of the necessary material for such replication, and we would be very interested to see others perform parallel evaluations so that quantitative inter-system comparisons may be reliably drawn.

In the same spirit, we intend to carry out further evaluations of our intonation rules, both by replicating the experiments of others and by extending our own experimental approaches. Despite the problems which we mentioned above regarding the methodologies of experiments such as van Bezooijen's and Benoit's, it would be both interesting and informative to evaluate our system along similar lines and allow both qualitative and quantitative comparisons. Similarly, it would be interesting to assess the affective information conveyed by various versions of synthetic prosody on criteria similar to Uldall's (1960) emotional and functional scales, or on the basis of the results in Ladd et al. (1985). A more telling evaluation metric, however, would be the performance of our rules on running text assessed on the basis of acoustic output: this would combine the best of both of the symbol-based and acoustic-based formal evaluation techniques introduced in the present thesis. Unfortunately, the problems mentioned here and in Section 3.1 currently preclude any meaningful assessment along those lines: we must first devise rigorous assessment criteria (by defining naturalness, acceptability, and similar notions) and then produce intonation which can claim to take account of text structure and meaning. Until we have done this, we must be content with partial evaluations.

Chapter 4

Summary and Conclusions

> it would be misleading to suggest that TTS synthesis should no longer be considered as a valid research topic. None of the currently available systems can as yet be mistaken for human speech.
> Moulines et al. (1990:310)

4.1 Summary

In the preceding three chapters we presented a set of rules for generating intonational specifications from unrestricted text. We set out the theoretical assumptions which underlie our work, and we discussed the performance of our rules in a variety of formal and informal evaluation experiments. The development of our rules was presented as a gradual refinement of default specifications on the basis of the results of these evaluations, such that the shortcomings of the implementation motivated improvements to the theoretical model. This section summarises the main points of this development, and presents the resultant model.

4.1.1 Intonation from Text

Chapter 1 introduced the problem of generating intonation in a text-to-speech conversion system, and discussed the importance of good-quality prosodic output (in particular intonation) for systems producing synthetic speech. We outlined the CSTR TTS system, presenting the phonetic and phonological models of intonation in some detail (Section

1.3), and discussed numerous other systems' theories and implementations of intonation generation. The nature of the TTS system developed at CSTR, and the findings of the various authors discussed in Section 1.4, led us to constrain the task of our rules and to adopt certain strategies for accomplishing that task.

It was clear from the literature reviewed in Section 1.4 that we could not extract enough information from text to produce the most appropriate intonation in all cases: indeed, such consistent performance is beyond most human readers. We therefore restricted our rules to producing one of the many **acceptable** intonational specifications, and specifically to producing a **neutral** specification which did not manifest any special understanding of the implications of the text to be read out. Such a Neutral rendering explicitly excluded unusual emphasis, contrastive intonation or stylistic variation, but was not more closely defined at this stage.

Even with these restrictions on the coverage of our rules, the requirement of producing an acceptable intonation from unrestricted text in a real-time TTS system was still a very ambitious one. Not only is the linguistic information available from text very limited in automatic systems, but the theory of how linguistic information maps onto intonational realisations is at best vague and equivocal. We therefore adopted a strategy of **default specification**: this allowed us to implement those ideas which were generally accepted or which were relevant to the information which was available to our rules, while ignoring other factors until the problems with our partial implementation were known. In addition, in view of the scarcity of reliable higher-level linguistic information to drive our rules, we decided to make use of any and all information readily available from text if it appeared to correlate with a particular prosodic treatment.

As a final restriction, we limited the intonational phenomena which our rules would specify to two types of phonological event: accents and boundaries. These were the phenomena which appeared from the literature to be most problematic for automatic systems, and they were also the phenomena whose acceptability and interpretation was the least ambiguous: other phenomena, such as peak height, contour choice and excursion size, were considered to be less "basic" to the intonational specification and much more difficult to predict or interpret.

4.1.2 Accents & Boundaries

Default rules for assigning accents and boundaries to unrestricted text on the basis of crude lexical and syntactic information were presented in Chapter 2, and the development of more specific rules to handle a variety of exceptional cases was discussed in Chapter 3. There are essentially three components to our generation of accents and boundaries: accent assignment, the Rhythm Rule, and domain assignment.

Our accent-assignment rules use lexical information to assign varying degrees of intonational prominence to the lexically-stressed syllables in their input. The default treatment is to assign primary (potentially nuclear) accents to nouns and proper nouns, secondary (non-nuclear) accents to verbs, adverbs and adjectives, and no accent to other forms (i.e. function words). There are thus three degrees of accent assigned by these rules. No distinction is made between primary and secondary lexical stress, except that our definition of secondary lexical stress requires that it correspond to a syllable which undergoes stress-shift: this definition allows us to incorporate the phenomenon of stress-shift into our rules very simply.

The accent-assignment rules assign more accents than are normally realised in fluent speech, and so this **over-accented** output generally requires to be reduced. Our Rhythm Rule applies to the output of the accent-assignment rules and reduces or deletes some of the accents: there are various stages to this rule, allowing it to produce a range of accentuations corresponding to increasing reduced, casual or fast speech. The output of the Rhythm Rule may therefore be varied according to the speech style which is required, but in the default case all stages are applied. The basic principles governing the operation of our Rhythm Rule are those of rightmost prominence and rhythmic alternation: the rightmost primary accent is identified as the nucleus of the current domain, and all other primaries are reduced to secondaries; these secondaries are then selectively deleted to produce an alternating pattern of accents and deletions. (Post-nuclear secondaries are all deleted, since the nucleus is by definition the final accent in the domain.) Syllables whose accents are deleted by the Rhythm Rule still retain more prominence than syllables which were never assigned an accent: these syllables are marked with a 'd' diacritic to indicate a **deaccented** syllable, and are assigned extra durational prominence but no pitch prominence in our current system.

The output of the Rhythm Rule encodes four degrees of prosodic prominence per syllable, and the same four degrees of prominence per word: unaccented, deaccented, accented, and nuclear. These distinctions form the basis of duration and vowel quality assignment in the CSTR TTS system, as well as the assignment of F_0 contours, but the last two degrees are collapsed for all purposes other than F_0 generation because we do not yet know how to make use of the distinction between them.

The input to the Rhythm Rule is a single domain from the lowest level of our domain hierarchy. This hierarchy expresses the relations between intonational domains, and is represented in terms of a sequence of boundaries whose phonetic interpretation involves manipulating the height of the current **register** and assigning **boundary tones** in certain cases. There are three levels of domain which are assigned in our default treatment: tg(2), corresponding to a full text sentence; tg(1), corresponding to a syntactic clause; and tg(0), corresponding to certain types of major syntactic phrase. The syntactic basis of these definitions is the result of constraints on the information available to our rules: intonational domains are not syntactic in nature, and there are many exceptions to these syntactic defaults, but syntactic structure is the best approximation to intonational structure which most TTS systems produce and the correspondence between syntactic and intonational domains holds in the majority of cases.

There is no reason why our hierarchy of domains should be limited to three levels, and in fact more levels would be required to handle the structuring of sentences into paragraphs, topics, and larger structures. The principle of embedded register settings (illustrated in Figure 7) is extendable to include these additional levels, although their realisation appears to be more complex than that of intra-sentential domains (Silverman 1987). There is also no reason why our domains should not be redefined to reflect different granularities of structure: finer structures seem to correspond to slower speech rates, and coarser ones to faster rates. Such redefinition, scaling all domains down or up, together with the flexibility of our Rhythm Rule allows us to produce a very large variety of different specifications for the same text: a wide range of styles is therefore available in the output of our rules. However, the interaction of accents and domains is quite complex and not all combinations may be acceptable.

There are numerous regular exceptions to our default rules which have been identified and whose special treatment is described in Chapters 2 and 3.[1] Exceptions to our accent-assignment rules include some of the **anomalies** discussed in Section 3.3, certain adverbs and conjunctions (examples (6a-f) above), and most of the cases of anaphora in Section 3.4. The treatment of these items involves the deletion of the accents assigned to them or the assignment of the 'd' diacritic directly by the accent-assignment rules: occasionally, as in the case of abbreviations, the accent-assignment rules are bypassed altogether.

Exceptions to our Rhythm Rule are quite rare, since its underlying principles of rhythmicality and domain-final nuclei are very general in their applicability. The only domains to which the Rhythm Rule does not apply are those consisting solely of dates: these anomalies are assigned rhythmic accents by special rules and always constitute a domain in their own right, as discussed in Section 3.3.3, and there is therefore no need for any further rhythmic adjustment. In all other cases, the Rhythm Rule applies to domains to determine accent placement on the basis of rhythmic alternation and degrees of accent.

There are several exceptions to our default domain-assignment rules: as we stated in Chapter 2, our syntax-prosody interface INTERFIX is the point where most of the exceptions to our regular rules are identified and given special treatment, and domain assignment is the heart of INTERFIX. The major exceptions discussed above are the sentence-final PPs which form separate tg(0) domains, the various anomalies discussed in Section 3.3 which all trigger tg(0) boundaries, and the interpretation of punctuation either as the special cases of lists and parentheticals (Figure 8) or as reinforcers of an existing domain boundary. The current version of INTERFIX handles all these exceptions in the same left-to-right fashion, splitting any input into a linear sequence of domains which is then processed by the more regular rules: our interface program is

[1] Some more common (and more easily remedied) exceptions are described in Ladd & Monaghan (1987).

thus doing precisely what we intended, interpreting whatever information is available from text and producing regular output where exceptional cases have either been marked for special treatment or converted to a form where no special treatment is required.

All the exceptions to our default rules are triggered by the availability of more information than is allowed for in those rules. As yet more information becomes available, and more exceptional cases are identified, so there will be more work for INTERFIX to do and perhaps more exemptions from our accent rules: however, all our default rules still apply in the majority of cases and we see no reason why they should not continue to act as the foundations of our intonation rules. Similarly, the WFCs presented in Section 2.5 continue to apply to all domains regardless of their source: empty domains and accentless domains are handled appropriately by these constraints, as are the initial and final accents of a tg(2). Although it is possible that additional WFCs would improve our output, we do not foresee any cases where our existing WFCs will not apply: if the definition of a tg(2) were modified to produce a different speech rate, for instance, the constraints on initial and final accents would still apply; and if a new source of domains (such as postal addresses or mathematical formulae, for example) were incorporated in our rules we would still expect such domains to be subsumed by neighbouring domains in cases where they were assigned no accents.

It appears, then, that our specification of intonation in terms of accents and boundaries allows us to provide appropriate treatments for all the cases which we have so far encountered, and that the defaults which we have specified provide an **acceptable** intonational specification for the majority of running text. In addition, it is clear that our strategy of defining a default treatment and then allowing for increasing numbers of exceptions to that treatment is sufficiently robust and flexible to handle almost all cases of **neutral** intonation in the texts we have processed, since none of the exceptions listed above have required us to modify this strategy.

4.1.3 Evaluation

Evaluation, both formal and informal, has directed most of the development and refinement of our rules described above. As we pointed out in Chapter 1, the output of our rules provides the best indications of their shortcomings: all our efforts have been directed towards correcting observed errors in that output. Although we have attempted to tackle our system's errors in a principled and linguistically-informed fashion, and to discover why certain approaches were successful while others were not, the overriding criterion which determined whether a rule was added or modified was the question of whether it improved our output.

In the absence of established evaluation procedures for intonation in TTS systems, we have developed two formal paradigms which both suited our purposes and are applicable to a range of different TTS systems and to various aspects of synthetic speech. Section 3.1 discusses the problems involved in assessing synthetic intonation and the requirements which must be met by any evaluation procedure, and presents a procedure for evaluating symbolic representations of intonation for running text. The results of this procedure were shown to be useful for both documentary and diagnostic purposes, and we demonstrated that expert judges could consistently assign scores to symbolic representations. The same procedure was applied to isolated sentences and phrases in Section 3.3.7, with even more consistent results.

The final section of Chapter 3 presents a second formal evaluation experiment, this time assessing the relative "naturalness" of acoustic output from both human speakers and automatic systems. The methodology used in this experiment produced indications of both the relative and the absolute naturalness of different methods of producing intonation contours: this procedure did not require expert judges, so that large numbers of subjects could be used, but it was restricted to the comparison of isolated utterances rather than larger units of text.

The results of all the evaluation experiments presented above have shown that the output of our rules has a very high degree of acceptability and naturalness. The experiment in Section 3.1 showed that our default rules produced acceptable accent placements for unrestricted running text in 68% of cases, with less than 4% of cases producing seriously unacceptable accentuations: this was very encouraging, since running text is the most stringent test of synthetic intonation. The main causes of errors in this experiment were addressed in the remainder of Chapter 3, and formal and informal evaluations discussed there indicated that considerable improvements had been made to our rules:

in particular, the assessment of our rules for handling anomalies (Section 3.3.7) showed that this group of problematic phenomena which accounted for five of the errors in Section 3.1 was now assigned acceptable accentuations by our improved rules in more than 94% of cases.

The results of the final evaluation experiment presented in Section 3.6 show that in isolated sentences our automatic intonation contours are not reliably distinguishable from natural intonation approximately 85% of the time. Although this evaluation is limited to single-sentence comparisons for methodological reasons, it is clear that the quality of our synthetic intonation can reasonably be compared with the intonation of natural read speech. There are various problems and uncertainties with our original methodology used in Section 3.6, but despite these the results are very clear and extremely positive. If the shortcomings of the phonetic model are taken into account, it is reasonable to claim even higher performance for our phonological rules: however, the interactions between the various aspects of prosody in acoustic output make it impossible to isolate the effects of this model.

In summary, the work presented above has produced significant contributions to both the generation and the evaluation of synthetic intonation, and has demonstrated the validity of our general approach to intonation in TTS systems. We have produced and implemented a set of rules which generate highly naturalistic phonological specifications of intonation automatically from running text, without the assumption of unrealistic amounts of linguistic analysis. Our rules are flexible and robust, and could be easily incorporated in most current TTS systems: their implications for future work in the areas of intonation and the development of TTS systems are discussed below.

4.2 Conclusions

The main purpose of the work presented in this thesis was to investigate the potential of a particular approach to producing intonation in a TTS system. This approach involved combining the insights of linguistic theory with the performance-based criteria applicable to a working system. In our development strategy, considerations of theory and performance worked in parallel to produce an implementation of a theoretical model whose errors were then analysed and led to modifications of the model. The implications of both our model and our working system for theoretical and practical work on generating intonation, and for the general characteristics of TTS systems, are discussed in this section.

4.2.1 Intonation

The term "intonation" can cover several aspects of the phonetic realisation of speech, including local and global pitch characteristics, timing, speaker-dependent factors such as height and range of the voice, and physiological factors such as microintonation. The relation of some of these factors to the work described in this thesis is discussed in Section 1.3, but they are for the most part ignored in the present work in favour of the phonological representation of intonation. We argue that, at least in cases of neutral intonation, a phonological representation solely in terms of **accents** and **boundaries** is sufficient to characterise the phonetic output, and we have demonstrated that such a representation is capable of producing highly natural-sounding intonation in the output of a TTS system. Although there are obviously aspects of intonational phonology, such as tune choice and emphasis, which we have ignored in our specifications, these aspects do not seem to play an important part in neutral readings of text.

The model of neutral intonation which we propose above takes the function of intonational accents and boundaries to be the indication of **focus**. Breadth of focus is indicated by the position of the main or nuclear accent in a domain: every utterance has some [+focus] constituent, and the size and number of such constituents determine the focus structure of the utterance. Some constituents, such as non-restrictive relatives, intransitive predicates, parentheticals and topicalised items, have domain boundaries at various hierarchic levels associated with them: however, these boundaries may only be realised if the constituent is [+focus]: the presence or absence of such boundaries is therefore a further clue to the focus structure of an utterance. Within any [+focus] domain, the position of the nucleus is determined by the relative positions and accentability of the various constituents: the rightmost of the most accentable constituents will be assigned the nuclear accent. Other accents in the domain are then optionally assigned to pre-nuclear accentable items, with the condition that these accents fit a regular rhythmic pattern for the entire domain.

For any utterance in a particular context, there is a predictable or **unmarked** focus structure which dictates a similarly **unmarked** accentuation and division into domains. This focus structure is determined both by contextual factors and by the interpretations associated with particular lexical items (such as *even* and *only*, or pronouns and reflexives) and syntactic structures (such as fronting, clefts and inversion): focus structure is therefore determined by choices at various linguistic levels, producing incidental correlations between these levels and the realisation of focus structure as intonation. Speakers' assignment of focus is a completely free choice: **marked** focus structures are assigned to convey particular emphasis or attitudes, and it is only on the basis of the deviation from the **unmarked** case that listeners are able to interpret such uses.

Although speakers have a number of choices as to how they indicate focus structure, such as the choice of deaccenting an item or of employing a particular syntactic construction, the intonational choices at least are binary: the treatment of a particular item, be it a syllable or an entire domain, is either the default [+focus] treatment or the converse [-focus] one, and there are no other options. There may be several such choices to be made for each item (such as whether to pronominalise an NP, whether to deaccent all or part of it, and whether to move the whole constituent), so that the effects of different choices may interact and produce a less clear picture, but it appears that the intonational choices are strictly limited.

Speakers may choose a **marked** intonational realisation for any number of reasons, but the realisation in itself does not convey the specific reason: the listener must deduce the reason on the basis of other information. We take the view that in general the use of a **marked** intonation draws attention to the constituent whose accentuation is affected, and informs the speaker that the usual assumptions regarding this constituent do not hold. Thus, in (106) above speaker B has chosen to give *man* a marked intonational realisation. The most obvious interpretation of this realisation is that John's qualities as a man contrast with his qualities in some other rôle, but this "contrastiveness" is not conveyed directly by the intonation: there is in fact no necessary difference between the intonation in (106) and the same intonation in a context where it constitutes an **unmarked** realisation (Ladd 1980:75ff.). Indeed, it has been shown (Eady & Cooper 1986:402) that such "contrastive" uses of intonation are phonetically identical to more "normal" uses. What has happened in (106), then, is simply the assignment of a marked focus structure: *man* should be [-focus] for a neutral accentuation in this context, but the speaker has chosen to make it [+focus]. The fact that this produces a marked realisation of *man* forces the listener to construct an explanation for the speaker's choice of focus structure, and in the absence of other information the most obvious explanation is the "contrastive" one mentioned above. This "contrastive" interpretation does not, however, involve the basic **man/woman** contrast or any similar primitive contrast, as this would clearly be quite implausible in the context and it is the context (rather than the semantics of the intonation contour) which provides the interpretation. If we change the context by adding more information, *à la* Bolinger, we can produce a different interpretation and in so doing illustrate the importance of context rather than accentuation: if we know, for instance, that John is a werewolf then the same accentuation as in (106) produces quite a different interpretation (a **man/animal** contrast which was impossible in (106)):

(107) A:What kind of a man is John?
B: oh, he's a WONderful MAN, but when there's a FULL MOON
he's DREADful!

In our rules, all relations between domains are completely determined by their position in the domain hierarchy which is in turn determined by their constituents and the current speech rate. We have found no reason to vary the realisations of the boundaries between domains to reflect particular semantic or other properties of their relations, despite the evidence (e.g. Ladd 1988) that such properties may affect the realisation of domain boundaries: nor have we found any reason to construct metrical or other hierarchic representations within domains to account for accent placement. It seems to us that the principles of focus and accentability are quite sufficient to produce natural-sounding neutral intonation, and that the realisation of metrical structure within or between domains is largely optional and redundant. It is our impression that prosodic structure, like syntactic structure, need only be partially specified above the foot or phrase level: such underspecification certainly suffices for our purposes in generating intonation, although our default specifications may be inadequate for less neutral realisations.

For present purposes, then, we have found that a phonological specification of intonation in terms of accents and boundaries produces very good results. In our system, the placement of these intonational events is determined by two sets of rules. The first set looks for exceptions to our default assumption that all input is [+focus]: these rules can only identify the exceptions discussed in Chapter 3, and they mark such exceptions for deaccenting. The second set chooses between our default accentuation strategy (discussed in Chapter 2) or our deaccenting rules (Chapter 3) on the basis of the decisions made by the first set. In cases where our rules can reliably predict [-focus] items, our output is highly acceptable: in the cases where there are errors in the accentuations which are finally produced, these are generally due to mismatches between the actual focus structure and the system's predictions of deaccenting. We feel, therefore, that intonation can be generated solely from our general principles of accentability and rhythmicality, together with a knowledge of focus structure[2]: there is no need for detailed syntactic information. Furthermore, it is possible to restrict intonational choices to the binary opposition of [+focus] and [-focus] realisations, and to simulate an **unmarked** focus structure on the basis of contextual information. These claims are based on our observations of synthetic intonation for written monologue, and may not be applicable to spoken dialogue or other styles of speech, but they constitute a coherent model of intonation for synthetic speech which we feel is both useful and interesting as the basis of a theory of intonation.

4.2.2 Text-to-Speech Conversion

In developing our rules as part of the CSTR TTS system, we have attempted to combine the best aspects of the many TTS systems discussed in Chapter 1. Our approach has

[2] Our rules neither construct nor make use of any representation of focus: the focus structure is deduced by the listener from the presence or absence of accents and boundaries.

been founded on linguistic theory wherever possible, but we have also attempted to make use of any less principled correspondences between text and speech revealed by our own work and that of others. Our rules constitute a coherent whole and reflect certain theoretical principles and assumptions, but we have attempted to make use of any and all sources of information which might improve the quality of our output: this seems to us to be the only reasonable basis for tackling the problem of TTS.

As we stated in Chapter 1, there are two major aspects of the output of a TTS system: segmentals and supra-segmentals. The task of such a system is to produce natural-sounding realisations of both aspects, and theoretical considerations must definitely come second in any TTS system's priorities. In our view, therefore, the purpose of all the rules in a TTS system is to contribute to the quality of these two aspects: if a theoretically-interesting lexicon hampers the construction of a segmental realisation, or if the output of a sophisticated parser is not what the prosodic rules require, there is little point in incorporating such modules in a working system. On the other hand, if a lexicon of only 200 items such as that in the PROS system (Kager & Quené 1989) produces the required information then that is what the system should use and the fact that such a lexicon is adequate may have interesting theoretical implications.

Our approach to generating intonation in TTS has resolved many of the problems which TTS systems as a whole have found problematic: the treatment of anomalies and anaphora described above is achievable by most TTS systems, and constitutes a significant improvement on previous work in these areas. There are, however, still serious problems which remain to be solved and whose proper treatment is beyond the scope of this thesis. The information required to produce non-neutral realisations of text, for instance, is still well beyond the reach of current text analysis: G. Hirst (1981:31-2) concludes that the knowledge which an automatic system would require to produce accurate semantic and pragmatic analyses of text is so large that "a solution may not exist." A more tractable problem which was mentioned above and which we have not addressed in any detail involves the interaction of the different acoustic parameters which realise prosodic prominence: we have concentrated on F_0, and touched on duration, but the interaction of these together with amplitude and segmental quality still requires a great deal of basic research. The duration rules applied by the CSTR TTS system

(Campbell 1989, 1990) produce very natural-sounding output (Monaghan 1991a), but we currently exercise no control over amplitude although it is known (Lieberman & Michaels 1972) that variations in amplitude contribute significantly to the perception of prosody. We intend to implement basic amplitude variation in our system, producing smooth amplitude envelopes which taper at the beginning and end of an utterance and which expand at intonational peaks, but such gross variation has been shown (Richter 1984) to account for only part of the amplitude variation found in natural speech. The fine control of segmental quality is currently impossible in a waveform-concatenation system such as the CSTR TTS system, and it will therefore be some time before any TTS system can control all these prosodic variables appropriately.

There is therefore much work still to be done in producing a TTS system which might attempt the Turing test, and although much of that work will affect the perceived quality of synthetic intonation the research which is required lies more in the areas of machine understanding and of basic phonetics than in that of intonational phonology. Nevertheless, we can justifiably claim to have successfully addressed the problem of assigning accents and boundaries to running text which Akers & Lennig (1985) identified as the major problem with synthetic intonation.

Appendix A

Symbolic Evaluation Data

This appendix contains the orthography and accent patterns of the four texts which were assessed in Section 3.1. The tg(0) boundaries are indicated by the vertical bar |.

A.1 Text 1

```
  2      -    -  d     -    2     d  -    -    2      -
Cricket: On the fourth and final day of their match against

d        -  1      |  d    1   |
England at Rajkot, | West Zone |

    d    -     2   d      -    2        d   -   1
declared at three hundred and ninety three for seven.

1        |   -  -    2    d      -    2    d     -      2
England, | who were four hundred and fifty eight for three

-    -    d    1        |  -    d   2    -  -    d   -   1
in their first innings, | were fourteen without loss at lunch.
```

```
-   -    -  2      d   -    2         -  d   1      |
And in the second test in Brisbane, the West Indies |

 -     d  2      -    d       2   -   d   -   2
were two hundred and sixty three for five at tea,

-      d  -      2       d    2        d    -  d    2
in reply to Australia's first innings total of one hundred

-    d       1
and seventy five.

 2      1         | -  2    d      -     1     -  -
Richie Richardson | is one hundred and twelve not out.

 2       |   -  1          -
Finally, | the headlines again.

-  2      1     |  -    2     -     d
Mr Enoch Powell | has praised Mrs Thatcher

 -    2       d   -    - 2     d      1
for standing firm at the Anglo-Irish summit.

  - 2    d      1         -       | -   2           d
The overseas development minister | is visiting Ethiopia

  -   1      |  -  d    -  2        d   d   1
this weekend | to see the famine relief operation.

 2    2    -    d     1   |  -    -    d       2        d      -
High winds and heavy seas | have been causing further problems in
```

```
  -  2        d   -    1      |
the southern part of Britain, |

 d       2      d     -    1      -
leaving homes flooded and roads blocked.

-     -  2    d     -   1       |
And the main news this morning: |

-   d       1     |  -    d   -  1     | -
A thousand people | were led to safety | after

 -      d     -  -  2   -    -  d     2            d    1
being trapped by a fire in the London Underground last night.

 -    -   -  2     -     -   2    -   -  d        1
Many had to walk along the track to the nearest station.

2d2  d     1
BBC Radio News.

-     2   d   -   1
It's now ten past eight.
```

A.2 Text 2

```
   -  2    -   -     2      d    -   d       2
The leader of the Alliance Party in Northern Ireland,

-   d    1         |  -  d       -  2       d        2
Mr John Kushnahan, | has asked the Northern Ireland secretary,

-   d   1    |  -     2     -  d       -    -
Mr Tom King, | to suspend the business of the

 2       d          1
Northern Ireland assembly.

-   1        |  2   |   -    1     |  -   -      d
Mr Kushnahan | says | the assembly | has been effectively

 2        -   - d         1
hijacked by the Unionist parties.

   -  1        |  -   d       -    2   -  -       d
The Government | has welcomed a report by an Australian

 2        d      -    -    2    -    d
Royal Commission on the effects of Britain's

  2     d    2        d       -    -      2      d     -
atomic bomb testing programme in the Australian desert in

  -  2      -   d      1
the fifties and early sixties.
```

```
  -    2     d        -   d        1    |  -     d     -  1     |
The Defence Minister, Mr Norman Lamont, | has accepted an offer |

 -  2    d    -    -  2        -        d        1
to hold talks on the findings with Australian officials.

  -  2     -   -    d    2          d     1     |
The nephew of Miss World organiser Julia Morley |

 -     2     -   d    -   1      |    d    -    2
has appeared in court in London, | accused of blackmailing

 -   d      2     d      -    2      d       1
her husband Eric Morley for twenty thousand pounds.

2        2     -  d      2          d       -   -  1       |
Edward Crozier, a former personal assistant to the Morleys, |

  -  d      -   2       -   d     1      |
who comes from Sydenham in South London, |

 -     2     -   d    -  -  1
was remanded on bail for a month.

 2  |   1    -
Now | share news.

  -    2       2    d     -   1   |
The Financial Times index at noon |
 -   -    d   -     2   -    d     2      -    d    -     1
was down ten point nine at eleven hundred and four point nine.
```

```
-     -  1
And the weather.
```

```
 2       1     |  -    -     2    d         -     1      |
Northern areas | will have bright intervals and showers, |
```

```
  -    -    -  d    -    1
which will be heavy in places.
```

```
 2   -   -    d       1     |  -     d     2   |
Rain in some southern areas | will clear away, |
```

```
 -   d       1   | -   d        1
but further rain | is likely tomorrow.
```

```
-     -      -  2   -   2    d       -   1
And that's the news at five minutes past one.
```

A.3 Text 3

```
    2      1         |  -      d     -  2       d
Professor Neugebauer | has suggested a general method
```

```
 -   2       d      1        |
for doubling unit fractions, |
```

```
  -    -   2    -    -   d     -      2      d   -
which may well have been used for computing some of
```

```
  - d       -    - 2    d  1
the entries in the 2-to-N table.

-  -    2    -     -  1       -    |     2     -  |
As an example of the method, we'll | suppose that |

 -  d    - 2   -   -     d
we want to use it to compute

  2   d    1
twice one-fifth.

 -   2  -       d     -    2   -    -  d  -  -  d
We try to represent the result as the sum of a natural

  2      -  d    2    -    -   -     d      1
fraction of one-fifth and some other unit fraction.

2     |    2           -   d   2    -  d    1    |
After | experimenting with one-half of one-fifth |

 2     -     d   -  1       |  -   2 d     1    |
fails to provide an answer, | we try one-third, |

-     -   2    d    2  d       -  -    1       |
and this gives one-fifteenth plus a remainder, |

  -    -   -  2     -    d      2   -     d   2
which has to make the remaining one and two thirds
```

```
-  d    1     |   -  -     2  d     -  1
of one-fifth, | that is, two-fifths in all.
```

```
 -  2   -   -  d         -   1     | 1   |
We do this by counting the thirds: | one |
```

```
    d    -     2    d     -     2   d    -    1
consists of three thirds, and two thirds of two.
```

```
   2  -     1 | -     d     -   2  -       -  d
Three and two | are written in red under the symbols
```

```
 -  2   -     d   1     | d   -      d   -    -  1     -
for one and two thirds, | as we've shown in the radio notes.
```

A.4 Text 4

```
 2       d    - -     2          -      d       -     1
Let's return to our philosophes, in particular to Voltaire.
```

```
 2 |     1    |  -   - -     2      -  - d         1
Now | Voltaire | was not a Christian in any orthodox sense.
```

```
-   -   2      -  2         d      -    2
In his view, the rituals, priests, and doctrines
```

```
-     d   1     |  -   2        d     -       2
of Christianity | had fostered hatred and extremism
```

```
 -       -      d      -        1
rather than compassion and toleration.
```

```
 2     -  |  -   2    -   d    2    d     1       -  |
Yet until | his death in seventeen seventy eight, he |

   2     -    d   -   1   |
retained a belief in God, |

  d        d       2     -  d   - -  1      - -   |
though Voltaire's God, the God of a deist, was one |

  -   d      1       |  -      d       -    1
that most Christians | would scarcely have recognised.

 1     |  -    1     |   -    2     -   d -  -    1      |
Deists | were people | who believed in God as a creator, |

 -  d  2   |   -   d       -  2       d
but unlike | the theists - a similar name

 -  -  2    d           1      -     |    d         1
but a very different school - deists | rejected revelation.

 2       -    1      |  -  -   1
Newton, for example, | was a theist.

    1    |  -  -  1
Voltaire | was a deist.

-  -  -      2         d   -    1        |
It was a continuing belief in Providence |
  -       2         d       1
which sustained Voltaire's deism.

    2    |  -   2  |   -    -  2        -   d    -    -  1
However, | we know | that the suffering and evil in the world

 2   -      d         2         -   -   -      d     -     1
made it increasingly difficult for him to maintain this belief.

  2    2   -  |  1     - |   d    -   2      d        1    |
Quite late in | life, he | wrote his famous novel Candide; |

 -     -    2       -  d          1   |
here, he rejected the Leibnizian view |

  -    -   -   -  d  -   -    d        1      |
that this is the best of all possible worlds, |

-      d    -   -  d         1       - |  -   d   1
even though in his earlier writings he | had felt otherwise.
```

Appendix B

Final Evaluation Data

This appendix contains the texts, transcriptions and parses of the twenty sentences used in the final evaluation experiment reported in Section 3.6. These sentences are numbers 006, 009, 011, 027, 028, 068, 080, 082, 087, 090, 106, 112, 113, 117, 124, 127, 133, 141, 151 and 152 in the ATR 200-sentence database.

The transcriptions use our accent symbols introduced above, and boundaries are indicated by one vertical bar | for a tg(0) boundary and two vertical bars for a tg(1) boundary. The transcriptions given are those for the Version B utterances.

The parses are those which were used to produce the Version A utterances, and use standard abbreviations.

Sentence 006

```
 2    2            |     2       2           1
John could lend him the latest draft of his work
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_proper([])],
[verb_phrase(_), syn_verb([aux(modal)]),
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_pronoun([])],
[noun_phrase,
[noun_phrase, syn_det([]), [ap, syn_adj([decap])], syn_noun([])],
[pp, syn_prep([of]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]]]].
```

Sentence 009

```
     3     1 ||          3              1
The bulb blew  when he switched on the light
```

```
[sentence, [clause(_), [clause(_), [clause(_),
[noun_phrase, syn_det([]), syn_noun([])],
[verb_phrase(_), syn_verb([main, gen])]], [conj, syn_conj([])],
[clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]), syn_prep([]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]]].
```

Sentence 011

```
      2             1    ||     2        1    |            1
They launched into battle  with all the forces they could muster
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]), [pp, syn_prep([]),
[noun_phrase, syn_noun([])]]]], [pp, syn_prep([]),
[noun_phrase,
[noun_phrase, syn_adv([]), syn_det([]), syn_noun([])],
[clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([aux(modal)]),
[verb_phrase(_), syn_verb([main, gen])]]]]]]]].
```

Sentence 027

```
       2          2      1    |            2    1
I'll draft those new proposals before the next meeting
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_pronoun([]), [ap, syn_adj([])], syn_noun([])]],
[pp, syn_prep([]),
[noun_phrase, syn_det([]), [ap, syn_adj([])], syn_noun([])]]]]].
```

Sentence 028

```
     2              1   ||        1      ||
The mud squelched loudly  and he realised
          3     1   |      1
that his suede boots were doomed
```

```
[sentence, [clause(_), [clause(_),
[noun_phrase, syn_det([]), syn_noun([])], [advp,
[verb_phrase(_), syn_verb([main, gen])], syn_adv([])],
[conj, syn_conj([])], [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[clause(_), [comp, syn_comp([])],
[noun_phrase, syn_det([]), [ap, syn_adj([])], syn_noun([])],
[verb_phrase(_), syn_verb([aux(modal)]), [ap, syn_adj([])]]]]]]]].
```

Sentence 068

```
      2                    1      ||    2     1
He remembered he needed a passport  to get a visa stamp
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]), [clause(_),
[noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])],
[clause(_), syn_own([to]), syn_verb([aux(modal)]),
[noun_phrase, syn_det([]), [n, syn_noun([]), syn_noun([])]]]]]]]]].
```

Sentence 080

```
      1      |      1       ||            3        1
The ceremony overwhelmed me   and I was moved to tears
```

```
[sentence, [clause(_), [clause(_), [clause(_),
[noun_phrase, syn_det([]), syn_noun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_pronoun([])]]], [conj, syn_conj([])],
[clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([aux(modal)]), syn_adj([]),
[pp, syn_prep([]),
[noun_phrase, syn_noun([])]]]]]]].
```

Sentence 082

```
 3 | 2           1  || 2           2              1
Bob milked the cows    after he'd gathered the chickens' eggs
```

```
[sentence, [clause(_), [clause(_), [clause(_),
[noun_phrase, syn_proper([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])]]],
[conj, syn_conj([])], [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), [ap, syn_adj([])],
[n, syn_noun([])]]]]]]].
```

Sentence 087

```
     2            1           |            2            1
He glimpsed the traffic warden out of the corner of his eye
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), [n, syn_noun([]), syn_noun([])]],
[pp, syn_prep([]), syn_prep([]),
[noun_phrase,
[noun_phrase, syn_det([]), syn_noun([])], [pp, syn_prep([of]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]]]].
```

Sentence 090

```
          2           1      ||         2   |              1
We were plunged into darkness  as the clouds engulfed the moon
```

```
[sentence, [clause(_), [clause(_), [clause(_),
[noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([aux(modal)]), syn_adj([]),
[pp, syn_prep([]),
[noun_phrase, syn_noun([])]]]], [conj, syn_conj([])],
[clause(_), [noun_phrase, syn_det([]), syn_noun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]].
```

Sentence 106

```
  3  2       |      3           1  |         1
This ointment will soothe the graze on your heel
```

```
[sentence, [clause(_), [clause(_),
[noun_phrase, syn_det([]), syn_noun([])],
[verb_phrase(_), syn_verb([aux(modal)]), syn_verb([main, gen]),
[noun_phrase,
[noun_phrase, syn_det([]), syn_noun([])], [pp, syn_prep([]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]]].
```

Sentence 112

```
     2              1    |             1    |
The walkers took a detour through the fields
     2              1
to avoid the busy thoroughfare
```

```
[sentence, [clause(_), [clause(_),
[noun_phrase, syn_det([]), syn_noun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])], [pp, syn_prep([]),
[noun_phrase, syn_det([]), syn_noun([])]],
[clause(_), syn_own([to]), syn_verb([main, gen]),
[noun_phrase, syn_det([]), [ap, syn_adj([])], syn_noun([])]]]]]].
```

Sentence 113

```
 2         1      | 2              2        1     |
Mary and Elizabeth both aim to be company directors
                1
by the age of thirty
```

```
[sentence, [clause(_), [clause(_), [noun_phrase,
[noun_phrase, syn_proper([])], [conj, syn_conj([])],
[noun_phrase, syn_proper([])]],
[verb_phrase(_), syn_adv([]), syn_verb([main, gen]),
[verb_phrase(_), syn_own([to]), syn_verb([aux(modal)]),
[noun_phrase, [ap, syn_adj([])], syn_noun([])]]],
[advp, [pp, syn_prep([]),
[noun_phrase,
[noun_phrase, syn_det([]), syn_noun([])], [pp, syn_prep([of]),
[noun_phrase, syn_noun([])]]]]]]]].
```

Sentence 117

```
 2       1      |    2                   1
Bulldog terriers yap almost as much as Chows

[sentence, [clause(_), [clause(_),
[noun_phrase, [ap, syn_adj([])], [n, syn_noun([])]],
[verb_phrase(_), syn_verb([main, gen]),
[advp, syn_adv([]), syn_det([]), syn_adv([]), syn_det([]),
[noun_phrase, syn_noun([])]]]]]].
```

Sentence 124

```
     3        1   ||         2              1
He grabbed a towel  and then answered the phone

[sentence, [clause(_), [clause(_),
[noun_phrase, syn_pronoun([])],
[verb_phrase(_),
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])]],
[conj, syn_conj([]), syn_conj([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]].
```

Sentence 127

```
    2                2       2       1   ||
We need to buy some more embroidery silks
   2           2          1
before we can finish the garment
```

```
[sentence, [clause(_), [clause(_), [clause(_),
[noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[verb_phrase(_), syn_own([to]), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_det([]), [n, syn_noun([]),
syn_noun([])]]]]], [conj, syn_conj([])],
[clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([aux(modal)]), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]].
```

Sentence 133

```
       2        1      ||            1   | 2            1
I'm obliged to tell you  that most women loathe their husbands
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[verb_phrase(_), syn_own([to]), syn_verb([main, gen]),
[noun_phrase, syn_pronoun([])], [clause(_), [comp, syn_comp([])],
[clause(_), [noun_phrase, syn_adj([]), syn_noun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]]]]].
```

Sentence 141

```
  1   |          3      1 ||
Amelia went to Chester Zoo
              3     1           2           1
and saw some tufted owls and a rare giant sloth
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_proper([])],
[verb_phrase(_),
[verb_phrase(_), syn_verb([aux(modal)]), [pp, syn_prep([]),
[noun_phrase, syn_proper([]), syn_proper([])]]],
[conj, syn_conj([])], [verb_phrase(_), syn_verb([main, gen]),
[noun_phrase,
[noun_phrase, syn_det([]), [ap, syn_adj([])],
[n, syn_noun([])]], [conj, syn_conj([])],
[noun_phrase, syn_det([]), [ap, syn_adj([]), syn_adj([])],
[n, syn_noun([])]]]]]]]].
```

Sentence 151

```
    3         1    |         2             1
He caught a glimpse of what looked like a badger
```

```
[sentence, [clause(_), [clause(_), [noun_phrase, syn_pronoun([])],
[verb_phrase(_), syn_verb([main, gen]),
[noun_phrase, syn_det([]), syn_noun([])], [pp, syn_prep([of]),
[noun_phrase, [clause(_), [comp, syn_comp([])],
[clause(_), [verb_phrase(_), syn_verb([main, gen]), syn_adv([]),
[noun_phrase, syn_det([]), syn_noun([])]]]]]]]]].
```

Sentence 152

```
     2      1       |     2         1      |
The family heirloom is a turquoise necklace
           2        1
made by a Bedouin tribe
```

```
[sentence, [clause(_), [clause(_),
[noun_phrase, syn_det([]), [ap, syn_adj([])], [n, syn_noun([])]],
[verb_phrase(_), syn_verb([aux(be)]),
[noun_phrase,
[noun_phrase, syn_det([]), [ap, syn_adj([])],
[n, syn_noun([])]], [ap, syn_adj([]), [pp, syn_prep([]),
[noun_phrase, syn_det([]), [ap, syn_adj([])],
[n, syn_noun([])]]]]]]]]].
```

Appendix C

Summary of Rules

This appendix is an explicit account of the rules used by INTERFIX and the accent-placement module. The first section gives the rules which were applied in the evaluation experiment in Section 3.1, and the second section gives the changes and additions implemented for the evaluation experiment in Section 3.6.

C.1 Rules from Symbolic Evaluation

C.1.1 INTERFIX

This takes the syntactic analysis as input and performs the following operations:

1) Assign tg(1) boundaries at clause boundaries marked in the syntactic analysis.

2) Assign tg(0) boundaries at the start of the first NP in a clause, and at the start of the first VP in a clause.

3) Interpret commas as per Section 2.1.6.

4) Delete all syntactic structure other than word-class information and interpreted commas.

C.1.2 Accent Assignment

These rules take the word-class information between each pair of TG boundaries in the output of INTERFIX, and process it without reference to any other information:

5) Assign a primary accent to each lexically-stressed syllable in a noun or proper noun, except in the cases listed in (8) and (9) below.

6) Assign a secondary accent to each lexically-stressed syllable in an adjective, adverb or verb.

7) Assign a 'd' to each lexically-stressed syllable in a subordinating conjunction.

8) Assign a 'd' to each lexically-stressed syllable in any noun which immediately follows a primary-accented noun, i.e. with no intervening words or prosodic events such as boundaries or register shifts. This handles multi-noun compounds, accenting the first noun in two-noun cases and producing a rhythmic alternation in longer compounds. This rule is merely an approximation of the appropriate treatment of compounds, and there are numerous exceptions to it, but no better treatment of these cases has been found (see e.g. Sparck Jones (1985), Sproat & Liberman 1987).

9) Assign a 'd' to each lexically-stressed syllable in any noun which is immediately followed by *of*, unless the *of* is followed by a definite determiner or a proper noun. We have no good explanation for the behaviour of *of*, but these rules reflect the tendency of nouns to be deaccented before *of* when there is a following indefinite NP. The word *of* is by far the most frequent preposition in our data, and it appears to affect accent location in a principled way. Other prepositions in the data (e.g. *in, between, by*) do not affect intonation in the same manner, suggesting that this is not a consequence of syntactic boundaries: moreover, the effect of *of* is dependent upon the "definiteness" of the following NP: in *the very concrete models of phonology current at the time* no accent is usually associated with *models*, whereas in *an indispensable part of any adequate theory* the word *part* is rarely unaccented. This pattern is repeated throughout the data, and in cases where *of* is followed by a determiner or a proper noun ,i.e. a "definite" NP, the word before it conforms to general rhythmic principles while in all other cases it is deaccented.

C.1.3 Rhythm Rule

This rule takes the output of the Accent Assignment rules, one domain at a time. The version applied in Section 3.1 performed the following reductions:

10) In domains containing no primary accent, temporarily promote the last secondary to a primary and demote it again after the Rhythm Rule has been applied.

11) In domains containing no accents at all, make no changes.

12) Delete all accents to the right of the rightmost primary (the nucleus).

13) Reduce all accents to the left of the nucleus to secondaries.

14) Starting from the nucleus, delete every odd secondary to the left.

C.1.4 WFCs

These constraints apply to the accents and boundaries for the entire utterance, as follows:

15) The first accentable item in an utterance must be accented: if it is a 'd', promote it to a secondary accent.

16) The last accent in an utterance must be a primary: if it is a secondary, promote it.

17) Every TG must contain an accent: remove any TG boundary which is not separated from the preceding TG boundary by an accent.

C.2 Rules from Final Evaluation

The following changes and additions to the rules applied in Section 3.1 were implemented before the evaluation described in Section 3.6.

C.2.1 INTERFIX

18) In the final domain of any sentence, the final preposition and anything which follows it is assigned to a separate tg(0) domain, as discussed in Section 3.2.

19) Anomalies identified by the preprocessor are treated as per the rules in Section 3.3.

C.2.2 Accent Assignment

20) Deaccentuation indicators marked in the lexicon are assigned primary accents and trigger deaccenting of all following nouns up to the next domain boundary. The items so marked are *alternative, another, final, first, former, latter, other,* and ordinals up to *sixth*.

21) Demonstratives are assigned a ‘d’, and trigger deaccenting of all following nouns up to the next domain boundary.

22) Content words marked (in the lexicon or by rule) as pragmatically deaccented are assigned ‘d’s.

C.2.3 Rhythm Rule

There have been no changes to this rule, except that the generation of ‘d’s by other rules has been allowed for as follows:

23) In Clause 3 of the Rhythm Rule, if a ‘d’ is encountered the next accent is preserved. This maintains the rhythmic alternation.

C.2.4 TGs

24) The TG in (15) was revised to promote sentence-initial ‘d’s to tertiaries instead of secondaries.

References

Abercrombie, D. (1967): *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

Adams, D. (1979): *The Hitch-Hiker's Guide to the Galaxy.* London: Pan.

Ainsworth, W. & J. Holmes (eds) (1988): *Speech 88: Proceedings of the 7th FASE Symposium.* Edinburgh: IOA.

Akers, G. & M. Lennig (1985): "Intonation in Text-to-Speech Synthesis: Evaluation of Algorithms." *JASA* 77, pp. 2157-2165.

Alderson, P., W. N. Campbell, J. B. Pickering & A. M. Trudgeon (1988): "Testing an Algorithm for the Automatic Generation of Natural-Sounding Intonation Contours." In Ainsworth & Holmes (eds) 1988, vol. 4 pp. 1225-1231.

Allen, J. M., S. Hunnicutt & D. H. Klatt (1987): *From Text to Speech: The MITalk System.* Cambridge: CUP.

Allerton, D. J. & A. Cruttenden (1979): "Three Reasons for Accenting a Definite Subject." *Journal of Linguistics* 15, pp. 49-53.

Altenberg, B. (1987): *Prosodic Patterns in Spoken English.* Lund: University Press.

Altmann, G. T. (ed.) (1990): *Cognitive Models of Speech Processing.* London: MIT Press.

Aubergé, V. (1990): "Semi-Automatic Constitution of a Prosodic Contour Lexicon for the Text-to-Speech Synthesis." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 215-218.

Baart, J. L. G. (1987): *Focus, Syntax & Accent Placement: Towards a Rule System for the Derivation of Accent Patterns in Dutch as Spoken by Humans and Machines.* Doctoral Dissertation, Rijksuniversiteit Leiden.

Baart, J. L. G. & J. S. Heemskerk (1988): "The Problem of Ambiguity in Morphological Analysis for a Dutch Text-to-Speech System." In Ainsworth & Holmes (eds) 1988, vol. 3 pp. 959-965.

Bäckström, M., K. Ceder & B. Lyberg (1989): "Prophon — An Interactive Environment for Text-to-Speech Conversion." In Tubach & Mariani (eds) 1989, vol. 1 pp. 144-147.

Bailly, G. (1986): "Multiparametric Generation of French Prosody from Unrestricted Text." Proceedings of ICASSP 1986, pp. 2419-2422.

Barber, S., B. Granström & P. Toutati (1988): "French Prosody in a Rule-Based Text-to-Speech System." In Ainsworth & Holmes (eds) 1988, vol. 3 pp. 967-974.

Barber, S., R. Carlson, P. Cosi, M. G. Di Benedetto, B. Granström & K. Vagges (1989): "A Rule-Based Italian Text-to-Speech System." In Tubach & Mariani (eds) 1989, vol. 2 pp. 517-520.

Bard, E. G. (1990): "Competition, Lateral Inhibition and Frequency." In Altmann (ed.) 1990, pp. 185-210.

Batliner, A. & E. Nöth (1989): "The Prediction of Focus." In Tubach & Mariani (eds) 1989, vol. 1 pp. 210-213.

Bauer, L. (1983): *English Word-Formation*. Cambridge: CUP.

Beckman, M. & J. B. Pierrehumbert (1986): "Intonational Structure in Japanese & English." *Phonology Yearbook* 3, pp. 255-309.

Bell, A. D. (1987): "Towards Assigning Prosodic Patterns in Speech Synthesis." In Laver & Jack (eds) 1987, vol. 2 pp. 169-172.

Benoit, C. (1989): "Intelligibility Test for the Assessment of French Synthesisers using Semantically Unpredictable Sentences." Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, September 1989, pp. 1.7.1-1.7.4.

Benoit, C. (1990): "An Intelligibility Test using Semantically Unpredictable Sentences: Towards the Quantification of Linguistic Complexity." *Speech Communication* 9, pp. 293-304.

Berman, A. & M. Szamosi (1972): "Observations on Sentential Stress." *Language* 48, pp. 304-325.

Berthelin, J. B., J. P. Fournier & B. Grau (1989): "Processing Non-Expected Language." In Tubach & Mariani (eds) 1989, vol. 1 pp. 550-552.

Bierwisch, M. (1968): "Two Critical Problems of Accent Rules." *Journal of Linguistics* 4, pp. 173-178.

Bing, J. M. (1979a): *Aspects of English Prosody*. Doctoral Dissertation, University of Massachusetts at Amherst.

Bing, J. M. (1979b): "A Reanalysis of Obligatory 'Comma' Pause in English." *University of Massachusetts at Amherst Occasional Papers in Linguistics* 5, pp. 1-23.

Bing, J. M. (1983): "Contrastive Stress, Contrastive Intonation and Contrastive Meaning." *Journal of Semantics* 2, pp. 141-156.

Bing, J. M. (1984): "A Discourse Domain Identified by Intonation." In Gibbon & Richter (eds) 1984, pp. 10-19.

Blauert, J. E. & E. Schaffert (1985): *Automatische Sprachein- u. Ausgabe*. Dortmund: Bundesanstalt für Arbeitsschutz.

Boisson, C. (1985): "L'Accentuation des Composés Anglais." *2ème Colloque d'Avril sur l'Anglais Oral: le Suprasegmental*, pp. 165-178. Paris: APLV.

Bolinger, D. (1951): "Intonation — Levels v. Configurations." *Word* 7, pp. 199-210.

Bolinger, D. (1958): "A Theory of Pitch Accent in English." *Word* 14, pp. 109-149.

Bolinger, D. (1961): "Contrastive Accent and Contrastive Stress." *Language* 37, pp. 83-96.

Bolinger, D. (1972a): "Accent is Predictable (if you're a mind-reader)." *Language* 48, pp. 633-644.

Bolinger, D. (ed.) (1972b): *Intonation*. Harmondsworth: Penguin.

Bolinger, D. (1983): "Where does Intonation Belong?" *Journal of Semantics* 2, pp. 101-120.

Bolinger, D. (1985): "Two Views of Accent." *Journal of Linguistics* 21, pp. 79-123.

Bolinger, D. (1986): *Intonation and its Parts*. Stanford: University Press.

Bolinger, D. (1987): "More Views on 'Two Views of Accent'." In Gussenhoven et al. 1987, pp. 124-146.

Bolinger, D. (1989): *Intonation and its Uses*. Stanford: University Press.

Booth, B. (1987): "Text Input and Pre-Processing: Dealing with the Orthographic Form of Texts." In Garside et al. (eds) 1987, pp. 97-109.

Brazil, D. (1984): "The Intonation of Sentences Read Aloud." In Gibbon & Richter (eds) 1984, pp. 46-66.

Brazil, D., M. Coulthard & C. Johns (1980): *Discourse Intonation and Language Teaching*. London: Longmans.

Bredvad-Jensen, A-C. (1981): "Tonal Interaction between Attitude and Grammar." In Fretheim (ed.) 1981, pp. 51-62.

Bresnan, J. (1971): "Sentence Stress and Syntactic Transformations." *Language* 47, pp. 257-281.

Bresnan, J. (1972): "Stress and Syntax: A Reply." *Language* 48, pp. 326-342.

Bresnan, J. (1982): "Control and Complementation." *Linguistic Inquiry* 13, pp. 343-434.

Brown, G. (1983): "Prosodic Structure and the Given/New Distinction." In Cutler & Ladd (eds) 1983, pp. 67-77.

Brown, G. & G. Yule (1983): *Discourse Analysis.* Cambridge: CUP.

Bruce, G. (1977): *Swedish Word Accents in Sentence Perspective.* Lund: Gleerup.

Bruce, G. (1981): "Tonal and Temporal Interplay." In Fretheim (ed.) 1981, pp. 63-74.

Bruce, G. (1982): "Textual Aspects of Prosody in Swedish." *Phonetica* 39, pp. 274-287.

Bruce, G., B. Granström & D. House (1990): "Prosodic Phrasing in Swedish Speech Synthesis." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 125-128.

Burton-Roberts, N. (1986a): "Thematic Predicates and the Pragmatics of Non-Descriptive Definition." *Journal of Linguistics* 22, pp. 41-66.

Burton-Roberts, N. (1986b): "Implications of the Pragmatics of Non-Descriptive Definition." *Journal of Linguistics* 22, pp. 311-329.

Butterfield, S, & A. Cutler (1990): "Intonational Cues to Word Boundaries in Clear Speech?" Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 1 pp. 87-94.

Campbell, W. N. (1987): "A Search for Higher-Level Duration Rules in a Real Speech Corpus." In Laver & Jack (eds) 1987, vol. 1 pp. 285-288.

Campbell, W. N. (1989): "Syllable-Level Duration Determination." In Tubach & Mariani (eds) 1989, vol. 2 pp. 698-701.

Campbell, W. N. (1990): "Evidence for a Syllable-Based Model of Speech Timing." Proceedings of ICSLP 1990, vol. 1 pp. 9-13.

Campbell, W. N., S. D. Isard, A. I. C. Monaghan & J. Verhoeven (1990): "Duration, Pitch and Diphones in the CSTR TTS System." Proceedings of ICSLP 1990, Kobe, Japan, November 1990, pp. 825-828.

Carlson, R. & B. Granström (1973): "Word Accent, Emphatic Stress & Syntax in a Synthesis by Rule Scheme for Swedish." *STL-QPSR* 2-3, pp. 31-35.

Carlson, R. & B. Granström (1986): "Linguistic Processing in the KTH Multi-Lingual Text-to-Speech System." Proceedings of ICASSP 1986, pp. 2403-2406.

Carlson, R., B. Granström & S. Hunnicutt (1990): "Multi-Lingual Text-to-Speech Development & Applications." In W. A. Ainsworth (ed.), *Advances in Speech, Hearing & Language Processing* vol. 1, pp. 269-296. London: JAI Press.

Carter, D. (1987): *Interpreting Anaphors in Natural Language*. Chichester: Ellis Horwood.

Cericola, D., M. Danieli, M. J. Mollo & D. Voltolini (1989): "Morpho-Syntactic Tools for Speech Processing." In Tubach & Mariani (eds) 1989, vol. 1 pp. 386-389.

Chafe, W. (1976): "Givenness, Contrastiveness, Definiteness, Subjects, Topics, and Points of View." In C. Li (ed.), *Subject and Topic*, pp. 25-56. New York: Academic Press.

Chafe, W. L. (1974): "Language and Consciousness." *Language* 50, pp. 111-133.

Charpentier, F. & E. Moulines (1989): "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis using Diphones." In Tubach & Mariani (eds) 1989, vol. 2 pp. 13-19.

Chomsky, N. & M. Halle (1968): *The Sound Pattern of English*. New York: Harper & Row.

Chomsky, N. (1970): "Remarks on Nominalisation." In R. A. Jacobs & P. S. Rosenbaum (eds), *Readings in English Transformational Grammar*, pp. 184-221. Boston: Ginn.

Chomsky, N. (1980): *Rules and Representations*. New York: Columbia.

Choppy, C. (1979): "La Ponctuation, Indicateur Prosodique pour la Synthèse à partir du Texte: Etude de la Virgule." *10èmes Journées d'Etude sur la Parole*, Grenoble, 30 May - 1 June 1979, pp. 183-191.

Choppy, C. & J. S. Liénard (1977): "Prosodie Automatique pour la Synthèse par Diphonèmes." *8èmes Journées d'Etude sur la Parole*, Aix-en-Provence, 25-27 May 1977, pp. 211-217.

Clifton, C. & P. Odom (1966): "Similarity Relations among Certain English Sentence Constructions." *Psychological Monographs* 80.

Collier, R. (1989): "Intonation Analysis: The Perception of Speech Melody in Relation to Acoustics and Production." In Tubach & Mariani (eds) 1989, vol. 1 pp. 38-44.

Collier, R. (1990): "Multi-Lingual Intonation Synthesis: Principles and Applications." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 273-276.

Collier, R., A. de Zitter & J. Terken (1989): "On the Perceptual Salience of Melodical Variations and its Consequences for Intonation Synthesis." In Tubach & Mariani (eds) 1989, vol. 2 pp. 108-111.

Collier, R. & J. Terken (1987): "Intonation by Rule in Text-to-Speech Applications." In Laver & Jack (eds) 1987, vol. 2 pp. 165-8.

Cooper, W. E. & J. P. Paccia-Cooper (1980): *Syntax & Speech*. Cambridge, Mass.: Harvard University Press.

Cooper, W. E. & J. M. Sorensen (1977): "Fundamental Frequency Contours at Syntactic Boundaries." *JASA* 62, pp. 683-692.

Cooper, W. E. & J. M. Sorensen (1981): *Fundamental Frequency in Sentence Production*. New York: Springer.

Couper-Kuhlen, E. (1986): *An Introduction to English Prosody*. Tübingen: Niemeyer.

Crystal, D. (1966): *Studies in the Prosodic Features of Educated Spoken British English, with Special Reference to Intonation*. Ph.D. Thesis, University of London.

Crystal, D. (1969): *Prosodic Systems & Intonation in English*. Cambridge: CUP.

Crystal, D. (1975): *The English Tone of Voice*. London: Arnold.

Cutler, A. & C. J. Darwin (1981): "Phoneme Monitoring Reaction Time and Preceding Prosody: Effects of Stop Closure Duration and of Fundamental Frequency." *Perception & Psychophysics* 28, pp. 217-224.

Cutler, A. & D. R. Ladd (eds) (1983): *Prosody: Models & Measurements*. Berlin: Springer.

Daneš, F. (1960): "Sentence Intonation from a Functional Point of View." *Word* 16, pp. 34-54.

Daneš, F. (1972): "Order of Elements and Sentence Intonation." In Bolinger (ed.) 1972b, pp. 216-232.

Datta, A. K., N. R. Ganguly & B. Mukherjee (1990): "Intonation in Segment-Concatenated Speech." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 153-156.

Delin, J. L. (1989): *Cleft Constructions in Discourse*. Ph.D. Thesis, University of Edinburgh.

Dell, F., D. Hirst & J-R. Vergnaud (1984): *Formes Sonores du Langage*. Paris: Hermann.

de Pijper, J. R. (1979): "Close-Copy Stylizations of British English Intonation Contours." *IPO Annual Progress Report* 14, pp. 66-71.

de Pijper, J. R. (1980): "A Melodical Model of British English Intonation." *IPO Annual Progress Report* 15, pp. 54-58.

de Pijper, J. R. (1983): *Modelling British English Intonation.* Dordrecht: Foris.

Di Cristo, A. & M. Rossi (1977): "Propositions pour un Modèle d'Analyse de l'Intonation." *8èmes Journées d'Etude sur la Parole*, Aix-en-Provence, 25-27 May 1977.

Di Cristo, A. (1981): "L'Intonation est Congruent à la Syntaxe: une Confirmation." In Rossi et al. 1981, pp. 272-289.

Dirksen, A. & H. Quené (1991): *Prosodic Analysis: The Next Generation.* IPO Manuscript 800.

Downing, B. (1970): *Syntactic Structure and Phonological Phrasing in English.* Doctoral Dissertation, University of Texas at Austin.

Downing, B. (1973): "Parenthisisation Rules and Obligatory Phrasing." *Papers in Linguistics* 6, pp. 108-128.

Eady, S. J. & W. E. Cooper (1986): "Speech Intonation and Focus Location in Matched Statements & Questions." *JASA* 80, pp. 402-415.

Emerard, F. & C. Benoit (1988): "Base de Données Prosodiques pour la Synthèse de la Parole." *Journal d'Acoustique* 1, pp. 303-307.

Faber, D. (1987): "The Accentuation of Intransitive Sentences in English." *Journal of Linguistics* 23, pp. 341-58.

Fallside, F. & S. J. Young (1978): "Speech Output from a Computer-Controlled Water-Supply Network." *Proceedings of the IEE* 125, pp. 157-161.

Faulkner, A. (1986): "Discrimination of Speech Intonation Contour: Evidence for Tonetic Categories?" Proceedings of IOA Autumn Conference, Windermere, November 1986, vol. 1 pp. 45-52.

Firbas, J. (1980): "Post-Intonation-Centre Prosodic Shade in the Modern English Clause." In S. Greenbaum, G. Leech & J. Svartvik (eds), *Studies in English Linguistics for Randolph Quirk*, pp. 125-133. London: Longmans.

Firth, J. R. (1948): "Sounds and Prosodies." In Firth 1957, pp. 121-138.

Firth, J. R. (1957): *Papers in Linguistics 1934-1951*. London: OUP.

Fitzpatrick, E. & J. Bachenko (1989): "Parsing for Prosody: What a Text-to-Speech System Needs from Syntax." Proceedings of the Annual AI Systems in Government Conference, pp. 188-194 Washington DC: IEEE Computer Society Press.

Fodor, J. A., T. G. Bever & M. F. Garrett (1974): *The Psychology of Language*. New York: McGraw-Hill.

Fretheim, T. (ed.) (1981): *Nordic Prosody II: Proceedings of the Second Symposium on Prosody in the Nordic Languages*. Trondheim: TAPIR.

Fuchs, A. (1984): "'Deaccenting' and 'Default Accent'." In Gibbon & Richter (eds) 1984, pp. 134-164.

Fudge, E. (1984): *English Word Stress*. London: Allen & Unwin.

Fujisaki, H. & K. Hirose (1983): "Modelling the Dynamic Characteristics of Voice Fundamental Frequency with Applications to Analysis and Synthesis of Intonation." Preprints of papers, working group on intonation, 13th International Congress of Linguists, Tokyo, pp. 57-70.

Fujisaki, H. & S. Nagashima (1969): "A Model for the Synthesis of Pitch Contours of Connected Speech." *Tokyo University Engineering Res. Inst. Annual Report* 28, pp. 53-60.

Fujisaki, H. & H. Sudo (1971): "Synthesis by Rule of Prosodic Features of Connected Japanese." Proceedings of the 7th International Congress of Acoustics, vol. 3 pp. 133-136.

Gårding, E. & G. Bruce (1981): "A Presentation of the Lund Model for Swedish Intonation." In Fretheim (ed.) 1981, pp. 33-39.

Gårding, E. (1983): "A Generative Model of Intonation." In Cutler & Ladd (eds) 1983, pp. 11-25.

Garside, R., G. Leech & G. Sampson (eds) (1987): *The Computational Analysis of English: A Corpus-Based Approach*. London: Longmans.

Gartenberg, R. & I. Hertrich (1988): "Predicting the Timing of F_0 Peaks in Partly Controlled Dialogue Situations." In Ainsworth & Holmes (eds) 1988, vol. 3 pp. 997-1004.

Gartenberg, R. & I. Hertrich (1989): "Speaker Responses to F_0 Manipulations in Partly Controlled Simulated Dialogues." In Tubach & Mariani (eds) 1989, vol. 2 p. 112.

Gazdar, G., E. Klein, G. Pullum & I. Sag (1985): *Generalized Phrase Structure Grammar*. Oxford: Blackwell.

Gee, J. P. & F. Grosjean (1983): "Performance Structures: A Psycholinguistic and Linguistic Appraisal." *Cognitive Psychology* 15, pp. 411-458.

Gibbon, D. & H. Richter (eds) (1984): *Intonation, Accent & Rhythm: Studies in Discourse Phonology*. Berlin: De Gruyter.

Gillott, T. J., M. C. Hall & S. Macgregor (1990): "Towards High Quality Automatic Announcement Systems." Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 1 pp. 111-116.

Giustiniani, M., A. Falaschi & P. Pierucci (1990): "Automatic Inference of a Syllabic Prosodic Model." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 197-200.

Goldsmith, J. (1976): *Autosegmental Phonology*. Doctoral Dissertation, MIT.

Goldsmith, J. (1982): "Accent Systems." In van der Hulst & Smith (eds) 1982, pp. 47-64.

Granström, B. (1990): "Development of a Multi-Lingual Text-to-Speech System." Proceedings of Verba 90, Rome, 22-24 January 1990, pp. 303-309.

Granström, B., P. M. Hansen & N. G. Thorsen (1987): "A Danish Text-to-Speech System using a Text Normalizer Based on Morph Analysis." In Laver & Jack (eds) 1987, vol. 1 pp. 21-4.

Gretter, R., G. A. Mian, R. Rinaldo & M. Salmasi (1990): "Linguistic Processing for an Italian Text-to-Speech System." Proceedings of Verba 90, Rome, 22-24 January 1990, pp. 334-342.

Grice, H. P. (1975): "Logic and Conversation." In P. Cole & J. Morgan (eds), *Syntax & Semantics 3: Speech Acts*, pp. 41-58. New York: Academic Press.

Grice, M. (1989): Syntactic Structures & Lexicon Requirements for Semantically Unpredictable Sentences in a Number of Languages." Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, September 1989, pp. 1.5.1-1.5.4.

Grishman, R. (1987): *Computational Linguistics: An Introduction*. Cambridge: CUP.

Guaitella, I. (1990): "Propositions pour une Méthode d'Analyse de l'Intonation en Parole Spontanée." Proceedings of the 1st French Conference on Acoustics, February 1990, pp. 515-518.

Guaitella, I. & S. Santi (1990): "Contribution of the Analysis of Punctuation to Improving the Prosody of Speech Synthesis." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 177-180.

Gunter, R. (1972): "Intonation and Relevance." In Bolinger (ed.) 1972b, pp. 194-215.

Gussenhoven, C. (1983a): "Focus, Mode and the Nucleus." *Journal of Linguistics* 19, pp. 377-417.

Gussenhoven, C. (1983b): "A Three-Dimensional Scaling of Nine English Tones." *Journal of Semantics* 2, pp. 183-203.

Gussenhoven, C. (1984): *On the Grammar and Semantics of Sentence Accents.* Dordrecht: Foris.

Gussenhoven, C. (1985): "Two Views of Accent: A Reply." *Journal of Linguistics* 21, pp. 125-138.

Gussenhoven, C. (1987): "More Views on 'Two Views of Accent': A Reply." In Gussenhoven et al. 1987, pp. 147-161.

Gussenhoven, C., D. Bolinger & C. E. Keijsper (1987): *On Accent.* Bloomington, Ind.: IULC.

Gussenhoven, C. & A. C. M. Rietveld (1987): "Perceived Speech Rate and Intonation." *Journal of Phonetics* 15, pp. 273-285.

Gussenhoven, C. & A. C. M. Rietveld (1988): "Fundamental Frequency Declination in Dutch: Testing Three Hypotheses." *Journal of Phonetics* 16, pp. 355-369.

Halliday, M. A. K. (1967a): "Notes on Transitivity and Theme: Part II." *Journal of Linguistics* 3, pp. 199-244.

Halliday, M. A. K. (1967b): *Intonation & Grammar in British English.* The Hague: Mouton.

Halliday, M. A. K. (1970): "Language Structure & Language Function." In J. Lyons (ed.), *New Horizons in Linguistics* vol. 1, pp. 140-165. Harmondsworth: Penguin.

Hazan, V. & M. Grice (1989): "The Assessment of Synthetic Speech Intelligibility using Semantically Unpredictable Sentences." Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, September 1989, pp. 1.6.1-1.6.4.

Hertrich, I. & R. Gartenberg (1988): "Signalling Accents by the use of Different Intonation Patterns in German — Some Potential Perceptual Ambiguities." In Ainsworth & Holmes (eds) 1988, vol. 3 pp. 989-97.

Hertrich, I. & R. Gartenberg (1989): "A New Method in Intonation Research using Partly Controlled, Simulated Dialogues." In Tubach & Mariani (eds) 1989, vol. 1 pp. 51-54.

Hirschberg, J. (1990a): "Using Discourse Context to Guide Pitch Accent Decisions in Synthetic Speech." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 181-184.

Hirschberg, J. (1990b): "Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech." Proceedings of AAAI 1990, pp. 952-957.

Hirschberg, J. & J. B. Pierrehumbert (1986): "The Intonational Structuring of Discourse. " Proceedings of the 24th Annual Meeting of the ACL, pp. 136-144.

Hirst, D. (1981): "Intonation et Interprétation Sémantique." In Rossi et al. 1981, pp. 303-318.

Hirst, D. (1983): "Interpreting Intonation: A Modular Approach." *Journal of Semantics* 2, pp. 171-181.

Hirst, G. (1981): *Anaphora in Natural Language Understanding: A Survey*. Berlin: Springer.

Holmes, J., I. Mattingly & J. Shearme (1964): "Speech Synthesis-by-Rule." *Language & Speech* 7, pp. 127-143.

Horne, M. (1987): "Towards a Discourse-Based Model of English Sentence Intonation." *Lund University Dept. of Linguistics Working Papers* 32.

House, J. (1987): "Enlivening the Intonation in Text-to-Speech Synthesis: An 'Accent-Unit' Model." Proceedings of the 11th International Congress of Phonetic Sciences, Tallinn, pp. 134-137.

House, J. (1989): "Syllable Structure Constraints on F_0 Timing." Poster presented at the Second Conference on Laboratory Phonology, Edinburgh, July 1989.

House, J. & M. Johnson (1986): "Modelling Nuclear Tones for Speech Synthesis by Rule." *Speech, Hearing and Language: Work in Progress at UCL* 2, pp. 85-101.

House, J. & N. Youd (1990): "Contextually Appropriate Intonation in Speech Synthesis." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 185-188.

House, S. A., C. E. Williams, M. H. Hecker & K. D. Kryter (1965): "Articulation Testing Methods: Consonantal Differentiation with a Closed-Response Set." *JASA* 37, pp. 158-166.

Hunnicutt, S. (1987): "Acoustic Correlates of Redundancy and Intelligibility." *STL-QPSR* 2-3, pp. 7-14.

Inkelas, S., W. R. Leben & M. Cobler (1991): "The Phonology of Intonation in Hausa." In J. Kingston & M. Beckman (eds), *Papers in Laboratory Phonology* 1. Cambridge: CUP.

Isard, S. D. & M. Pearson (1988): "A Repertoire of British English Intonation Contours for Speech Synthesis." In Ainsworth & Holmes (eds) 1988, vol. 4 pp. 1233-40.

Jensen, K & J-L. Binot (1987): "Disambiguating Prepositional Phrase Attachments by using On-Line Dictionary Definitions." *Computational Linguistics* 13, pp. 251-260.

Jespersen, O. (1909): *A Modern English Grammar, Part I*. Heidelberg: Winter.

Johnson, M. & J. House (1986): "An Accent-Unit Model of Intonation for Text-to-Speech Synthesis." Proceedings of IOA Autumn Conference, Windermere, November 1986, vol. 3 pp. 409-416.

Johnson, M. & M. Grice (1990): "Some Phonetic Correlates of Stylisation in the Step-Down Contour." Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 1 pp. 71-78.

Joshi, A. K. (1990): "Phrase Structure and Intonational Phrases." In Altmann (ed.) 1990, pp. 513-532.

Kager, R. & H. Quené (1987): "Deriving Prosodic Sentence Structure Without Exhaustive Syntactic Analysis." In Laver & Jack (eds) 1987, vol. 1 pp. 243-246.

Kager, R. & H. Quené (1989): "A Sentence Accentuation Algorithm for a Dutch Text-to-Speech System." In H. Bennis & A. van Kermenade (eds), *Linguistics in the Netherlands* 6, pp. 101-109. Dordrecht: Foris.

Kaisse, E. M. (1985): *Connected Speech: The Interaction of Syntax and Phonology.* New York: Academic Press.

Keijsper, C. E. (1987): "Two Views of Accent: A Third Opinion." In Gussenhoven et al. 1987, pp. 162-201.

King, R. W. (1989): "Layout Processing, User Control and Prosody Insertion in an On-Line Synthetic Speech System." In Tubach & Mariani (eds) 1989, vol. 1 pp. 121-124.

Kingdon, R. (1958a): *The Groundwork of English Stress.* London: Longmans.

Kingdon, R. (1958b): *The Groundwork of English Intonation.* London: Longmans.

Klatt, D. H. (1979): "Synthesis by Rule of Segmental Durations in English Sentences." In B. Lindblom & S. Öhman (eds), *Frontiers of Speech Communication Research*, pp. 287-300. New York: Academic Press.

Klatt, D. H. (1987): "Review of Text-to-Speech Conversion for English." *JASA* 82, pp. 737-793.

Klatt, D. H. & W. E. Cooper (1975): "Perception of Vowel Duration in Sentence Contexts." *JASA* 57, pp. 47-57.

Knowles, G. (1984): "Variable Strategies in Intonation." In Gibbon & Richter (eds) 1984, pp. 226-242.

Knowles, G. (forthcoming): "From Text Structure to Prosodic Structure." In Knowles & Alderson (eds).

Knowles, G. & L. Lawrence (1987): "Automatic Intonation Assignment." In Garside et al. (eds) 1987, pp. 139-148.

Knowles, G. & P. Alderson (eds) (forthcoming): *Working with Speech.* London: Longmans.

Kohler, K. (1987): "The Linguistic Functions of F_0 Peaks." Proceedings of the 11th International Congress of Phonetic Sciences, Tallinn, pp. 149-52.

Kohler, K. (1988): "An Intonation Model for a German Text-to-Speech System." In Ainsworth & Holmes (eds) 1988, vol. 4 pp. 1241-7.

Kohler, K. (1990): "Improving the Prosody in German Text-to-Speech Output." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 189-192.

Koopmans-van Beinum, F. J. & D. R. van Bergem (1989): "The Role of 'Given' and 'New' in the Production and Perception of Vowel Contrasts in Read Text and in Spontaneous Speech." In Tubach & Mariani (eds) 1989, vol. 2 pp. 113-116.

Kruyt, J. G. (1985): *Accents from Speakers to Listeners.* Doctoral Dissertation, Rijksuniversiteit Leiden.

Kugler-Kruse, M. & R. Posmyk (1987): "Methods for the Simulation of Natural Intonation in the 'Syrub' Text-to-Speech System for Unrestricted German Text." In Laver & Jack (eds) 1987, vol. 2 pp. 177-180.

Kulas, W., M. Kugler-Kruse, U. Jekosch & M. Kesselheim (1986): "Ergonomische Gesichtspunkte bei Vorleseautomaten: Benutzerschnittstelle und Prosodiesteurung im Bochumer SYRUB-Programmsystem zur Umsetzung von deutschen Schrifttext in Lautschrift mit Prosodieinformation." *NTG-Fachberichte* 94 (Sprachkommunication), pp. 198-202.

Lacerda, F., G. Bruce, D. House & L. Eriksson (1988): "Prosodic Parsing of Swedish — Status Report: Segmentation Strategy." In Ainsworth & Holmes (eds) 1988, vol. 4 pp. 1187-1195.

Ladd, D. R. (1979a): *Basic Bibliography on English Intonation*. Bloomington, Ind.: IULC.

Ladd, D. R. (1979b): "Light and Shadow: A Study of the Syntax and Semantics of Sentence Accents in English." In L. Waugh & F. van Coetsem (eds), *Contributions to Grammatical Studies: Semantics and Syntax*, pp. 93-131. Leiden: E. J. Brill.

Ladd, D. R. (1980): *The Structure of Intonational Meaning: Evidence from English*. Bloomington: Indiana University Press.

Ladd, D. R. (1981): "On Intonational Universals." In Myers et al. (eds) 1981, pp. 389-397.

Ladd, D. R. (1983a): "Even, Focus, and Normal Stress." *Journal of Semantics* 2, pp. 157-170.

Ladd, D. R. (1983b): "Levels vs. Configurations, Revisited." In F. B. Agard, G. Kelley, A. Makkai & V. B. Makkai (eds), *Essays in Honor of Charles F. Hockett*, pp. 49-59. Leiden: E. J. Brill.

Ladd, D. R. (1984a): "English Compound Stress." In Gibbon & Richter (eds) 1984, pp. 253-266.

Ladd, D. R. (1984b): "Declination: A Review and some Hypotheses." *Phonology Yearbook* 1, pp. 53-74.

Ladd, D. R. (1986): "Intonational Phrasing: The Case for Recursive Prosodic Structure." *Phonology Yearbook* 3, pp. 311-340.

Ladd, D. R. (1987): "A Phonological Model of Intonation for Use in Speech Synthesis by Rule." In Laver & Jack (eds) 1987, vol. 2 pp. 21-24.

Ladd, D. R. (1988): "Declination 'Reset' and the Hierarchical Organization of Utterances." *JASA* 84, pp. 530-544.

Ladd, D. R. (1991): "Metrical Representation of Pitch Register." In J. Kingston & M. Beckman (eds), *Papers in Laboratory Phonology* 1. Cambridge: CUP.

Ladd, D. R. (forthcoming): "Introduction to Intonational Phonology." To appear in G. Docherty & D. R. Ladd (eds), *Papers in Laboratory Phonology* 2. Cambridge: CUP.

Ladd, D. R. & C. Johnson (1987): "'Metrical' Factors in the Scaling of Sentence-Initial Accent Peaks." *Phonetica* 44, pp. 238-245.

Ladd, D. R. & A. I. C. Monaghan (1987): "Modelling Rhythmic and Syntactic Effects on Accent in Long Noun Phrases." In Laver & Jack (eds) 1987, vol. 2 pp. 29-32.

Ladd, D. R., K. Silverman, F. Tolkmitt, G. Bergmann & K. R. Scherer (1985): "Evidence for the Independent Function of Intonation Contour, Pitch Range, and Voice Quality." *JASA* 78, pp. 435-444.

Lakoff, G. (1972): "The Global Nature of the Nuclear Stress Rule." *Language* 48, pp. 285-303.

Larreur, D., F. Emerard & F. Marty (1989): "Linguistic and Prosodic Processing for a Text-to-Speech Synthesis System." In Tubach & Mariani (eds), vol. 1 pp. 510-513.

Laver, J. & M. Jack (eds) (1987): *European Conference on Speech Technology*, vols. 1 & 2. Edinburgh: CEP.

Leben, W. R. (1973): *Suprasegmental Phonology*. Doctoral Dissertation, MIT.

Leech, G. (1990): "Existing Text Corpus Resources for Speech and NL Processing." Paper presented at the SALT Workshop on Corpus Resources, Oxford, 3-4 January 1990.

Lehiste, I. (1970): *Suprasegmentals*. London: MIT Press.

Lehiste, I. (1975): "The phonetic structure of paragraphs." In A. Cohen & S. G. Nooteboom (eds), *Structure and Process in Speech Perception*, pp. 195-203. Heidelberg: Springer.

Levelt, W. J. M. & A. Cutler (1983): "Prosodic Marking in Speech Repair." *Journal of Semantics* 2, pp. 205-217.

Levi, J. N. (1978): *The Syntax and Semantics of Complex Nominals.* London: Academic Press.

Liberman, M. Y. (1975): *The Intonation System of English.* Ph.D. Dissertation, Massachusetts Institute of Technology.

Liberman, M. Y. & A. Prince (1977): "On Stress and Linguistic Rhythm." *Linguistic Inquiry* 8, pp. 249-336.

Liberman, M. Y. & R. W. Sproat (1987): *The Stress and Structure of Modified Noun Phrases in English.* AT&T Bell Laboratories Technical Memo.

Lieberman, P. (1965): "On the Acoustic Basis of Perception of Intonation by Linguists." *Word* 21, pp. 40-54.

Lieberman, P. (1967): *Intonation, Perception and Language.* Cambridge, Mass.: MIT Press.

Lieberman, P. & S. B. Michaels (1972): "Some Apsects of Fundamental Frequency and Envelope Amplitude as Related to the Emotional Content of Speech." In Bolinger (ed.) 1972b, pp. 235-249.

Liénard, J. S. & D. Teil (1970): "Les Eléments Phonétiques et la Traduction Automatique du Message Ecrit en message Parlé." *Automatisme* 10.

Lucas, S. M. & R. J. Damper (1990): "Syntactic Neural Nets for Speech Technology." Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 2 pp. 299-306.

Luce, P. A., T. C. Feustel & D. B. Pisoni (1983): "Capacity Demands in Short-Term Memory for Synthetic and Natural Word Lists." *Human Factors* 25, pp. 17-32.

Majewski, W., C. Basztura & W. Myslecki (1988): "Relation between Speech Intelligibility and Subjective Scale of Speech Transmission." In Ainsworth & Holmes (eds) 1988, vol. 2 pp. 719-726.

Marcus, M. & D. Hindle (1990): "Description Theory and Intonation Boundaries." In Altmann (ed.) 1990, pp. 483-512.

Marslen-Wilson, W. D. (1973): "Linguistic Structure and Speech Shadowing at Very Short Latencies." *Nature* 244, pp. 522-523.

Marslen-Wilson, W. D. (1975): "Sentence Perception as an Interactive Parallel Process." *Science* 189, pp. 226-228.

Marslen-Wilson, W. D. (1990): "Activation, Competition and Frequency in Lexical Access." In Altmann (ed.) 1990, pp. 148-172.

Martin, P. (1981): "Pour une Théorie de L'Intonation: l'Intonation est-elle une Structure Congruente à la Syntaxe?" In Rossi et al. 1981, pp. 234-271.

Martin, P. (1982): "Phonetic Realisations of Prosodic Contours in French." *Speech Communication* 1, pp. 283-294.

Matsumoto, T. & Y. Yamaguchi (1990): "A Multi-Language Text-to-Speech System using Neural Nets." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 269-272.

Matthews, P. (1974): *Morphology*. Cambridge: CUP.

Mattingly, I. G. (1966): "Synthesis by Rule of Prosodic Features." *Language & Speech* 9, pp. 1-13.

McAllister, J. & L. Shockey (1986): "The Edinburgh University CSTR Text-to-Speech System." *Edinburgh University Dept. of Linguistics Work in Progress* 19, pp. 36-44.

McAllister, M. (1989): "The Problems of Punctuation Ambiguity in Fully Automatic Text-to-Speech Conversion." In Tubach & Mariani (eds) 1989, pp. 538-541.

Miller, G. A. (1962): "Some Psychological Studies of Grammar." *American Psychologist* 17, pp. 748-762.

Milligan, S. (1972): *Adolf Hitler: My Part in his Downfall.* Harmondsworth: Penguin.

Monaghan, A. I. C. (1987a): *Automatic Intonation from Text.* M.A. Dissertation, University of Edinburgh.

Monaghan, A. I. C. (1987b): "A System for Left-to-Right Intonation Specification from Text." In Laver & Jack (eds) 1987, vol. 2 pp. 25-28.

Monaghan, A. I. C. (1988): "Generating Intonation in the Absence of Essential Information." In Ainsworth & Holmes (eds) 1988, vol. 4 pp. 1249-1256.

Monaghan, A. I. C. (1989): "Phonological Domains for Intonation in Speech Synthesis." In Tubach & Mariani (eds) 1989, vol. 1 pp. 502-505.

Monaghan, A. I. C. (1990d): "Automatic Accent Placement in Anomalous Text Sequences." Proceedings of the IOA Autumn Conference, Windermere, November 1990, pp. 79-86.

Monaghan, A. I. C. (1990a): "Rhythm & Stress Shift in Speech Synthesis." *Computer Speech & Language* 4, pp. 71-78.

Monaghan, A. I. C. (1990c): "Treating Anaphora in the CSTR TTS System." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, September 1990, pp. 113-116.

Monaghan, A. I. C. (1990e): "The Design of a Spoken Corpus for Deriving Prosodic Rules." Proceedings of the IOA Autumn Conference, Windermere, November 1990, pp. 103-110.

Monaghan, A. I. C. (1990b): "A Multi-Phase Parsing Strategy for Unrestricted Text." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, September 1990, pp. 109-112.

Monaghan, A. I. C. (1991a): "Evaluation of the Naturalness of Prosody Generated by the CSTR TTS System." Proceedings of Eurospeech 1991, September 1991, Genoa, pp. 883-886.

Monaghan, A. I. C. (1991b): "Heuristic Strategies for the Higher-Level Analysis of Unrestricted Text." To be published in C. Benoit & G. Bailly (eds), *Talking Machines.* Amsterdam: Elsevier.

Monaghan, A. I. C. & D. R. Ladd (1989): "Evaluating Intonation in the CSTR Text-to-Speech System." Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, September 1989, pp. 3.6.1-3.6.4.

Monaghan, A. I. C. & D. R. Ladd (1990a): "Speaker-Dependent and Speaker-Independent Parameters in Intonation." Proceedings of the ESCA Workshop on Speaker Characterisation, Edinburgh, June 26-28 1990, pp. 167-174.

Monaghan, A. I. C. & D. R. Ladd (1990b): "Symbolic Output as the Basis for Evaluating Intonation in Text-to-Speech Systems." *Speech Communication* 9, pp. 305-314.

Monaghan, A. I. C. & D. R. Ladd (1991): "Manipulating Synthetic Intonation for Speaker Characterisation." Proceedings of ICASSP 1991, vol. 1 pp. 453-456.

Mortamet, L., F. Emerard & L. Miclet (1990): "Attempting Automatic Prosodic Knowledge Acquisition Using a Database." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 209-213.

Morton, K. (1990): "Naturalness in Synthetic Speech." Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 1 pp. 125-132.

Moulines, E., F. Emerard, D. Larreur, J. L. Le Saint Milon, L. Le Faucheur, F. Marty, F. Charpentier & C. Sorin (1990): "A Real-Time French Text-to-Speech System Generating High-Quality Synthetic Speech." Proceedings of ICASSP 1990, vol. 1 pp. 309-312.

Moulines, E., C. Sorin & F. Charpentier (1990): "New Approaches for Improving the Quality of Text-to-Speech Systems." Proceedings of Verba 90, Rome, 22-24 January 1990, pp. 310-319.

Mousel, P., J. M. Pierrel & A. Roussanaly (1989): "Cooperation and Representation of Syntactic-Semantic and Pragmatic Knowledge in a Natural Language Task Oriented Spoken Dialogue System." In Tubach & Mariani (eds) 1989, vol. 1 pp. 183-186.

Murray, I. R. & J. L. Arnott (1990): "Evaluation of a Synthetic Speech System which Emulates Vocal Emotion by Rule." Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 1 pp. 117-123.

Murray, I. R., J. L. Arnott & A. F. Newell (1988): "HAMLET — Simulating Emotion in Synthetic Speech." In Ainsworth & Holmes (eds) 1988, vol. 4 pp. 1271-1223.

Murray, I. R., J. L. Arnott, A. F. Newell, G. Cruickshank, K. E. P. Carter & R. Dye (1990): "Experiments with a Full-Speed Speech-Driven Word-Processor." Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 2 pp. 459-466.

Myers, T., J. Laver & J. Anderson (eds) (1981): *The Cognitive Representation of Speech.* Amsterdam: North-Holland.

Nebbia, L. (1990): "Text-to-Speech Synthesis System for Italian: An Overview." Proceedings of Verba 90, Rome, 22-24 January 1990, pp. 326-333.

Nespor, M. & I. Vogel (1982): "Prosodic Domains of External Sandhi Rules." In van der Hulst & Smith (eds) 1982, pp. 225-255.

Newman, S. S. (1946): "On the Stress System of English." *Word* 2, pp. 171-187.

Niedermair, G. Th. (1989): "The Use of a Semantic Network in Speech Dialogue." In Tubach & Mariani (eds) 1989, vol. 1 pp. 26-29.

Niedermair, G. Th., M. Streit & H. Tropf (1990): "Linguistic Processing Related to Speech Understanding in SPICOS II." *Speech Communication* 9, pp. 565-586.

Nooteboom, S. G. & J. G. Kruyt (1987): "Accents, Focus Distribution, and the Perceived Distribution of Given and New Information: An Experiment." *JASA* 82, pp. 1512-1524.

Nooteboom, S. G., T. Kruyt & J. Terken (1981): "What Speakers and Listeners Do with Pitch Accents: Some Explorations." In Fretheim (ed.) 1981, pp. 9-32.

O'Connor, J. D. & G. F. Arnold (1961): *Intonation of Colloquial English.* London: Longmans.

O'Connor, J. D. & G. F. Arnold (1973): *Intonation of Colloquial English*, second edition. London: Longmans.

Oehrle, R., E. Bach & D. Wheeler (eds) (1988): *Categorial Grammars & Natural Language Structure*. Dordrecht: Reidel.

Olaszy, G. & G. Gordos (1987): "On the Speaking Module of an Automatic Reading Machine." In Laver & Jack (eds) 1987, vol. 1 pp. 25-28.

Olaszy, G. (1989): "Multivox — A Flexible Text-to-Speech System for Hungarian, Finnish, German, Esperanto, Italian and Other Languages for IBM-PC." In Tubach & Mariani (eds), 1989, vol. 2 pp. 525-528.

Olaszy, G., G. Gordos & G. Németh (1990): "Phonetic Aspects of the Multivox Text to Speech System." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 277-280.

Olive, J. P. & L. H. Nakatani (1974): "Rule Synthesis of Speech by Word Concatenation: A First Step." *JASA* 55.

Ouado, M., A. Rajouani, M. Najim, M. Zyoute & P. Baylou (1987): "Arabic Text-to-Speech: Single Board." In Laver & Jack (eds) 1987, vol. 2 pp. 83-86.

Palmer, H. (1922): *English Intonation, with Systematic Exercises.* Cambridge: Heffer.

Pasdeloup, V. (1990a): *Modèle de Règles Rhythmiques du Français Appliqué à la Synthèse de Parole.* Doctoral Thesis, University of Aix-en-Provence.

Pasdeloup, V. (1990b): "Multi-Style Prosodic Model for French Text-to-Speech Synthesis." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 193-196.

Peters, P. S. & R. W. Ritchie (1973): "Context-Sensitive Immediate Constituent Analysis: Context-Free Languages Revisited." *Mathematical Systems Theory* 6, pp. 324-333.

Phillips, M. S. (1985): "A Feature-Based Time Domain Pitch Tracker." *JASA* 77, pp. S9-S10.

Pickering, B. (forthcoming): "An Analysis of Transcriber Differences in the SEC." In Knowles & Alderson (eds).

Pierrehumbert, J. B. (1979): "The Perception of Fundamental Frequency Declination." *JASA* 66, pp. 363-369.

Pierrehumbert, J. B. (1980): *The Phonology and Phonetics of English Intonation.* Doctoral dissertation, Massachusetts Institute of Technology.

Pierrehumbert, J. B. (1981): "Synthesizing Intonation." *JASA* 70, pp. 985-995.

Pierrehumbert, J. B. & M. Beckman (1988): *Japanese Tone Structure.* London: MIT Press.

Pisoni, D. B. (1987): "Some Measures of Intelligibility and Comprehension." In Allen et al. (eds) 1987, pp. 151-171.

Pisoni, D. B., B. G. Greene & J. S. Logan (1989): "An Overview of Ten Years of Research on the Perception of Synthetic Speech." Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, September 1989, pp. 1.1.1-1.1.4.

Pisoni, D. B., L. M. Manous & M. J. Dedina (1987): "Comprehension of Natural and Synthetic Speech: Effects of Predictability on the Verification of Sentences Controlled for Predictability." *Computer Speech & Language* 2, pp. 303-320.

Pollard, C. J. (1988): "Categorial Grammar and Phrase Structure Grammar: An Excursion on the Syntax-Semantics Frontier." In Oehrle et al. (eds) 1988.

Pols, L. C. W. (1990): "Assessing the Speech Quality of Text-to-Speech Synthesisers." Proceedings of Verba 90, Rome, 22-24 January 1990, pp. 295-302.

Pratt, R. L. (1986): "On the Intelligibility of Synthetic Speech." Proceedings of IOA Autumn Conference, Windermere, November 1986, vol. 2 pp. 183-192.

Pratt, R. L. & J. P. Newton (1988): "Quantifying Text-to-Speech Synthesiser Performance: An Investigation of the Consistency of Three Speech Intelligibility Tests." In Ainsworth & Holmes (eds) 1988, vol. 2 pp. 453-458.

Prince, E. F. (1981): "Towards a Taxonomy of the Given/New Distinction." In P. Cole (ed.) 1981, *Radical Pragmatics*. London: Academic Press.

Quazza, S. & E. Vivalda (1987): "Contextual Syntactic Analysis for Text-to-Speech Conversion." In Laver & Jack (eds) 1987, vol. 1 pp. 389-392.

Quazza, S., G. Varese & E. Vivalda (1989): "Syntactic Pre-Processing for High Quality Text-to-Speech." In Tubach & Mariani (eds) 1989, vol. 1 pp. 506-509.

Quené, H. & A. Dirksen (1990): "A Comparison of Natural, Theoretical and Automatically Derived Accentuations of Dutch Texts." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 137-140.

Quené, H. & R. Kager (1989): "Automatic Accentuation and Prosodic Phrasing for Dutch Text-to-Speech Conversion." In Tubach & Mariani (eds) 1989, vol. 1 pp. 214-217.

Quené, H. & R. Kager (1990): *PROS—Automatic Sentence Analysis, Accentuation and Phrasing for Dutch Text-to-Speech Conversion.* Analysis & Synthesis of Speech Report 17. Utrecht: University Press.

Quirk, R. & J. Svartvik (1978): "A Corpus of Modern English." In H. Bergenholtz & B. Schaeder (eds), *Empirische Testwissenschaft.* Frankfurt: Lang.

Richter, H. (1984): "An Observation Concerning Intensity as a Predictable Feature of Intonation." In Gibbon & Richter (eds) 1984, pp. 283-310.

Riley, M. D. (1990): "Tree-Based Modelling for Speech Synthesis." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 229-232.

Rodriguez-Crespo, M. A. & J. G. Escalada-Sardina (1990): "Text Analysis System with Automatic Letter to Allophone Conversion for a Spanish Text to Speech Synthesizer." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 105-108.

Rossi, M. (1981): "L'Intonation n'est pas Congruente à la Syntaxe: Une Explication." In Rossi et al. 1981, pp. 290-296.

Rossi, M., A. Di Cristo, D. Hirst, P. Martin & Y. Nishinuma (1981): *L'Intonation: de l'Acoustique à la Sémantique.* Paris: Klincksieck.

Rühl, H-W. (1981): "Über ein Wortanalyse-Verfahren für Vorleseautomaten." *Acustica* 48, pp. 143-148.

Rühl, H-W., D. Dreissig & W. Kulas (1984): "Sprachausgabe: Die Ansteurung von Phonemsynthetisatoren." *Nachrichtentechnischezeitung Archiv* 6, pp. 243-248.

Russi, T. (1990): "A Framework for Morphological and Syntactic Analysis and its Application in a Text-to-Speech System for German." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 117-120.

Sakai, S. & K. Muraki (1990): "From Interlingua to Speech: Generating Prosodic Information from Conceptual Representation." Proceedings of ICASSP 1990, vol. 1 pp. 329-332.

Santi, S. & M. Grenié (1990): "Individual Strategies in Synthetic Speech Evaluation." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 265-268.

Sato, H. (1990): "Pitch Frequency Characteristics in Japanese Words Related to Phonemes." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 145-48.

Schank, R. C. (1975): "Using Knowledge to Understand." In Schank & Nash-Webber (eds) 1975.

Schank, R. C. & B. L. Nash-Webber (eds) (1975): *Issues in Natural Language Processing*. Cambridge, Mass.: MIT Press.

Scharpff, P. J. & V. J. van Heuven (1988): "Effects of Pause Insertion on the Intelligibility of Low Quality Speech." In Ainsworth & Holmes (eds) 1988, vol. 1 pp. 261-268.

Scheffers, M. T. M. (1988): "Automatic Stylisation of F_0 Contours." In Ainsworth & Holmes (eds) 1988, vol. 3 pp. 981-987.

Scherer, K. R., D. R. Ladd & K. Silverman (1984): "Vocal Cues to Speaker Affect: Testing Two Models." *JASA* 76, pp. 1346-1356.

Schmerling, S. (1976): *Aspects of English Sentence Stress*. Austin: Texas University Press.

Schnabel, B. & H. Roth (1990): "Automatic Linguistic Processing in a German Text-to-Speech Synthesis System." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 121-124.

Selkirk, E. O. (1981a): "On the Nature of Phonological Representation." In Myers et al. (eds) 1981, pp. 379-388.

Selkirk, E. O. (1981b) "On Prosodic Structure and its Relation to Syntactic Structure." In Fretheim (ed.) 1981, pp. 111-140.

Selkirk, E. O. (1984): *Phonology and Syntax: The Relation between Sound and Structure*. Cambridge, Mass.: MIT Press.

Shi, B. (1989): "A Chinese Text-to-Speech System." In Tubach & Mariani (eds) 1989, pp. 521-524.

Silverman, K. (1984): "F_0 Perturbations as a Function of Voicing of Prevocalic and Postvocalic Stops and Fricatives, and of Syllable Stress." Proceedings of IOA Autumn Conference, Windermere, pp. 445-452.

Silverman, K. (1987): *The Structure and Processing of* F_0 *Contours*. Doctoral Thesis, University of Cambridge.

Sorin, C., D. Larreur & R. Llorca (1987): "A Rhythm-based Prosodic Parser for Text-to-Speech Systems in French." Proceedings of the 11th International Congress of Phonetic Sciences, Tallinn, vol. 1 pp. 125-128.

Sparck Jones, K. (1985): "Compound Noun Interpretation Problems." In F. Fallside & W. A. Woods (eds), *Computer Speech Processing*, pp. 363-381. London: Prentice Hall.

Sperber, D. & D. Wilson (1986): *Relevance*. Oxford: Blackwell.

Spiegel, M., M. J. Altom, M. Macchi & K. Wallace (1989): "A Monosyllabic Test Corpus to Evaluate the Intelligibility of Synthesised and Natural Speech." Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, September 1989, pp. 1.2.1-1.2.4.

Spiegel, M., M. J. Altom, M. Macchi & K. Wallace (1990): "Comprehensive Assessment of the Telephone Intelligibility of Synthesised and Natural Speech." *Speech Communication* 9, pp. 279-292.

Sproat, R. W. (1990): "Stress Assignment in Complex Nominals for English Text-to-Speech." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 129-132.

Sproat, R. W. & M. Y. Liberman (1986): *Stressing English Compounds Correctly.* AT&T Bell Laboratories Technical Memo.

Sproat, R. W. & M. Y. Liberman (1987): "Toward Treating English Nominals Correctly." Proceedings of the 25th Annual Meeting of the ACL, pp. 140-146.

Steedman, M. J. (1985): "Dependency and Co-ordination in the Grammar of Dutch and English." *Language* 61, pp. 523-568.

Steedman, M. J. (1987): "Combinatory Grammars and Parasitic Gaps." *Natural Language & Linguistic Theory* 5, pp. 403-439.

Steedman, M. J. (1989): "Intonation and Syntax in Spoken Language Systems." Expanded version of papers in the proceedings of the AAAI Symposium on Spoken Language Systems and the DARPA Meeting on Spoken Language Systems.

Steedman, M. J. (1990): "Syntax and Intonational Structure in a Combinatory Grammar." In Altmann (ed.) 1990, pp. 457-482.

Stockwell, R. P. (1960): "The Place of Intonation in a Generative Grammar of English." *Language* 36, pp. 360-367.

Stockwell, R. P. (1972): "The Role of Intonation: Reconsiderations and Other Considerations." In Bolinger (ed.) 1972b, pp. 87-109.

Strawson, P. F. (1959): *Individuals*. London: Methuen.

Streit, M. (1989): "Presuppositions and Anaphora in a Question Answering Speech System." In Tubach & Mariani (eds) 1989, vol. 1 pp. 175-178.

Summerfield, A. Q. & M. Haggard (1972): "Articulatory Rate v. Acoustical Invariants in Speech." *JASA* 52, pp. 113-131.

Sydeserff, H. A., R. J. Caley, S. D. Isard, M. A. Jack, A. I. C. Monaghan & J. Verhoeven (1991): "Evaluation of Speech Synthesis Techniques in a Comprehension Task." Proceedings of Eurospeech 1991, September 1991, Genoa, pp. 277-280.

Tatham, M. (1990): "Preliminaries to a New Text-to-Speech Synthesis System." Proceedings of IOA Autumn Conference, Windermere, November 1990, vol. 1 pp. 233-240.

Teil, D. (1975): *Conception et Réalisation d'un Terminal à Réponse Vocale*. Doctoral Thesis, Université de Paris VI.

Terken, J. M. B. (1980): "The Distribution of Pitch Accents in Descriptive Language as a Function of Informational Variables." *IPO Annual Progress Report* 15, pp. 48-53.

Terken, J. M. B. (1982): "The Role of Accentuation in Comprehension: A First Test." *IPO Annual Progress Report* 17, pp. 57-62.

Terken, J. M. B. (1985): *Use and Function of Accentuation: Some Experiments*. Doctoral Dissertation, Rijksuniversiteit Leiden.

Terken, J. M. B. (1989a): "Automatic Synthesis of Dutch Intonation for Reading Machines: Tune, Prominence, Declination." Poster presented at the Second Conference on Laboratory Phonology, Edinburgh, July 1989.

Terken, J. M. B. (1989b): "Reaction to C. Gussenhoven and A. C. M. Rietveld: 'Fundamental Frequency Declination in Dutch: Testing Three Hypotheses'." *Journal of Phonetics* 17 pp. 357-364.

Terken, J. M. B. & R. Collier (1990): "Designing Algorithms for Intonation in Synthetic Speech." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 205-208.

Terken, J. M. B. & G. Lemeer (1988): "Effects of Segmental Quality and Intonation on Quality Judgements for Texts and Utterances." *Journal of Phonetics* 16, pp. 453-457.

Terken, J. M. B. & S. G. Nooteboom (1987): "Opposite Effects of Accentuation and Deaccentuation on Verification Latencies for Given and New Information." *Language & Cognitive Processes* 2, pp. 145-167.

't Hart, J. (1979): "Explorations in Automatic Stylization of F_0 Curves." *IPO Annual Progress Report* 14, pp. 61-65.

't Hart, J. (1984): "A Phonetic Approach to Intonation: From Pitch Contours to Intonation Patterns." In Gibbon & Richter (eds) 1984, pp. 193-202.

't Hart, J. & A. Cohen (1973): "Intonation by Rule: A Perceptual Quest." *Journal of Phonetics* 1, pp. 309-327.

't Hart, J. & R. Collier (1975): "Integrating Different Levels of Intonation Analysis." *Journal of Phonetics* 3, pp. 235-255.

Thomassen, J. M. E. W. (1979): "Melodic Accent in Computer Composed Metrical Tone Sequences." *IPO Annual Progress Report* 14, pp. 43-50.

Thomassen, J. M. E. W. (1980): "Melodic and Temporal Accentuation Combined." *IPO Annual Progress Report* 15, pp. 59-64.

Thompson, H. S. (1980): *Stress & Salience in English: Theory & Practice*. Palo Alto: Xerox PARC.

Thorsen, N. (1985): "Intonation and Text in Standard Danish." *JASA* 77, pp. 1205-1216.

Traber, C. (1990): "F_0 Generation with a Database of Natural F_0 Patterns and with a Neural Network." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 141-144.

Trager, G. & H. L. Smith (1951): *An Outline of English Structure*. Norman, Ok.: Battenburg Press.

Tropf, H. S. (1989): "Syntax in the Spoken Dialogue System SPICOS-II." In Tubach & Mariani (eds) 1989, vol. 1 pp. 30-33.

Tubach, J. P. & J. J. Mariani (eds) (1989): *European Conference on Speech Communication and Technology*, vols. 1 & 2. Edinburgh: CEP.

Turing, A. M. (1950): "Computing Machinery and Intelligence." *Mind* 59, pp. 433-460.

Tyler, L. K. (1990): "The Relationship between Sentential Context and Sensory Input." In Altmann (ed.) 1990, pp. 315-323.

Uhman, S. (1987): *Fokussierung und Intonation*. Ph.D. Dissertation, University of Konstanz.

Uldall, E. (1960): "Attitudinal Meanings Conveyed by Intonation Contours." *Language & Speech* 3, pp. 223-234.

Uldall, E. (1962): "Ambiguity: Question or statement? or 'Are you asking me or telling me?'" Proceedings of the 4th International Congress of Phonetic Sciences, Helsinki, pp. 779-783.

Uldall, E. (1964): "Dimensions of Meaning in Intonation." In D. Abercrombie et al. (eds), *In Honour of Daniel Jones*, pp. 271-279. London: Longmans.

Uldall, E. (1982): "A Footnote on Intonation and Attitude." *Edinburgh University Dept. of Linguistics Work in Progress* 15, p. 26.

Vaissière, J. (1971): *Contribution à la Synthèse par Regles du Français*. Ph.D. Thesis, Université de Grenoble.

Vaissière, J. (1983): "Language-Independent Prosodic Features." In Cutler & Ladd (eds) 1983, pp. 53-66.

van Bergem, D. R. & F. J. Koopmans-van Beinum (1989): "Vowel Reduction in Natural Speech." In Tubach & Mariani (eds) 1989, vol. 2 pp. 285-289.

van Bezooijen, R. (1985): *Characteristics and Recognizability of Vocal Expressions of Emotion*. Ph.D. Dissertation, Catholic University of Nijmegen.

van Bezooijen, R. (1989a): *Evaluation of an Algorithm for the Automatic Assignment of Sentence Accents in Written Text*. Analysis & Synthesis of Speech Report 9. Amsterdam: Institute of Phonetic Sciences.

van Bezooijen, R. (1989b): "Evaluation of the Suitability of Dutch Text-to-Speech Conversion for Application in a Digital Daily Newspaper." Proceedings of the ESCA Workshop on Speech Input/Output Assessment and Speech Databases, Noordwijkerhout, September 1989, pp. 6.3.1-6.3.4.

van Bezooijen, R. & L. C. W. Pols (1989): "Evaluation of a Sentence Accentuation Algorithm for a Dutch Text-to-Speech System." In Tubach & Mariani (eds) 1989, vol. 1 pp. 218-221.

van den Berg, R., C. Gussenhoven & T. Rietveld (forthcoming): "Downstep in Dutch: Implications for a Model." To appear in G. Docherty & D. R. Ladd (eds), *Papers in Laboratory Phonology* 2. Cambridge: CUP.

van der Hulst, H. & N. Smith (eds) (1982): *The Structure of Phonological Representations (Part I)*. Dordrecht: Foris.

van Riemsdijk, H. & E. Williams (1986): *Introduction to the Theory of Grammar*. Cambridge, Mass.: MIT Press.

Vanderslice, R. & P. Ladefoged (1972): "Binary Suprasegmental Features and Transformational Word-Accentuation Rules." *Language* 48, pp. 819-838.

Voiers, W. D. (1983): "Evaluating Processed Speech using the Diagnostic Rhyme Test." *Speech Technology* 1, pp. 30-39.

Weizenbaum, J. (1966): "ELIZA." *Communications of the ACM* 9, pp. 36-45.

Werth, P. (1979): "If Linear Order Isn't in the Base, then Where Is it?" In J. Meisel & M. Pam (eds), *Linear Order & Generative Theory*, pp. 187-251. Amsterdam: John Benjamins B. V.

Wheeler, D. (1988): "Consequences of some Categorially Motivated Phonological Assumptions." In Oehrle et al. (eds) 1988.

Willems, N. J. (1977): "Some Experiments on the Perception of Rhythmic Peaks of Prominence in Dutch." *University of Utrecht Institute of Phonetics Progress Report* 2, pp. 56-66.

Willems, N. J. (1978): "Discriminability of Dutch and English Intonation Contours." *University of Utrecht Institute of Phonetics Progress Report* 3, pp. 3-17.

Willems, N. J. (1979): "Perceptual Tolerances of Some Properties of Pitch Movements in English." *University of Utrecht Institute of Phonetics Progress Report* 4, pp. 71-91.

Willems, N. J. (1983): "STEP: A Model of Standard English Intonation Patterns." *IPO Annual Progress Report* 18, pp. 37-42.

Willems, N. J., R. Collier & J. 't Hart (1988): "A Synthesis Scheme for British English Intonation." *JASA* 84, pp. 1250-1261.

Willemse, R. & L. Boves (1991): "Context Free Wild Card Parsing in a Text-to-Speech System." Proceedings of ICASSP 1991, vol. 2 pp. 757-760.

Willemse, R. & L. Boves (1989): "Context Free Wild Card Parsing in a Speech-to-Text System." *University of Nijmegen Dept. of Language & Speech (Phonetics) Proceedings* 13, pp. 65-81.

Williams, B. J. & P. Alderson (1986): *Synthesising British English Intonation using a Nuclear Tone Model.* IBM UKSC Report 154.

Winograd, T. & F. Flores (1986): *Understanding Computers and Cognition.* New York: Addison-Wesley.

Witten, I. H. (1977): "A Flexible Scheme for Assigning Timing and Pitch to Synthetic Speech." *Language & Speech* 20, pp. 240-260.

Wothke, K. (1990): "From Orthography to Phonetic Transcription in the German Text-to-Speech System TETOS." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 219-223.

Xiang, Z. & G. Bi (1990): "A Neural Network Model for Chinese Speech Synthesis." Proceedings of Verba 90, Rome, 22-24 January 1990, pp. 417-420.

Yamashita, Y., N. Mizutani & R. Mizoguchi (1990): "Concept Description for Synthetic Speech Output System." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 241-244.

Yannakoudakis, E. J. & P. J. Hutton (1987): *Speech Synthesis & Recognition Systems.* Chichester: Ellis Horwood.

Yiourgalis, N. & G. Kokkinakis (1990): "Text Normalisation and Intonation in Text-to-Speech Synthesis of Greek." Proceedings of Verba 90, Rome, 22-24 January 1990, pp. 409-416.

Youd, N. & F. Fallside (1987): "Generating Words and Prosody for Use in Speech Synthesis." In Laver & Jack (eds) 1987, vol. 2 pp. 17-20.

Youd, N. & F. Fallside (1989): "Driving a Speech Synthesizer from Conceptual Input in the Context of a Voice Dialogue System." In Tubach & Mariani (eds) 1989, vol. 1 pp. 514-517.

Young, S. J. & F. Fallside (1980): "Synthesis by Rule of Prosodic Features in Word Concatenation Synthesis." *International Journal of Man-Machine Studies* 12, pp. 241-258.

Zinglé, H. (1982): *Traitement de la Prosodie Allemande dans un Système de Synthèse de la Parole*. Ph.D. Thesis, Université de Strasbourg II.

Zinglé, H. (1990): "Morphological Segmentation and Stress Calculus in German with an Expert System." Proceedings of the ESCA Workshop on Speech Synthesis, Autrans, 25-28 September 1990, pp. 133-136.

Zonnefeld, W. (1976): "Destressing in Halle's English Stress Rules.", *Linguistic Inquiry* 7, pp. 520-525.

Zwicky, A. M. (1970): "Auxiliary Reduction in English." *Linguistic Inquiry* 1, pp. 323-336.