Controlling Intonational Variation Using Escape Sequences in the Bell Laboratories Text-to-Speech System

Julia Hirschberg

June 22, 1995

Abstract

A current problem for text-to-speech systems is their failure to produce natural-sounding intonational variation. This memo documents the use of escape sequences to control intonational variation in the Bell Laboratories Text-to-Speech System (TTS).

1 Introduction

Text-to-speech systems, like the Bell Laboratories Text-to-Speech System (TTS), take unrestricted text as input and produce a synthetic spoken version of that text as output (Hirschberg et al., 1990). In the course of transforming one into the other, such systems typically must 1) identify words and determine their pronunciations; 2) decide how such items as abbreviations and numbers should be pronounced; 3) determine which words should be made prominent in the output, where pauses should be inserted, and what the overall shape of the intonational contour should be; 4) compute appropriate durations and amplitudes for each of the words that will be synthesized; 5) determine how the overall intonational contour will be realized for the text to be synthesized; 6) identify which acoustic elements will be used to produce the spoken text; and, 7) put all these decisions together to produce the synthetic utterance. Of all these problems, perhaps the one least successfully addressed by research or commercial text-to-speech systems is the third, the determination of the intonation which the synthetic utterance should be assigned.

No existing text-to-speech system, for English or other languages, controls intonational variation entirely successfully. For most English synthesizers, long sentences that lack commas are uttered without 'taking a breath', so that it is almost impossible to remember the beginning of the sentence by the end. Words that should be emphasized are not, while others may be much too prominent. And the basic intonational contour of a sentence is varied only by reference to its final punctuation; sentences ending with '.', for example, are always produced with the same contour, contributing to a numbing sense of monotony.

The newest version of TTS uses a variety of corpus-based techniques to improve default intonation: Prominence (or, PITCH ACCENT ASSIGNMENT) is varied according to several types of INFORMATION STATUS, which are inferred from the textual input (Hirschberg, 1992b); INTONATIONAL PHRASING decisions are based on syntactic and other information also inferred from text (Hirschberg, 1991; Hirschberg, 1992a); PITCH RANGE and timing variation are controlled via analysis of the input text's orthographic characteristics and certain lexical cues (Davis and Hirschberg, 1988; Sproat, 1992). However, older versions of TTS make no such attempts at variation; and the methods implemented in the newer versions are not always successful.

One aspect of TTS which distinguishes it from many text-to-speech systems is the degree of control over aspects of intonational variation which it provides to users. This control is specifiable at a fairly high level, via intonational ESCAPE SEQUENCES which can be used to change prosodic features such as prominence, phrasing, rate, range, and overall contour from the default specifications. Most of these escape sequences take the form **\!X**, where **\!** identifies the string as a special command to TTS and 'X' is a character or sequence of characters identifying the type of prosodic feature to be altered, and, in some cases, specifying the precise way in which it is to be altered. Every escape sequence must be followed by at least one blank space. These escape sequences can be inserted in the input by hand, or, for applications where the input has some predictability, specified via pre-processors (Hirschberg, 1990); such TTS pre-processors have been used, inter alia, for assigning appropriate intonation for driving directions (Davis and Hirschberg, 1988) and for pronouncing names and addresses retrieved from a database in a reverse telephone directory look-up. However, there is virtually no documentation on how and when to use these escape sequences, beyond a brief summary in (Liberman and L., 1985) which is condensed and updated in the current manpages.

This paper is intended to remedy this deficiency, by providing a general reference document on how to control intonational variation in TTS and some guidelines on when this intervention is useful. It will also describe how to change word pronunciations and (briefly) how to change various text analysis modes using TTS's front-end text analyzer, FREND. Note that the discussion below will distinguish between the current version (distributed in November 1992 and described in (Sproat, 1992)) and earlier versions (May 1991 and earlier, described in (Liberman and L., 1985)), since the chief differences in intonational defaults occur between these. The November 1992 version will be referred to as NewTTS, and all earlier versions will be referred to as OldTTS, when such a distinction is important. Otherwise, the system will simply be referred to as TTS. Appendix B provides a quick reference to the use of escape sequences in NewTTS and OldTTS. Additional forms of variation provided via escape sequences will also be discussed, such as specifying a particular word pronunciation in the input (Section 3). A brief summary of mode options provided by TTS' front end (FREND) will also be presented in Appendix D; fuller discussion of these modes is available in (Pavlovcik, 1991). Technical terms mentioned in this document are defined in Appendix A.

2 Changing Default Intonation in TTS

Intonational contours in TTS represent a partial implementation (Anderson, Pierrehumbert, and Liberman, 1984) of Pierrehumbert's description of English intonation (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986; Pierrehumbert and Beckman, 1988). A brief discussion of this system provides an introduction to the intonational phenomena that can be varied using TTS escape sequences, and to how they will be described in the remainder of this paper.

2.1 TTS' Intonational Model

TTS' intonational model is based on Pierrehumbert's model of English intonation (Pierrehumbert, 1980). In this system, and in TTS, intonational contours are represented as sequences of high (H) and low (L) targets in the FUNDAMENTAL FREQUENCY (f0). These contours are defined over prosodic units known as INTONATIONAL PHRASES in Pierrehumbert's system and as MINOR PHRASES in TTS documentation. Intonational phrases are composed of one or more smaller phrasal units called INTERMEDIATE PHRASES, plus a high or low boundary tone (H% or L%), which occurs at the right edge of the phrase. (INITIAL BOUNDARY TONES may also be specified in TTS, to control the pitch at the left edge of an intonational phrase.) Utterances are modeled as sequences of these phrases, and, for historical reasons, are sometimes referred to as MAJOR PHRASES in TTS documentation. So, in TTS documentation '*I like apples a lot, but I really prefer oranges.*' would represent a single major phrase, composed of two minor phrases '*I like apples a lot*' and '*but I really prefer oranges*'. In Pierrehumbert's system, such an utterance would simply be represented as two intonational phrases. Intermediate phrases are themselves composed of one or more PITCH ACCENTS, or intonational prominences, from an inventory of six possible types (**H***, **L***, **H*+L**, **H+L***, **L*+H,** or **L+H***),[1] plus a high (**H**) or low (**L**) PHRASE ACCENT which controls the pitch from the last pitch accent to the end of the phrase. Figure 1 illustrates the composition of legal intonational phrases in Pierrehumbert's system as a regular grammar.

Figure 1: Pierrehumbert's Description of English Intonation

Human speakers use pitch contours to convey part of the overall meaning of their speech. Pitch contours have as their domain the intonational phrase, and can be specified in Pierrehumbert's system as sequences of pitch accents, phrase accents, and boundary tones, as constrained by the grammar presented in Figure

[1]In theory, every intermediate phrase must contain at least one pitch accent. However, there have been observations that some TAG phrases — attributions like '*he said*' — can be uttered as separate intermediate phrases with no perceptible pitch accent.

1. For example, TTS produces '*The meeting's scheduled for eleven.*' as a statement, with the **H* L L%** contour shown in Figure 2. There are three high pitch accents (**H***) in this contour, followed by a low

Figure 2: A Declarative Contour: H* L L%

PHRASE ACCENT (**L**) and a low BOUNDARY TONE (**L%**) at the edge of the intonational phrase. The contour is represented abstractly in terms of the type of pitch accents, and the phrase accent and boundary tone, as **H* L L%** — in Pierrehumbert's system and in TTS. The **H* L L%** contour is the standard declarative contour in American English, and can be used to produce assertions or imperatives, inter alia.

In contrast, TTS assigns a 'yes-no-question' contour to the same sentence if it ends in '*?*'. TTS implements yes-no-question contours as shown in Figure 3, with a sequence of low pitch accents (**L***), followed by a high phrase accent (**H**) and a high boundary tone (**H%**) at the end of the phrase. In Pierrehumbert's system, this contour is represented as **L* H H%**.

Still other intonational contours can be produced either by altering individual components of intonational phrases separately — i.e., by specifying the tones to be assigned to accented words and to the ends of intermediate and intonational phrases directly — or by using one of a small set of overall contour escape sequences. Variation of pitch accent location and type will be described next, in Section 2.2. Variation in phrasing in TTS will then be described, in Section 2.3. Then variation of overall contour, by changing the defaults of individual components or by using contour-specific escape sequences, will be discussed, in Section 2.4.

2.2 Pitch Accent Variation

In English, pitch accents make items intonationally prominent. They usually appear as peaks or valleys in the f0 contour, as shown in Figures 2 and 3, and are modeled as such in TTS. Pitch accents are aligned with

Figure 3: A Yes-No-Question Contour: L* H H%

the lexically stressable syllable of a word — the syllable that seems rhythmically most prominent in that word when the word is spoken normally. For example, when we say the word *telephone*, its first syllable seems most prominent; when *telephone* is accented, its pitch accent is aligned with this first syllable, as shown in TTS' production of '*Buy a telephone.*' in Figure 4. While every word in English has a lexically stressable syllable, not every stressable syllable bears a pitch accent. We call such words 'deaccented' or 'unaccented'. An unaccented version of *telephone* is shown in Figure 5, for the same sentence, '*Buy a telephone*'.

In addition to f0 excursions, accented words in human speech are usually louder and longer than their unaccented counterparts, although TTS currently varies only f0 and duration to distinguish between accented and unaccented items. Accent type and prominence can be specified in TTS via the escape sequence **\!***xy, where x specifies a tone type from the set of implemented pitch accent types (**H*, L*, L+H*, L*+H, H+L*** — see Section 2.2.2) and y specifies a real-valued positive f0 multiplier on the most prominent f0 peak in the intonational phrase. This most prominent f0 peak in a minor phrase in TTS often coincides with what is termed the NUCLEAR ACCENT or NUCLEAR STRESS of a prosodic phrase — the most perceptually prominent accent in the phrase. Except in cases of DOWNSTEPPED pitch contours (see Section 2.10), or **L*** pitch accents, the nuclear accent usually contains the f0 peak of the phrase, and, in all cases, it is believed to fall on the last accented word in the intonational phrase.

For example, the default nuclear accent for a declarative utterance like the one in Figure 2 could be specified explicitly in TTS as **\!*H*1**, while an accent with half the prominence of this accent and with a **L+H*** accent could be specified via **\!*L+H*.5**. The input illustrated in Example (1a), perhaps in response to the query '*Is that cat under the car again?*, would produce an f0 peak on the preposition *on* which is 80% as high as the peak TTS specifies for *mat*, which bears nuclear stress in this utterance.

(1) a. The cat is **\!*H*.8** on the mat.

Figure 4: *telephone* Accented

Figure 5: *telephone* Unaccented

b. The cat is **!*** on the mat.

c. The cat is **!!** on the mat.

d. The cat is **!!L+H*2** on the mat.

e. The cat is on the **!-** mat.

Both type specification and prominence level can be omitted. Example (1b), for example, specifies simply that *on* be accented, leaving it to TTS to determine accent type (based upon default accent for the utterance type) and default prominence (based on position of word in phrase and prominence of other accents). And a shorthand method for changing the location of a phrase's nuclear stress is illustrated in Example (1c), where **!!** instructs TTS to make *on* the nuclear stressed element of its intonational phrase, thus both making *on* itself prominent and making all subsequent items in the phrase unaccented, such as *mat*. **!!** can also take pitch accent type and prominence specifications. In Example (1d), for instance, the escape sequence **!!L+H*2** tells TTS to change the default accenting of the sentence in three ways: First, *on* is to receive the utterance's nuclear stress; second, its peak is to be twice as prominent as TTS defaults would have normally specified for the nuclear accent of this utterance; third, *on* is to bear **L+H*** pitch accent. Finally, words can be made intonationally **un**-prominent (unaccented) using the escape sequence **!-**, as shown in Example (1e), in which *mat* is unaccented.

2.2.1 Varying Placement of Pitch Accent

Changing the accent pattern of an utterance by accenting some words and failing to accenting others can change its meaning dramatically. For example, the utterances in (2b) and (2c) show how the text in (2a) can be spoken in different ways to convey different meanings, depending upon the location of the phrase's nuclear stress.

Accent and Focus

(2) a. John only introduced Mary to Sue.

b. John only introduced MARY to Sue.

c. John only introduced Mary to SUE.

In (2b), Mary is the only person John introduced to Sue; in (2c), Sue is the only person John introduced Mary to. (2b) is false if John introduced Bill, as well as Mary, to Sue; (2c) is false if John introduced Mary to Bill, as well as to Sue.

The differences illustrated in Example (2) are called differences in FOCUS. An utterance's focus is most easily explained by asking 'To what question(s) is the utterance with this specified accent pattern a felicitous answer?' For example, (2b) is a felicitous response to the question '*Whom did John introduce to Sue?*', while (2c) is an appropriate response to the question '*To whom did John introduce Mary?*' In each case the focussed information is the information being requested, and is the most prominent information in the utterance.

By default, TTS produces the orthographic input shown in (2a) with nuclear stress on *Sue*, the last content word in the utterance, producing the effect of focus on *Sue*. This is equivalent to the explicit specification of nuclear accent shown in Example (3a). However, the alternative version, with nuclear stress on *Mary*, shown in (2b), can be produced by any one of the methods illustrated in (3b)-(3d); all of these inputs produce the same TTS output.

(3) a. John only introduced Mary to **!*H*1** Sue.

b. John only introduced Mary to **!-** Sue.

c. John only introduced **!*H*1** Mary to **!-** Sue.

d. John only introduced **!!** Mary to Sue.

In (3b), **!-** removes the default accent TTS would otherwise assign to *Sue*, leaving *Mary* as the last accented word in its intonational phrase, which TTS will thus assign nuclear stress. Thus, the specification of accent location, type, and prominence on *Mary* in (3c) is redundant in this circumstance. The same effect can be produced by **!!** in (3d), which assigns nuclear stress to *Mary* explicitly, making all subsequent items in the phrase unaccented, as described above.

The sentences in Examples (2)-(3) all contain the FOCUS-SENSITIVE OPERATOR, *only*, which interacts with the intonational prominence of pitch accents to produce the different interpretations of the sentence discussed above. That is, interpretation of a sentence contain such an operator will vary, depending upon the location of the sentence's nuclear accent. Other focus-sensitive operators include *all*, *most*, *some*, *sometimes*, *must*, and *It's odd that*, among other items. Like the sentences illustrated in Examples (2) and (3) above, utterances containing such items have different meanings, depending upon which words are intonationally focussed.

Accent and Pronominal Reference There are other circumstances when the TTS escape sequences which control accent placement can be used to choose the desired pronunciation of an ambiguous sentence from among several acceptable pronunciations to convey the intended meaning of the text. For example, pronouns are most often unaccented in English; when they are accented, it is often to convey a 'marked' interpretation — one distinct from the statistically most likely. In Example (4), the REFERENTS (persons being referred to) of the pronouns *he* and *him* will be different in (4a) and (4b), because the accent strategies are different (Lakoff, 1971).

(4) a. John called Bill a Republican and then he insulted him.

b. John called Bill a Republican and then HE insulted HIM.

In (4a), *he* and *him* are unaccented, and most people will understand that the speaker is asserting that John did two things: call Bill a Republican and then insult him. In (4b), with both pronouns accented, most listeners will understand that the speaker wishes to convey that John did one thing, which was to call Bill a Republican, and that this was tantamount to insulting Bill; in return, then, Bill insulted John.

In another case of interaction between REFERENCE RESOLUTION and pitch accent, shown in Example (5), the meaning of one clause can be affected by the intonational pattern of the preceding clause.

(5) a. John likes his colleagues and so does Sue.

b. John likes HIS colleagues and so does Sue.

In (5a), with no accent on *his*, listeners are likely to understand that 'John likes his colleagues and Sue also likes John's colleagues'. But in (5b), with *his* accented, listeners are likely to understand that 'John likes his own colleagues and Sue likes her own colleagues'.

Discourse Markers Other ambiguities arise in the use of DISCOURSE MARKERS, or CUE PHRASES (Hirschberg and Litman, 1991). Cue phrases are words and phrases such as *now*, *well*, *in the first place*, which can either serve as explicit indicators of discourse structure or can have a SENTENTIAL reading, often as adverbials. Example (6) illustrates the dual role that *now* plays, as a temporal adverbial or cue phrase.

(6) a. Now, Bill is a vegetarian.

b. Now Bill is a vegetarian.

c. NOW Bill is a vegetarian.

Variation in intonational phrasing and pitch accent are correlated with the distinction between 'discourse' and sentential uses of cue phrases (Hirschberg and Litman, 1991). Tokens interpreted as 'discourse' uses are commonly produced either as separate phrases (6a) or as part of larger phrases but unaccented (6b) or uttered with a low pitch accent (see Section 2.2.2). However, when cue phrases are produced with high prominence, they tend to be interpreted as temporal adverbs (6c). So (6a)-(6b) are likely to be interpreted as starting a new subtopic in a discourse, while (6c) is likely to be interpreted as a temporal statement: '*Now Bill is a vegetarian, although he wasn't before.*' And (6a)-(6b) convey no such assertion.

Accent and the Given/New Distinction The ability to override TTS defaults by specifying which words are to be accented and which unaccented is a very useful tool. In addition to allowing users to choose which among several interpretations of an utterance they wish to convey, this ability is also useful when TTS makes accenting 'errors', producing output that no human speaker would produce in a given context. Some common and easily corrected mistakes, such as making too many items prominent in a given text, are described below.

Human speakers commonly do **not** accent FUNCTION WORDS (e.g., prepositions, articles) and items that represent old, or GIVEN information in a discourse (Prince, 1981). For example, consider (7a) in the discourse below:

(7) a. the SENATE BREAKS for LUNCH at NOON, so i HEADED to the CAFETERIA to GET my STORY.

b. there are SENATORS, and there are THIN senators

c. For SENATORS, LUNCH at the cafeteria is FREE. For REPORTERS, it's not.

d. But CAFETERIA food is CAFETERIA food.

(7a) shows a simple pattern of unaccented function words and accented CONTENT WORDS (e.g., nouns, verbs, modifiers). However, in (7b), while human speakers are likely to accent the content word *senators* on first mention, they tend **not** to accent it on subsequent mention, when it represents 'old', or given, information. But in (7c), while *senators* still represents given information, speakers are likely to give it CONTRASTIVE STRESS (i.e., to use a particularly prominent pitch accent), contrasting *senators* with *reporters*. *Cafeteria* in this utterance is unaccented, since it has been previously mentioned in (7a), and is thus 'given' information. Finally, in (7d), this same 'given' item, *cafeteria*, is more likely to be accented, in part because of the stress pattern of the COMPLEX NOMINAL '*cafeteria food*' of which it is a part, and in part because all items in the utterance appear to be in some sense 'given' in this context. Since, in general, every intermediate phrase must contain at least one pitch accent, '*cafeteria*' is accented in this utterance.

By default, OldTTS accents content words and does not accent function words. NewTTS implements a more sophisticated accent assignment algorithm, described in detail in (Hirschberg, 1992b; Sproat, Hirschberg, and Yarowsky, 1992a; Sproat, 1992). This algorithm makes use of information inferred from text analysis to avoid accenting words representing given information and to accent words deemed 'contrastive' or 'focussed'. It also incorporates into its accent decisions information about how complex nominals are stressed. The same effects can be produced by hand for OldTTS by annotating the input, as illustrated in Example (8).

(8) The senate breaks for lunch at noon, so I headed to the cafeteria to get my story.
There are senators, and there are thin **!-** senators.
For senators, lunch at the **!-** cafeteria is free. For reporters, it's not.

Accenting Complex Nominals NewTTS implements a complex nominal parser, **np**, (Sproat, 1990; Sproat, Hirschberg, and Yarowsky, 1992a) to handle the accenting of noun-noun sequences such as '*cafeteria food*' in (7d) above. While OldTTS accents both elements of all complex nominals, with the more prominent accent on the right element of the nominal. NewTTS tries to determine for each individual complex nominal whether that nominal should have its major prominence on the rightmost or on the leftmost element, using a combination of general rules and specific lexical entries (Sproat, 1990; Sproat, Hirschberg, and Yarowsky, 1992b). For example, while OldTTS would make *food* the most prominent item in the nominal '*cafeteria food*' in (7d), using its default rule to accent all content words, NewTTS makes only the left element of '*cafeteria food*' prominent, (i.e., *CAFETERIA food*), which is the way most native speakers would produce it.

While **np** performs quite well, agreeing with human judgments about 86% of the time, it sometimes makes mistakes. (It should be noted that humans agree among themselves on stress assignment for complex nominals only about 91% of the time (Sproat, Hirschberg, and Yarowsky, 1992a).) For example, the complex nominals in Example (9) are produced even in NewTTS with fairly unlikely stress patterns:

(9) FRYING PAN
PART timers
HEART WARMING
LEGAL force
bite SIZE pieces
FULL COURT game
BOTTOM FEEDERS
THINK TANK
school BOARD meeting

These stress patterns can be changed using **!*** and **!-**, as in Example (10), to produce more acceptable output.

(10) frying **!-** pan
part **!*** timers
heart **!-** warming
legal **!*** force
!* bite **!-** size **!*** pieces

full \!- court \!* game
bottom \!- feeders
think \!- tank
\!! school board meeting

Improved stress patterns can of course also be assigned to complex nominals as necessary in OldTTS, just as illustrated in (10).

Accenting Phrasal Verbs NewTTS also improves over OldTTS by recognizing that PHRASAL VERBS (including VERB-PARTICLE and VERB-VERBAL PREPOSITION constructions) are accented differently from VERB-PREPOSITION sequences. For example, in simple verb-prepositions sequences, the verb is more prominent than the preposition, as in the sentences in Example (11).

(11) a. LIVE up the ROAD
b. MEET on the STREET
c. LEAVE in the LIMO

This is the default accenting strategy for phrasal verbs followed in OldTTS. But the same strategy used with the phrasal verbs in (12) sounds odd.

(12) a. BACK up the DISK
b. TURN on the LIGHT
c. LEAVE in the TYPO

In these phrases, the preposition or particle (*on*, *up*, *in*) **should** be uttered with more prominence than the preceding verb. NewTTS handles this problem in general by simple table look-up: likely phrasal verbs are accented so that the verbal preposition or particle is more prominent than the verb, as in Example (13):

(13) a. back UP the DISK
b. turn ON the LIGHT
c. leave IN the TYPO

However, clearly, for cases such as (11c), this simple approach will produce the wrong results, since NewTTS will produce '*leave in*' in (11c) the same way it produces '*leave in*' in (12c). The solution here is to explicitly remove the default TTS accent from the preposition ('*leave* \!- *in the limo*'), to produce the desired result.

Note that prepositions that are not part of such phrasal verbs may also be accented, to convey focus or contrast, as illustrated in Example (14).

(14) I didn't shoot \!* at him, I shot \!* past him.

Similarly, the unannotated sentences corresponding to those in Example (15) would sound odd in NewTTS because a final verb has not been accented by TTS default rules. These sentences can be made to sound more natural by accenting these items explicitly:

(15) But what if they **!*** were?
And here I **!*** am.
A lot already **!*** has.
How much power do staffers **!*** have?
But hardness there **!*** was.
I don't understand what a moderate smoker or drinker **!*** is.

The important point to remember in variation of pitch accent placement from a practical point of view is to be alert to the possibilities for ambiguity. It is usually not necessary to identify all the possible meanings of a sentence in order to avoid producing synthetic speech which conveys an unintended reading. It is usually enough to listen to the synthesized output to flag prominences that seem unnatural, given what one is trying to say. Then, one can simply override the default accent decisions by removing default accents from some items (**!-**) and accenting others (**!***), to produce something closer to the patterns one would use oneself in producing the sentence.

2.2.2 Varying Type of Pitch Accent

Accent type is commonly varied by human speakers to convey shades of meaning such as uncertainty or incredulity. This variation is important in text-to-speech systems, to convey naturalness and to eliminate monotony.

TTS implements five of Pierrehumbert's six pitch accents: **H***, **L***, **H*+L**, **L*+H,** and **L+H***. As noted in Section 2.2, the escape sequences **!*** and **!!** allow users to vary the type of pitch accent of the subsequent word by specifying one of these five accent types (with an optional prominence level specification), e.g. **!*L+H*.8**. However, the **H*+L** accent type is specified somewhat differently, as described below.

Simple Accents The simple **H*** and **L*** accents were illustrated above in Figures 2 and 3. The **H*** accent is the most common accent in American English. It is modeled as a simple peak in the f0 contour, as illustrated in Figure 2; this peak is aligned with the word's stressable syllable. **H*** accents are commonly used to convey that the accented item should be treated as 'new' information in the discourse, and is part of what is being asserted in an utterance (Pierrehumbert and Hirschberg, 1990). **H*** accents are the TTS default for prominent words in sentences other than yes-no-questions, where the default is the **L*** accent.

L* accents are modeled as valleys in the f0, as shown in Figure 3. **L*** accents have been described as conveying that the accented item should be treated as *salient* but is not part of what is being asserted as fact (Pierrehumbert and Hirschberg, 1990). In addition to their use in yes-no-questions, **L*** accents are often employed to make initial prepositions (16a) or adverbs (16b) prominent. They can also be used to favor a 'discourse' over a 'sentential' reading of cue phrases (16c)-(16d), as described above in Section 2.2.1.

(16) a. **!*L*1** To George, it seemed absurd.
b. **!*L*1** Yesterday the defense rested.
c. **!*L*1** Now we don't want to worry you.
d. **!*L*1** Well you never know.

L* accents are also used in the so-called 'surprise-redundancy' contour, described in Section 2.6.

Complex Accents **L+H*** accents can be used to produce a pronounced 'contrastive' effect, as in Example (17).

(17) John only introduced \\!!**L+H*** Mary to Sue.

(17) conveys more of a sense of contrast, for example, than (3b)-(3d), which are produced with the default **H*** accent. **L+H*** accents are appropriate in contexts like that presented in Example (18).

(18) a. \\!***L+H*** Never reboot your machine without syncing your disks.

b. . . . Not four \\!!**L+H*** nine oh five. Four \\!!**L+H*** one oh five.

Note also that in (18b), parts of a correction that have been said before (i.e., the uncorrected parts) can be made unaccented for a more natural effect (cf. Section 2.2.1).

L*+H accents can be used to produce the effect of uncertainty or deference in a smaller pitch range, or incredulity, when they are used in a larger pitch range and faster rate. These variations are illustrated in Examples (44a) and (44b) in Section 2.11.

H+L* accents occur rarely in English, but can be used in TTS to good effect when greater intonational variation is desired. They generally convey a rather playful air but their meaning is less well understood than other English accents. They appear to convey that the information being provided is something the hearer should have already been aware of. Some examples of appropriate use are presented in the exchanges in Example (19).

(19) a. A: There are no native speakers of German in our center.
B: \\!***H+L*1** Juergen's from Germany.

b. A: Why is that child so fussy?
B: She's \\!***H+L*** teething.

The other complex pitch accent in Pierrehumbert's system, **H*+L**, is not explicitly represented in the TTS implementation of this system. Instead, the effect of this accent is produced when DOWNSTEP is invoked via the specifications \\!_ or \\!@; see Section 2.4 for a description of contours involving **H*+L** accents and their meanings.

2.3 Phrasing Variation

As noted above, in Pierrehumbert's theory, there are two levels of phrasing in English, the intonational phrase and the intermediate phrase, a sub-unit. Both are identified by changes in f0, and are also frequently associated with other acoustic and prosodic cues, such as PHRASE-FINAL LENGTHENING (a lengthening of the syllable preceding the phrase boundary), glottalization ('creaky voice') over the last syllable or syllables in the phrase, and some amount of pause. Not all these features accompany all phrase boundaries in human speech, but, when they do, they tend to be more marked for intonational phrases than for intermediate phrases.

Intuitively, prosodic phrases divide an utterance into meaningful 'chunks' of information (Bolinger, 1989). Variation in phrasing can change the meaning hearers assign to a sentence. For example, the interpretation of a sentence like *Bill doesn't drink because he's unhappy* is likely to change, depending upon whether it is uttered as one phrase or two. Uttered as a single phrase, as in Figure 6, this sentence is commonly interpreted

Figure 6: One Intonational Phrase

as conveying that Bill does indeed drink — but the cause of his drinking is not his unhappiness. Uttered as two phrases (*Bill doesn't drink* | *because he's unhappy*), as in Figure 7, it is more likely to convey that Bill does **not** drink — and that unhappiness is the reason for his abstinence. In effect, variation in phrasing influences the scope of negation in the sentence. When the sentence is uttered as a single phrase the negative is interpreted as having wide scope — over the entire phrase, and, thus, the entire sentence. When *Bill doesn't drink* is separated from the second clause by a phrase boundary, the scope of negation is limited to just the first clause.

2.3.1 Phrasing Variation in TTS

TTS currently implements a partial version of Pierrehumbert's model of English phrasing: Pierrehumbert's intonational phrase is the basic unit of analysis for TTS' intonational model, as noted above. Intonational phrase boundaries are inserted automatically by OldTTS where non-final punctuation (*, ; :*) appears in the input — and only at such locations. It is thus simple to alter this level of phrasing in the input; just insert a comma to produce a phrase boundary. The input text in (20a) produces Figure 6, as a single intonational phrase, while the input in d (20b) produces Figure 7, as two intonational phrases.

(20) a. Bill doesn't drink because he's unhappy.

b. Bill doesn't drink, because he's unhappy.

In NewTTS, however, phrase boundaries insertion is only partially influenced by the presence of punctuation in the input. Phrasing procedures in NewTTS are trained on a large corpus of AP news text, and have been tested at better than 95% correct in comparison with human decisions (Hirschberg, 1992a).

Figure 7: Two Intonational Phrases

Intermediate Phrasing In English, intonational phrases may contain more than one intermediate phrase; in TTS, only one intermediate phrase is allowed per intonational phrase. So, in effect, only a single level of phrasing from the Pierrehumbert system is implemented. In consequence, it is impossible to produce an utterance like that illustrated in Figure 8 directly, with minor prosodic boundaries within the intonational phrase, in either OldTTS or NewTTS. However, in OldTTS a level of phrasing called FUNCTION-WORD GROUPING **is** implemented, which produces at least the tonal characteristics — the f0 change — of intermediate phrasing. This level of phrasing, which inserts a FLOATING PHRASE ACCENT before non-initial function words in the text to be synthesized, is done automatically in OldTTS, but has not beend implemented in NewTTS. Floating phrase accents can be inserted in OldTTS or NewTTS by hand, using **!f***xy*, where *x* can be either L (default) or H, and **y** is a real number controlling prominence. The tonal changes occur at the point in the input text where the escape sequence occurs, as illustrated in Examples (21b)-(21c) below. The normal range of prominence for such accents is 0 to 1.

In OldTTS and in NewTTS, an effect similar to that of Pierrehumbert's intermediate boundary, which can sometimes be useful in increasing naturalness, can in fact be simulated by using **!f***xy*, as described above, in conjunction with other escape sequences which vary phrase-final lengthening (**!**}*y* (where *y* specifies degree of lengthening as a real multiplier on the default lengthening for syllables ending minor phrases, i.e. **!**}**1**, (Rowley and Silverman, 1989)) and pause (**(** ***[***y***]** **)**) (where *y* specifies amount of pause in centiseconds. For example, to produce the effect of an intermediate — instead of an intonational — phrase boundary between city and state for the input text (21a), the escape sequences shown in (21b) can be substituted for the comma.

(21) a. He comes from Little Rock, Arkansas.

b. He comes from Little Rock **!fL1** **!**}**.5** **(*[4]** **)** Arkansas.

Figure 8: An Intonational Phrase with two Intermediate Phrases

c. He comes from Little Rock **\!fH1 \!}.5 \(*[4] \)** Arkansas.

\!fL1 inserts a low phrase accent with prominence '1' after *Rock*; **\!}.5** lengthens the syllable preceding it (again, *Rock*) to half the duration it would have preceding a full intonational boundary, and **\(*[4] \)** inserts four centiseconds of silence. A similar effect is produced in (21c) with a high phrase accent.

In natural speech, a low phrase accent tends to produce more of a sense of juncture between conjuncts than a high phrase accent. For example, when (22) is uttered with a low phrase accent before *or*, it conveys more of a sense that there may be other juices available; when uttered with a high phrase accent, the two juices tend to be viewed as an exhaustive set.

(22) Would you like apple juice | or orange juice?

NewTTS will eventually implement intermediate phrase boundaries in a more principled way. However, the general effect of a lesser phrase boundary is very useful currently, especially when intonational variation is to be controlled from script.

Examples (21b)-(21c) demonstrate how lesser intermediate boundaries can be substituted for full boundaries in TTS. Other common phrasing problems in OldTTS' and NewTTS' default phrasing algorithms include:

- inserting infelicitous function group boundaries (OldTTS only);
- inserting infelicitous intonational phrase boundaries;
- failing to insert intonational phrase boundaries;

Function Word Group Boundaries The problem of TTS inserting unwanted function word group boundaries arises in OldTTS when two adjacent function words which seem to group naturally together are assigned by OldTTS to different function word groups. The intervening boundary is perceived as a slight but disconcerting rapid f0 change, which can produce a 'burbling' effect. Examples of this are found in OldTTS' production of Example (23), where the default function word group boundaries are indicated by '|'.

(23) a. Go up | to the window and place your bet.
b. Give | me all your money.

This problem can be handled in OldTTS using the escape sequence \!> to turn off function word grouping entirely. It is then possible to insert floating phrase accents by hand, as noted above, using \!f*xy*. So, the sentences in Example (23) can be improved in OldTTS by annotating the input as shown in (24), in effect shifting the location of the function group boundary to a more appropriate spot in the utterance — or eliminating such boundaries entirely.

(24) a. \!> Go up to the window \!**fL1** and place your bet.
b. \!> Give me all your money.

Since function group boundaries are chiefly apparent to TTS listeners only when they are inappropriately assigned, a safe alternative in OldTTS is simply to eliminate all function word group intonation, by prefacing every input sentence with \!> and insert no new function group boundaries at all.

Intonational Phrase Boundaries Problems of inserting or omitting full intonational boundaries infelicitously are general problems for all text-to-speech systems. As noted above, it is simple to add or delete major boundaries in OldTTS, since such boundaries are completely determined by orthographic punctuation. The same control can be obtained currently in NewTTS by specifying an alternative phrasing algorithm which emulates phrasing in OldTTS and then inserting or deleting internal punctuation as in OldTTS. The next version of the NewTTS phrasing algorithm will provide a better method for overriding phrasing defaults.

The advantage of decoupling phrasing decisions from orthography can be seen in Example (25), a sentence chosen at random from all sentences in the AP news of October 8, 1992, that were over 60 tokens in length.

(25) For his part , Clinton told reporters in Little Rock , Ark. , on Wednesday that the pact can be a good thing for America if we change our economic policy to rebuild American industry here at home and if we get the kind of guarantees we need on environmental and labor standards in Mexico and a real plan to help the people who will be dislocated by it.

When (25) is fed to OldTTS it is almost incomprehensible: The lack of phrase boundaries, due to the scarcity of internal punctuation marks, means the last fifty-seven words are uttered as if in a single breath. And the punctuation provided in the original text in fact should *not* be realized with a full intonational boundary, as discussed above with respect to the sentences in (21a). NewTTS produces this sentence with the phrase boundaries marked in (26).

(26) For his part | Clinton told reporters in Little Rock , Ark. , on Wednesday | that the pact can be a good thing for America | if we change our economic policy | to rebuild American industry here at home | and if we get the kind of guarantees we need on environmental and labor standards in Mexico | and a real plan | to help the people | who will be dislocated by it.

2.3.2 Disambiguation via Phrasing

Above we noted one way in which phrasing affects meaning, by helping to favor one reading of negation over another in ambiguous sentences like those presented in (20). There are many other constructions in which phrasing appears to influence disambiguation, and it is useful to keep these in mind for synthesis applications in which the text is known in advance. Some are shown in Examples (27)-(35), where boundaries are again marked by |:

(27) a. Anna frightened the woman | with the gun.
[Anna held the gun]

b. Anna frightened | the woman with the gun.
[the woman held the gun]

(28) a. Mary knows many languages you know.
[Mary knows many languages that you also know]

b. Mary knows many languages | you know.
[as you are aware, Mary knows many languages]

(29) a. The animal that usually fights the lion is missing.
[the lion's normal opponent is missing]

b. The animal that usually fights | the lion | is missing.
[the lion is missing]

(30) a. My brother who is a writer needs a new job.
[I have at least one other brother]

b. My brother | who is a writer | needs a new job.
[I may or may not have other brothers]

(31) a. John laughed | at the party.
[John laughed while he was at the party]

b. John laughed at | the party.
[John made fun of the party]

(32) a. If you need me | when you get there call me.
[if you need me, call me when you arrive]

b. If you need me when you get there | call me.
[if you need me when you arrive, call me]

(33) a. This collar is dangerous to younger | dogs and cats.
[the collar may be dangerous to younger dogs and younger cats]

b. This collar is dangerous to younger dogs | and cats.
[the collar may be dangerous to younger dogs and all cats]

(34) a. Stir in rice wine | and seasonings.
[stir in two ingredients]

b. Stir in rice | wine | and seasonings.
[stir in three ingredients]

(35) a. We only suspected | they all knew that a burglary had been committed.
[we only suspected that they all knew that a burglary had been committed]

b. We only suspected | they all knew | that a burglary had been committed.
[they all knew that we only suspected that a burglary had been committed]

c. We only suspected | THEY all KNEW | that a burglary had been committed.
[we only suspected but they all knew that a burglary had been committed]

2.4 Overall Contour Variation

TTS assigns a default 'declarative' (**H* L L%**) contour to input sentences ending in sentence final punctuation other than a question-mark (i.e. '.' or '!') and a default 'yes-no-question' contour to sentences ending in '?' — unless the latter are deemed to represent *wh*-questions. TTS infers that sentences ending in '*?*' but containing initial *wh*-words (*who*, *what*, *why*, *when*, or *how*) represent *wh*-questions, which, in English, are commonly uttered with declarative intonation. So, questions like '*Who says that's all?*' and '*Where is the meeting?*' are pronounced by default with a **H* L L%** contour.

Intonational contours different from TTS defaults can be produced using a set of escape sequences that alter overall contour type, overriding inferences based on text analysis.

\!? produces a yes-no-question contour;

\!/ produces a CONTINUATION RISE, conveying a sense that there is 'more to come';

\!_ produces a DOWNSTEPPED CONTOUR, producing a sequence of compressed pitch accents resembling a staircase in the f0.

Intonational contours may also be varied by varying the inividual components of contours — pitch accents, phrase accents, and boundary tones — as described in Sections 2.2 and 2.3.

2.4.1 The Declarative Contour

As noted in Section 2.1, TTS' default intonational contour for utterances ending in '.' or '!' is the **H* L L%** contour, one or more **H*** accents followed by a low phrase accent and low boundary tone; this is the standard 'declarative' contour in American English. It is also commonly used with *wh*-questions. This contour is illustrated in Figure 2. It may be produced explicitly in TTS by specifying **H*** accents, a low phrase accent, and a low boundary tone. For example, the input '*The \!*H* meeting's \!*H* scheduled for \!*H* eleven \!pL \!{bL .* ' would produce the default contour illustrated in Figure 2 explicitly.

2.5 The Interrogative Contour

For utterances ending in '?' which are not identified as *wh*-questions, the default contour in TTS is the YES-NO-QUESTION CONTOUR, described as **L* H H%** in Pierrehumbert's system; this contour may also be produced via the escape sequence **!?**, placed anywhere in the intonational phrase, or by explicitly assigning **L*** pitch accents to accented words and by specifying a high phrase accent (**!pL1**) and high boundary tone (**!bH1**) for the intonational phrase. This contour is illustrated in Figure 3, and could be produced explicitly in TTS by the following input: '*The \\!*L* meeting's \\!*L* scheduled for \\!*L* eleven \\!pH \\!{bH .* '

2.6 Surprise-Redundancy

L* accents can also be combined with **H*** accents to produce the so-called SURPRISE-REDUNDANCY contour, as in Example (36); the phrase accent and boundary tone are both low, the TTS defaults, and thus do not need to be specified explicitly

(36) a. The **!*L*1** blackboard's painted **!*H*1** orange.

b. Your **!*L*1** bank **!-** account is **!*H*1** empty.

The contour TTS produces for (36a) is illustrated in Figure 9.

Figure 9: The Surprise/Redundancy Contour

2.7 Continuation Rise

The CONTINUATION RISE contour (**!/**), which is represented by a low phrase accent and high boundary tone (**L H%**) in Pierrehumbert's system, can be used to convey that there is 'more to come', as in Example (37).

(37) a. The number is \!/ : 555-1212.

b. Open the carton \!/ . Now remove the monitor carefully.

c. Open \!/ the carton . Now remove the monitor carefully.

Note that the \!/ escape sequence can be located anywhere within an intonational phrase, but will produce the desired effect at the end of that phrase. That is, the output of (37b) and (37c) will be identical.

The contour produced by TTS for (37a) is shown in Figure 10.

Figure 10: A Contour Ending in Continuation Rise

2.8 Rise-Fall-Rise Contour

The RISE-FALL-RISE contour (one or more **\!*L*+H** accents plus a low phrase accent and high boundary tone) can be used to indicate either uncertainty or incredulity, depending upon the speaking rate and pitch range (See Section 2.11). In (38a), **L*+H L H%** is produced to indicate uncertainty; in (38b), it is produced to convey incredulity.

(38) Did you finish those slides?

a. **\!r1.3** \!{**T125** **\!*L*+H1** Sort of \!/ .

b. **\!r.8** \!{**T150** **\!*L*+H1** Sort of \!/ .

When the rise-fall-rise contour conveys uncertainty (38a), it is produced in a narrower pitch range and more slowly than when it conveys incredulity (38b) The contour TTS produces for the input shown in (38a) is illustrated in Figure 11, and that for the input in (38b) in Figure 12.

Figure 11: The Rise-Fall-Rise Contour Conveying Uncertainty

Figure 12: The Rise-Fall-Rise Contour Conveying Incredulity

2.9 Plateau Contour

A PLATEAU contour (**H* H L%** conveys the sense that the speaker is talking about an 'open-ended set', as in Example (39).

(39) The Johnsons are solid citizens. They **!*H*** pay their **!*H*** taxes **!pH1** **!bL1** **!{F1.0** . They **!*H*** attend **!*H*** PTA **!*H*** meetings **!pH1** **!bL1** **!{F1.0** . They're just good people.

The contour TTS produces for '*They* **!*H*** *pay their* **!*H*** *taxes* **!pH1** **!bL1** **!{F1.0** . ' is shown in Figure 13.

Figure 13: Plateau Contour

2.10 Downstepped Contours

Finally, a type of downstepped contour can produce a variety of effects in TTS, including the opening or closing of a topic, or a sense that the speaker is instructing the hearer. This **H*+L L L%** contour, of the downstepped contours described in Pierrehumbert's system, is implemented in TTS in two ways: by **!_**, which compresses the pitch range for the high tone of every non-initial pitch accent in the minor phrase (by a factor called the DOWNSTEP RATIO, implemented as **!{K***y*, where *y* is a real number, normally ranging between 0 and 1) or by **!@**, which compresses the pitch range only of the next pitch accent in the intonational phrase. In OldTTS, **!<** may be used to downstep the first accent of every non-initial function word group in a minor phrase. Some examples of the appropriate use of **H*+L L L%** appear in Example (40).

(40) a. **!_** **!{T155** **!{K.6** **!*H*1** Today we're **!*H*1** going to **!*H*1** look at the **!*H*1** population of Ghana.

b. Hit the **\!@** hint **\!-** key.

c. **\!_ \!{R85 \!{K.6 \!{T140 \!*H*1** Here is **\!*H*1** how to **\!*H*1** inflate the raft.

The contour produced from the input in Example (40a) is illustrated in Figure 14. This contour is produced

Figure 14: A Downstepped Contour

most effectively in TTS if the pitch range is initially increased from the default value, and if all of the pitch accents to be compressed are originally assigned a prominence of '1'. It may also be desirable to change the REFERENCE LINE (see Section 2.11).

Speakers produce many other varieties of contour in English, whichd can be mimicked in TTS. Any combination of pitch accents, phrase accents, and boundary tone may be used to convey different shades of meaning.

2.11 Varying Timing and Pitch Range

Two parameters that can be varied in TTS to provide a variety of effects are timing and pitch range. These may be varied separately or together to convey a variety of effects.

2.11.1 Variation in Timing

Speaking Rate Variation in speaking rate can be accomplished via the escape sequence **\!r**y, where y is either an integer (OldTTS) or a continuous-valued time multiplier (NewTTS) (van Santen and Sproat, 1992). **\!R**y changes rate for the invocation of the program. In OldTTS, the default speaking rate is '2', with '0' and '1' available to specify faster speaking rates (41b) and '3' and '4' specifying slower rates (41a).

(41) a. \!**r1** This is a fast utterance.

b. \!**r4** This is a slow utterance.

In NewTTS, the default speaking rate is '1', and factors greater than the default specify slower rates as real multipliers on the default (42a), and factors smaller specify faster rates (42b).

(42) a. \!**r1.5** This is a slow utterance.

b. \!**r.7** This is a fast utterance.

Varying Interphrasal Pause Another aspect of timing, pausal duration, can be varied via the escape sequences \!**si**y and \!**sf**y. Each of these escape sequences can be given a value in centiseconds (y) to produce initial or final silence, respectively, within an intonational phrase. Regardless of their location in the input string, each produces silence at the beginning phrase (\!**si**y) or at the end (\!**sf**y) of the current minor phrase. So, in Example (43), ninety centiseconds of silence will be produced at the beginning and end of the phrase.

(43) In fact, there is silence at \!**si90** \!**sf90** the beginning and at the end of this phrase, she said.

2.11.2 Variation in Pitch Range

Variation in aspects of pitch range as well as rate can change the meaning of particular intonational contours, such as the rise-fall-rise contour (**L*+H L H%**), as noted in Section 2.8. In Example (44a), for instance, a smaller pitch range and slower rate combine with the rise-fall-rise contour to convey an impression of uncertainty.

(44) a. \!{**T115** \!**r1.5** I \!***L*+H1** think I should connect you to the clothing department.

b. \!{**T140** \!**r.5** I can't \!***L*+H1** believe you \!!**L*+H1** did that.

In (44a), the synthesizer conveys that it is not sure it is connecting the caller to the correct department. In (44b), however, the same contour is used with an expanded pitch range and a faster rate to convey incredulity.

Range variation can also convey differences in degree of speaker 'involvement', or communicate the topic structure of a text (Hirschberg and Pierrehumbert, 1986; Pierrehumbert and Hirschberg, 1990; Grosz and Hirschberg, 1992).

Size of Overall Range In general, new topics are begun in a wider pitch range and end in a narrower range. Subsequent expansion of pitch range tends to be associated with a topic shift. So, for example, the structure of the text in Example (45) can be conveyed by range variation; range variation is accomplished by varying the TOPLINE of the current phrase, via the escape sequence \!{**T**y, where y represents the height of the most prominent pitch accent in the intonational phrase in Hz.

(45) \!{**T140** Poor communication on the bridge, a high rate of speed and incomplete navigational charts contributed to last summer's grounding of the Queen Elizabeth 2, the Coast Guard said Thursday. \!{**T125** The National Transportation Safety Board said earlier this week that poor communication among the crew was the main reason the ship hit an uncharted rock Aug. 7 off Cape Cod.

\!{**T115** The NTSB ruled that poor charts and faulty assumptions about the ship's draft were only contributing factors.
\!{**T125** The Coast Guard didn't place the bulk of the blame on any one factor.
\!{**T115** It said the captain and harbor pilot didn't properly discuss the vessel's intended route...

The default topline in the male version of TTS is 115 Hz; that of the female is 195 Hz. Two other range parameters can also be varied similarly, the BASELINE (\!{**B***y*) and the REFERENCE LINE (\!{**R***y*). The former represents the lowest point a speaker reaches in his/her pitch range, across all utterances; prominent **L*** accents approach this baseline. The default for the male speaker is 75 Hz; for the female, 149 Hz. The reference line is used to calculate high and low tones with respect to the topline and baselin; it is sometimes useful to vary this parameter explicitly, when it is desirable to shift an entire utterance up in the pitch range. The male speaker's reference line is set by default at 86 Hz, halfway between the baseline and the default topline; the female speaker's reference line is set at 165 Hz. Note that each range parameter described above can be set by \!{*X* for the current minor phrase; range parameters may be set for the current invocation of the program via \!{**!**. So, for example, \!{**!T150** changes TTS' topline to 150 for the entire invocation of the program.

Other discourse phenomena can be conveyed by variation of these range parameters. For example, parentheticals are generally uttered in a compressed pitch range and with a faster speaking rate than other phrases (Grosz and Hirschberg, 1992), as illustrated in Example (46).

(46) \!{**T130** The crowd, \!{**T105** \!**r.7** which had been waiting for hours in the snow, \!{**T130** \!**r1** became unruly.

Attributive tags (e.g. '*he said*') are also characterized by a compressed pitch range (Grosz and Hirschberg, 1992), as illustrated in Example (47).

(47) My dog Millie knows more about foreign policy than those two bozos, \!{**T100** he said.

Range variation is also reliably associated with the beginning of direct quotations and with attributive tags (Grosz and Hirschberg, 1992). The first phrase of a direct quotation is uttered in an expanded pitch range, as illustrated in Example (48a).

(48) a. Bush said, \!{**T150** my dog Millie knows more about foreign policy than those two bozos.
b. Bush said my dog Millie knows more about foreign policy than those two bozos.

Example (48b) sounds much more like an indirect quotation than Example (48a).

Final Lowering It is also possible to vary another range parameter, FINAL LOWERING, using \!{**F***y*; here *y* is a multiplier for the f0 values which would otherwise obtain at the end of the phrase in the absence of final lowering. Final lowering compresses the pitch range during the last half second of a major phrase, or utterance, and can convey a sense that the speaker has finished his/her 'turn'.

2.12 Varying Timing and Range Together

When final lowering, overall range, and pausal duration are varied together, the result sounds reasonably natural for the synthesis of longer texts, as in Example (49).

(49) ...
\!{T145 \!{F.95 \!si90 \!/ The only way to learn a new programming language is by writing programs in it.
\!{T130 \!{F.95 \!si60 \!/ The first program to write is the same for all languages:
\!{T115 \!{F.95 \!si30 \!/ Print the words 'Hello world'.
...
\!{T100 \!{F.85 \!si15 With these mechanical details mastered, everything else is comparatively easy.

The beginning phrase of this paragraph is marked by an expanded pitch range and a relatively long pause at the beginning of the phrase; there is little final lowering at the end of this phrase — the final lowering ratio compresses the overall pitch range at the end of the phrase to 95% of what it otherwise would be, instead of the default. Within the text, phrases are given somewhat narrower pitch ranges, final lowering continues to be minimal, and there are shorter pauses between phrases. At the end of the text, to indicate that the topic is concluding, final lowering is increased, the pitch range is quite compressed, and there is little pause between this final phrase and the preceding one. A subsequent introduction of a new topic would be characterized by another pitch range expansion, longer intervening pause, and minimal final lowering. These features mimic behavior observed in (Grosz and Hirschberg, 1992), although the absolute values assigned in synthetic speech are chosen for maximum effect.

In NewTTS, pitch range is varied systematically depending upon the paragraph structure of a text; paragraph initial phrases are produced in an expanded range and subsequent phrases are uttered in a smaller one. In OldTTS, similar variation can be produced by rule, as noted in (Hirschberg, 1992b; Grosz and Hirschberg, 1992). Appendix E gives some simple scripts which can be used to vary TTS parameters by rule.

3 Changing Default Word Pronunciations

All text-to-speech systems mispronounce words from time to time, either because the word is unfamiliar or because the same orthographic string has several pronunciations and the system has chosen the wrong one. For example, a proper name may not be listed in the system's dictionary and may require pronunciation rules different from those used in the base language of the synthesizer. Telephone books in the United States contain many personal names derived from varied non-English origins; often the pronunciation of these names may be Anglicized by those bearing the name or those who refer to them, so it is not always easy to decide what the best pronunciation should be for a text-to-speech system. For example, if Tony Perez' surname were pronounced by a native Spanish-speaker (*PERez*), it might not be easily recognized by millions of baseball fans used to the Anglicized form *perEZ*. Similarly, residents of Lima, Ohio, do not pronounce the name of their city like residents of Lima, Peru, do. In TTS, the guiding principle in pronouncing proper names has been to emulate a native speaker of American English with no special knowledge of other languages; so, TTS pronounces 'Perez' with stress on the second syllable. However, for some applications, clearly, a different choice will be more appropriate. For example, if names are used in applications in which a user will hear his or her own name, the user will probably want that name pronounced as he or she would normally pronounce it. So, it is often necessary to customize the pronunciation of proper names.[2]

3.1 Specifying Input Phonetically

TTS provides a mechanism for explicitly specifying which sounds should be concatenated to produce a particular word. The general format of such phonetic input is \\(...**X** ...\\), where 'X' is a string taken from a phonetic symbol set. In older versions of TTS, either of two symbol sets, JPOBET or CHCBET, may be used to specify input phonetically; in NewTTS, chcbet has been selected as standard, to provide uniformity with the dictionary representation. So, altering the default TTS pronunciation of *Lima* in Example (50a) is accomplished in OldTTS by either (50b) or (50c) and by (50d) in NewTTS.

(50) a. John Rudolph, president of Rudolph Foods Co. in Lima, Ohio, said his company, which specializes in pork rinds and cracklings and controls 35 percent of the national market for those products, has received a lot of publicity. (AP 11-22-88)

b. OldTTS (jpobet): ....\\(**lai&muh** \\), Ohio ...

c. OldTTS (chcbet): ...\\(**C “lIm&** \\), Ohio ...

d. NewTTS (chcbet): ...\\(**“lIm&** \\), Ohio ...

3.2 Specifying Lexical Stress

Lexical stress placement can be specified explicitly in either symbol set. In jpobet, syllables with main stress are marked by the symbol '**&**' and syllables with secondary stress by '**+**' — both marked at the end of the appropriate syllable. The jpobet transcription of *telephone* is '\\(**tel&efon+** \\)', for example. Lexical stress in chcbet dis marked at the beginning of syllables: '**"**' indicates a primary stressed syllable and '**.**' indicates a secondary stress, e.g., \\(**"lI.m&** \\). In some cases, only the stress placement need be changed to override the desired TTS default, such as *OBject* and *obJECT*.

3.3 Homographs

Some words (called HOMOGRAPHS) have identical spellings but different pronunciations and meanings (Sproat, Hirschberg, and Yarowsky, 1992a). In many cases, particularly those in which one member of the homographic pair is a noun and the other a verb (e.g. object, desert, permit), this pronunciation difference is one of primary stress placement. The noun has primary stress on the first syllable and the verb, on the second. In many cases TTS infers the correct part-of-speech of the word, and thus produces the correct pronunciation. However, in cases such as (51a), either the correct part-of-speech **or** the correct pronunciation must be provided:

(51) a. Permit access is strictly enforced.

b. Permit$n access is strictly enforced.

c. Object$v as much as you like.

[2]For more information on proper name pronunciation in TTS, see (Schulz, 1987; Buchsbaum, 1986; Church, 1985; Church and Golding, 1987; Coker, Church, and Liberman, 1990).

d. Begin$C took office later.

e. chcbet: \(**"pRmit** \) access is strictly enforced.

f. Permit$1 access is strictly enforced.

In example (51b), the correct part-of-speech (noun) is given in the input via **$n**; other part-of-speech specifiers include **$v** for 'verb' **$j** for 'adjective' and **$C** for 'proper noun'. Some sample uses of these are shown in Examples (51c)-(51d). Or, the correct pronunciation may be specified explicitly, as in (51e). A non-default pronunciation may also be specified by number (51f), based upon the order in which pronunciations for the string are listed in TTS' dictionary, as in Example (51f). At present, these orderings can only be found by testing explicitly using the pronunciation program *pron*; for example, the following procedure identifies the two pronunciations of *read*:

```
$/usr/local.1122/bin/tts.nov92/pron
read
read "rEd  _W_  7  read
read$2
read "red  _W_  7  read
```

3.4 Dictionary Cache

Mispronounced words with no entry in TTS' dictionary may be added to the TTS CACHE, a list of words and their pronunciations which TTS checks before general dictionary look-up (Sproat, 1992). Example (52) shows some sample cache entries:

(52) abcs "A"bE"sEz
aids "Adz//acronym
associate &"sO-SE-&t//_julia
associates &"sO-SE-&ts//_julia
dos "d>s//use_dos$2
infiniti in"fin-&-tE//the_car
paris p&"rE//for_francophiles
refuse rE"fyuz/1 "refyuz/2
refusing rE"fyUziN//_julia
resume r&"zUm/v "rezu.mA/n//_use_$v...
scsi "sk^zE//example_of_acronyms_in_cache
whet "hwet//_julia
within .wiT"in//_dyad_problem_in_/D-i/

Column one in (52) represents the orthographic entries in cache; this column must be alphabetized and entered in lower case. Column two for each entry contains a chcbet pronunciation of the item, as discussed above in Section 3.1 and in Appendix C. An optional comment may be added to the pronunciation field flagged by '//'; this comment must contain no spaces. An entry may have multiple pronunciation fields, separated by a single space, as in the entry for *resume*. Single or multiple-pronunciation entries may have a part-of-speech specified from the set $\{v,n,j,C\}$ (verb, noun, adjective, or proper noun) following a '/', as in the *resume* entry. They may also include integer tags following the '/', to indicate alternative pronunciations, which then can be referenced in text by the user, as described above and illustrated in the entry for *refuse*. Note that cache entries must represent exact matches in all but case to items in the text. For example, to change the pronunciation of a word and its plural form, such as *associate* and *associates* two entries in cache are required. The cache is intended either as a temporary extension of the dictionary (words in it presumably would eventually be entered in TTS' dictionary) or as a means of customizing the lexicon for particular users or applications. However, at the moment there is no provision in TTS for allowing individual users to maintain their own caches.

4 Improving Intonation by Script for Limited Domain Applications

Many applications can be made to sound more natural by a few annotations to the 'canned' portion of text to be synthesized, or by writing simple shell scripts to modify TTS defaults; see Appendix E for a simple example. Particularly when the domain itself is restricted, preprocessors can be written to take advantage of knowledge about the general form of the information to be communicated to produce remarkably natural sounding intonation (Syrdal, 1988; May, Pavlovcik, and Winter, 1989); such preprocessors have been shown to actually improve transcription accuracy in such trials (Silverman et al., 1993). Applications such as reverse telephone directories, in which names and addresses for input telephone numbers are looked up in a database and synthesized over the phone, provide a good example of limited domain applications in which enough information is available in advance about the text to be synthesized to permit the writing of simple filters to improve naturalness. The knowledge that TTS is synthesizing a name and address, for example, allows us to write simple scripts which take the various fields of this type of text as input and insert escape sequences appropriately to mimic a desired speaking style for this domain; an example of the output of such a preprocessor is show in Example (53):

(53) **\!{T145** The address is **\!/** . **\!{T130 \!si90** Julia B. Hirschberg **\!{F.95 \!/** . **\!{T115 \!si60** 2D-450 **\!{F.95 \!/** . **\!{T115 \!si30** AT&T Bell Laboratories **\!{F.95 \!/** . **\!{T115 \!si30** 600 Mountain Avenue **\!{F.95 \!/** . **\!{T100 \!si15** Murray Hill NJ 07974 **\!{F.85** .

TTS' text-preprocessor, FREND, also provides a set of modes, which can be turned off and on by other escape sequences to tell the preprocessor how to handle ambiguous input like digit sequences and abbreviations in subsequent input. For example, FREND escape sequences can tell TTS to favor 'telephone-number' over other interpretations of ambiguous digit strings, and to pronounce these strings accordingly when they are recognized in input text. FREND modes are fully documented in (Pavlovcik, 1991) and are listed and described briefly in Appendix D for easy reference.

Acknowledgments

Thanks to Joe Olive, Jim Rowley, Beki Sciacca, Richard Sproat and Diana Winter for helpful comments on an earlier draft of this TM.

References

Anderson, Mark D., Janet B. Pierrehumbert, and Mark Y. Liberman. 1984. Synthesis by rule of English intonation patterns. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 2.8.1–2.8.4, San Diego. ICASSP84.

Beckman, Mary and Janet Pierrehumbert. 1986. Intonational structure in Japanese and English. *Phonology Yearbook*, 3:15–70.

Bolinger, Dwight. 1989. *Intonation and Its Uses: Melody in Grammar and Discourse.* Edward Arnold, London.

Buchsbaum, Adam. 1986. Namsa: A system for pronouncing proper names. Technical Memorandum TM-11227-860822-12, AT&T Bell Laboratories.

Church, K. W. and A. R. Golding. 1987. Pronouncing names by analogy. Technical Memorandum TM-11225-870821-11, AT&T Bell Laboratories.

Church, Kenneth. 1985. Stress assignment in letter to sound rules for speech synthesis. In *Proceedings of the 23rd Annual Meeting*, pages 246–253, Chicago. Association for Computational Linguistics.

Coker, C., K. W. Church, and M. Y. Liberman. 1990. Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis. Technical Memorandum TM-11227-901207-16, AT&T Bell Laboratories.

Davis, J. and J. Hirschberg. 1988. Assigning intonational features in synthesized spoken directions. Technical Memorandum TM-11225-880209-02, AT&T Bell Laboratories.

Grosz, B. and J. Hirschberg. 1992. Some intonational characteristics of discourse structure. Technical Memorandum TM-11222-92-1113-12, AT&T Bell Laboratories.

Hirschberg, J. 1990. Accent and discourse context: Assigning pitch accent in synthetic speech. Technical Memorandum TM-11222-900426-07, AT&T Bell Laboratories.

Hirschberg, J. 1991. Using text analysis to predict intonational boundaries. Technical Memorandum TM-11222-910606-07, AT&T Bell Laboratories.

Hirschberg, J. 1992a. Automatically trained phrasing rules for *newtts*. Technical Memorandum TM-11222-921117-18, AT%T Bell Laboratories.

Hirschberg, J. 1992b. Pitch accent in context: Predicting intonational prominence from text. Technical Memorandum TM-11222-921113-15, AT&T Bell Laboratories.

Hirschberg, J. and D. Litman. 1991. Empirical studies on the disambiguation of cue phrases. Technical Memorandum 11222-911022-23TM, AT%T Bell Laboratories, Murray Hill NJ.

Hirschberg, J. and J. Pierrehumbert. 1986. The intonational structuring of discourse. In *Proceedings of the 24th Annual Meeting*, pages 136–144, New York. Association for Computational Linguistics.

Hirschberg, J., S. Riederer, J. Rowley, and A. Syrdal. 1990. Voice response systems: Technologies and applications. *AT&T Technical Journal: Speech Technologies*, 65(5):42–51, September/October.

Lakoff, George. 1971. Presupposition and relative well-formedness. In *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*. Cambridge University Press, Cambridge UK, pages 329–340.

Liberman, Mark Y. and Adam L. 1985. Structure and usage of current Bell Labs text to speech programs. Technical Memorandum TM-11225-85-0731-11, AT&T Bell Laboratories.

May, Delano B., Peter J. Pavlovcik, and Diana S. Winter. 1989. Text-to-speech synthesis trail with rjr/nabisco. Technical Memorandum TM-55236-890125-01, AT&T Bell Laboratories.

Pavlovcik, Peter J. 1991. Frend reference manual, issue 3. Technical Memorandum TM-55430-910711-01, AT&T Bell Laboratories.

Pierrehumbert, Janet and Julia Hirschberg. 1990. The meaning of intonational contours in the interpretation of discourse. In P. Cohen, J. Morgan, and M. Pollack, editors, *Intentions in Communication*. MIT Press, Cambridge MA.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation*. Ph.D. thesis, Massachusetts Institute of Technology, September. Distributed by the Indiana University Linguistics Club.

Pierrehumbert, Janet B. and Mary E. Beckman. 1988. *Japanese Tone Structure*. MIT Press, Cambridge MA.

Prince, E.F. 1981. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*. The Academic Press, New York, pages 223–255.

Rowley, J. E. and K. Silverman. 1989. A TTS boundary escape controlling lengthening at boundaries. Technical Memorandum TM-11222-890803-05, AT&T Bell Laboratories.

Schulz, D. E. 1987. Surname exception dictionary for at&t bell labs experimental text-to-speech system. Technical Memorandum TM-55236-870318-01, AT&T Bell Laboratories.

Silverman, Kim, Ashok Kalyanswamy, Julie Silverman, Sara Basson, and Dina Yashchin. 1993. Synthesiser intelligibility in the context of a name-and-address information service. In *Proceedings of the 3rd European Conference on Speech Communication and Technology*, volume 3, pages 2169–2172, Berlin. Eurospeech93.

Sproat, R. 1992. NewTTS: A user's manual. Technical Memorandum 11222-921102-09TM, AT&T Bell Laboratories.

Sproat, R., J. Hirschberg, and D. Yarowsky. 1992a. A corpus-based synthesizer. Technical Memorandum TM-11222-921117-19, AT&T Bell Laboratories.

Sproat, R., J. Hirschberg, and D. Yarowsky. 1992b. A corpus-based synthesizer. In *Proceedings of the International Conference on Spoken Language Processing*, Banff, October. ICSLP.

Sproat, Richard. 1990. Stress assignment in complex nominals for English text-to-speech. In *Proceedings of the European Speech Communication Association Workshop on Speech Synthesis*, pages 129–132, Autrans, France, September. European Speech Communication Association.

Syrdal, A. K. 1988. Plans for improving the quality of text-to-speech synthesis of names and addresses. Internal memorandum, AT&T Bell Laboratories.

van Santen, Jan and Richard Sproat. 1992. Rate escapes for **newtts**. Technical Memorandum TM-11222-921231-32, AT&T Bell Laboratories.

A Glossary

baseline The lowest point a speaker realizes in his or her range, over all utterances; constant for each speaker. This feature can be varied in TTS by the escape sequence **!{B***y*, where *y* is a value in Hz.

boundary tone The high (**H%**) or low (**L%**) tone occurring at the right edge of an intonational phrase in Pierrehumbert's system. Final boundary tones may be specified in TTS by the escape sequence **!p***xy*, where *x* is a tone and *y* is a prominence specification relative to the f0 peak of the nuclear stress of the phrase. Initial boundary tones may also be specified in TTS, with the sequence **!i***xy*, where *x* may be high (H) or low (L) and *y* specifies the tone's prominence.

cache In TTS, a list of words and their pronunciations which TTS checks before general dictionary look-up.

complex nominal A sequence of nouns such as *city hall*, *parking lot*, or *city hall parking lot*, which pose particular difficulties for prominence assignment.

content words Words such as nouns, verbs, modifiers.

continuation rise The end of an intonational contour produced by the escape sequence **!/** in TTS and represented as **L H%** in Pierrehumbert's system; it may be used with varied pitch accents. Produces the effect of there being "more to come".

contrastive stress Particular intonational prominence serving to pick out items as unusually salient for some purpose, which may not in fact 'contrast' with anything.

cue phrases Words and phrases which may be used to explicitly indicate discourse structure, such as *now*, *well*, *first*.

declarative contour In English, a falling contour, described in Pierrehumbert's system as **H* L L%**.

discourse markers See cue phrases.

downstep The compression of the pitch range for pitch accents following any complex accent; a downstepped contour is produced in TTS by either \!_ (to downstep all non-initial pitch accents in the minor phrase) or \!@ (to downstep the next pitch accent). In OldTTS the escape sequence \! < downsteps the first accent of every non-initial function word group in the minor phrase.

downstep ratio In TTS, the relationship between the f0 of a complex pitch accent and a subsequent downstepped accent, whose range has been compressed by its predecessor. Invoked by \!{**K**y, where y is a ratio.

escape sequence A command such as\!{**T130**, which can be specified within input text to control various prosodic parameters in TTS.

final lowering Compression of the pitch range during the last half second or so of speech; the final lowering ratio (compression of range relative to current range) can be varied in TTS via \!{**F**y, where y is a ratio.

focal accent See FOCUS.

focus The information that a speaker makes most salient in an utterance, either intonationally (INTONATIONAL FOCUS or FOCAL ACCENT) or by other means.

focus-sensitive operator Words such as *only* whose interpretation interacts with the intonational focus of the phrase.

floating phrase accent A very minor boundary inserted between function word groups in OldTTS automatically; may be produced in both OldTTS and NewTTS by \!**f**xy, where x is a tone indicator (H or L) and y indicates prominence ($y \geq 0$).

function word Words such as prepositions, articles, determiners.

function-word grouping In OldTTS, a level of phrasing which automatically creates very minor boundaries (see floating phrase accent) between adjacent function words (or certain other specified words). In OldTTS, this feature can be turned off for the following minor phrase by \!>.

fundamental frequency (f0) The physical correlate of pitch.

given information Information that is treated as 'old' or known to the hearer in a discourse.

homograph Words which are spelled the same but differ in meaning — and sometimes in pronunciation — such as *bass* (a fish)/*bass* (a musical instrument) and *desert* (a barren wasteland)/*desert* (to leave behind)

information status Given/new or focal status in a discourse.

initial boundary tone See BOUNDARY TONE.

intermediate phrase A smaller prosodic phrase in Pierrehumbert's system; each intermediate phrase can be defined as a sequence of one or more pitch accents plus a high (H) or low (L) phrase accent. The effect of an intermediate phrase boundary can be produced in TTS by: **!f***xy* **!}.5** **(** *****[*z*] **)**, where *x* may be H or L, *y* represents a prominence specification (usually $0 \leq y \leq 1$), and *z* represents silence in centiseconds — 3-5 centiseconds is usually about right.

intonational phrase The largest prosodic phrasal unit in Pierrehumbert's system; intonational phrases are in turn made up of one or more intermediate phrases plus a final boundary tone.

major phrase An entire utterance (sentence) in TTS.

minor phrase The TTS equivalent of Pierrehumbert's INTONATIONAL PHRASE; a prosodic phrase smaller than an utterance.

new information Information that is treated as 'new' or novel to the hearer in a discourse.

nuclear accent Also called NUCLEAR STRESS. The most prominent pitch accent in an intermediate phrase, although not necessarily the f0 peak. It is generally the last accent in the intermediate phrase.

nuclear stress See NUCLEAR ACCENT.

phrasal verbs Verb-particle or verb-verbal preposition sequences such as *throw away* and *turn off*, which are general accented differently from similar-appearing verb-preposition sequences such as *throw to the first baseman* or *turn on a dime*.

phrase accent The high (H) or low (L) tone in Pierrehumbert's system) controlling the pitch in an intermediate phrase from the last pitch accent in the phrase to the end of the phrase.

phrase-final lengthening Lengthening of the syllable preceding an intermediate or intonational phrase boundary; an important cue to boundary perception. Can be varied in TTS via the escape sequence **!**}*y*, where *y* represents a fraction of the default lengthening value for the syllable preceding a minor phrase boundary.

pitch accent Intonational prominence associated with the stressable syllable of words in English. In Pierrehumbert's system there are six types of pitch accent, two simple accents (**H***, **L***) and four complex accents (**H*+L**, **H+L***, **L*+H**, or **L+H***). The '*' indicates which tone is associated with the stressable syllable. Intermediate phrases typically have at least one pitch accent.

pitch range The difference between a speaker's baseline and the highest point the speaker reaches in an intermediate phrase. When speakers use an expanded pitch range they produce the effect of 'speaking up'.

plateau contour Represented as **H* H L%** in Pierrehumbert's system; produces the effect of an 'open-ended list'.

reference line In TTS, a line from which pitch accent heights are calculated; can be set explicitly with **!**{**R***y*, where *y* is an absolute value in Hz.

reference resolution The process by which the referents of pronouns and other referring expressions are identified.

referent Person, thing, or concept being referred to in a discourse, as, by a pronoun.

rise-fall-rise contour Represented in Pierrehumbert's system as **L*+H L H%**; indicates uncertainty in a smaller pitch range (and at a slower rate) or incredulity, when produced in an expanded range (and at a faster rate).

surprise-redundancy contour Described as **H* L* L L%** in Pierrehumbert's system.

topline In TTS, the highest point reached in a minor phrase; can be varied using **!**{**T**y, where y represents the pitch maximum desired in Hz (any positive integer).

wh-**quesion** A question asking *who*, *what*, *when*, *where*, *why* or *how*. In English, typically uttered with a declarative (falling) contour.

yes-no-question contour In English, a rising contour, specified in Pierrehumbert's notation as **L* H H%**; this is automatically invoked in TTS for input sentences ending in '?' which are **not** determined to be *wh*-questions. It may be explicitly evoked by the sequence **!?**.

B TTS Escape Sequences

Express understands the following escape sequences within input text. Note that each entire sequence must be surrounded by some white space (or the beginning or end of line) and not directly adjacent to punctuation or other text (e.g. **avoid** '...the end **!{F.96**.') Note also that a value of, e.g., 60% is denoted 0.60:

!xy: For x = r or R. In OldTTS, y is a rate value, from 0 (fast) to 4 (slow). The default is 2. In NewTTS, the rate value is a multiplier on the default value of 1; smaller values increase the rate and larger values decrease it. Each rate setting is effective until the next one, or for the duration of the major phrase if x is r, or through the end of the current input file if x is R.

!{xy: y is a value for some F0 phrasal parameter denoted by x. Each of these escape sequences is effective only for the current phrase. However, if the escape is given as **!{!**xy, it will change the default parameter values for the current invocation of the program.

x	Description	Default Value		
		NOV92 Male	NOV92 Female	MAY91 Male
T	minor phrase top line	115 Hz	195 Hz	115 Hz
R	minor phrase reference line	96 Hz	165 Hz	95 Hz
X	alternating prominences ...	60%	60%	60%
Y	... among function groups	–	–	80%
P	alternating prominences ...	50%	50%	50%
Q	... within a function group	–	–	70%
B	major phrase base line	75 Hz	140 Hz	75 Hz
K	major phrase downstep ratio	60%	60%	70%
F	major phrase final lowering ratio	90%	90%	85%
C	major phrase rate scalar	75%	75%	75%
D	ratio for H* legs	40%	40%	40%
E	major phrase final raising ratio	120%	120%	120%
G	end effect time	50%	50%	50%
U	time fraction for H-	30%	30%	30%

Table 1: Phrasal Parameters in TTS

!zxy: For each minor phrase, a phrase accent (z = p), initial boundary tone (z = i), and final boundary tone (z = b) can be set. A floating phrase accent at the point of the escape (z = f) can also be set. x is H (high) or L (low), and y, which is optional, is a floating point number denoting prominence, such that, generally, $0 \leq y \leq 1$.

!sxy: For each minor phrase, an initial (x = i) and final (x = f) silence can be added, with its duration given in centiseconds by y.

!Mmy: Used in OldTTS to mark a minor phrase with a non-negative integer y. Can occur anywhere within a phrase. Synthesis programs can detect these markers to report when portions of the text are reached during audio conversion.

!-: Do **not** accent the following word.

\!*xy Accent the following word, with optional tone x (one of **L***, **H***, **L+H***, **L*+H**, or **H+L*** and optional prominence y. (The effect of the **H*+L** accent is produced by \!_ — over all but the last accented word in a minor phrase — or \!@ — for the previous accented word.)

\!!xy: Like \!*xy, but also removes any default accents from subsequent words in the minor phrase.

\!c: Cliticize the following word: Remove any default accent it bears, remove any duration influencing stresses, and don't put in a word boundary after it.

\!#y: Set the emphasis of the following word to y.

\!|y: Set the emphasis of the boundary following the preceding word to y.

\!}y: Insert a variable-duration boundary effecting only duration. \!}.1 gives little segmental lengthening, and \!}1. gives the same durational lengthening as a minor phrase boundary. \!} is equivalent to \!}1.

\!_: Downstep every non-initial pitch accent in the minor phrase.

\!<: Downstep the first accent of every non-initial function word group in the minor phrase.

\!@: Downstep the next word.

\!fxy: Insert a function group boundary at the point of the escape with tone x (L or H) and prominence y.

\!>: Turn off function group phrasing in the current minor phrase.

\!∧: Invert the default free-floating phrase accent pattern.

\!?: Put a yes/no question intonation pattern, essentially the opposite of the declarative pattern, on the current minor phrase.

\!/: Cause a continuation rise at the end of the current minor phrase.

\!(: Treat the phrase as a parenthetical.

\!%y: Scale the pitch from this point until the next scaling by y (as a floating point number) or until the end of the current major phrase. Scaling is absolute: $y = 1.0$ will restore normal pitch.

\!%%y: Like \!%N, only sets the default pitch scaling throughout the current input file.

\u$xz$: Set speaker gender to male ($z =$ m) or female ($z =$ f). Set x to 's' to change voice for the current minor phrase, or 'S' to change voice for the remainder of the invocation of the program.

\!uxz: Control utter. x is a flag telling what to change, and z is an argument. x may be f, to scale the front of the vocal tract by $1.0/y$, or b, to scale the back of the vocal tract by $1.0/y$. Each escape is effective for the current minor phrase only. Specifying the x flag as a capital letter (F or B) changes the default for the rest of the input file. In OldTTS, x and y may also take on the following values: $x =$ e, to specify an excitation method of the current minor phrase, ($y \in \{$0 (fixed pulse) – default, 1 (single pulse), or 2 (multi pulse)$\}$. $x \in \{$s – use speaker k (Keith – default), m (Mark)$\}$. Again, specifying x as a capital letter causes the specification to hold for the remainder of the invocation of the program.

C CHCBET and JPOBET

Two types of phonemic transcription may be used in OldTTS, JPOBET and CHCBET. Only CHCBET can be used in NewTTS, and its use is somewhat different than in OldTTS.

In OldTTS, jpobet-style phonemic transcriptions can be inserted between **\\(** and **\\)**. So, to say "this is a test", with an unusually long '/e/', one can type:

```
This is a \( t e[50]\& s t \) .
```

In jpobet transcriptions, a duration can be assigned to a phoneme by placing its desired value in centiseconds between brackets ([]) following the phoneme. If the phoneme is aspirated, then the value represents closure*100+aspiration. For example, a 30-centisecond-long '/e/' can be produced by the input: '**\\(e[30]& \\)** . '. To give the initial '/t/' in *test* 4 centiseconds of closure and 2 of aspiration, one would say: '**\\(t[402] e& s t \\)** .'.

In OldTTS, chcbet phonemic transcriptions can be inserted between **\\(C** and **\\)** in OldTTS. To say "this is a test", one types:

```
This is a \(C "test \) .
```

One cannot assign durations to phonemes in chcbet transcriptions in OldTTS. In NewTTS, however, chcbet is the only phonemic transcription scheme recognized; in NewTTS, chcbet behaves likes jpobet in OldTTS, in that phonemes are inserted between parentheses, without the '**C**' flag, and phonemes takes optional durations in centiseconds, enclosed between '[]'. So, the following produces "this is a test" in NewTTS with a long '/e/':

```
This is a \( ''te[50]st \) .
```

Remember that some space must separate the '.' from the **\\)**. A full list of phonemes and diacritical marks follows:

CHCBET	JPOBET	Example
A	ei	b**ai**t
a	ae	b**a**t
E	ii	b**ea**t
e	e	b**e**t
I	ai	b**uy**
i	i	b**i**t,roses
O	o	b**oa**t
o	a	B**o**b
>	c	b**ough**t
W	au	d**ow**n
Y	ci	b**oy**
∧	uh	b**u**t
U	uu	b**oo**t
u	u	b**oo**k
&	uh	**a**bout
R	er	b**ir**d
b	b	**b**et
C	ch	**ch**ur**ch**
d	d	**d**og
f	f	**f**og
g	g	**g**ot
h	h	**h**ot
J	j	**j**ump
k	k	**k**it,**c**ot
l	l	**l**ook
m	m	**m**ug
n	n	**n**od
N	ng	si**ng**
p	p	**p**al
r	r	**r**at
s	s	**s**it
S	sh	**sh**ut
t	t	**t**ell
D	dh	**th**at
T	th	**th**ick
v	v	**v**at
w	w	**w**on
y	y	**y**ou
z	z	**z**oo
Z	zh	a**z**ure
	*	silence,glottal stop

D Frend Modes

Nine text classes are detected by FREND: addresses, dates, fractions, money, telephone numbers, time, ordinal numbers, measurements and names. Detetion for each class can be set in one of the three Text Class Detection Modes: 'off', 'conservative' and 'risky'. Initially, all detectors are set to conservative mode. This setting can be changed using escape sequences with the general format **!ncm**, where '**c**' is one of the following text classes:

c	Class
a	address class
d	date class
f	fraction class
i	measurements class
m	money class
n	proper names class
o	ordinal numbers class
p	telephone numbers class
t	time class

Table 2: Text Class Specification in **!ncm**

m	Mode
o	turn the class detector off
c	set the class detector to conservative mode (default)
r	set the class detector to risky mode

Table 3: Mode Specification in **!ncm**

Some examples of the use of this escape sequence are given in Table 4:

Escape Sequence	Function
!nar	switch address detector to risky mode
!ndc	switch date detector to conservative mode
!npo	switch telephone numbers detector off

Table 4: Sample Uses of Frend Escapes

Some other escape sequences related to the text class detection modes are presented in Table 5. Special translation modes are also controlled via frend escape sequences, as identified in Table 6:

Escape Sequence	Function
\!co	show current setting of text class detectors
\!de	set all detectors to default (conservative mode)

Table 5: Miscellaneous Frend Escapes

Escape Sequence	Function
\!sb	spell mode begin
\!se	spell mode end
\!mb	math mode begin
\!me	math mode end
\!rb	raw mode begin
\!re	raw mode end
\!pb	proofread mode begin
\!pe	proofread mode end
\!ab	begin acronym spell mode (default)
\!ae	end acronym spell mode
\!eb	begin abbreviation expansion mode (default)
\!ee	end abbreviation expansion mode
\!hnv	set numerical hyphen verbose mode (default)
\!hnt	set numerical hyphen terse mode
\!hav	set alphabetic hyphen verbose mode
\!hat	set alphabetic hyphen terse mode (default)

Table 6: Special Translation Modes

E Sample Scripts which Vary TTS Parameters

The script shown in Figure 15 changes the default f0 contour in TTS to one in which the final word in each sentence is downstepped, if accented, and pitch range, pausal duration, and final lowering are varied according to the paragraphing of the input text. Initial phrases are produced in a larger pitch range, preceded

```
#!/usr/local.1122/bin/ksh
#change default f0 contour and vary topline, pause between phrases,
#and final lowering according to paragraph structure
case "$#" in 0) echo "Usage: $0 <topline> <refline> <maxsil> <infiles>";exit 1;;
esac
ran=$1
ref=$2
sil=$3
shift 3
#preserve blanklines indicating new para
nawk '
/^#/ {next}
NF<1 {print "PARA.";next}
{print}
' $* |
apsent1 |
nawk -f lively.awk range="$ran" refline="$ref" silence="$sil"
```

Figure 15: Script to Change Default f0 and Range, Pauses, and Final Lowering

by longer pauses, with less final lowering. The shell script initially runs the input through an end-of-sentence detector, which produces one sentence of input text per line (*apsent1*). The awk script *lively.awk*, shown in Figure 16 is then called on the output.

```
BEGIN{
  ran=range;rangedec=15
    sil=silence;sildec=30
      finlow=.96;fl=finlow;fldec=.06;lastfl=".78"
ds="\\!{R" refline " \\!{K.6 \\!@ "
  nsent=sprintf("\t")
}

/PARA/ {
  ran=range
    sil=silence
      fl=finlow
nsent=sprintf("%s\t",nsent)
  if(NF==1)next
    else gsub(/PARA./,"")
}
/^%/{print;next}
NF>0 {
  printf "%s", nsent
    $NF=ds $NF
      nsent=sprintf("\\!{!T%d \\!si%d \\!{!F%f %s\n",ran,sil,fl,$0)
if(ran>120)ran=ran-rangedec
  if(sil>30)sil=sil-sildec
    if(fl>.78)fl=fl-fldec
}
END{
  gsub(/F[0-9.][0-9.]*/,"F" lastfl,nsent)
    print nsent
}
```

Figure 16: Script ‘lively.awk’

A sample input to this text — a randomly chosen paragraph from the AP newswire) is show in Example (54).

(54) Jo Hackman never expected to be sitting down for Thanksgiving dinner in her home. The house in Hartsburg, Mo., was a gutted shell after the Missouri River jumped its banks this summer.

Output from this input text produced by the script shown in Figure 15 is shown in Example (55).

(55)
```
\!{!T150 \!si120 \!{!F0.960000 Jo Hackman never expected to be sitting down for Thanksgiving dinner in her \!{R85 \!{K.6 \!@ home.
\!{!T135 \!si90 \!{!F.78 The house in Hartsburg, Mo., was a gutted
shell after the Missouri River jumped its banks this \!{R85 \!{K.6 \!@
summer.
```