Pitch Accent in Context: Predicting Intonational Prominence from Text

Julia Hirschberg
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974
julia@research.att.com

June 21, 1995

Abstract

Explaining speakers' choice of which items to emphasize or de-emphasize intonationally has been an important topic in theoretical linguistics, as well as in applications such as speech synthesis, where accent decisions affect the naturalness as well as interpretation. Heretofore, most researchers have assumed that detailed syntactic, semantic, and discourse-level information must be available in order for accent assignment to be predicted successfully. However, a series of recent experiments on corpora of recorded (read) speech and spontaneous (elicited) speech suggest that it is indeed possible to model human accent strategies with fair success (80-98% correct) for unrestricted text — with only the tools for automatic text analysis currently available. The algorithm developed from these experiments is currently used to assign pitch accent in the Bell Laboratories Text-to-Speech System. (This appears is *Artificial Intelligence* 63(1-2), 1993.)

1 Introduction

How speakers decide which items to emphasize intonationally and which to de-emphasize has been a popular but vexing question in studies of intonation and discourse. While most researchers today accept Bolinger's pessimistic view that "accent is predictable (if you're a mind-reader)" (1972), mind-reading attempts continue, fueled in particular by the need to assign acceptable accentuation in speech synthesis. Most current text-to-speech systems generally use simple word-class information for input text to determine which items to accent and which to deaccent. In such systems, FUNCTION, or CLOSED CLASS words, such as prepositions and articles, are deaccented while CONTENT, or OPEN CLASS words, such as nouns and verbs, are accented. It is widely believed that more detailed syntactic, semantic, and discourse-level information is needed to model human performance accurately, particularly for the synthesis of longer texts, but the problem of obtaining such information for unrestricted text appears unlikely to be solved in the foreseeable future.

Yet considerably more detailed syntactic information than the simple function/content word distinction is already available for unrestricted text through improvements in part-of-speech tagging. And simple sorts of 'higher level' information can also be inferred about a text, based upon results of a number of psycholinguistic studies of information status, taken in conjunction with proposals advanced by computational linguists for modeling discourse structure. In short, fairly minimal sorts of text analysis can provide much richer sources of information to guide prosodic prediction than are currently being exploited.

This paper describes the development of procedures for the assignment of pitch accent automatically in synthetic speech from a minimal analysis of unrestricted text. These procedures are based on results of a series of experiments in the modeling of pitch accent assignment from analysis of prosodic features of recorded (read) speech in several speech corpora. In these experiments, both hand-derived models and automatically-derived models were generated, the latter derived using Classification and Regression Tree analysis (CART) (Brieman et al., 1984) techniques. The experiments focused upon discovering: What sorts of information which can be derived currently from text analysis are useful in the prediction of pitch accent assignment? What sort of discourse representation permits the most accurate modeling of human accent strategies? How does prediction from automatically obtainable information means compare with prediction from the fuller and presumably more accurate information obtained by hand?

The hand-crafted rules were developed on a corpus of radio speech and modeled human accent decisions on this data with 82.4% accuracy; this performance compares quite favorably with previous results reported by Altenberg (1987) of 57% (92% correct with 62% coverage) on partly scripted monologue. However, when tested on a corpus of citation-format utterances, the hand-written rules described below performed at 98.3% accuracy, a quite impressive success rate, albeit on a simpler task. Procedures developed automatically via CART techniques produced decision trees which predicted accent in the radio speech with a (cross-validated) success rate of only 76.5%, but whose performance improved to 85.1% when trained on larger corpora. The procedures for assigning pitch accent developed as a result of these experiments have been implemented in the Bell Laboratories Text-to-Speech System (TTS) (Olive and Liberman, 1985; Hirschberg, 1990a; Hirschberg, 1990b) which varies intonational features based upon simple text analysis in order to generate more natural-sounding synthetic speech.

2 Identifying Pitch Accent in Natural Speech

In natural speech, some words appear more intonationally prominent than others. Such words are said to be stressed, or to bear PITCH ACCENTS. While, in English, each word has a characteristic (lexical) stress pattern, not every word is accented. *Telephone*, for example, has three syllables and a '1-0-2' stress pattern; that is, its first syllable exhibits primary stress, its second no stress, and its third, secondary stress. When uttered, *telephone* may be accented or not. If accented, the pitch accent is associated with *telephone*'s (primary) stressed syllable, *tel*. Thus, LEXICAL STRESS may be distinguished from pitch accent. For example, note that the nominal and verbal homonyms of *object* are distinguished in speech by differences in lexical stress, but either may be uttered with or without a pitch accent with this difference preserved. Consider the sentences in (1), where intonational prominence is indicated orthographically by capitalization:

(1) a. the ALIEN OBject is MOVing.

b. We must STOP the object before it deSTROYS the WORLD.

c. I obJECT to your PLAN.

d. THEY object to it TOO.

In (1a), *object* is a noun and thus is uttered with primary stress on the first syllable (Figure 1). In (1b) *object* is also a noun and thus has the same (lexical) stress pattern. However, while in (1a) the noun is uttered with a pitch accent, in (1b) it is uttered with less prominence or, DEACCENTED; the stressable syllable of this noun is **not** made intonationally prominent (Figure 2). In both (1c) and (1d), *object* is used as a verb and thus has primary stress on its second syllable. However, in (1c) the verb is accented (Figure 3), while in (1d) it is deaccented (Figure 4). Note particularly that, whether *object* is a noun or a verb, it may be accented or deaccented. Its stress pattern is independent of a speaker's decision to accent it or not.

Figures 1-4 go about here.

Although pitch accent is a perceptual phenomenon, words that hearers identify as accented tend to differ from their deaccented versions with respect to some combination of pitch, duration, amplitude, and spectral characteristics. Accented words are usually identifiable in the FUNDAMENTAL FREQUENCY CONTOUR (f0) as local maxima or minima, aligned with the word's stressed syllable; deaccented items do **not** exhibit such pitch excursions. Words perceived as accented also tend to be somewhat longer and louder than their deaccented counterparts. The vowel in the stressed syllable of a deaccented word is often reduced from the full vowel of the accented version. Words that are deaccented may or may not be CLITICIZED; that is, they may lack adjacent word boundaries as well as exhibit vowel reduction (e.g., '*to* in '*gotta*' vs. '*got to*').

How humans decide which words to accent and which to deaccent — what constrains accent placement and what function accent serves in conveying meaning — is an open research question in linguistic and speech science. While syntactic structure was once believed to determine accent placement (Bresnan, 1971; Quirk et al., 1964; Crystal, 1975; Cruttenden, 1986), it is now generally believed that syntactic, semantic, and discourse/pragmatic factors are all involved in accent decisions (Bolinger, 1989). Currently, theoretical and empirical investigations of accent include word class, syntactic constituency, and surface position, as well as less easily defined phenomena falling into the broad category of INFORMATION STATUS, including CONTRASTIVENESS (Bolinger, 1961), FOCUS (Jackendoff, 1972; Rooth, 1985), and the GIVEN/NEW distinction (Chafe, 1976; Halliday and Hassan, 1976; Clark and Clark, 1977; Prince, 1981b).

However, this new consensus has been slow to influence accent assignment in text-to-speech systems. Identifying a larger set of potential predictors is only a step toward understanding how these predictors interact to produce individual accent decisions. And obtaining the syntactic, semantic and pragmatic analyses required under most accounts for appropriate accent assignment a) automatically, b) in real time, and c) from analysis of unrestricted text, presents major obstacles to current natural language technologies. More successful accent assignment is possible in 'message-to-speech' systems, which synthesize speech from an abstract representation of the message to be conveyed (Young and Fallside, 1979; Danlos et al., 1986; Davis and Hirschberg, 1988). But most experimental text-to-speech systems, and all commercial text-to-speech systems, rely upon the simple function/content word distinction mentioned above, in which function words are deaccented and content words are accented. Since this approach tends to assign prominence to too many words, such systems commonly provide some limited capability for users to override incorrect system defaults in particular cases by annotating the input text. However, such mechanisms do not address the fundamental problem of assigning accent automatically via improved text analysis. This is the problem we address below.

3 Speech Corpora Used for Testing and Training

Several speech corpora of American English were examined in the course of developing the accent prediction procedures discussed below, including citation-form sentences spoken by each of three speakers, news stories read by a single speaker, multi-speaker broadcast radio speech, and multi-speaker elicited spontaneous speech. Such variety permitted comparison of prediction performance for single vs. multi-speaker speech, sentence vs. discourse length speech, and planned vs. spontaneous speech.

3.1 Citation-Form Sentences

The citation-form sentences examined for this study represent a portion of the utterances elicited from each of three speakers (one male and two female) for the acoustic inventory used in concatenative synthesis by the Bell Labs synthesizer, TTS. The 1380 utterances selected for analysis below consist of fairly short sentences and phrases, which are nevertheless all over four words long, with an average length of 5.5 words per sentence. Samples appear in (2).

(2) a. Plow a young man.

b. The large cow accidentally brushed against the fence.

While some of these sentences may be difficult to interpret without constructing a rather specialized context, none of them appeared to present the speakers with prosodic challenges. The productions are regular and monotonous, without hesitations or disfluencies. The total duration for this subset is approximately 45 minutes for each of the three speakers.

3.2 The Audix Speech Corpus

The single speaker corpus of read news stories, is the Audix Speech Corpus, consisting of ten news stories recorded by a female professional newscaster under laboratory conditions. The stories were selected from the AP newswire and produced by the speaker (after familiarizing herself with the text) largely as written, with minor omissions and substitutions. While this database was collected for potential use in concatenative speech synthesis, the productions are quite natural for professional radio speech. A sample from one of the stories appears in (3). (Words that were accented by the speaker appear in upper case.)

(3) SUN MICROSYSTEMS INC., the UPSTART COMPANY that HELPED LAUNCH the DESKTOP COMPUTER industry TREND TOWARD HIGH powered WORKSTATIONS, was UNVEILING an ENTIRE OVERHAUL of its PRODUCT LINE TODAY.

SOME of the new MACHINES, PRICED from FIVE THOUSAND NINE hundred NINETY five DOLLARS to seventy THREE thousand nine HUNDRED dollars, BOAST SOPHISTICATED new graphics and DIGITAL SOUND TECHNOLOGIES, HIGHER SPEEDS AND a CIRCUIT board that allows FULL motion VIDEO on a COMPUTER SCREEN.

SCOTT MCNEALY, SUN'S PRESIDENT, SAID in a prepared STATEMENT, the REVAMPED LINE is ANOTHER STEP toward bringing MORE POWERFUL DESKTOP COMPUTERS into everyday BUSINESS and PERSONAL USE.

This corpus currently has a total duration of approximately half an hour, with each story averaging about five minutes in length. Sentences (from the materials read) averaged 23.4 words in length. There are virtually no hesitations or disfluencies in the speech analyzed, since disfluent material was re-recorded.

3.3 The FM Radio Newscasting Database

The multi-speaker broadcast radio speech used in this study is a portion of the FM Radio Newscasting Database, a series of studio recordings of newscasts and other material provided by National Public Radio Station WBUR in association with Boston University, which is currently being collected by SRI International (Patti Price), Boston University (Mari Ostendorf), and MIT (Stefanie Shattuck-Hufnagel).[1] Two news stories were used in this sample, representing approximately nine minutes of speech, produced by a total of nine speakers. An orthographic transcription of a sample paragraph is presented in (4), with accented words appearing in upper case.

(4) WANTED. CHIEF JUSTICE of the MASSACHUSETTS SUPREME COURT. IN APRIL the sjc's CURRENT leader EDWARD HENNESSY REACHES the MANDATORY RETIREMENT age of SEVENTY and a SUCCESSOR is EXPECTED to be NAMED in MARCH. It MAY BE the MOST IMPORTANT APPOINTMENT Governor MICHAEL DUKAKIS MAKES during the REMAINDER of his ADMINISTRATION AND ONE of the TOUGHEST....

While, again, these stories appear typical of fluent 'radio speech', the inclusion of 'sound bites' from interviewed speakers makes them somewhat less useful as exemplars of coherent read text. There are some minor disfluencies in the recordings.

3.4 The ATIS Corpus

The multi-speaker spontaneous speech database investigated here included 298 sentences (representing 24 minutes of speech by 26 speakers, with an average length in words of 11.5) from the DARPA Air Travel Information Service database. These utterances were taken from the early Texas Instruments portion of the ATIS corpus, and consisted of over 770 sentences elicited in a wizard-of-Oz simulation by Texas Instruments to provide a common database for the DARPA Spoken Language Systems task (Hemphill et al., 1990). To simulate interactions between a caller and a travel agent, subjects were given a travel scenario and asked to make travel plans accordingly, providing spoken input and receiving visual output at a terminal. They were told that their utterances were being recognized by machine; vocabulary constraints were enforced by error messages.

The quality of the results is quite varied. Speakers ranged in fluency from close to isolated-word speech to fluency. Many utterances contain hesitations and other disfluencies, as well as long pauses (greater than three seconds in some cases). The simulation introduced certain obstacles to the goal of gathering one side of a 'natural' discourse, notably, the long turn-around time between utterance and response and the need for subjects to hunt through large tables retrieved from a relationship database to discover the answer to

[1] The corpus will eventually comprise several hours of prosodically labeled speech; the labeling used for the current analysis was, however, done at Bell Laboratories. A sample of the corpus is being distributed through the Association for Computational Linguistics Data Collection Initiative (Mark Liberman, Department of Linguistics, University of Pennsylvania, Don Walker, Bell Communications Research).

their queries. A sample from one of the recording session logs in Figure 5 illustrates the difficulty (System translations have been removed from this log and tables abbreviated.)

Figure 5 goes here.

First, note that the pace of the interactions was rather slow for a simulation of human-human interaction. In Figure 5, system response time can consume over a minute. In the 864 interactions in the entire ATIS sample examined here, mean elapsed time between the system's transcription of an utterance (not its production) and the generation of a response was 21.8 seconds; median response time was thirteen seconds. Since the nature of the task required subjects to absorb and use the information they retrieved, considerable time might also elapse between system responses and a subject's next utterance. Mean and median elapsed times for the entire corpus were 80.4 seconds and 57 seconds. So, interactions did not approximate the desired goal of caller and travel agent. This feature of the interactions appears likely to have affected speaker prosody in a way significant for this study, since it appears that items which might well have been deaccented in fluent speech due to prior context were less likely to be deaccented in this corpus. Of course, the fact that subjects believed that they were interacting with a machine may also have influenced their intonation.

The corpus of citation sentences was used for testing only, for reasons explained below in Section 4.3; portions of the other three were used for training and for testing. For these three, prosodic labels were assigned by hand for all of the speech, using speech analysis software (WAVES) developed by David Talkin (1989). Labels were assigned by ear and from pitch tracks and spectrograms.

3.5 Prosodic Analysis

The prosodic analysis of this data, which serves as the basis for the discussion below, employs Pierrehumbert's (Pierrehumbert, 1980; Beckman and Pierrehumbert, 1986) method of description of English intonation. In this system, intonational contours are described as sequences of low (**L**) and high (**H**) tones in the f0 contour. A phrase's TUNE is represented as a sequence of one or more pitch accent(s), a PHRASE ACCENT, and a BOUNDARY TONE. For Pierrehumbert, there are six types of pitch accent in English, two simple tones – high and low – and four complex ones. The high tone, the most frequently used accent, comes out as a peak on the accented syllable and is represented as '**H***'; the '**H**' indicates a high tone, and the '*' that the tone is aligned with a stressed syllable. Low (**L***) accents occur much lower in the pitch range than **H*** and are phonetically realized as local f0 minima. Complex accents have two tones, one of which is aligned with the stress. Using the diacritic '*' to indicate this alignment, these accents can be represented as **L*+H**, **L+H***, **H*+L**, and **H+L***. A well-formed INTERMEDIATE PHRASE consists of one or more pitch accents, plus a simple **H** or **L** tone which characterizes the phrase accent. The phrase accent spreads over the material between the last pitch accent of the current intermediate phrase and the beginning of the next — or the end of the utterance. INTONATIONAL PHRASES are composed of one of more such intermediate phrases plus a BOUNDARY TONE, which may also be **H** or **L** and is indicated by '%'. It falls exactly at the phrase boundary. Note that, in Pierrehumbert's system, no distinction is made between accents based on relative prominence or phrasal position; however, the most perceptually prominent accent in an intermediate phrase is termed its NUCLEAR STRESS.

A sample of prosodic labeling for the FM Radio database appears in Figure 6. Orthographic text is aligned with prosodic symbols below the text. In addition to the tonal symbols introduced above, for our purposes, '**-**' will be used to denote a word that is deaccented but not cliticized; '**cl**' will identify a cliticized item. While the labels indicate not only presence or absence of pitch accent, but type of accent — as well as presence and type of intonational boundaries — most of the analysis below will distinguish only three categories of pitch accent: accented, deaccented but not cliticized, and cliticized items.

Figure 6 goes here.

4 Predicting Pitch Accent

As noted above, syntactic, semantic, and pragmatic/discourse factors have been proposed as predictors of pitch accent. However, not all the types of information proposed in the literature can be obtained automatically from text the requisite detail and accuracy with current natural language processing capabilities. Complete and accurate information about syntactic structure is an obvious example of information believed essential to 'natural' pitch accent assignment which is not availablecomputationally for unrestricted text. However, the goal of the current study is to determine what sort of success can be achieved in predicting pitch accent in natural speech simply from information which *can* be obtained automatically from text with current computational techniques — and which types of information are most useful in this endeavor.

4.1 Variables for Analysis

As noted above, part-of-speech information is perhaps the most commonly employed predictor of pitch accent. Most text-to-speech systems assign accent purely based on simple function word/content word distinction, with the consequence of sometimes accenting words human speakers would deaccent, especially in the synthesis of longer texts. Also, items that are ambiguous with respect to the function-content distinction, as in preposition (commonly deaccented)/verb particle (commonly accented) ambiguities, are rarely disambiguated, since text-to-speech parsers and taggers tend to be extremely simple. Finer-grained distinctions among word classes have, however, been explored empirically, with rather limited results to date. Altenberg (1987), for example, examined the extent to which pitch accent (as a binary feature) could be predicted in a 48-minute partly read, partly extemporaneous public talk by a non-professional speaker. From hand-tagged data, Altenberg found that accent could be predicted on the basis of word class in this (training) data only 57% of the time (with 92% correct and 62% coverage of the corpus). While part-of-speech alone clearly cannot account for all accent behavior, however, it is nonetheless an important factor. For the current study, part-of-speech information was obtained via Church's (1988) tagger, which has a success rate of around 98%.

A special problem for assigning pitch accent from part-of-speech information is the problem of stressing complex nominals, items such as those in (5), which occur frequently in English text.[2] (5).

(5) a. city HALL
b. TAX office
c. CITY hall TAX office

For example, while both (5a) and (5b) are commonly analyzed in citation form as noun-noun sequences, (5a) commonly has its most prominent accent on its right-hand element, while (5b) has its on the left.[3] But when

[2] (Liberman and Sproat, Forthcoming) estimate 70,000 tokens per million words from the Brown Corpus.
[3] (Liberman and Sproat, Forthcoming) finds that 75-95% of complex nominals are stressed on the left, depending upon genre.

the two are combined to form the complex nominal in (5c), stress on *hall* is normally shifted, so that *city* is now more prominent than *hall*. So, while knowing that the items in (5) are tagged as sequences of nouns is necessary to identifying them as complex nominals, this information does not tell us what their likely accent pattern will be. Stress patterns in complex nominals appear to be predicted in part by the semantic roles filled by the nominal's components, but is also to some extent idiosyncratic. In the current study, citation form stress patterns of complex nominals are obtained from a complex nominal analyzer, NP, developed by (Sproat, 1990), which employs semantic rules together with table lookup.

Surface order is also mentioned in the literature as a factor in predicting pitch accent. While this feature is most frequently appealed to in attempts to predict the likely location of nuclear stress, it has also been observed informally that items which are preposed will be accented more frequently than the same item in non-initial position. Compare (6a), for example, with (6b):

(6) a. TODAY we will BEGIN to LOOK at FROG anatomy.
 b. We will BEGIN to LOOK at FROG anatomy today.

Even prepositions in preposed PPs will sometimes receive a **L*** accent, although they are rarely accented when not preposed. In certain genres, too, the importance of surface order in combination with word class in determining pitch accent has been noted. For example, in broadcasters' speech, Bolinger (1982; 1989) has noted the tendency of broadcasters to place nuclear stress as close to the end of a phrase as possible, even when this decision leads to accenting items that, in ordinary speech, would probably be deaccented.

The remaining predictors of pitch accent commonly noted in the literature may be broadly characterized under the rubric INFORMATION STATUS. In this category we include an item's status in the discourse as GIVEN or NEW, that is, whether the item represents 'old' information, which a speaker is entitled to believe is shared with a hearer, or not (Prince, 1981b), as well as whether the item is being linguistically FOCUSED or whether the item is being used CONTRASTIVELY.

Experimental evidence indicates that speakers tend to introduce 'new' items in utterances by accenting them, and tend to deaccent items representing 'given' information (Brown, 1983).[4] Perception studies also indicate that hearers appear sensitive to this association of givenness with deaccenting and newness with accent (Terken and Nooteboom, 1987). For example, *lawyers* may be accented in its initial occurrence in (7a), but it is likely to be deaccented in its second occurrence.

(7) a. There are LAWYERS, and there are GOOD lawyers.
 b. EACH NATION DEFINES its OWN national INTEREST.
 c. The SLEEPY SOLDIER SLEPT for an HOUR.
 d. JOHN TRIED to be HELPFUL, but ONLY SUCCEEDED in being UNhelpful.
 e. I LIKE GOLDEN RETRIEVERS, but MOST dogs LEAVE me COLD.

But how items come to be perceived as 'given' — and how they lose that 'givenness' — are open research questions. For example, a concept need not have been explicitly evoked in prior discourse to be taken as

[4] Using a taxonomy based on Prince's (1981b) classification of given/new information, (Brown, 1983) found that 87% of BRAND NEW entities (not assumed known by the hearer) and 79% of NEW INFERRABLE entities (assumed by the speaker to be inferrable by reasoning from previously evoked entities) were in fact accented by subjects, while 96-100% of EVOKED entities (those already mentioned or situationally salient) were deaccented.

given; in (7b) the mention of *nation* can make the subsequent mention of *national* deaccentable as given information. However, mention of *sleepy* does not seem to make *slept* deaccentable in (7c). Mention of one item in a contrast set can license the 'givenness' of other members of a set, although sometimes with different consequences for accentuation; for example, *unhelpful* is made given by mention of *helpful* in (7d), and, in consequence, only the negative prefix is stressed. Mention of subordinate categories can make superordinates given and, hence, deaccentable, as illustrated in (7e), where mention of *golden retrievers* can make *dogs* deaccentable.

While rules encapsulating the full complexities of givenness may be far in the future, it is nonetheless possible to model the simpler aspects of this phenomenon. Silverman (1987) proposed using a FIFO queue of word roots of mentioned open-class items in a text-to-speech system, to determine whether subsequent items should be deaccented, clearing the queue at paragraph boundaries.[5] In contrast to this linear approach, the current study makes use of Grosz and Sidner's (1986) hierarchical notion of a discourse's ATTENTIONAL STRUCTURE to define the domain of givenness. In this model of discourse, discourse structure itself comprises three structures, a LINGUISTIC STRUCTURE, which is the text/speech itself; an ATTENTIONAL STRUCTURE, which includes information about the relative salience of objects, properties, relations, and intentions at any point in the discourse; and an INTENTIONAL STRUCTURE, which relates the intentions underlying the production of speech segments to one another. The attentional structure is represented as a stack of FOCUS SPACES, which is updated by pushing items onto or popping them from the stack.[6] For the current study, no attempt has been made to model (Grosz and Sidner, 1986)'s attentional structure entirely. In particular, no aspect of intentional structure is included, and objects, properties and relations are represented here by their roots rather than by some more abstract conceptual representation. At any point in the discourse, the current focus space contains some representation of items that are currently salient in the discourse, or, in LOCAL FOCUS (Sidner, 1983; Grosz et al., 1983; Brennan et al., 1987). The relationship between items in a particular focus space — as well as the relationship between items in different focus spaces — is not well understood; however, some hierarchical relationship between focus spaces is usually assumed in the literature. Specifically, items in a focus space that is associated with a portion of the discourse that is understood as superordinate to another portion (e.g., a super-topic perhaps), are generally assumed to be available during the processing of the subtopic. Side by side with this notion of local focus is a notion of GLOBAL FOCUS, in which concepts central to the main purpose of the discourse remain salient throughout the discourse.

Local focus is modeled in the current experiments as a stack of focus spaces, each of which contains the roots of some set of lexical items contained in the (orthographic) phrase currently being processed; these roots substitute for the concepts which would ideally be represented in the focus stack. Hierarchical relationships between the discourse segments associated with these focus spaces — and, thus, hierarchical relationships between the focus spaces themselves — are inferred from orthographic cues, such as paragraphing and punctuation, and lexical cues, such as CUE PHRASES[7]; each cue phrase has a 'push' or 'pop' operation associated with it for this purpose (Litman and Hirschberg, 1990). In the current study, roots of members of certain word classes mentioned in the discourse are added to the focus space; the classes that are to be

[5] This proposal was not actually tested by Silverman, although certain aspects of the proposal appear to be supported by experimental findings of the current study discussed below.

[6] The notion of 'focus' in this tradition as 'what is currently being talked about' should not be confused with the linguistic notion of 'focus' as 'what the speaker wishes to make specially prominent'.

[7] Cue phrases are words such as *now*, *well*, and *by the way*, which provide explicit information about the structure of a discourse.

considered were varied experimentally, to see which strategy best supported observed accent behavior. While local focus is continually updated during the discourse, items in global focus are fixed early in the analysis of the text. The procedures by which local focus is updated and the point at which global focus is fixed were also varied experimentally. Positional relationships between items in local focus was also examined. And, the content of both global and local focus spaces were also varied systematically by word class.

Linguistically 'focused' or 'contrastive' items[8] commonly receive special intonational prominence (Lakoff, 1971; Schmerling, 1976; Bolinger, 1961; Jackendoff, 1972; Rooth, 1985; Horne, 1987). In some cases, such prominence is the sole means by which 'contrastiveness' can be inferred. Suggestions for how to predict this status have been many and varied, including the identification of preposed constituents (Prince, 1981a; Ward, 1985) and of semantic or syntactic parallelism. In addition, a notion of PROPER-NAMING, involving the use of proper names for individuals who might otherwise be differently referenced (e.g. by common nouns or pronouns) and derived from experimental results discussed in (Sanford, 1985), provides another method for predicting contrastive stress. The discourse model described above has been employed to model this phenomenon. For example, items in global focus but not currently in local focus were frequently observed to be uttered with special emphasis, as if these items were being reintroduced into the discourse. This observation was tested empirically by keeping track of items in both local and global focus and comparing their status along both dimensions.

4.2 The FM Radio Database

The first stage of analysis involved data from the FM Radio database, described in Section 3.3. The original rules for accent assignment were developed by hand and implemented in Quintus Prolog in a program which takes unrestricted text as input and assigns accent based upon values of the variables discussed in Section 4.1. This rule-based system was employed to analyze transcriptions of recorded, prosodically labeled speech, and predictions were compared with hand-labeled observations to retrain the system. Experimentation with this paradigm included varying both the text analysis and the structure of the decision procedure to minimize error rates when predictions were compared with observed accent values in the prosodically labeled speech. So, for example, composition of broad word classes was varied systematically, and so was the use of the part-of-speech variable in the decision procedure. Best results for each of the FM news stories examined (79% correct pitch accent prediction for one and 85% for the other) occurred for the same variations of individual variable treatment and decision tree composition; combined results, with a success rate of 82.4%, are presented in Table 1, where the procedure's decision is based on the feature specified.

The association between word class and accent decision was initially modeled from Altenberg (1987)'s data, translating from Altenberg's tag set to that of (Church, 1988) and filling in gaps in Altenberg's coverage where necessary based upon the FM corpus. Such observation had previously suggested that a simple division by part-of-speech into open and closed classes was insufficient for making accurate pitch accent predictions. So, from observed pitch accent decisions, finer-grained distinctions were developed. Optimal predictive power was obtained for this corpus when four broad classes were assumed: open class, closed_cliticized, closed_deaccented (but not cliticized) and closed_accented. Membership in the four classes defined was not always possible from the tag set used in (Church, 1988), so exceptions were further specified by lexical item. In the final version of the hand-crafted rule set, 'closed_accented' items include the negative article, negative

[8] To avoid confusion with the concepts of local and global focus discussed above, these will be termed 'contrastive', although this term too has its problems. In particular, it is often impossible to find a contrast set for many items receiving so-called CONTRASTIVE STRESS.

Total correct 1000 (82.4%)			Total incorrect 214 (17.6%)		
Correctly predicted by		Percent of total error	Incorrectly predicted by		Percent of total error
new	418	(42%)	*new*	63	(30%)
given	32	(3%)	*given*	35	(16%)
closed_accented	60	(6%)	*closed_accented*	11	(5%)
closed_cliticized	283	(28%)	*closed_cliticized*	30	(14%)
closed_deaccented	119	(12%)	*closed_deaccented*	36	(17%)
compound_accented	0	0%	*compound_accented*	0	(0%)
compound_deaccented	19	(2%)	*compound_deaccented*	13	(6%)
cue_phrase	16	(2%)	*cue_phrase*	6	(3%)
contrast	59	(6%)	*contrast*	18	(8%)

Table 1: Accent Prediction for the FM Radio Database (n=1214)

modals, negative *do*, most nominal pronouns, most nominative and all reflexive pronouns, pre- and post-qualifiers (e.g. *quite*), pre-quantifiers (e.g. *all*), post-determiners (e.g. *next*), nominal adverbials (e.g. *here*), interjections, particles, most *wh*-words, plus some prepositions (e.g. *despite, unlike*). 'Closed_deaccented' items include possessive pronouns (including *wh*-pronouns), coordinating and subordinating conjunctions, existential *there*, *have*, accusative pronouns and *wh*-adverbials, some prepositions, positive *do*, as well as some particular adverbials like *ago*, nominative and accusative *it* and nominative *they*, and some nominal pronouns (e.g. *something*). And 'closed_cliticized' items include the definite and indefinite articles, many prepositions, some modals, copular verbs and *have* auxiliaries. Other items (adjectives, adverbials, nouns, verbs) are deemed 'open'.

Forty-six percent of correct accent predictions for the FM Radio corpus were made simply by distinguishing which closed-class items were likely to be accented and which were not: 86% of closed class items were correctly predicted. Most (83%) errors for closed class items involved instances of function words which were deaccented in the majority of cases but which were sometimes accented by the speakers, such as coordinate conjunctions, copular verbs, and some prepositions and determiners. The simple division into broad word classes would in fact correctly predict 77% of observed pitch accents in the corpus. Note, however, that the simple traditional association between function word status and deaccenting and content word and accenting would correctly predict only 68% of items in the corpus. So, the finer-grained model employed here appears to produce an improvement.

Citation-form stress patterns predicted by NP account for only 2% of correct accent predictions and 6% of incorrect predictions; on balance, 59% of predictions based on citation-form compound stress assignment proved to be correct. Errors in every case arose from incorrectly predicting deaccenting of an element; however, in half these cases the algorithm predicted accent strategies which were in fact clearly acceptable to native speakers, even though not chosen by the speakers in the corpus.

The collection and manipulation of the attentional state representation was also varied experimentally in the following ways: both global and local focus representations were manipulated independently such that the global focus space was set, and the local focus spaces updated, by the orthographic phrase, the sentence, or the paragraph. So, for example, the global space was considered to be defined after the (items mentioned in the) first phrase, first sentence or first paragraph of a text. The local stack was updated independently at the end of each phrase, sentence or paragraph – although cue phrases could also motivate a push or a pop of the stack under any of the experimental conditions. For the FM database, the best results were obtained when the global space was defined as the first full sentence of the text and the local attentional stack updated by paragraph; the latter finding to some extent appears to bear out Silverman's proposal for updating a set of roots at paragraph boundaries (page 9), although the updating for the current study was also sensitive to the presence of cue phrases. The potential content of both global and local focus spaces was also varied, so that, for example, all open-class words, nouns only, or nouns plus some combination of verbs and modifiers were allowed to affect and be affected by changes in the attentional state representation. For the FM data, best results were obtained when focal spaces included roots of all content words (nouns, verbs and modifiers) rather than some subset. However, overall, distinguishing between given and new for this corpus proved to be of limited use: While 42% of total correct predictions were due to associating new information with accent, only 3% of correct predictions were those associating 'given' items with lack of accent. Thirty percent of the total error was due to predicting incorrectly that 'new' items were accented, and 16% to predicting incorrectly that 'given' items were deaccented. Overall, 87% of items labeled 'new' were correctly predicted to be accented, but only 48% of 'given' items were correctly predicted to be deaccented. Incorrect accent predictions based on inference of given/new status account for 31.6% of the overall error on this corpus.

It is possible that the inaccuracy of given/new predictions might be due to the fact that interview portions of these news stories have been spliced into the newsreaders' commentary. So, it is not clear how information status should be assigned, when some speakers are not in fact participants in the conversation in which their speech appears; that is, for the newsreader, what is said in an interview presumably becomes given information; but clearly what represents given information for the interviewee is not always discernible from a fragment of the interview. It is also possible of course that givenness was not correctly modeled, or that other factors should have been considered in evaluating the influence of this type of information status on pitch accent. These questions will be pursued below.

Finally, an attempt was made to infer local and global focus and contrastiveness from surface position, part-of-speech, and the state of the local and global focus spaces. In earlier experiments it was noted that, in the FM corpus, speakers often placed particular intonational prominence on proper names which had previously been introduced into the text, and which thus seemed suitable for deaccentuation as 'given'. The referential strategy of PROPER-NAMING (Sanford, 1985), in which the use of proper names was noted as a mechanism to focus attention, was hypothesized as a means of accounting for this tendency. That is, it was conjectured that such referential behavior might indicate the speaker's attempt to refocus attention upon recently-mentioned persons when other persons had been mentioned even more recently. It was further hypothesized that this re-focusing accounted for the special intonational prominence which appeared to be assigned to these items. While there was not enough data to test this hypothesis in much detail, in the FM corpus it was found to predict correctly in all cases where a proper name was employed to identify a person who name was already represented in local focus. Similarly, it was noted through experimentation that speakers tended to accent items that were predicted to be in global but not in local focus. Again, the general notion of reintroducing an entity mentioned earlier in the text, when the discussion had shifted to other topics, seemed related to the assignment of intonational prominence, although there were relatively few instances in which to test the hypothesis. Other approaches to modeling contrastive stress included the prediction of accent on preposed adverbials and prepositional phrases and the proposal that in complex nominals, when several elements are given and at least one is new, the new item(s) receives special prominence. In general, accent for items identified as contrastive was correctly predicted 77% of the time. However, such items account for only 6% of all items in the data.

The hand-crafted rules as currently implemented (Hirschberg, 1990a; Hirschberg, 1990b) operate on text which is first tagged for part-of-speech, and in which standard orthographic indicators of paragraph and sentence boundaries are preserved. Tables are maintained of word classes and individual items divided into four broad classes — closed-cliticized, closed-deaccented, closed-accented, and open — where distinctions among the first three groups are based upon frequency distributions in the training data. As processing of the labeled text proceeds from left to right, the following information is added to a record maintained for each item: (1) Preposed adverbials are identifed from surface position and part-of-speech, as are fronted *PP*s, and labeled as *preposed*. (2) Cue phrases are identified from surface position and part-of-speech as well, and their accent status is predicted following findings in (Hirschberg and Litman, 1987; Litman and Hirschberg, 1990) (3) Verb-particle constructions are identified from table look-up. (4) Local focus is implemented in the form of a stack of roots of all nouns, verbs, and modifiers in the previous text. New items are pushed on the stack as each phrase is read. Individual cue phrases trigger either push or pop operations, roughly as identified in (Grosz and Sidner, 1986). Paragraph boundaries cause the entire stack to be popped. As new items are read, those whose roots appear in local focus are marked as 'given', while others are labeled 'new'. Additionally, items with prefixes such as *un*, *il*, and so on whose roots appear in local focus are marked as 'prefixed'. (5) Global focus is defined as simply the set of all content words in the first sentence of the first paragraph of the text. Nouns, verbs and modifiers whose roots appear in global focus are so marked. (6) Potential complex nominals are identified from part-of-speech and shipped to Sproat's complex nominal parser (Sproat, 1990); their stress pattern in citation form is stored with the nominal for subsequent processing. Other information collected for these nominals for subsequent evaluation includes whether or not they are inferred to be proper (e.g., street names, personal names that include titles, acronyms, and so on). (7) Finally, possible contrastiveness within a complex nominal is inferred by comparing the presence of roots of elements of the nominal in local focus; roughly, if some items are 'given' and others 'new', the new items are marked as potentially contrastive.

When each record is complete, a simple decision procedure determines whether the item will be predicted as cliticized, deaccented (but not cliticized), or accented. The algorithm is presented in Figure 7.

Figure 7 goes here.

In summary, experimentation with a hand-crafted rule set for pitch accent prediction on a corpus of recorded radio speech revealed the following: A significant improvement in accent prediction can be made by a more fine-grained use of word class information. Where a simple, function-content distinction correctly predicts 68% of pitch accents observed in this sample, more sophisticated distinctions permit correct prediction of 77% of observed accents. Disappointingly, modeling of the given/new distinction and of contrast adds only 5.4% to the overall score of 82.4% correct. However, while the given/new predictions are often incorrect, predictions of accent from inferred contrastive status are generally correct.

4.3 The Citation Sentences

The second corpus examined in developing rules for pitch accent prediction was the corpus of citation-form sentences. The original purpose in examining this corpus was not to test the success of the accent assignment algorithm specifically, but rather to see whether it could supply prosodic information which was not available from other hand labelings of the corpus for an independent analysis of segmental durations (Van Santen, 1992). There is some empirical evidence that a word's accent status plays a significant role in the duration at

	Accented	Deaccented	Cliticized
Speaker I (m)	168 msec	142 msec	115 msec
Speaker II (f)	184 msec	137 msec	112 msec
Speaker III (f)	159 msec	130 msec	105 msec

Table 2: Mean Duration for Stressable Syllables

least of its stressable syllable (Cooper et al., 1985; Eefting, 1991). So, using the hand-crafted rules described in Section 4.2, three levels of accent were predicted for these utterances: accented, deaccented but not cliticized, and cliticized. To obtain a measure of performance on this unlabeled corpus, these predictions were checked by hand for 302 of the utterances for one speaker, including a total of 1756 words. Results from this evaluation indicate that accent predictions were correct in 98.3% of cases, a quite phenomenal success rate — especially in comparison with the earlier results on the data for which these rules were originally developed. The validity of this evaluation received further indirect confirmation by results of the durational analysis subsequently performed on the entire set of 1380 sentences for each of the three speakers, as shown in Table 4.3, based on data from (Van Santen, 1992). Note that stressable syllables in words the accent prediction rules predicted as accented were found to be 18-26% longer than stressable syllables in words predicted to be deaccented but not cliticized, and 32-39% longer than similar syllables in words predicted to be cliticized.

It seems likely that the high success rate of accent prediction in citation-form sentences is due to the nature of the data: In these sentences, intonation is monotonous and regular, and each sentence has clearly been uttered with little influence from the semantic content of prior utterances. The 302 sentences checked indicate little effect of a given/new distinction, and almost no evidence of focal or contrastive accent. Most of the success of the prediction rules, then, appears to have come from the finer-grained distinctions made between word class and pitch accent, as discussed above. Thus, the usefulness of this simple refinement in word class use appears very promising for assigning pitch accent at the level of isolated sentences in text-to-speech.

4.4 Automatic Classification of Pitch Accent

The derivation of prediction rules by hand which was done for the FM Radio corpus is a labor-intensive and time-consuming process, with results whose generalizability to new corpora is difficult to predict. The rule set described in Section 4.2 took several months to construct and refine, for example. And the difference in performance of these rules on the Citation Sentence data and the FM data provides some evidence that different text genres require different rule sets, despite the interesting result that rules derived from one data set in fact performed considerably better on another. To address these concerns, further experimentation was performed using Classification and Regression Tree (Brieman et al., 1984) (CART) techniques to generate and cross-validate decision trees automatically. It was highly desirable to see whether procedures derived automatically could perform as well or better than hand-crafted rules.

4.4.1 CART Techniques

CART techniques allow one to produce decision trees from sets of continuous and discrete variables, provided to CART as feature vectors associating variables with values, and independent with dependent variables. Sample trees produced by CART appear in Figures 8-11. Producing a particular decision tree for a given data set involves the operation of sets of splitting rules, stopping rules, and prediction rules to control the production of internal nodes, the height of the tree, and the labeling of terminal nodes, respectively. At each internal node, CART determines from the input feature vectors which single feature should govern the forking of two paths from that node, and which values of that feature to associate with each path. Ideally, the splitting rules should choose the factor and value split which minimizes the prediction error rate. The splitting rules in the implementation employed for this study (Riley, 1989) approximate optimality by choosing at each node the split which minimizes the prediction error rate on the training data. In this implementation, all these decisions are binary and local, based upon consideration of each possible binary partition of values of categorical variables and consideration of different cut-points for values of continuous variables.

Stopping rules terminate the splitting process at some internal node. To determine the best tree, this implementation uses two sets of stopping rules. The first set is extremely conservative, resulting in an overly large tree, which usually lacks the generality necessary to account for data outside of the training set. To compensate, the second rule set forms a sequence of subtrees. Each tree for this study was grown on 90% of the training data and tested on the remaining 10%. This step is repeated until the tree has been grown and tested on all of the data. The stopping rules thus have access to cross-validated error rates for each subtree. The subtree with the lowest error rates then defines the stopping points for each path in the full tree. Decision trees presented or described below all represent cross-validated data; full trees in every case thus account for more of the data correctly but are less reliable in their prediction of new test data.

The prediction rules add the necessary labels to the terminal nodes. For continuous variables, the rules calculate the mean of the data points classified together at that node. For categorical variables, the rules choose the class that occurs most frequently among the data points. The success of these rules can be measured through estimates of deviation. The deviation for the categorical dependent variable examined here, pitch accent, is simply the number of misclassified observations.

4.4.2 Automatic Classification of the FM Database

To compare performance of the hand-crafted rules with those which could be produced automatically, the same variables referenced by those rules (described in Section 4.1) were used to create input for CART classification. Data from both FM stories was pooled for this analysis, for a total of 1214 accent decisions for testing and training, a rather small corpus for automatic prediction purposes as it turned out. In fact, the prediction tree produced for this data set was quite degenerate, consisting of only a single split, on the function/content word distinction. The cross-validated success rate for this tree was 76.5%, the same percent correctly predicted by this variable for the hand-crafted rules. Testing on a larger corpus produced more interesting results, suggesting that the classification procedure simply had too little data to make generizable predictions for this corpus.

4.4.3 The ATIS and Audix Corpora

Analysis of the larger Audix corpus involved the same variables used above, for 2854 datapoints — more than twice the number in the FM Radio sample. Since the performance of the 'given/new' distinction in the analysis of the FM Radio database had been somewhat disappointing, it was desirable to determine whether the modeling of this feature was flawed in general, or whether a larger corpus would provide a better testbed. This larger database also provided an opportunity to examine several variations on the discourse modeling strategy described in Section 4.1. In addition to varying the units used in updating the local focus stack and in modeling global focus, position of a lexical root within the local focus space was noted when the current word was found to be given. The notion here was that items more remote in local focus (i.e., made given earlier rather than more recently) might be less likely to impart givenness to the current item.

In analyzing the ATIS corpus, several new variables were considered as well, capitalizing upon prosodic information which had been identified for a separate analysis of intonational phrasing (Wang and Hirschberg, 1991; Wang and Hirschberg, 1992). Since it seemed possible that the position of an item within its intonational phrase, as well as the overall length of this phrase, might be related to its likely accent status, these additional features were examined. At a minimum, most phonological theories predict that items alone in their intonational phrase will be accented. In addition, since a large number of speakers are represented in the ATIS corpus, and since accent strategy may indeed be speaker dependent, a speaker identity variable was added. Finally, to capture the intuition that a larger word class context may play a role in accent decisions, the part-of-speech of both left and right contexts of the current items, as well as the part-of-speech of the item itself, were examined.

The results of using CART techniques to model pitch accent predictions automatically for the ATIS and Audix corpora are a set of cross-validated decision trees, some of which are displayed in Figures 8-11. Table 3 identifies the variable names used in these figures. Of the part-of-speech tags mentioned in this table, all but one (PART, or, 'particle') are those used in (Church, 1988), which, in turn, represent a subset of those in (Francis and Kucera, 1982); they are defined in Table 4. As noted in Section 4.4.1, nodes in the trees are labeled with the dependent variable value for the majority of data points classified at that node. So, in Figure 8 the root node is labeled *deacc*, or 'deaccented', since the majority of nodes classified by the tree are classified by it as deaccented. The proportion of nodes **correctly** so classified is indicated by the fraction under the box or circle; e.g., for the root in Figure 8 1495 or 2854 items are correctly classified by the tree as deaccented. Arcs in the tree are labeled by the feature controlling the split from the parent node, as well as the values of the feature which the split is made on. The first split from the root node in Figure 8 is made on the feature *cl*, or 'class'. Values *clcl* and *clde* (closed_cliticized and closed_deaccented) of the class feature are grouped together here, both being used to predict deaccenting on the left branch, while features *clac* (closed_accented) and *op* are collapsed to form the right branch of the tree. The total number of correct preditions for a full tree may be ascertained by summing over the numerators of fractions labeling leaf nodes; however, as discussed in Section 4.4.1, it should be noted again that these numbers are somewhat higher than the cross-validated scores for these trees.

The cross-validated success rates for automatic classification are similar to those obtained for the hand-written rules described in Section 4.2 without cross-validation — about 80% of pitch accents in the Audix data are correctly predicted and 85% of the ATIS accents. The best full decision trees for both corpora cover 92-93% of the data.

Figure 8 shows a decision tree for the Audix corpus, which makes use of the same variables used in the hand-written rules created for the FM corpus. This tree successfully classifies 80% of the data. In this tree, as in the remainder of the predictions for the Audix corpus, the procedure over-predicts accent; in this tree, two thirds of the errors occur when an item which the speaker has deaccented is predicted to be accented. The (cross-validated) tree itself is quite a simple one, however, and makes a good deal of intuitive sense. The first split in the tree occurs on the partition of items by part-of-speech into open or closed, with the latter class divided into 'closed_cliticized', 'closed_deaccented', and 'closed_accented'; 76.7% (2190/2854) of the data can be correctly classified on this distinction alone — virtually identical to the proportion of items in the FM database that could be so identified. Here, 90.5% (929/1027) of items predicted to be deaccented or cliticized by virtue of part-of-speech assignment are indeed deaccented or cliticized, while 69% (1261/1827) of items predicted by class to be accented are accented. However, the second split in the tree is on the given/new variable, which, it will be recalled, did not prove a very successful predictor of accent for the FM corpus. Here, open class and closed_accented items are further subdivided into given and new (or not applicable); of 'given' items at this node, 69.6% (87/125) are deaccented, while 71.9% (1223/1702) of 'new' items are accented. The third split in the tree is on this set of new items, which is divided according to accent predictions arising from citation stress for complex nominals; 73% (1242/1702) of these items are correctly predicted by this feature. The final split in the tree separates nominative pronouns (both singular and plural) from all other parts-of-speech: these pronouns tend to be deaccented (27/33, or 81.8%).

So, the decision procedure illustrated in Figure 8 can be summarized as follows: If an item's part-of-speech indicates it is 'closed_cliticized' or 'closed_deaccented', or if it represents given information, or if it is (predicted to be) deaccented in complex nominal citation form, or if it is a nominative pronoun, predict that it is deaccented; otherwise, predict accented. This same simple result — and almost identical predictive success — appears even when the method of updating local and global focus is varied, when a 'focal space distance' measure is included, or when overall length of the current phrase and the current sentence are considered. While this result thus seems reliable, it does not suggest avenues for improving predictive power. However, a look at the full classification tree suggests that obtaining more data might reduce the errors in classification, where the addition or manipulation of variables will not. In many cases there are too few representatives of a potentially disambiguating word class to justify disambiguation on grounds of class membership, even though the existing members seem to be behaving predictably. In fact, this classification procedure does not make use of word identity, so there are no lexical exceptions to accenting likelihoods based upon part-of-speech, as there are in the hand-crafted rules.

Results from automatic classification of pitch accent for the ATIS corpus sample proved similar to results obtained for the Audix corpus — but improved considerably when additional features were examined. Performance based on the original variable set as described in Section 4.1 is illustrated by the tree in Figure 9, in which 81.9% of pitch accents are correctly predicted. However, note that, while the same variables are available to CART, the tree in Figure 9 makes use of a wider variety of features than any of the trees produced by CART for the Audix corpus. In particular, the initial split into three classes of 'closed' and 'open' is still the single most important predictor in the tree (correctly identifying 76.5% of pitch accents in the database), accent prediction from citation-form complex nominal defaults also accounts for a large share of the data, and whether or not the current item is a pronoun accounts for a small portion of the 'deaccented' predictions. What distinguishes the tree in Figure 9 from analyses of the Audix data is the important role allocated to various measures of distance and to the part-of-speech of the right and left contexts in the tree.

First, note that the first split in Figure 9 separates words whose class suggests they will be cliticized (articles and some pronouns, for the most part) from the other broad word classes. This class itself is then split again by distance from end of utterance, with items appearing at the end of an utterance likely to be accented, while the vast majority are likely to be deaccented. This result may capture the likelihood that prepositions ending an utterance may in fact be particles — which are more likely to be accented. Moving from the root through its right daughter, we can also see some evidence that a combination of an item's own word class together with knowledge of its neighbors' has considerable predictive power. For example, items not predicted by class to be cliticized, not predicted by complex nominal citation form to be deaccented, and that are not nominative or accusative pronouns, can be predicted with fair success (87.2%) based upon the part-of-speech of their left and right neighbors, their own distance from the beginning of the utterance, and the total number of words in the utterance. The rightmost branch of this tree reveals similar predictive power for items predicted by class to be accented, whose accent status depends upon whether their left neighbor is a *wh*-determiner and whether they themselves are common nouns at the end of an utterance or not. In short, while the tree depicted in Figure 9 is somewhat more complex than any obtained for the Audix corpus, the addition of new context and distance variables appears to improve predictive power. More significant improvements come when additional prosodic and speaker-identity variables are included as well.

Figure 10 shows a decision tree very similar to that of Figure 9; however, this tree successfully predicts 85.1% of the ATIS sample. In addition to the variables provided for CART analysis in Figure 9, distance with respect to preceding or succeeding intonational boundary is provided, as well as the type (major or minor) of that boundary. Comparing Figures 9 and 10, we notice that relative distance to boundary appears to subsume relative distance to beginning or end of utterance — understandably, since utterance edges are themselves boundaries; where relative position with respect to utterance beginning or end appears in Figure 9, relative position with respect beginning or end of intonational phrase appears in Figure 10. However, this feature provides additional discriminatory power in Figure 10: For example, in Figure 9, items predicted by word class to be 'closed_deaccented' are found to indeed be deaccented about three quarters of the time (152/200); however, in Figure 10 this prediction is further refined, with items appearing at the beginning of intonational phrases further distinguished by boundary type and part-of-speech of left context. In this way 84% (168/200) of these 'closed_deaccented' items are correctly predicted. Since the actual intonational phrasing of an utterance may not itself be reliably recoverable from text, it may seem optimistic to include observed values as predictors here. Current work on phrasing prediction from transcription, however, has achieved success rates of just over 90% (Wang and Hirschberg, 1991). Future investigations will explore the use of these predicted values in place of the observed values employed here.

Finally, the usefulness of one additional variable was tested for the ATIS data — speaker identity. In Figure 11, this variable helps to discriminate cases identified by part-of-speech context or distance from phrase boundary in Figure 10. Note that both Figures 10 and 11 classify approximately the same number of data points successfully, however. And speaker identify is no substitute for, say, the boundary-related variables employed in Figure 10. So, while speaker identity can serve as an alternative to automatically inferrable variables used in the preceding analysis, it does not significantly improve predictive power. The speaker-independent decision tree presented in Figure 10 is, in fact, the most successful classification of the ATIS sample.

5 Discussion

The experiments described above were designed to develop procedures for modeling human pitch accent behavior in several different types of speech corpora. The most successful attempt involved predictions for citation-form sentences, where a prediction rate of 98.3% was achieved. Less successful, although still improving over the simple function/content word prediction employed by most text-to-speech systems, were attempts to predict accent location in larger contexts. Rule systems developed by hand and others generated through CART automatic classification techniques attained success rates of 80-85% on these corpora.

While, throughout the experiments, it was clear that part-of-speech plays a major role in the prediction of pitch accent, this feature alone reliably predicts only about three quarters of observed pitch accents. To improve predictive power, it is thus necessary to explore additional predictive features, none of which individually may be responsible for a large number of successful predictions, but which may, together, improve predictive accuracy. In exploring new sources of automatic accent prediction, a hierarchical representation of the attentional structure of the discourse was employed, which supported inference of local and global focus, as well as contrast. Additional sources of information tested included: word class context, citation-form compound nominal stress (derived from Sproat's NP parser), a variety of relative distance measures within intonational phrase and utterance, as well as utterance length and speaker identity. While not enough data have been examined to assess fully the usefulness of any of these additional information sources, it is notable that different collections of features can produce roughly similar success rates for a given corpus, suggesting a certain amount of redundancy in predictors of pitch accent. Larger databases will have to be analyzed in order to tease apart relationships among the variables examined, and to improve our ability to model pitch accent strategies successfully.

To develop larger labeled speech corpora in support of this endeavor, the accent prediction procedures developed here are being employed, together with automatic segmentation and labeling software, to speed the prosodic labeling of speech corpora. Accent hypotheses generated currently from the hand-written rules described above are aligned with stressable syllables in the (digitized) speech; these syllables are located via automatic segmentation and labeling programs written by David Talkin and by Andrej Ljolje to produce a prosodic label hypothesis, which can then be post-edited. The result is a speed-up in labeling performance by a factor of five or better.

While much work remains to be done to improve our ability to assign pitch accent naturally for text-to-speech applications, the results of the experiments described here have been used in this endeavor with considerable success. The algorithm described above is in fact currently used to assign pitch accent in TTS from text analysis.

The experiments described above represent a step toward the goal of predicting accent placement reliably from information automatically obtainable via text analysis. More and more varied types of speech data must be analyzed and additional sources of predictive information explored. However, the current results of 98% for citation-form sentences and 80-85% correct prediction for other genres do suggest that current text analysis technology can be mined more thoroughly than it yet has been to support better modeling of human prosodic performance.

6 Acknowledgments

Thanks to Ken Church, Don Hindle, Andrej Ljolje, Michael Riley, Richard Sproat, and David Talkin, all of whom provided software used in the course of the experiments described above. Thanks also to Jill Burstein, Nava Shaked, and Michelle Wang for their work on prosodic labeling of several of the corpora, to Mari Ostendorf, Patti Price, and Stefanie Shattuck-Hufnagel for providing the FM Radio Newscasting data and transcripts, and to Mark Liberman and Joe Olive for the Audix data.

References

Bengt Altenberg. 1987. *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, volume 76 of *Lund Studies in English.* Lund University Press, Lund.

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

Dwight Bolinger. 1961. Contrastive accent and contrastive stress. *Language*, 37:83–96.

Dwight Bolinger. 1972. Accent is predictable (if you're a mindreader). *Language*, 48:633–644.

Dwight Bolinger. 1982. The network tone of voice. *Journal of Broadcasting*, 26:725–728.

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

Susan E. Brennan, Marilyn W. Friedman, and Carl J. Pollard. 1987. A centering approach to pronouns. In *Proceedings of the 25th Annual Meeting*, pages 155–162, Stanford. Association for Computational Linguistics.

Joan Bresnan. 1971. Sentence stress and syntactic transformations. *Language*, 47:257–281.

Leo Brieman, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1984. *Classification and Regression Trees.* Wadsworth & Brooks, Monterrey CA.

Gillian Brown. 1983. Prosodic structure and the given/new distinction. In D. R. Ladd and A. Cutler, editors, *Prosody: Models and Measurements*, pages 67–78. Springer Verlag, Berlin.

Wallace Chafe. 1976. Givenness, contrastiveness, definiteness, subjects, topics, and point of view. In C. Li, editor, *Subject and Topic*, pages 25–55. Academic Press, New York.

K. W. Church. 1988. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin. Association for Computational Linguistics.

H. H. Clark and E. V. Clark. 1977. *Psychology and Language.* Harcourt, Brace, Jovanovich, Inc.

W. E. Cooper, S. J. Eady, and P. R. Mueller. 1985. Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America*, 77:2142–2155.

A. Cruttenden. 1986. *Intonation.* Cambridge University Press, Cambridge UK.

D. Crystal. 1975. *The English Tone of Voice: Essays in Intonation, Prosody, and Paralanguage.* Edward Arnold, London.

L. Danlos, E. LaPorte, and F. Emerard. 1986. Synthesis of spoken messages from semantic representations. In *Proceedings of the 11th International Conference on Computational Linguistics*, pages 599–604. International Conference on Computational Linguistics.

James R. Davis and Julia Hirschberg. 1988. Assigning intonational features in synthesized spoken directions. In *Proceedings of the 26th Annual Meeting*, pages 187–193, Buffalo. Association for Computational Linguistics.

W. Eefting. 1991. The effect of 'information value' and 'accentuation' on the duration of dutch words, syllables, and segments. *Journal of the Acoustical Society of America*, 89:412–424.

W. Nelson Francis and Henry Kucera. 1982. *Frequency Analysis of English Usage*. Houghton Mifflin, Boston.

B. Grosz and C. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

B. Grosz, A.K. Joshi, and S. Weinstein. 1983. Providing a unified account of definite noun phrases in discourse. In *Proceedings of the 21st Annual Meeting*, pages 44–50, Cambridge MA, June. Association for Computational Linguistics.

M. A. K. Halliday and Ruquaiya Hassan. 1976. *Cohesion in English*. Longman.

Charles T Hemphill, John J. Godfrey, and George R.. Doddington. 1990. The atis spoken language systems pilot corpus. In *Proceedings of the DARPA Speech and Natural Language Workshop*, June.

J. Hirschberg and D. Litman. 1987. Now let's talk about *now*: Identifying cue phrases intonationally. In *Proceedings of the 25th Annual Meeting*, pages 163–171, Stanford University. Association for Computational Linguistics.

Julia Hirschberg. 1990a. Assigning pitch accent in synthetic speech: The given/new distinction and deaccentability. In *Proceedings of the Seventh National Conference*, pages 952–957, Boston. American Association for Artificial Intelligence.

Julia Hirschberg. 1990b. Using discourse context to guide pitch accent decisions in synthetic speech. In *Proceedings of the European Speech Communication Association Workshop on Speech Synthesis*, pages 181–184, Autrans, France.

Merle Horne. 1987. Towards a discourse-based model of English sentence intonation. Working Papers 32, Lund University Department of Linguistics.

Ray S. Jackendoff. 1972. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge MA.

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

M. Liberman and R. Sproat. Forthcoming. The stress and structure of modified noun phrases in English. In I. Sag, editor, *Lexical Matters*. University of Chicago Press.

Diane Litman and Julia Hirschberg. 1990. Disambiguating cue phrases in text and speech. In *Papers Presented to the 13th International Conference on Computational Linguistics*, pages 251–256, Helsinki, August. International Conference on Computational Linguistics.

J. P. Olive and M. Y. Liberman. 1985. Text to speech – an overview. *Journal of the Acoustic Society of America, Suppl. 1*, 78(Fall):s6.

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

E.F. Prince. 1981a. Topicalization, focus-movement, and Yiddish-movement: A pragmatic differentiation. In *Proceedings of the Seventh Annual Meeting*, pages 249–264. Berkeley Linguistics Society.

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

R. Quirk, J. Svartvik, A. P. Duckworth, J. P. L. Rusiecki, and A. J. T. Colin. 1964. Studies in the correspondence of prosodic to grammatical features in English. In *Proceedings of the Ninth International Congress*, pages 679–691. International Congress of Linguists.

Michael D. Riley. 1989. Some applications of tree-based modelling to speech and language. In *Proceedings.* DARPA Speech and Natural Language Workshop, October.

Mats Rooth. 1985. *Association with Focus.* Ph.D. thesis, University of Massachusetts, Amherst MA.

Anthony J. Sanford. 1985. Aspects of pronoun interpretation: Evaluation of search formulations of inference. In G. Rickheit and H. Strohner, editors, *Inferences in Text Processing*, pages 183–204. North-Holland, Amsterdam.

Susan F. Schmerling. 1976. *Aspects of English Sentence Stress.* University of Texas Press, Austin. Revised 1973 thesis, University of Illinois at Urbana.

C. L. Sidner. 1983. Focusing in the comprehension of definite anaphora. In M. Brady, editor, *Computational Models of Discourse*, pages 267–330. MIT Press, Cambridge MA.

K. Silverman. 1987. *The Structure and Processing of Fundamental Frequency Contours.* Ph.D. thesis, Cambridge University, Cambridge UK.

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

David Talkin. 1989. Looking at speech. *Speech Technology*, 4:74–77, April-May.

J. Terken and S. G. Nooteboom. 1987. Opposite effects of accentuation and deaccentuation on verification latencies for given and new information. *Language and Cognitive Processes*, 2(3/4):145–163.

Jan P. H. Van Santen. 1992. Contextual effects on vowel duration. Manuscript.

Michelle Q. Wang and Julia Hirschberg. 1991. Predicting intonational boundaries automatically from text: The ATIS domain. In *Proceedings.* DARPA Speech and Natural Language Workshop, February.

Michelle Wang and Julia Hirschberg. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6. To appear.

Gregory Ward. 1985. *The Semantics and Pragmatics of Preposing.* Ph.D. thesis, University of Pennsylvania, Philadelphia PA.

S. J. Young and F. Fallside. 1979. Speech synthesis from concept: A method for speech output from information systems. *Journal of the Acoustic Society of America*, 66(3):685–695, September.

sw	distance from beginning of utterance
ew	distance from end of utterance
totw	total words in utterance
bp	distance in words from prior boundary
bs	distance in words from next boundary
bpt	type of prior boundary (n:none,f:minor,i:major,s:utterance)
bst	type of next boundary (n:none,f:minor,i:major,s:utterance)
p.o.s. for prior (p), subsequent (s), and current (c) word:	
v	DO HV HVD HVZ MD SAIDVBD VB VBD VBG VBN VBZ NA
b	BE BED BEDZ BEG BEM BEN BER BEZ NA
m	ABN DT DTI DTS DTX JJ QL RB NA
f	* AT CC CD CS EX IN PART TO UH NA
n	NN NNS NP PN NA
p	PPL PPO PPS PPSS NA
w	WDT WPS WRB NA
cl	class clac clcl clde op
cmpacc	whether np predicts accented or deaccented (ac de NA)
comp	complex nominal or not smp cmp NA
cmp	s(imple),c_a (complex, accent predicted)
	c_d (complex, predict deaccented) NA
info	given new NA
accent	deacc acc

Table 3: Key to Node Labels in Figures 8-11

'*'	*not, n't*
ABN	pre-quantifier (e.g., *all,half*)
AT	article
BE	*be*
BED	*were*
BEDZ	*was*
BEG	*being*
BEM	*am, 'm*
BEN	*been*
BER	*are, 're*
BEZ	*is, 's*
CC	coordinate conjunction
CD	cardinal number
CS	subordinate conjunction
DO	*do*
DT	singular determiner
DTI	singular/plural determiner (e.g., *some,any*)
DTS	plural determiner
DTX	determiner for double conjunction
EX	existential *there*
HV	*have*
HVD	*had,'d*
HVZ	*has,'s*
IN	preposition
JJ	adjective
MD	modal auxiliary
NN	singular or mass noun
NNS	plural noun
NP	singular proper noun
PART	verb particle
PN	nominal pronoun
PPL	singular reflexive personal pronoun
PPO	objective personal pronoun
PPS	third person singular nominative pronoun
PPSS	other nominative pronoun
QL	qualifier (e.g., *very*)
RB	adverb
SAIDVBD	*said*
TO	infinitive marker
UH	interjection
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle
VBN	verb, past participle
VBZ	verb, third person singular, present tense
WDT	*wh*-determiner
WPS	possessive *wh*-pronoun
WRB	*wh*-adverbial

Table 4: Part-of-Speech Labels in Figures 8-11

Figure 1: Accented Noun ‘Object’

Figure 2: Deaccented Noun 'Object'

Figure 3: Accented Verb 'Object'

Figure 4: Deaccented Verb 'Object'

```
Scenario: A small gymnastic team from overseas will be touring several
cities in this country.  Pick the cities they will visit and plan
their itinerary.  Please be sure: 1) the team will not miss meals, 2)
there will be ground transportation to and from downtown, and 3) they
do not travel on weekends, and 4) they fly on large (presumably safe)
aircraft.  :End Scenario

Show me all the nonstop flights from Atlanta to Philadelphia.

[Sent transcription for utterance 0 on Wed May 23 10:26:24 1990]

Result:
 FLT CODE FLT DAY FRM TO   DEPT  ARRV AL  FLT# CLASSES EQP MEAL STOP DC  DURA
--------- ------- --- --- ----- ----- -- ----- -------- --- ---- ---- -- -----
102766 1234567 ATL PHL   636   825 DL   296 FNYNBNMQ 72S B     0 N    109
102767 1234567 ATL PHL   740   929 US   258 FYBHQ D9S B     0 N    109
   ....

[Sent answer for utterance 0 on Wed May 23 10:26:51 1990]

Show me all the flights with breakfast from San Francisco to Denver.

[Sent transcription for utterance 1 on Wed May 23 10:34:00 1990]

Result:
 FLT CODE FLT DAY FRM TO   DEPT  ARRV AL  FLT# CLASSES EQP MEAL STOP DC  DURA
--------- ------- --- --- ----- ----- -- ----- -------- --- ---- ---- -- -----
133430 1234567 OAK DEN   625   949 UA   982 FYBMQ 733 B     0 N    144
144143 1234567 SFO DEN   620   946 UA   194 FYBMQ D10 B     0 N    146
   ....

[Sent answer for utterance 1 on Wed May 23 10:35:18 1990]

Show me all ground transportation in San Francisco.

[Sent transcription for utterance 1.5 on Wed May 23 10:37:07 1990]

Result:
CITY AIRPORT TRANSPORT GROUND FARE
---- ------- --------- -----------
SSFO OAK     L      $5.00
SSFO SFO     A
...

[Sent answer for utterance 1.5 on Wed May 23 10:37:44 1990]
```

Figure 5: Sample Log from ATIS Session

```
Wanted.     Chief Justice of the Massachusetts Supreme Court.     In April the
H*    LL%   H*+L  H*+L    cl cl L+H*           H*+L    H*   LL%   L* H*    cl
SJC's current leader Edward Hennessy     reaches the mandatory retirement
-     H*      -      H*     H*      LH%  H*       cl H*         H*
age of  seventy   and a  successor is expected to be named in  March.
-   cl  L+H*   L  cl  cl L+H*      -  H*       -  -  H*    cl  H*+L LL%
It may  be   the most important appointment Governor Michael Dukakis  makes
-  L+H* H*+L cl  H*   L+H*      H*          -        H*      H*       L+H* L
during the remainder of his  administration  and    one of the toughest.
-      cl  L+H*      cl cl   H*            L L+H* L L*  cl cl  H+L*    LH%
```

Figure 6: Prosodic Labels for an FM Corpus Fragment

For each item w_i labeled with part-of-speech p_i:

If w_i is a phrasal verb, deaccent;
Else if p_i is classified 'closed-cliticized', cliticize;
Else if p_i is classified 'closed-deaccented', deaccent;
Else if w_i is marked 'contrastive', 'prefixed', or 'preposed', assign it emphatic accent;
Else if w_i is part of a proper nominal
    If w_i's status is 'given', assign emphatic accent,
    else assign a simple pitch accent;
Else if w_i is in global focus but not in local focus ('given'), assign emphatic accent;
Else if w_i is classified 'closed-accented', accent;
Else if w_i is in local focus ('given'), deaccent;
Else if w_i is part of a (common) complex nominal
    If w_i is predicted to be accented in citation form, accent
    else deaccent;
Else accent w_i.

Figure 7: Accent Assignment Algorithm Developed by Hand

Figure 8: Pitch Accents Predicted from the Audix Corpus (80% Correct)

Figure 9: Pitch Accents Predicted from the ATIS Database (81.9% Correct)

Figure 10: ATIS Predictions with Boundary Information (85.1% Correct)

Figure 11: ATIS Predictions with Speaker Information (85.1% Correct)