# THE CONTRIBUTION OF PARSING TO PROSODIC PHRASING IN AN EXPERIMENTAL TEXT-TO-SPEECH SYSTEM

*Joan Bachenko*
*Eileen Fitzpatrick*
*C. E. Wright*

AT&T Bell Laboratories
Murray Hill, New Jersey 07974

## ABSTRACT

While various aspects of syntactic structure have been shown to bear on the determination of phrase-level prosody, the text-to-speech field has lacked a robust working system to test the possible relations between syntax and prosody. We describe an implemented system which uses the deterministic parser Fidditch to create the input for a set of prosody rules. The prosody rules generate a prosody tree that specifies the location and relative strength of prosodic phrase boundaries. These specifications are converted to annotations for the Bell Labs text-to-speech system that dictate modulations in pitch and duration for the input sentence.

We discuss the results of an experiment to determine the performance of our system. We are encouraged by an initial 5 percent error rate and we see the design of the parser and the modularity of the system allowing changes that will upgrade this rate.

## INTRODUCTION

We describe an experimental text-to-speech system that uses a deterministic parser and prosody rules to generate phrase-level pitch and duration information for English input. This information is used to annotate the input sentence, which is then processed by the text-to-speech programs currently under development at Bell Labs. In constructing the system, our goal has been to test the hypotheses (i) that information available in the syntax tree, in particular, grammatical functions such as subject-predicate and head-complement, is by itself useful in determining prosodic phrasing for synthetic speech, and (ii) that it is possible to use a syntactic parser that specifies grammatical functions to determine prosodic phrasing for synthetic speech.

Although certain connections between syntax and prosody are well-known (e.g. the influence of part of speech on stress in words like *progress*, or the setting off of parenthetical expressions) very little practical knowledge is available on which aspects of syntax might be connected to prosodic phrasing. In many studies, investigators have sought connections between constituent structure and prosody (e.g. Cooper and Paccia-Cooper 1980, Umeda 1982, Gee and Grosjean 1983) but, with the exception of Selkirk (1984), they tend to neglect the representation of grammatical functions in the syntax tree. Moreover, previous work has not been specific enough to provide the basis for a full system implementation. Based on our study of prosodic phrasing in recorded human speech, we decided to emphasize three aspects of structure that relate to phrasing: syntactic constituency, grammatical function, and constituent length. These findings, which we will discuss in detail, have been implemented as a collection of prosody rules in an experimental text-to-speech system.

Two important features characterize our system. First, the input to our prosody system is a parse tree generated by a version of the deterministic parser Fidditch (Hindle 1983). The left-corner search strategy of this parser and, in particular, its determinism, give Fidditch the speed that makes online text-to-speech production feasible.[1] In building a parse tree, Fidditch identifies the core subject-verb-object relations but makes no attempt to represent adjunct or modifier relations. Thus relative clauses, adverbials, and other non-argument constituents have no specified position in the tree and no specified semantic role. Second, the rules in the prosody system build a prosody tree by referring both to the syntactic structure and to earlier stages of prosodic structure. The result is a hierarchical representation that supports the view, also proposed in Selkirk (1984), that grammatical function information is related to prosodic phrasing, but indirectly, through different levels of processing.

Informal tests of the system show that it is capable of producing a significant improvement in the prosodic quality of the resulting synthesized speech. Our investigations of the system's problems, which we describe, have not revealed any serious counterexample to our basic approach. In many cases, it appears that problems with the current version can be resolved by taking our approach a step further, and including lexical information required by the parser as another factor in the determination of prosodic phrasing.

## TEXT-TO-SPEECH

Most text-to-speech systems comprise two components: pronunciation rules and a speech synthesizer. Pronunciation rules convert the input text into a phonetic transcription; this information may also be supplemented by a dictionary that provides information about the part of speech, stress pattern, and phonetic makeup of particular words. The speech

[1] With a grammar of about 600 rules and a lexicon of about 2400 words, Fidditch parses the 25 sample sentences of Robinson (1982), averaging 7 words per sentence and chosen for their structural diversity, at an average rate of .405 seconds per sentence on a Symbolics 3670. The rate is approximately proportional to the number of words in a sentence.

synthesizer then converts this phonetic transcription into a series of speech parameters which are subsequently processed to produce digitized speech.

While these systems tend to perform quite well on word pronunciation, they fall short when it comes to providing good prosody for complete sentences. Current text-to-speech systems have no access to the syntactic and semantic properties of a sentence that influence phrase-level prosody. Hence rules for sentence prosody, when they are provided at all typically depend on superficial aspects of text (e.g. punctuation) and on heuristics that vary widely in sophistication. Although such techniques often add a more natural quality to the resulting synthetic speech, they can fail in important ways, for example, by ignoring the prosodic event between a lengthy subject and a predicate, so that there is no clear prosodic boundary between *right* and *mark* in *The characters on the right mark the salient features.*[2]

Several authors (e.g. Allen 1976; Elovitz *et al.* 1976; Luce *et al.* 1983) have suggested that prosodic differences between synthetic and natural speech are the primary, unaddressed factor leading to difficulties in the comprehension of fluent synthetic speech. The relation between phrase-level prosody and its sources, however, is so poorly understood that we have no good sense of the degree to which different levels of explanation--syntactic, semantic, or pragmatic--are applicable. We currently have reasonable tools for automatic syntactic analysis of a text, but there is nothing equivalently well-developed for semantic or pragmatic textual analysis. Thus an obvious goal is to explore the extent to which phrase-level prosody can be explained by the syntax tree and develop a detailed description of that relation. A further goal is to convert the resulting insights about this relation into a system that can work with a speech synthesizer. This allows us to test our description more adequately and perhaps also produce something that will further text-to-speech technology.

## SYNTACTIC STRUCTURE AND PROSODIC PHRASING

Certain relations between syntax and prosody, especially at the word level, are well-known. For example, the syntactic category of a word may affect its phonetic realization, as in the verb/adjective distinction of *separate, approximate*, and the verb/noun distinction of *house, wind, lives*. Likewise, syntactic category affects word stress, so that verbs such as *progress, insert, object*, and *rebel* receive final stress, whereas the corresponding nouns receive penultimate stress.

Beyond the word level, however, there has been little investigation of systematic connections between syntactic structure and prosodic phrasing. The psycholinguistic and acoustic investigations of Cooper and Paccia-Cooper (1980), Umeda (1982) and Gee and Grosjean (1983) and the prosodic theory of Selkirk (1984) are among the more notable studies and represent the two main approaches to syntax/prosody relations. In Cooper and Paccia-Cooper (1980) and Umeda (1982), the connection from syntax to prosodic phrasing is unmediated by any filtering process, i.e., they propose that the details of prosodic phrasing can be determined directly from syntactic structure by associating particular syntactic nodes (or constituent boundaries) with a phonetic value, either pausing, segmental lengthening, or the blocking of the cross-word conditioning of phonological rules. By contrast, Gee and Grosjean (1983) and Selkirk (1984) believe that the syntax-prosody relation is indirect: prosodic phrasing is derived by rules that refer to left-to-right ordering, length (or branching patterns), and, in the case of Selkirk, grammatical function, as well as constituent membership in order to infer a hierarchical prosodic structure. But while their respective positions are quite clear, none of these studies is conclusive. All lack a syntactic framework sufficiently detailed and formalized to allow extensive testing, and most consider only a small number of sentences and sentence types.[3]

To develop our analysis, we first examined prosodic phrasing in the speech of one of us reading prose from various texts, including four instruction manuals. These texts were later augmented by a professional reading of a prose story. The boundaries between prosodic phrases were identified and then classed according to their syntactic context and semantic function.

Our results, which are outlined below, indicate an organization of the prosodic phrases that supports the 'indirect relationship' approach of Gee and Grosjean (1983) and Selkirk (1984). We found that, in our corpus, prosodic phrasing depends on three aspects of structure: the breakdown into syntactic constituents, the grammatical function of a constituent, and constituent length. Let us review each of these factors.

### Syntactic Constituency.

The possible constituents recognized by our parser are Noun Phrase (NP), Verb Phrase (VP), Adjective Phrase (AdjP), Adverb Phrase (AdvP), and Prepositional Phrase (PP). In general, we found that syntactic constituency is particularly important for predicting points at which a prosodic phrase boundary is not produced, i.e., the words within a syntactic constituent cohere. For example, the italicized phrases in (1)-(5) had no perceptible boundaries at the locations indicated by #:

(1) *Left-hand # power unit* is connected ...

(2) This procedure *shows # you* ...

(3) An *extremely # narrow* opening ...

(4) To spread powerload *more # evenly*

(5) ... *next # to* any powered di-group

The single exception to word cohesion within syntactic

[2] Note that without a syntactic analysis that correctly identifies grammatical functions, it is impossible to determine whether the word *mark* is a noun ending the subject phrase or the verb of the predicate phrase. Simple 'surface' parsers, such as that described in Umeda and Teranishi (1974), will still fail to identify the prosodic boundary correctly.

[3] Gee and Grosjean (1983) use a corpus of 14 sentences. Umeda (1982) considers a large corpus but, like Gee and Grosjean, does not distinguish among grammatical functions. Although Selkirk cites many examples in her discussions of phrasal stress and word-level prosody, her description of prosodic phrasing focusses on only a single example.

constituents involved boundaries between the verb and its first or second object when the object in question was lengthy. We discuss this exception below.

**Grammatical Functions.**

Our sample indicated that phrase boundaries are also determined by the grammatical relations among the syntactic constituents, i.e. the argument structure of the sentence. Four grammatical relations concern us:

(a) subject-predicate, as in *The 48-channel module -- has two di-groups.*

(b) head-complement, where the head can be a noun, verb, or adjective and may have one complement, e.g. *has -- two di-groups,* or two complements, e.g. *shows -- you -- how to fly your kite.*

(c) sentence-adjunct, as in *Insert unit into correct shelf location -- per detail instructions.*

(d) head-modifier, where the head can be a noun, verb, adverb, or adjective and the modifier can be one of several things, depending on the head (e.g., for nouns, the modifier can be a relative clause; for verbs, it can be a prepositional phrase; for adjectives and adverbs, the modifier can be a comparative).

We observed a hierarchy among these relations with respect to the strength, or perceptibility, of a prosodic boundary, with the boundary between sentence and adjunct receiving the highest potential boundary strength, followed by the subject-predicate boundary, then the head-complement and head-modifier boundaries. Thus in (6), there is a strong boundary between subject and predicate, whereas in (7), due to the strong boundary between adjunct and core sentence, the subject-predicate boundary diminishes. (Dashes indicate the location of the boundary being discussed.)

(6) The name of the character -- is not pronounced.

(7) When this switch is off -- the name of the character is not pronounced.

**Constituent Length.**

While we may view each boundary as having an intrinsic strength based on constituency and grammatical function, the determination of actual strengths appears to depend on the interaction of the intrinsic strength of a boundary with the strengths of other boundaries in the sentence, as well as the distance between these boundaries. The most salient of the interactions we observed was between the placement of a boundary at the subject-predicate junction and the placement of a boundary following the verb-complement junction. The mediating factor in this interaction was the relative length of the subject with respect to the length of the verb's complements. Thus a sentence such as (8), with both a short subject and a single short object generally is produced without a boundary in either position.

(8) You have completed the task.

But if, as in (9), the subject is long relative to the object, then a break occurs between the subject and predicate. Conversely, if the subject is short relative to the object, then a break will occur between the verb and the object, as in (10). Or, if there are two objects and the first is simple, the break will occur between them, as in (11).

(9) The materials required -- are one kite kit.

(10) How shall we judge -- the goodness of an algorithm?

(11) This procedure shows you -- how to fly your kite.

## AN EXPERIMENTAL PROSODY SYSTEM

Our findings confirmed that syntactic structure plays a major role in determining prosodic structure, but the relationship is indirect--the exact influence of syntactic constituency varies according to the length and grammatical function of each constituent. To refine and test this idea, we implemented an experimental text-to-speech system in which rules apply to a parse tree to infer prosodic structure and then annotate the input string with phrasing information derived from the prosodic structure; this annotated input string is submitted to the Bell Labs text-to-speech programs, which convert it into a speech file. Our system comprises three components: a parser that builds syntactic structure, rules that derive prosody information from the syntactic structure, and the Bell Labs text-to-speech programs. The parser and speech programs are independent components. The prosody rules act as a filter between them, converting the syntactic information generated by the parser into prosodic information that can be supplied to the text-to-speech programs.

**Parsing.**

Our parser is a version of Fidditch (Hindle 1983), a moderate coverage parser based on the deterministic model described in Marcus (1980). To build syntactic structure, Fidditch uses a grammar that requires the representations produced by lexical and syntactic rules to be consistent with the (semantic) predicate-argument structure. The surface syntactic structures generated by the parser represent the argument structure of a phrase or sentence, i.e. the "core" constituents of a sentence (its subject (NP), modality (AUX), and predicate (VP)) and the complements of phrasal heads. The structure is determined, for the most part, by rules that refer to argument information that is specified in the lexicon for the content words (nouns, verbs, adjectives, adverbs), and by rules that insert null terminals such as the "trace" of *wh*-movement. In general, the grammar is consistent with the government and binding framework of Chomsky (1981), as adapted to the needs of a parser.

The input to the parser is a phrase or sentence (punctuation is optional). Its output is a surface structure tree in which the status of a constituent with respect to the predicate-argument structure of the sentence is indicated by the constituent's attachment to higher nodes in the tree. Thus only constituents that belong to the core are attached to the S node, and only complements of a phrasal head can become righthand sisters of the head. Adjuncts and modifiers, whose role depends on semantic and pragmatic information about the discourse domain, have no assigned position within a structure and so are represented as "orphan" nodes in the tree.

For example, Figure 1 shows the parse tree for *Left-hand power unit on each shelf in 48-channel module can power only the echo cancelers that are in that shelf.* [4] The structure in Figure 1 contains a single core sentence -- *unit can power the cancelers* -- with left-branching modifiers -- *left-hand, power,* and *echo.* The sentence also contains three modifiers -- the PPs *on each shelf* and *in 48-channel module,* and the adverb *only* -- which are unattached constituents. This is the significance of the unlabeled node dominating each of these constituents. The PPs are not attached because *unit* is not lexically marked to take a PP headed by *on* or *in* as a complement, and *shelf* is not lexically marked to take a PP complement headed by *in*. Nor is any constituent lexically marked to accept *only* as an argument.

Figure 1 also contains a relative clause, *that are in that shelf*. In the relative clause, T is a null terminal that stands for the trace of the relativized subject NP; the * in tense stands for a null Aux element. Because nouns do not select relative clauses as arguments (any noun can be relativized), the parser does not identify the relations of the modifier constituent to the elements of the core sentence. Hence the relative clause is not attached to any other syntactic node in the tree.

**Text-to-speech Synthesis.**

The programs that make up the speech component are described in Liberman and Buchsbaum (personal communication). These programs take English text as input and produce digitized speech output. By annotating the input text to this system, many aspects of its operation can be overridden or modified: e.g. the location of major and minor phrase boundaries, the stress given to words, the transcription of words and the boundaries between them, the timing of segments, and details of the pitch contour. As we will show, with our prosody system we are able to produce strings in which four boundary levels are identified and perceptually distinguished, using the current text-to-speech system annotations.

**Prosodic Phrasing.**

The prosody rules use information about constituent structure, grammatical role, and length to map a surface structure such as that in Figure 1 onto a prosody tree such as that in Figure 2. The prosody tree identifies the location of phrase boundaries (signified by the $\Phi$ nodes) and the relative strength of each boundary (signified by a number in the $\Phi$ node). It is this information that is used to annotate the input text with escape sequences that provide the text-to-speech system with instructions about prosodic phrasing.

In formulating our rules for building the prosodic structure, we began with the idea of simply implementing the model of Gee and Grosjean (1983). This model, initially proposed to predict a form of psychological data describing subjective sentence structure known as *performance structure,* determines prosodic boundaries from a syntactic tree, but assumes rather than explicitly presents a syntactic component. We were initially attracted to the Gee and Grosjean model because of its emphasis on relative boundary weighting, i.e., on the determination of the strength of a given boundary with respect to the other boundaries in the sentence. We found that in the data we had collected, this weighting played an important role. In fact, we incorporated directly into our system one method of doing this weighting, namely Gee and Grosjean's rule to determine the strengths of the prosodic phrase boundaries around a verb using relative length (as measured by terminal node count).

As we extended Gee and Grosjean's model to create an algorithm adequate for use in a general purpose system, our algorithm diverged from its starting point, reflecting our attempts to correct weaknesses and lacunae that we encountered in the Gee and Grosjean model. That we encountered these problems is not surprising given the difference between our goals and those of Gee and Grosjean.

The most important difference between the Gee and Grosjean model and our current algorithm involves the factors determining boundary weight. Gee and Grosjean assume that this weighting is dependent only on the number of syntactic nodes, their left-to-right ordering and, in the case of the verb phrase, on constituent length. In contrast, our data, in agreement with Selkirk's (1984) theoretical analysis, indicated that boundary strength is dependent on the grammatical functions that the constituents in a given sentence play. In particular, we observed a hierarchy among these functions with respect to boundary strength, as discussed below.[5]

In addition to incorporating grammatical function information into our system, we fleshed out the model of Gee and Grosjean to deal with syntactic structures that they do not explicitly consider. In particular, Gee and Grosjean's strictly left-to-right building of the

[5] As an example of the effect that grammatical functions have on prosodic phrasing, consider the sentence *Finally the strange young man left*. We view this sentence as consisting of two grammatical relations: subject-predicate and adjunct-sentence. In our hierarchy of grammatical relations, the boundary between the adjunct and the sentence is more salient than the boundary between the subject and the predicate. The system reflects this by assigning a stronger boundary following *Finally* than following *man*.

If we exclude any effects of grammatical functions and assume a simple left-to-right attachment of the three constituents *Finally, the strange young man* and *left*, to the prosody tree, we would assign a stronger boundary following *man* than following *Finally*. It is not clear that Gee and Grosjean make this left-to-right assumption in such examples. They view adverbial phrases like *Finally* as dominated by the complementizer node in the syntax tree, and it is difficult to determine whether they integrate the material in the complementizer with the material in the core sentence as they are analyzing the material in the core sentence or after that analysis is completed. If they integrate the complementizer with the core sentence, then they assume that *Finally* bundles with the sentence in a left-to-right manner and predict, incorrectly, that the stronger boundary occurs after *man*. If they complete the prosodic analysis of the core sentence before bundling the sentence with the complementizer, then they incorrectly predict that there is a strong boundary after *wh-* phrases in the complementizer. In particular, they would incorrectly predict that in sentences like *At the outset what problems did you expect* the most perceptible boundary would be after *problems*.

Furthermore, assuming that an adjunct in sentence-initial position is dominated by the complementizer node and in sentence-final position by S-bar creates an inconsistent description, which hampers the value of the model as an experimental tool.

prosodic tree left certain questions open. For example, their model does not deal with sentences embedded in the middle of a main sentence (as in *The notion [that he would refrain from such an act] was incorrect.)* We incorporate embedded sentences into the prosodic tree in a cyclic manner to insure that the material in the embedded sentence is processed before that in the main sentence.[6] In addition, Gee and Grosjean leave open the treatment of the multiple rightward embedding of non-sentential constituents, e.g., the NP embedding in *The destruction of the good name of his father.* Our approach is to handle these cases recursively, from the most deeply embedded phrase up, in order to preserve the prosodic cohesion of the entire NP.

Our adjunction rules are derived for the most part from Selkirk's account. We have also made use of the idea, which Gee and Grosjean (1983) take largely from the work of Selkirk, that certain syntactic heads mark off phonological phrase boundaries, and provide the basic prosodic constituents for higher level analysis.

Our prosody rules run in four independent stages. Each stage builds on the previous stage, so that the rules can refer to both syntactic and prosodic structure as they build successively higher levels of prosodic structure.

(i) *Adjunction Rules* combine orthographically distinct words into phonological constituents with no internal word boundary. They join a word to its left or right neighbor depending on (a) the category of the word, and (b) its structural relation to other words. In general, adjoinable words are the function words-- articles, complementizers, auxiliary verbs, conjunctions, prepositions and pronouns (except for the "strong" possessives, *mine, hers, theirs, yours, ours,* which are treated as regular NP's).

Adjunction occurs six times for the sentence in Figure 2 to create six multiple word groups, all right-adjoining: *on each, in 48-channel, can power, the echo, that are* and *in that*. These groups of adjoined words appear as terminals in the prosody tree in Figure 2. In subsequent processing the boundaries between the words in these groups are marked so that the text-to-speech system does not produce the prosodic indications of a word boundary. In addition, these groups are treated as single words in further analyses.

(ii) *Φ-phrasing Rules* construct phonological (or Φ) phrases, which are the building blocks of the prosody tree. These rules identify groups of words that cohere strongly in speech and thus should not be separated by phrase boundaries. In the present implementation, each Φ phrase is constructed by a left-to-right process that collects the words formed by adjunction until it reaches a noun or verb. At this point, a Φ phrase is created that consists of the collected words plus the noun or verb, which acts as head of the phrase. For example, *in that shelf*, in Figure 2, is a single Φ phrase consisting of two words.

In Figure 2, the Φ nodes marked with a syntactic category are the minimal phonological constituents with respect to later rules that build the prosodic phrases; these Φ phrases have an internal structure, but the structure plays no role in further processing. Note that neither adjectives nor adverbs are allowed to be the head of a Φ phrase, so that *three additional open slots* is a single Φ phrase consisting of four words. Examples such as *Someone tall walked into the room,* however, suggest that our treatment of these categories is not detailed enough and that, in future versions of the system, some adjectives and adverbs should act as Φ heads.

(iii) *Prosody-phrasing rules* use information about Φ phrases and syntactic structure to create a new organization of the sentence and to assign strength values to the boundaries between successive Φ phrases. The process of building the prosody tree starts with the sentence node (S or Sbar) that is most deeply embedded in the utterance, transforming it into a prosody subtree. This process continues through successively higher levels of sentence nodes until all top-level sentences have been transformed into prosody subtrees. All the processing of each successive sentence is done before the relation of the sentences to each other is considered.[7]

Within a sentence, the Φ phrases are processed from left to right. This stage of the analysis uses a window that allows access to three adjacent nodes. Pattern-action rules, which are described below, apply to the nodes in the window and build prosody subtrees that replace the syntax nodes. These subtrees are headed by a Φ node containing a number that represents node count; the number is determined by counting the number of nodes contained in the prosody subtree, plus 1 for the Φ node that heads the subtree.[8] In general, the prosody phrase rules do three things:

(a) Balance prosodic phrases by referring to constituent length. This rule only applies for building the prosody subtree that contains the verb. If the node count for subject plus verb is less than the node count of the verb's complement, then subject and verb are grouped together in a prosodic subtree; this gives the phrasing in *The characters on the right -- mark the salient features*. Otherwise, the verb is grouped with its complement in a prosodic subtree; an example of this grouping is the subtree for *can power only the echo cancelers* in Figure 2.

(b) Combine the Φ phrase daughters of the major constituents, excluding VP, into a prosodic subtree. At present, this rule only applies to NP and PP since adjectives and adverbs are currently not treated as Φ heads. For example, *the name of the character,* which forms two Φ phrases under NP *(the name* and *of the character),* become a single prosody phrase that replaces the NP.

[6] Having taken this strong approach, we now understand the limited exceptions to this mechanism, which we discuss below.

[7] We have found at least one class of phrases for which this order of processing appears inappropriate. In these, the head of the top-level phrase is epistemic -- e.g., *believe, know, belief, knowledge* -- and its complement is a sentence. In most cases, the current processing order for embedded sentences will produce a break between a head and a following embedded sentence. For this class of sentences, however, the break does not seem to be appropriate. While it would be straightforward to handle this as an exception, we are currently examining whether there is a more principled way to describe what must be done in these cases.

[8] Only the top-level Φ nodes, those which contain the head of the syntactic phrase, are counted in computing the node count. Thus, for example, in Figure 2, the sub-phrasal branching of *Left-hand* and *power unit* does not contribute to the node count.

(c) Bundle together prosodic constituents (Φ phrases) from left to right if no other rules apply. This rule integrates the constituents left unattached by the parser into the prosodic structure. It accounts for the prosodic structure of *left-hand power unit on each shelf in 48-channel module* in figure 2, which is formed by first bundling *left-hand power unit* with *on each shelf*, into Φ-3, and then bundling the result with *in 48-channel module* into Φ-5. The final application of bundling replaces the Sigma node with the top level prosody node, which is Φ-13 in Figure 2.

(iv) *Prosody conversion rules* map the boundary strength indices onto three phonological mechanisms. Boundary indices in the low range, e.g. the Φ-3 nodes in Figure 2, are realized as a phrase accent (Pierrehumbert 1980). Mid-range indices such as Φ-5 and Φ-9 in Figure 2 are realized as changes in pitch range. High indices are realized with modulations in both pitch range and duration. Thus the hierarchical organization of a structure such as that in Figure 2 can be reflected directly in the synthesized speech.

## PHENOMENA NOT TREATED

Several phenomena have been omitted from this preliminary version of the system. Some of these omissions arise from the fact that we concentrated on sentence analysis rather than discourse analysis. Others involve phenomena that characterize spoken English, and thus did not occur in our original corpus of technical repair manuals.

Contrastive stress is an example of prosodic phrasing based on discourse analysis. In our system's analysis, the phrase *from India* does not receive contrastive stress in (12).

(12) Passengers from several countries entered
the terminal.
Finally a man from India walked in.

In designing the current system, we have concentrated on the level of sentence analysis. Handling the contrasts involved in data like (12) necessitates an additional level of discourse analysis.

In addition, the system never explicitly manipulates segment durations or overall speech rate. For example, we have yet to explore whether lengthening of the segment before a mid-range boundary value is appropriate, or whether increasing the duration of constituents of the core sentence might enhance the natural sound of the system.

## RESULTS AND FUTURE RESEARCH

To date, our system has been tested systematically on a set of 39 sentences, and its performance has been observed less formally on a set of approximately 300 sentences.[9] The test corpus covers a repair manual for telephone switching systems and an introductory description of the Prose 2000 text-to-speech system. We added sentences cited in Umeda (1982) and sentences that we composed in order to extend the range of syntactic constructions represented in the test. In general, we have observed a significant improvement of prosodic quality in those test sentences where the parser and the prosodic component have returned acceptable results.

We have observed problems, however, especially in the formal test corpus, much of which we chose for its potential difficulty. Of the 39 test sentences, 38 parsed correctly. Of these, the prosodic component returned 26 sentences with a complete set of acceptable prosody markings. In terms of actual markings, the system marked 393 prosodic events, of which 21 markings were unacceptable. We can attribute errors in those sentences with unacceptable prosodic markings to three distinct problems discussed below.

### Complement Sentences.

Five of the errors that arose from the prosody system's treatment of the test corpus result from the fact that the system sets off all subordinate sentences, including complement sentences, from the main sentence. Informal testing of the productions of four informants on the relevant data indicated that this approach works correctly for complement sentences such as (13)-(16). (Complement sentences are italicized):

(13) Health services cautioned Western residents
-- *that they should ask where their watermelons come from before buying.*

(14) We have to satisfy people -- *that the crisis is past.*

(15) The vendors explained -- *that this is the result of illness among 281 people who ate pesticide-tainted watermelons.*

(16) Watermelon growers wonder -- *whether this will continue throughout the rest of the season.*

However, the informant test consistently indicated that the complement sentences in (17)-(19) are not set off by a comparable boundary:

(17) They believe *California sales are still off 75 percent.*

(18) They think *the Southeast is shipping half its normal load.*

(19) Growers and retailers claimed *the incident hurt sales across the USA.*

Cases like (17)-(19), in which no break is perceived between the verb and its complement sentence, form a syntactically distinct class in Fidditch. This class is characterized by the fact that the verbal head in each case is one that does not require that its complement sentence begin with a complementizer (either *that, for*, or a *wh-* word). The class includes epistemic verbs, like those in (17)-(19), as well as a wide range of verbs that take either tensed sentences, or various types of non-tensed sentences as complements.[10] The examples (20)-(26) demonstrate the range of this class (complement sentences are italicized):

[9] The 39 sentences are listed in the appendix to this paper.

[10] Fidditch, in following the outlines of Chomsky's (1981) Government and Binding theory, assumes that propositions, i.e., those elements that contain both a predicate and a perhaps null subject, are syntactically represented as sentences, regardless of tensing.

(20) We had *the ship's forces make temporary repairs.*

(21) We saw *the crew repairing the unit.*

(22) He wants *the units repaired by the ship's force.*

(23) The construction of the unit makes *detailed investigation impractical.*

(24) Try *to give the names of the characters in advance.*

(25) They will help *finish the job.*

(26) The new equipment will facilitate *making repairs.*

**Sentence-Final Constituents.**

Fifteen of the errors that arose from the system's treatment of the test corpus result from a high boundary value that sets final constituents off from the main sentence. The high value is due to the system's purely left-to-right attachment of syntactically unattached constituents (see rule iii.d above). The high boundary value is acceptable in sentences like (27)-(29). (The relevant final constituents in these examples are italicized).

(27) In these instances it may be desirable to use phoneme characters instead of text characters to represent a word -- *each time it appears in the input text.*

(28) Phonemic characters can also be used to handle syntactic data such as boundaries -- *which can improve speech quality.*

(29) We were unable to finish the work -- *due to equipment failure.*

However, the high boundary value sets the final constituent off unnaturally from the main sentence in data such as (30)-(32).

(30) The method by which you convert a word into phonemes is provided -- *in Chapter 7.*

(31) The experimenters instructed the informant to speak -- *naturally.*

(32) We discussed the techniques -- *we had implemented.*

In many cases it appears that the grammatical relation of the final constituent to the rest of the sentence determines the boundary value that sets off this constituent. In particular, sentence adjuncts, which bear no relation to any single item in a sentence, are set off by a minor phrase boundary, whereas final constituents that modify a particular item are less perceptibly set off. This is the distinction between the final constituents in (27)-(29), which are adjuncts, and those in (30)-(32), which are modifiers. However, while the distinction between the grammatical relations of the core sentence (complement and subject) and those of the periphery (adjunct and modifier) is fairly straightforward, and handled directly by the mechanisms of the Fidditch parser, the distinctions between the peripheral elements of adjunct and modifier are complex and require the addition of costly mechanisms.

The cost of adding adjunct/modifier distinctions is illustrated by the ambiguity that arises when both adjunct and modifier readings are possible. For example, on one reading of (31), *naturally* modifies the verb *speak*; i.e., the informants were to speak in a natural manner. On the other reading, *naturally* is an adjunct equivalent to *of course.* (To see this meaning more clearly, consider the rearrangement of this sentence with the adjunct at the beginning: *Naturally, they instructed the informants to speak.*) The context of speech analysis prefers the former reading. However, the net benefit of adding sophisticated contextual analysis to our system, if attainable, is, at best, unclear. The same may be said of adding selectional restrictions, or detailed information on logical form.

In contrast, a finer treatment of local syntactic constraints on boundary values preceding final constituents is within reach. From the data we have examined, it appears that the character of the prosodic event before the final constituent can be locally determined to a great extent. For the most part, this determination depends on the category type of the final constituent and on the contents of the leading edge of the constituent. For example, interjections (*however, moreover, therefore, alas, thus, of course,* etc.) and sentence adverbs (*apparently, generally, luckily* etc.) are uniformly set off by a high boundary value and should remain so. In contrast, the boundary value of final prepositional phrases, particularly those with a monosyllabic preposition (*in, on, at, to, with, for*) as the left edge of the phrase, should be reduced.[11] We are currently engaged in categorizing the constituent types and left-edge items that characterize final constituents with respect to the prosodic event that precedes them.

Alternatively, we are considering the play-it-safe approach of reducing the high boundary values that set off final constituents to mid-boundary values. Currently these values are converted to a downstepping feature. This approach may also be useful in conjunction with our local determination approach for those constituents whose status is either undecidable or ambiguous under the latter approach.[12]

[11] In this view, expressions such as *in principle, in general, in particular, in consideration of*, etc. must be treated like interjections.

[12] Reducing the final boundary value leaves ambiguities unresolved. For sentences such as (i) and (ii), below, we believe this lack of resolution is appropriate:

(i) John saw a girl in the park with a telescope.
[The telescope is with John or the girl, or it's in the park.]

(ii) I need a woman to fix the sink.
[I need a woman so that I can fix the sink.
I need a woman who can fix the sink.]

Our view, following Marcus and Hindle (p.c.) is that in normal, spoken English, such ambiguities are not processed unless the speaker or listener is directly questioned regarding the ambiguity. Likewise the prosodic events that might disambiguate are inappropriate unless such questioning occurs.

Other cases are less clear. For example, it is difficult to imagine that, in (28) the difference between the reading of the *which* clause as a sentence adjunct and as a noun phrase modifier on *boundaries* is not processed. We would hope that in such cases some local distinction, such as the presence or absence of the comma in (28), obtains.

**Sentence-Initial Constituents.**

When a sentence contains both sentence-initial and sentence-final adjuncts, the sentence-initial adjuncts will be less prominently set off than the sentence-final adjuncts due to the left-to-right attachment of adjuncts to the prosodic tree (see rule iii.b above). In data like (33), however, a more appropriate rendering would have the boundary after the adjunct *On a clear day* be strong relative to the boundary before the adjunct *as it rises over the mountains*.

(33) On a clear day you can see the sun as it rises over the mountains.

While it would be trivial to increase the value of the pertinent boundary, we are as yet unsure what the critical features are which require a more perceptible boundary. For example, while a higher boundary value after the prepositional phrase in (34) might be acceptable, it is not clear that it is necessary:

(34) In the morning John left.

Given the stylistically distinct nature of this data, we have not yet considered this question in detail.

**Summary.**

While we have systematically tested our system so far on a small set of examples, the number of prosodic events involved in those examples, 393, is high, due to the length of the sentences tested. We find the 5 percent error rate, representing 21 prosodic events, encouraging at this stage in the development of the system. In addition, we have delimited the problem areas of an approach that relies solely on information available in the syntax tree. Our initial investigation of these problems indicates that at least part of the necessary information about phrase-level prosody is conveyed in the lexicon per se. Additionally, due to the left-corner orientation of the Fidditch parser, which exists independently to optimize search strategies, the necessary lexical information is made easily available.

## CONCLUSIONS

We have described an on-line experimental system that uses prosody rules to infer prosodic phrasing from constituent structure, grammatical functions, and length considerations. The system contains three modules: a deterministic parser, a set of prosodic phrasing rules, and an algorithm to convert the output of the prosodic phrasing rules into signals for the Bell Labs text-to-speech system.

In developing the experiment, our intention was to build a working system that would allow us to test various hypotheses about the connections between syntax and prosodic phrasing in human speech and to upgrade the prosody of existing synthetic speech. The modularity of our system enables us to alter each module independently in order to test different hypotheses. For example, the parser can be altered to reflect the difference between verbs that require a complementizer before a sentential complement and those that do not.[13] This alteration is independent of the workings of the prosody system or the prosody conversion rules.

The existence of this prosody system makes the problem areas in the syntax-prosody relation more tractable by allowing online testing of a large body of data. For example, the prosodically different character of the two classes of complement sentences discussed above became apparent after several examples from each class were run through the system. We therefore feel we have built a tool that will aid in designing better approximations of sentence prosody as it relates to syntactic structure.

[13] Fidditch represents this as a difference in the level of the complement sentence. Verbs that require a complementizer take an S-bar complement, while verbs that do not require a complementizer take an S complement with an optional *that* preceding.

## REFERENCES

Allen, J. 1976. Synthesis of speech from unrestricted text. *Proceedings of the IEEE, 4*, 433-442.

Chomsky, N. 1971. *Lectures on government and binding*. Dordrecht: Foris Publications.

Cooper, W. and J. Paccia-Cooper. 1980. *Syntax and speech*. Cambridge, MA: Harvard University Press.

Elovitz, H., R. Johnson, A. McHugh, and J. E. Shore. 1976. Letter-to-sound rules for automatic translation of English text to phonetics. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 6*, 446-459.

Gee, J. P. and F. Grosjean. 1983. Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology, 15*, 411-458.

Hindle, D. 1983. User manual for Fidditch, a deterministic parser. NRL Technical Memorandum #7590-142.

Luce, P.A., Feustel, T.C., and Pisoni, D.B. 1983. Capacity demands in short-term memory for synthetic and natural speech. *Human Factors, 25*, 17-32.

Marcus, M. 1980. *A theory of syntactic recognition for natural language*. Cambridge, MA: MIT Press.

Pierrehumbert, J. B. 1080. The phonetics and phonology of English intonation. Ph.D. Dissertation, MIT.

Selkirk, E. O. 1984. *Phonology and syntax: the relation between sound and structure*. Cambridge, MA: MIT Press.

Umeda, N. 1982. Boundary: perceptual and acoustic properties and syntactic and statistical determinants. *Speech and Language, 7*, 333-371.

Umeda, N. and R. Teranishi. The parsing program for automatic text-to-speech synthesis developed at the Electrotechnical Laboratory in 1968. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 23*, 183-188.

## APPENDIX: TEST SENTENCES

1. THE NAME OF THE CHARACTER IS NOT PRONOUNCED.

2. LEFT-HAND POWER UNIT ON EACH SHELF IN FORTY-EIGHT
CHANNEL MODULE POWERS ONLY ECHO CANCELLERS IN THAT
SHELF.

3. THE CONNECTION MUST BE DETERMINED FOR THE LEFT-HAND POWER UNITS ON EACH SHELF.

4. THE CONNECTION MUST BE DETERMINED FOR THE LEFT-HAND POWER UNITS WHICH ARE ON EACH SHELF.

5. THE METHOD BY WHICH ONE CONVERTS A WORD INTO PHONEMES IS PROVIDED IN CHAPTER 7.[14]

6. WE DISCUSSED THE TECHNIQUES WE HAD IMPLEMENTED.

7. THE TECHNIQUES WE HAD IMPLEMENTED WERE TESTED ON A LARGER MACHINE.

8. THE MAN WHOM WE SAW YESTERDAY LIVES FAR AWAY FROM HERE.

9. THEY TOLD HIM TO WALK SLOWLY.

10. THE DESTRUCTION OF THE GOOD NAME OF HIS FATHER BOTHERED HIM.

11. LATELY HE HAD HAS CONTROL OVER THE SITUATION.

12. I NEED A WOMAN TO FIX THE SINK.

13. JOHN MET A WOMAN HE THOUGHT HE LIKED.

14. THE WOMAN I SAW CAME FROM HERE.

15. IN THESE INSTANCES IT MAY BE DESIRABLE TO USE PHONEME CHARACTERS INSTEADOF TEXT CHARACTERS TO REPRESENT A WORD EACH TIME IT APPEARS ON THE INPUT TEXT.

16. PHONEME CHARACTERS GIVE MORE CONTROL OVER THE PARTICULAR SOUNDS THAT ARE GENERATED.

17. THE MATERIALS REQUIRED ARE ONE KITE KIT.

18. PHONEMIC CHARACTERS CAN ALSO BE USED TO HANDLE SYNTACTIC DATA SUCH AS THE BOUNDARIES WHICH CAN IMPROVE SPEECH QUALITY.

19. IT MAY BE DESIRABLE TO GIVE JOHN A HAND.

20. AFTER THESE QUESTIONS, A DETAILED DESCRIPTION OF THE USE OF PHONEMES WILL BE
PROVIDED IN CHAPTER 7.

21. THE ENGLISH THAT IS SPOKEN IN AMERICA AT THE PRESENT DAY HAS RETAINED A GOOD MANY CHARACTERISTICS OF EARLIER BRITISH ENGLISH THAT DO NOT SURVIVE IN BRITISH ENGLISH TODAY.

22. PHONEMIC CHARACTERS CAN ALSO BE USED TO HANDLE SYNTACTIC DATA SUCH AS THE LOCATION OF THE ENDS OF PHRASES WHICH CAN IMPROVE SPEECH QUALITY.

23. THE STUDENTS CONSIDERED THE ASSUMPTION THAT A BREAK MIGHT OCCUR.

24. FINALLY YOU MUST ASSUME THAT YOUR CIGARETTES WILL BOTHER THE PASSENGERS.

25. TRY TO GIVE THE NAMES OF THE CHARACTERS TO JOHN.

26. I PREFER FOR HIM TO GIVE THE NAMES OF THE CHARACTERS TO JOHN.

27. I BELIEVE THOSE PEOPLE TO BE INTELLIGENT.

28. I PROMISED HIM THAT HE COULD COME.

29. THEY GAVE THE BOY A BOOK.

30. THEY GAVE HIM A BOOK.

31. THE 48-CHANNEL MODULE CAN HAVE ONLY TWO DI-GROUPS BUT CAN HAVE UP TO FOUR POWER UNITS IF BOTH DI-GROUPS ARE EQUIPPED WITH ECHO CANCELERS.

32. I TOLD HIM YESTERDAY TO CLEAN HIS ROOM.

33. MOVE THE POWER OPTION JUMPER PLUG SO THAT IT IS ADJACENT TO DI-GROUP ONE ON PRINTED WIRING BOARD.

34. I WANT A LOT MORE COOKIES.

35. THE MINUS-SIGN PRONUNCIATION SWITCH IS IN THE MIDDLE.

36. HE ASKED THE CHILDREN TO FINISH THE JOB.

37. HE ARGUED THAT IT WAS IMPOSSIBLE.

38. IS A MAN AT THE DOOR.

39. A DETAILED DESCRIPTION OF THE USE OF PHONEMES IS PROVIDED IN CHAPTER 7.

[14] Fidditch failed here on the relative clause with a PP left edge.

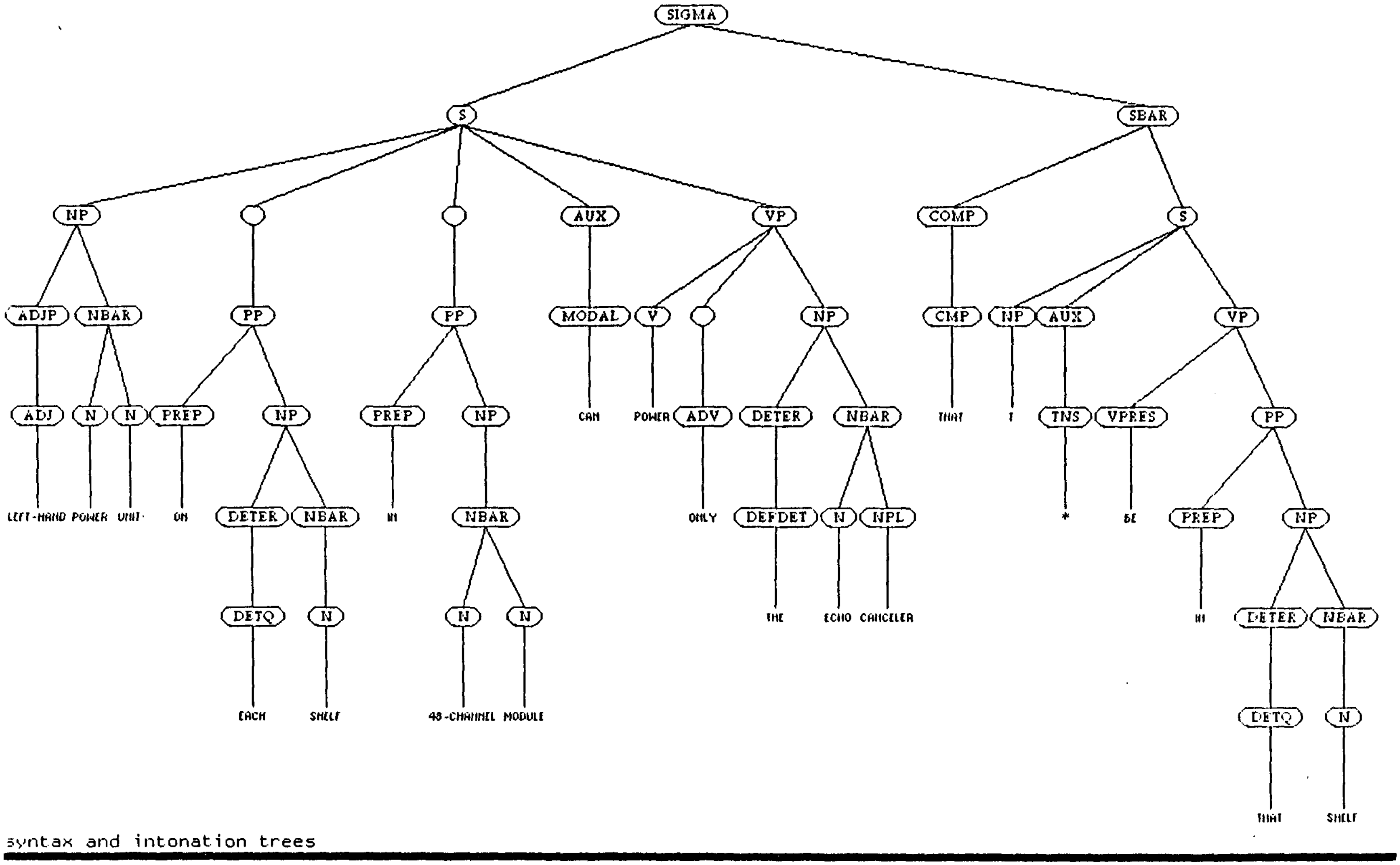

Figure 1. A sample syntactic tree produced by the Fidditch parser

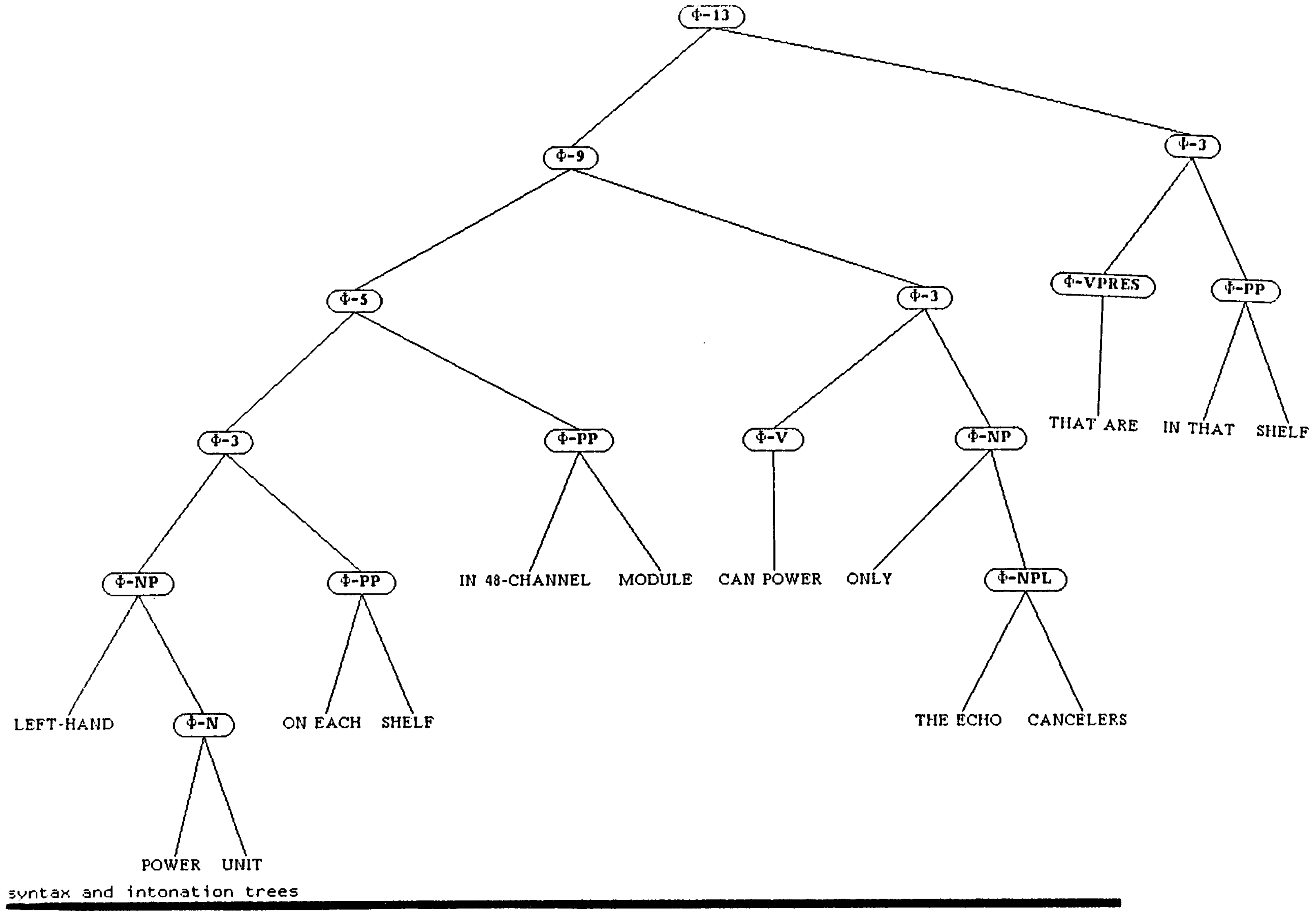

Figure 2. Prosody tree for the sentence shown in Figure 1