Specifying Intonation from Context for Speech Synthesis

Scott Prevost and Mark Steedman

Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104-6389, USA

To appear in *Speech Communication*.

Accepted April 26, 1994. Revised May 16, 1994.

Abstract

This paper presents a theory and a computational implementation for generating prosodically appropriate synthetic speech in response to database queries. Proper distinctions of contrast and emphasis are expressed in an intonation contour that is synthesized by rule under the control of a grammar, a discourse model, and a knowledge base. The theory is based on Combinatory Categorial Grammar, a formalism which easily integrates the notions of syntactic constituency, semantics, prosodic phrasing and information structure. Results from our current implementation demonstrate the system's ability to generate a variety of intonational possibilities for a given sentence depending on the discourse context.

Cet article vise a présenter une théorie et une réalisation informatique de la génération de la parôle synthèthique accompagnée d'intonation approprié, en réponse à des enquêtes apropos d'une base de données. Les distinctions appropriées de contraste et d'emphase sont marquées par l'intonation automatiquement synthesisé sous la gouvernance de la grammaire, un modèle du discours, et d'une réprésentation de la domaine cognitive. La théorie se fonde sur la Grammaire Categoriale Combinateurique, formalisme qui se prête à l'intégration directe de la syntaxe, la semantique, la structure prosodique, et le statut discursale de l'information. Les résultâts de nos expériences demontrent les capacités du système de générer plusieurs intonations différemment modulés selon le contexte du discours pour une phrase donnée.

Dieser Artikel präsentiert ein Modell zur Generierung prosodisch adäquater, synthetisierter Antworten auf Datenbankanfragen. Dabei werden die passenden Unterscheidungen zwischen Kontrast und Betonung in Bezug auf ein Diskursmodell und eine Wissensbasis vermittelt. Das Modell für die Generierung der Betonungen basiert auf Combinatory Categorial Grammar (Kombinatoriale Kategorial-Grammatiken), ein Formalismus, der die Verwendung von syntaktischen Konstituenten, prosodischer Phrasierung und Informationsstrukturen integriert. Resultate unserer Implementierung demonstrieren die Fähigkeit des Systems, eine breite Auswahl von Intonationsmöglichkeiten für einen gegebenen Satz in Abhängigkeit vom Diskurs-Kontext zu generieren.

1. Introduction

One source of unnaturalness in the output of many text-to-speech systems stems from the involvement of algorithmically generated default intonation contours, applied under minimal control from syntax and semantics. The intelligibility of the speech produced by these systems is a tribute to both the resilience of human language understanding and the ingenuity of the algorithms' inventors. It has often been noted, however, that the results frequently sound unnatural when taken in context, and may on occasion mislead the hearer.

It is for this reason that a number of discourse-model-based speech generation systems have been proposed, in which intonation contour is determined from context or the model. Work in this area includes an early study by Young and Fallside (1979), and studies by Terken (1984), Houghton (1986), Isard and Pearson (1988), Davis and Hirschberg (1988), Hirschberg (1990), and Zacharski *et al.* (1993), although the representations of information structure and its relation to syntax employed by these authors are rather different from those proposed here.

Consider the exchange shown in (1), which is an artificial example modeled on the domain of TraumAID, a medical expert system in the context of which we are investigating spoken language output.[1] This particular example is slightly unrealistic in that TraumAID acts purely as a critiquing device and does not possess such an interactive query system for its knowledge base; nor is it likely that such a query system would be of practical use in the trauma surgery. However, such examples are useful for present purposes since they force unambiguously contrastive contexts that motivate intonational focus and contrastive stress.

In example (1), capitals indicate stress and brackets informally indicate the intonational phrasing. The intonation contour is indicated more formally using a version of Pierrehumbert's notation (cf. Pierrehumbert 1980, Pierrehumbert and Hirschberg 1986).[2] In this notation, L+H* and H* are different high pitch accents. LH% (and its relative LH$) and L (and its relatives LL% and LL$) are rising and low boundaries respectively. The difference between members of sets like L, LL% and LL$ boundaries embodies Pierrehumbert and Beckman's (1986) distinction between intermediate phrase boundaries, intonational phrase boundaries, and utterance boundaries.[3] We shall skate over the former distinction here, noting only that utterance boundaries are distinguished from the others by a greater degree of lengthening and pausing.

The other annotations in (1) indicate that the intonational tunes L+H* LH% (or the related L+H* LH$) and H* L (or the related H* LL$) convey two distinct kinds of discourse information. First, both H* and L+H* pitch accents mark the word that

[1] The examples used throughout the paper are based on a the domain of TraumAID, which is currently under development at the University of Pennsylvania (Webber *et al.* 1992). The lay reader may find it useful to know that a *thoracostomy* is the insertion of a tube into the chest, and *pneumothorax* refers to the presence of air or gas in the pleural cavity.

[2] A brief summary of Pierrehumbert's notation can be found in Steedman (1991a).

[3] Since utterance boundaries always coincide with an intonational phrase boundary, this distinction is often left implicit in the literature, both being written with % boundaries. For purposes of synthesis, however, the distinction is important.

(1) Q: I know that a LEFT thoracostomy is needed for the SIMPLE pneumothorax,
(But what condition) (is a RIGHT thoracostomy needed for?)
L+H* LH% H* LL$

A:

(A	RIGHT	thoracostomy is needed for)	(the	PERSISTENT	pneumothorax.)
	L+H*	LH%		H*	LL$
ground	*focus*	*ground*	*ground*	*focus*	*ground*
	Theme			*Rheme*	

they occur on (or rather, some element of its interpretation) for "focus", which in the context of such simple queries as example (1) usually implies contrast of some kind. Second, the tunes as a whole mark the constituent that bears them (or rather, its interpretation) as having a particular function in the discourse. We have argued at length elsewhere that, at least in this same restricted class of dialogues, the function of the L+H* LH% and L+H* LH$ tunes is to mark the "theme" – that is, "what the participants have agreed to talk about". The H* L(L%/$) tune marks the "rheme" – that is, "what the speaker has to say" about the theme. This phenomenon is a strong one: the same intonation contour sounds quite anomalous in the context of a question that does not establish an appropriate theme, such as "which procedure is needed for the persistent PNEUMOTHORAX?". The advantage for present purposes of Pierrehumbert's system, like other autosegmental approaches, is that the entire tune can be defined independently of the particular string that it occurs with, by interpolation of pitch contour between the pitch-accent(s) and the boundary for those parts bearing no tonal annotation. It will be notationally convenient to speak of the latter as bearing "null tone". (Of course such elements may bear pitch and even secondary accent, and the specification of such details of the interpolated contour is by no means a trivial matter. However, we do not believe that anything hangs crucially on our use of this theory of intonation, rather than some other.)

2. Combinatory Prosody

From the example in the preceding section, it is clear that intonational units corresponding to theme or rheme need not always correspond to a traditional syntactic constituent. Since many problems in the analysis and synthesis of spoken language result from this apparent independence of syntactic and intonational phrase boundaries, we have chosen to base our system on Combinatory Categorial Grammar (CCG), a formalism that generalizes the notion of surface constituency, allowing multiple derivations and constituent structures for sentences, including ones in which the subject and verb of a transitive sentence can exist as a constituent, complete with an interpretation.

CCG (Steedman 1987, 1990a, 1990b, 1991a) is an extension of Categorial Grammar (CG). Elements like verbs are associated with a syntactic "category" which identifies them as *functions*, and specifies the type and directionality of their arguments and the type of their result. We use a notation in which a rightward-combining functor over a domain β into a range α is written α/β, while the corresponding leftward-combining functor is written $\alpha\backslash\beta$. α and β may themselves be function categories. For example, a transitive verb is a function from (object) NPs into predicates – that is, into functions from (subject) NPs into S, written as follows:

(2) $(S\backslash NP)/NP$

We also need the following two rules of functional application, where X and Y are variables over categories:

(3) FUNCTIONAL APPLICATION:

$$\begin{array}{lcccl} a. & X/Y & Y & \Rightarrow & X \quad (>) \\ b. & Y & X\backslash Y & \Rightarrow & X \quad (<) \end{array}$$

These rules allow the function category (2) to combine with arguments to yield context-free derivations of which (4) is a simple example:[4]

(4)
```
Traumaid   recommends   lavage
--------   ----------   ------
   NP      (S\NP)/NP      NP
           ------------------->
                   S\NP
-----------------------<
           S
```

The syntactic types in this derivation are simply a reflection of the corresponding semantic types, apart from the addition of directional information. If we expand the category (2) to express the semantics of the transitive verb, the same context-free derivation can be made to build a compositional interpretation, $(recommend'\ lavage')\ traumaid'$. One way of writing such an interpreted category that is particularly convenient for translating into unification-based programming languages like Prolog is the following:

(5) $(S : \mathit{recommend}'\ x\ y\backslash NP : y)/NP : x$

In (5), syntactic types are paired with a semantic interpretation via the colon operator, and the category is that of a function from NPs (with interpretation x) to functions from NPs (with interpretation y) to Ss (with interpretation $recommend'\ x\ y$). Constants in interpretations bear primes, variables do not, and there is a convention of left-associativity, so that $recommend'\ x\ y$ is equivalent to $(recommend'\ x)\ y$.

CCG extends this strictly context-free categorial base in two respects. First, all arguments, such as NPs, bear only *type-raised* categories, such as $S/(S\backslash NP)$. That is to say that the category of an NP, rather than being that of a simple argument, is that of

[4]It may be helpful for the reader to know that *lavage* refers to the therapeutic cleansing of an organ.

a function over functions-over-such-arguments, namely verbs and the like. Similarly, all functions into such categories, such as determiners, are functions into the raised categories, such as $(S/(S\backslash NP))/N$. For example, subject NPs bear the following category in the full notation:

(6) traumaid := $S : s/(S : s\backslash NP : traumaid')$

The derivation of the same simple transitive sentence using type-raised categories is illustrated in example (7) in the abbreviated notation.[5]

(7)
```
Traumaid  recommends        lavage
--------  ----------  ------------------
S/(S\NP)  (S\NP)/NP   (S\NP)\((S\NP)/NP)
          ------------------------------<
                     S\NP
-------------------------->
           S
```

Second, the combinatory rules are extended to include functional composition, as well as application:

(8) FORWARD COMPOSITION (>B):
$X/Y \quad Y/Z \quad \Rightarrow_{\mathbf{B}} \quad X/Z$

This rule allows a *second* syntactic derivation for the above sentence, as shown in example (9).[6]

(9)
```
Traumaid  recommends   lavage
--------  ----------  --------
S/(S\NP)  (S\NP)/NP   S\(S/NP)
-------------------->B
        S/NP
        ----------------------<
                 S
```

The original reason for making these moves was to capture the fact that fragments like *Traumaid recommends*, which in traditional terms are not regarded as syntactic constituents, can nevertheless take part in coordinate constructions, like (10)a, and form the residue of relative clause formation, as in (10)b.

[5] It is important to realize that the semantics of the type raised categories and of the application rules ensures that this derivation yields an S with the same interpretation as the earlier derivation (4), namely $recommend'\ lavage'\ traumaid'$. At first glance, it looks as though type-raising will expand the lexicon alarmingly. One way round this problem is discussed in Steedman (1991b).

[6] As before, it is important to realize that the semantics of the categories and of the new rule of functional composition guarantee that the S yielded in this derivation bears exactly the same interpretation as the original purely applicative derivation (4).

(10) a. You propose, and *Traumaid recommends*, lavage.
b. The treatment that *Traumaid recommends*

The full extent of this theory (which covers unbounded rightward and leftward "movement", and a number of other types of supposedly "non-constituent" coordination), together with the general class of rules from which the composition rule is drawn, and the problem of processing in the face of such associative rules, is discussed in the earlier papers, and need not concern us here. The point for present purposes is that the partition of the sentence into the object and a non-standard constituent ($S : recommend'\ x\ traumaid'/NP : x$) makes this theory structurally and semantically perfectly suited to the demands of intonation, as exhibited in exchanges like the following:[7]

(11) Q: I know that the surgeon recommends a left thoracotomy,
but what does Traumaid recommend?
A: (TRAUMAID recommends) (LAVAGE.)
L+H* LH% H* LL$

We can therefore directly incorporate intonational constituency in syntax, as follows (cf. Steedman 1990b, 1991a, 1991c). First, we assign to each constituent an autonomous prosodic category, expressing its potential for combination with other prosodic categories. Then we lock these two structural systems together via the following principle, which says that syntactic and prosodic constituency must be isomorphic:

(12) PROSODIC CONSTITUENT CONDITION:
Combination of two syntactic categories via a syntactic combinatory rule is only allowed if their prosodic categories can also combine via a prosodic combinatory rule.

One way to accomplish this is to give pitch accents the category of functions from boundaries to intonational/intermediate phrases. As in CCG, categories consist of a (prosodic) structural type, and an (information structural) interpretation, associated via a colon. The pitch accents have the following functional types:[8]

(13) L+H* := $p : theme/b : lh$
H* := $p : rheme/b : ll$

We further assume, following Bird (1991), that the presence of a pitch accent causes some element(s) in the translation of the category to be marked as focused, a matter which we will for simplicity assume to occur at the level of the lexicon. For example, when *recommends* bears a pitch accent, its category will be written as follows:

[7] A similar argument in a related categorial framework is made by Moortgat (1989).
[8] Here we are ignoring the possibility of multiple pitch accents in the same prosodic phrase, but cf. Steedman (1991a).

(14) $(S : *recommend' \; x \; y \backslash NP : y)/NP : x$

We depart from earlier versions of this theory in assuming that boundaries are not simply *arguments* of such functions, but are rather akin to *type-raised* arguments, as follows:[9]

(15)
$$\begin{array}{lll} \text{L} & := & p : rheme \backslash (p : rheme/b : ll) \\ \text{LL\$} & := & u : rheme \backslash (p : rheme/b : ll) \\ \text{LH\%} & := & p : theme \backslash (p : theme/b : lh) \\ \text{LH\$} & := & u : theme \backslash (p : theme/b : lh) \end{array}$$

These categories closely correspond to Pierrehumbert's distinction between various levels of phonological phrases. For example, the boundary L maps an H* pitch accent into an intermediate phrase rheme, $p : rheme$. The LH% boundary maps an L+H* pitch accent onto a full intonation phrase, which it is convenient for present purposes to write as $p : theme$. (In a fuller notation we would make the distinction between intermediate and intonational phrases explicit, but for present purposes it is irrelevant). The LH$ boundary maps the same L+H* pitch accent into an utterance-level thematic phrase, written $u : theme$.

The categories that result from the combination of a pitch accent and a boundary may or may not constitute entire prosodic phrases, since there may be prenuclear material bearing null tone. There may also be material bearing null tone separating the pitch accent(s) from the boundary. (Both possibilities are illustrated in (1)). We therefore assign the following category to the null tone, which can thereby apply to the right to any non-functional category of the form $X : Y$, and compose to the right with any function into such a category, including another null tone, to yield the same category:

(16) Ø := $X : Y/X : Y$

It is this omnivorous category that allows intonational tunes to be spread over arbitrarily large constituents, since it allows the pitch accent's desire for a boundary to propagate via composition into the null tone category, as in the earlier papers.

In order to allow the derivation to proceed above the level of complete prosodic phrases identifying themes and rhemes, we need the two unary category-*changing* rules shown in (17) and (18) to change the phonological category of complete themes and rhemes.[10]

(17)
$$\begin{array}{ccc} \Sigma & \Rightarrow & \Sigma \\ p : X & & utterance/utterance \end{array}$$

[9]Note again that $ boundaries are often conflated with % intonational phrase boundaries in the literature. These categories, which in some sense imply that boundaries are phonological heads, constitute a modification to previous versions of the present theory that brings it more closely into line with the proposals in Pierrehumbert and Hirschberg (1990). The idea that boundaries are functors has been independently proposed by Kirkeby-Garstad and Polgardi (p.c.).

[10]These rules represent another minor departure from the earlier papers.

(19)

```
         Traumaid                    recommends            lavage
           L+H*                          LH%                 H* LL$
-----------------------  --------------------------  -----------
S:s/(S:s\NP:*traumaid') (S:recommend'x y\NP:y)/NP:x  NP:*lavage'
     p:theme/b:lh          p:theme\(p:theme/b:lh)      u:rheme
------------------------SYN------------------------->B
------------------------PHON------------------------<
           S:recommend' x *traumaid'/NP:x
                      p:theme
           ============PHON=============            ====PHON===
           S:recommend' x *traumaid'/NP:x           NP:*lavage'
                 utterance/utterance                 utterance
           ------------------------SYN------------------------->
           ------------------------PHON------------------------>
                      S: recommend' *lavage' *traumaid'
                                 utterance
```

Theme: $S : recommend'\ z\ {*}traumaid'/NP : z$
Rheme: $NP : {*}lavage$

(18) $$\begin{array}{c} \Sigma \\ u : X \end{array} \;\Rightarrow\; \begin{array}{c} \Sigma \\ utterance \end{array}$$

These rules change the prosodic category either to $utterance$, or to an endocentric function over that category. These types capture the fact that the LL\$ and LH\$ boundaries can only occur at the end of a sentence, thereby correcting an overgeneration in some early versions of this theory noted by Bird (1991). The fact that $utterance$ is an atom rather than a term of the form $X : Y$ is important, since it means that it can unify only with another $utterance$. This is vital to the preservation of the intonation structure.[11]

The application of the above two rules to a complete intonational phrase should be thought of as precipitating a side-effect whereby a copy of the category Σ is associated with the clause as its theme or rheme. (We gloss over details of how this is done, as well as a number of further complications arising in sentences with more than one rheme).

In Steedman (1991a, 1991c), a related set of rules of which the present ones form a subset are shown to be well-behaved with a wide range of examples. Example (19) gives the derivation for an example related to (9).[12] Note that it is the identification of the theme and rheme at the stage *before* the final reduction that determines the information structure for the response, for it is at this point that discourse elements like the theme of the answer can be defined, and can be used in semantically-driven synthesis of intonation contour directly from the grammar.

[11] The category has the effect of preventing further composition into the null tone achieved in the earlier papers by a restriction on forward prosodic composition.

[12] Note that since the raised object category is not crucial, it has been replaced by NP for ease of reading comprehension. Also note the focus-marking effect of the pitch accents.

Of course, such effusively informative intonation contours are comparatively rare in normal dialogues. A more usual response to the question "What does Traumaid recommend?" in (11) would put low pitch – that is, the null tone in Pierrehumbert's terms – on everything except the focus of the rheme, *lavage*, as in (20).

(20) Traumaid recommends LAVAGE.
H* LL$

Such an utterance is of course ambiguous as to whether the theme is *traumaid* or *what traumaid recommends*. The earlier papers show that such "unmarked" themes, which include no primary pitch accent because they are entirely background, can be captured by a "Null Theme Promotion Rule", as follows:

(21)
$$\begin{array}{ccc} \Sigma & & \Sigma \\ X : Y/X : Y & \Rightarrow & p : theme \end{array}$$

This rule says that any sequence bearing the null tone can be regarded as an "unmarked" intermediate phrase theme.

3. Modeling Contrast

The preceding remarks about the ambiguity of unmarked themes should make it clear that in general the information structure of the response to a query cannot be identified on the basis of the question alone, but requires information from the discourse model as well, to which we now turn.[13]

This remark applies even more strongly to the assignment of focus and the corresponding pitch accents in the generation of the response, as Davis and Hirschberg (1988), and Hirschberg (1990), among others, have pointed out. That is, while it might appear as though pitch-accents could be assigned on some basis such as the mention or non-mention of the relevant words in the theme of the query, such an expedient will often break down. Consider the following example, which might be produced by such a strategem, since the words "left" and "thoracotomy" do not occur in the theme *Which incision*:[14]

(22) Q: Which incision does TRAUMAID prefer?
A: (TRAUMAID prefers) (a LEFT thoraCOTomy.)
L+H* LH% H* H* LL$

In some contexts, including the null context, this intonation contour will indeed be appropriate. However, in any context where thoracotomy procedures are already established as the set of procedures in question, the pitch accent on *thoracotomy* in the response will be inappropriate and perhaps even misleading.

[13] See Prevost and Steedman (1993a) for an investigation of how much one can get away with on the basis of the question alone.

[14] It may be helpful to point out that a *thoracotomy* is a surgical incision of the chest wall, and a *thoracostomy* is the insertion of a tube into the chest.

For example, in (23) below, the noun *thoracotomy* must remain unstressed while the adjective *left* must be accented in the response, despite having been explicitly mentioned in the text of the question.[15] Here the question itself establishes a contextual set. The fact that the entity that is referenced in the response must be contrasted with other alternatives in this set on the relevant property requires the assignment of a pitch accent to the corresponding word.

(23) Q: Does Traumaid prefer a LEFT thoracotomy or a RIGHT thoracotomy?
A: (Traumaid prefers) (a LEFT thoracotomy.)

The mere fact that alternatives are contrasted on a given property is not enough however to mandate the inclusion of a pitch accent on the corresponding linguistic material. The property in question must restrict contrastively *at the relevant point in the semantic evaluation*, before a pitch accent is forced. Thus, in a situation in which the choices include a left thoracotomy, a right thoracotomy, a left thoracostomy and a right thoracostomy, the response to question (24), in which the adjective is unstressed, is perfectly appropriate:[16]

(24) Q: Does Traumaid prefer a LEFT thoraCOTomy or a RIGHT thoraCOSTomy?
A: (Traumaid prefers) (a left thoraCOTomy).

This example suggests that the set that is being considered by the time the adjective is semantically evaluated is no longer the entire set including the left and right thoracotomy and thoracostomy procedures. In fact, it is not even the set containing only the left thoracotomy and right thoracostomy procedures, but rather the set containing only the left thoracotomy procedure, which by definition does not stand in contrast to any other thoracotomy procedure by virtue of the property of being performed on the left side. This set arises because the noun *thoracotomy* restricts over the set including the left thoracotomy and the right thoracostomy procedures.

To see this, consider the next exchange, uttered in the same situation.

(25) Q: Does Traumaid prefer a LEFT thoraCOTomy, a RIGHT thoraCOTomy or a LEFT thoraCOSTomy?
A: (Traumaid prefers) (a LEFT thoraCOTomy).

Here the set established by the question is restricted by the noun in the rheme of the answer to be a set of two thoracotomy procedures (both left and right). Since they are distinguished by the property *left*, the corresponding linguistic material must be accented.

[15] In using these examples to motivate the treatment of contrast in the system, we go beyond the class of discourses that are actually handled by the system as currently implemented. We are in fact glossing over a number of subtle problems concerning the theme-rheme structures that are involved, and the precise reflection of these information structures in intonation.

[16] That is not to claim that the adjective *cannot* carry a pitch accent, of course.

The algorithm for determining which items are to be stressed for reasons of contrast works as follows.[17] For a given object x, we associate a set of properties which are essential for constructing an expression that uniquely refers to x, as well as a set of objects (and their referring properties) which might be considered *alternatives* to x with respect to the database under consideration. The set of alternatives is restricted by properties or objects explicitly mentioned in the theme of the question. Then for each property of x in turn, we restrict the set of alternatives to include only those objects having the given property. If imposing this restriction decreases the size of the set of alternatives, then the given property serves to distinguish x from its alternatives, suggesting that the corresponding linguistic material should be stressed.

Besides determining the location of primary sentence stress, contrastive properties may also necessitate adopting non-standard lexical stress patterns. For example, in the following question/answer pair, the normal lexical stress on *thor* switches to *pneu* in *pneumothorax* because *pneumothorax* stands in contrast to *hemothorax*.

(26) Q: I know which procedure is recommended for the simple hemothorax.
But which condition is a left THORACOSTOMY recommended for?
A: A left THORACOSTOMY is recommended for the simple PNEUmothorax.

In the current implementation, such lexical stress shift is handled by identifying the lexical contrast properties in the alternative set representations and supplying separate pronunciations in the lexicon. However, when such properties are determined to stand in contrast to one another, the alternate pronunciation could in principle be generated by employing the methods described above within the lexicon.

4. The Implementation

The present paper is an attempt to apply the theories outlined in the preceding sections to the task of specifying contextually appropriate intonation for natural language responses to database queries. The architecture of the system (shown in Figure 1) identifies the key modules of the system, their relationships to the database and the underlying grammar, and the dependencies among their inputs and outputs.

The process begins with a fully segmented and prosodically annotated representation of a spoken query, as shown in example (27).[18] We employ a simple bottom-up shift-reduce parser, making direct use of the combinatory prosody theory described above, to identify the semantics of the question. The inclusion of prosodic categories in the grammar allows the parser to identify the information structure within the question as well, marking "focused" items with *, as shown in (28). For the moment, unmarked themes are handled by taking the longest unmarked constituent permitted by the syntax.

[17] We omit a more detailed description of the algorithm and its associated data structures for the sake of brevity. A more detailed account and numerous examples are given in Prevost and Steedman (1993c).

[18] We stress that we do *not* start with a speech wave, but a representation that one might obtain from a hypothetical system that translates such a wave into strings of words with Pierrehumbert-style intonation markings.

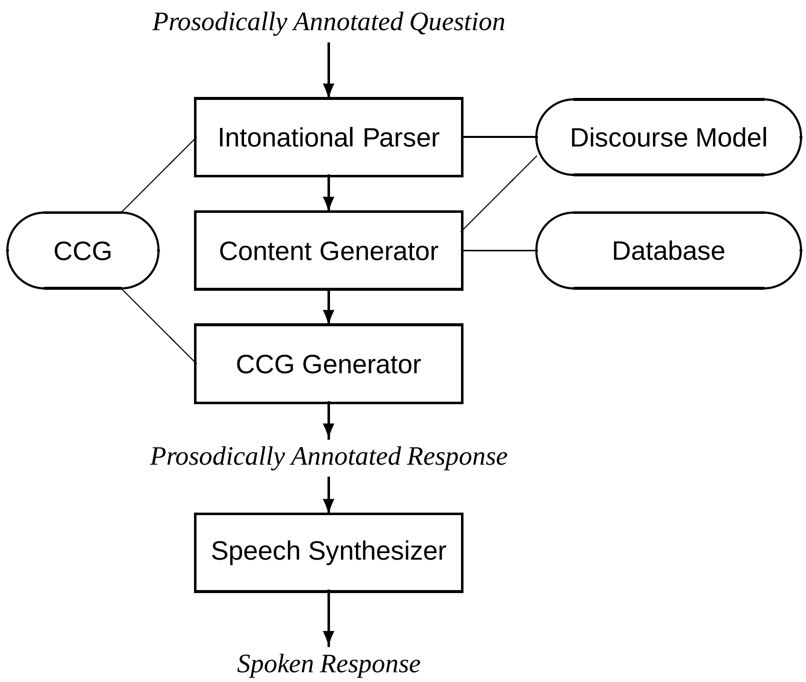

Figure 1: Architecture

(27) I know what the CAT scan is for,
but WHICH condition does URINALYSIS address?
L+H* LH% H* LL$

(28) Proposition:
$s : \lambda x[condition(x) \& address(*urinalysis, x)]$
Theme:
$s : \lambda x[condition(x) \& address(*urinalysis, x)]/$
$(s : address(*urinalysis, x)/np : x)$
Rheme:
$s : address(*urinalysis, x)/np : x$

The content generation module, which has the task of determining the semantics and information structure of the response, relies on several simplifying assumptions. Foremost among these is the notion that the rheme of the question is the sole determinant of the theme of the response, including the specification of focus (although the type of pitch accent that eventually marks the focus will be different in the response). The overall semantic structure of the response can be determined by instantiating the variable in the lambda expression corresponding to the *wh*-question with a simple Prolog query. Given the syntactic and focus-marked semantic representation for the response, along with the syntactic and focus-marked semantic representation for the theme of the response, a representation for the rheme of the response can be worked out from the CCG rules. The assignment of focus for the rheme of the response (i.e. the instantiated variable) must be worked out from scratch, on the basis of the alternative sets in the database, as described in section 3.

For the question given in (27), the content generator produces the following:

(29) Proposition:
$s : address(*urinalysis, *hematuria)$
Theme:
$s : address(*urinalysis, x)/np : x$
Rheme:
$np : *hematuria$

From the output of the content generator, the CCG generation module produces a string of words and Pierrehumbert-style markings representing the response, as shown in (30).[19]

(30) urinalysis@lhstar addresses@lh hematuria@hstarllb

The final aspect of generation involves translating such a string into a form usable by a suitable speech synthesizer. The current implementation uses the Bell Laboratories TTS system (Liberman and Buchsbaum 1985) as a post-processor to synthesize the speech wave itself.

5. Results

The system described above produces quite sharp and natural-sounding distinctions of intonation contour in minimal pairs of queries like those in examples (31)–(38), which should be read as concerning a single patient with multiple wounds. These examples illustrate the system's capability for producing appropriately different intonation contours for a single string of words under the control of discourse context. If the responses in these examples are interchanged, the results sound distinctly unnatural in the given contexts.[20]

[19] Full descriptions of the CCG generation algorithm are given in Prevost and Steedman (1993a, 1993c).

[20] The first line of each query is for reader assistance only, and is not processed by the system described here. The *waves* files corresponding to the examples in this section are available by anonymous ftp from ftp.cis.upenn.edu, under the directory */pub/prevost/speechcomm*.

Examples (31) and (32) illustrate the necessity of the theme/rheme distinction. Although the pitch accent *locations* in the responses in these examples are identical, occurring on *thoracostomy* and *simple*, the alternation in the theme and rheme tunes is necessary to convey the intended proposition in the given contexts.

Examples (32) and (34) show that the system makes appropriate distinctions in focus placement within themes and rhemes based on context. Although the responses in these two sentences possess the same intonational tunes, the pitch accent location is crucial for conveying the appropriate contrastive properties.

Examples (31)–(38) manifest the eight basic combinatorial possibilities for pitch accent placement and tune selection produced by our program for the given sentence. The inclusion of contrastive lexical stress shift increases the number of intonational possibilities even more, as exemplified in (39) and (40).

(31) Q: I know what's recommended for the PERSISTENT pneumothorax,
but which procedure   is recommended for the SIMPLE pneumothorax?
L+H*   LH%   H*   LL$
A: A left THORACOSTOMY   is recommended for the SIMPLE pneumothorax.
H* L   L+H*   LH$

(32) Q: I know what's recommended for the PERSISTENT pneumothorax,
but which pneumothorax   is a left THORACOSTOMY recommended for?
L+H*   LH%   H*   LL$
A: A left THORACOSTOMY is recommended for   the SIMPLE pneumothorax.
L+H*   LH%   H*   LL$

(33) Q: I know what's recommended for the PERITONITIS,
but which procedure   is recommended for the simple pneumoTHORax?
L+H*   LH%   H* LL$
A: A left THORACOSTOMY   is recommended for the simple pneumoTHORax.
H* L   L+H* LH$

(34) Q: I know what's recommended for the PERITONITIS,
but which condition   is a left THORACOSTOMY recommended for?
L+H*   LH%   H*   LL$
A: A left THORACOSTOMY is recommended for   the simple pneumoTHORax.
L+H*   LH%   H* LL$

(35) Q: A RIGHT thoracostomy is recommended for the PERSISTENT pneumothorax,
but which thoracostomy is recommended for the SIMPLE pneumothorax?
L+H* LH% H* LL$
A: A LEFT thoracostomy is recommended for the SIMPLE pneumothorax.
H* L L+H* LH$

(36) Q: A RIGHT thoracostomy is recommended for the PERSISTENT pneumothorax,
but which pneumothorax is a LEFT thoracostomy recommended for?
L+H* LH% H* LL$
A: A LEFT thoracostomy is recommended for the SIMPLE pneumothorax.
L+H* LH% H* LL$

(37) Q: A RIGHT thoracostomy is recommended for some condition,
but which thoracostomy is recommended for the simple pneumoTHORax?
L+H* LH% H* LL$
A: A LEFT thoracostomy is recommended for the simple pneumoTHORax.
H* L L+H* LH$

(38) Q: A RIGHT thoracostomy is recommended for some condition,
but which condition is a LEFT thoracostomy recommended for?
L+H* LH% H* LL$
A: A LEFT thoracostomy is recommended for the simple pneumoTHORax.
L+H* LH% H* LL$

(39) Q: I know which procedure is recommended for the simple hemothorax,
but which procedure is recommended for the simple PNEUmothorax?
L+H* LH% H* LL$
A: A left THORACOSTOMY is recommended for the simple PNEUmothorax.
H* L L+H* LH$

(40) Q: I know which procedure is recommended for the simple hemothorax,
but which condition is a left THORACOSTOMY recommended for?
L+H* LH% H* LL$
A: A left THORACOSTOMY is recommended for the simple PNEUmothorax.
L+H* LH% H* LL$

6. Conclusions

The results show that is possible to generate synthesized spoken responses with contextually appropriate intonational contours in a database query task. Many important problems remain, both because of the limited range of discourse-types and intonational tunes considered here, and because of the extreme oversimplification of the discourse model (particularly with respect to the ontology, or variety of types of discourse entities). Nevertheless, the system presented here has a number of properties that we believe augur well for its extension to richer varieties of discourse, including the types of monologues and commentaries that are more appropriate for the actual TraumAID domain. Foremost among these is the fact that the system and the underlying theory are entirely modular. That is, any of its components can be replaced without affecting any other component because each is entirely independent of the particular grammar defined by the lexicon and the particular knowledge base that the discourse concerns. It is only because CCG allows us to unify the structures implicated in syntax and semantics on the one hand, and intonation and discourse information on the other, that this modular structure can be so simply attained.

Acknowledgments

Preliminary versions of some sections in the present paper were published as Prevost and Steedman (1993a, 1993b). We are grateful to the audiences at those meetings, to AT&T Bell Laboratories for allowing us access to the TTS speech synthesizer, to Mark Beutnagel, Julia Hirschberg, and Richard Sproat for patient advice on its use, to Abigail Gertner for advice on Traumaid, to Janet Pierrehumbert for discussions on notation, and to the anonymous referees for many helpful suggestions. The usual disclaimers apply. The research was supported in part by NSF grant nos. IRI90-18513, IRI90-16592, IRI91-17110 and CISE IIP-CDA-88-22719, DARPA grant no. N00014-90-J-1863, ARO grant no. DAAL03-89-C0031, and grant no. R01-LM05217 from the National Library of Medicine.

Bibliography

M. Beckman and J. Pierrehumbert (1986), "Intonational Structure in Japanese and English", *Phonology Yearbook*, Vol. 3, pp. 255–310.

S. Bird (1991), "Focus and phrasing in Unification Categorial Grammar", *Declarative Perspectives on Phonology*, Working Papers in Cognitive Science 7, ed. by S. Bird (University of Edinburgh), pp. 139–166.

J. Davis and J. Hirschberg (1988), "Assigning Intonational Features in Synthesized Spoken Directions", *Proceedings of the 26th Annual Conference of the ACL*, Buffalo, pp. 187–193.

J. Hirschberg (1990), "Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech", *Proceedings of AAAI: 1990*, pp. 952–957.

G. Houghton (1986), *The Production of Language in Dialogue: a Computational Model*, unpublished PhD dissertation, University of Sussex.

S. Isard and M. Pearson (1988), "A Repertoire of British English Intonation Contours for Synthetic Speech", *Proceedings of Speech '88, 7th FASE Symposium*, Edinburgh, pp. 1233–1240.

M. Liberman and A.L. Buchsbaum (1985), "Structure and Usage of Current Bell Labs Text to Speech Programs", TM 11225-850731-11, AT&T Bell Laboratories.

M. Moortgat (1989), *Categorial Investigations* (Foris, Dordrecht).

R. Oehrle (1988), "Multi-dimensional Compositional Functions as a basis for Grammatical Analysis", in *Categorial Grammars and Natural Language Structures*, ed. by R. Oehrle, E. Bach and D. Wheeler (Reidel, Dordrecht), pp. 349–390.

J. Pierrehumbert (1980), *The Phonology and Phonetics of English Intonation*, PhD dissertation, MIT. (Dist. by Indiana University Linguistics Club, Bloomington, IN.)

J. Pierrehumbert and J. Hirschberg (1990), "The Meaning of Intonational Contours in the Interpretation of Discourse", in *Intentions in Communication*, ed. by P. Cohen, J. Morgan, and M. Pollack (MIT Press, Cambridge MA), pp. 271–312.

S. Prevost (1993), "Intonation, Context and Contrastiveness in Spoken Language Generation", dissertation proposal, University of Pennsylvania.

S. Prevost and M. Steedman (1993a), "Generating Contextually Appropriate Intonation", *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, pp. 332–340.

S. Prevost, and M. Steedman (1993b), "Using Context to Specify Intonation in Speech Synthesis", *Proceedings of the 3rd European Conference of Speech Communication and Technology (EUROSPEECH)*, Berlin, September 1993, pp. 2103–2106.

S. Prevost and M. Steedman (1993c), "Generating Intonation from Context Using a Combinatory Grammar", manuscript, University of Pennsylvania.

M. Steedman (1987). "Combinatory Grammars and Parasitic Gaps", *Natural Language and Linguistic Theory*, Vol. 5, pp. 403–439.

M. Steedman (1990a). "Gapping as Constituent Coordination", *Linguistics & Philosophy*, Vol. 13, pp. 207–263.

M. Steedman (1990b), "Structure and Intonation in Spoken Language Understanding", *Proceedings of the 25th Annual Conference of the Association for Computational Linguistics*, Pittsburgh, June 1990, pp. 9–17.

M. Steedman (1991a), "Structure and Intonation", *Language*, Vol. 68, pp. 260–296.

M. Steedman (1991b), “Type-raising and Directionality in Categorial Grammar”, *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, Berkeley, June 1991, pp. 71–78.

M. Steedman (1991c), “Surface Structure, Intonation, and ‘Focus’ ”, in *Natural Language and Speech, Proceedings of the ESPRIT Symposium, Brussels, 1991*, ed. by E. Klein and F. Veltman, pp. 21–38.

J. Terken (1984), “The Distribution of Accents in Instructions as a Function of Discourse Structure”, *Language and Speech,* Vol 27, pp. 269–289.

B. Webber, R. Rymon and J.R. Clarke (1992), “Flexible Support for Trauma Management through Goal-directed Reasoning and Planning”, *Artificial Intelligence in Medicine*, Vol. 4(2), pp. 145-163.

S. Young and F. Fallside (1979), “Speech Synthesis from Concept: a Method for Speech Output from Information Systems”, *Journal of the Acoustical Society of America*, Vol. 66, pp. 685–695.

R. Zacharski, A.I.C. Monaghan, D.R. Ladd and J. Delin (1993), “BRIDGE: Basic Research on Intonation in Dialogue Generation”, unpublished manuscript. HCRC, University of Edinburgh.