# Information Based Intonation Synthesis*

*Scott Prevost & Mark Steedman*

Computer and Information Science
University of Pennsylvania
200 South 33rd Street
Philadelphia, PA 19104-6389
(Internet: prevost@linc.cis.upenn.edu steedman@cis.upenn.edu)

## ABSTRACT

This paper presents a model for generating prosodically appropriate synthesized responses to database queries using Combinatory Categorial Grammar (CCG - cf. [22]), a formalism which easily integrates the notions of syntactic constituency, prosodic phrasing and information structure. The model determines accent locations within phrases on the basis of contrastive sets derived from the discourse structure and a domain-independent knowledge base.

## 1. Introduction

Previous work in the area of intonation generation includes an early study by Young and Fallside ([26]), and studies by Terken ([24]), Houghton, Isard and Pearson (cf. [11, 12]), Davis and Hirschberg (cf. [4, 10]), and Zacharski et al. ([27]). The present proposal differs from the earlier studies in the accent assignment rules, and in the representation of information structure and its relation to syntax and semantics. In the CCG framework, the information units that are delineated by intonation are directly represented, complete with semantic interpretations. These interpretations are utilized in making accent placement decisions on the basis of contrastive properties rather than *previous-mention* heuristics. While such heuristics have proven quite effective in the earlier studies, we believe the model-theoretic approach taken here will eventually lead to the development of similar heuristics for handling a wider range of examples involving contrastive stress.

The remainder of the paper discusses the contrastive stress model, describes the implemented system, and presents results demonstrating the system's ability to generate a variety of intonational possibilities for a given sentence depending on the discourse context.

*Preliminary versions of some sections in the present paper were published as [17] and [18]. 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, and to Janet Pierrehumbert for discussions on notation. 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.

## 2. Motivation

Meaning-to-speech systems differ from text-to-speech systems in the manner in which semantic and pragmatic information is exploited for assigning intonational features. Text-to-speech systems for unrestricted text are forced to rely on crude syntactic analyses and word classifications in making judgements about the accentability of words in an utterance, often using the strategy of *previous mention* whereby a word is de-accented if it (or perhaps its root) has previously occurred in some restricted segment of the text (cf. [10], [15]). The text can be divided into such meaningful discourse segments on the basis of cue phrases and paragraph boundaries.

Meaning-to-speech systems, on the other hand, have been employed in applications with limited, well-defined domains where semantic and discourse level knowledge is available. For these systems, the effectiveness of the previous mention strategy can be improved by considering semantic givenness in addition to lexical givenness when deciding if a word should be de-accented.

Such enhanced previous-mention heuristics, while proving quite effective in practice, have exhibited several deficiencies that have been noted by their proponents. Foremost among these is the inability of such strategies to model the seemingly contrastive nature of many accentual patterns in spoken language ([10]). In some cases, contrastive stress errors may sound unnatural and in the worst case may actually mislead the hearer. Another problem that has been attributed to previous-mention strategies is the tendency to include too many accents ([15]), potentially resulting in an inability for the hearer to determine the most important aspects of the speaker's intended message. The remainder of this section addresses these two problems and proposes explicitly modeling contrast in meaning-to-speech systems as a potential solution.

A previous-mention strategy might work as follows:

- Assign accents to open-class items (e.g. nouns, verbs, other content words)
- Do not assign accents to closed-class items (e.g. function words)
- De-accent any words that were already mentioned in the local discourse segment.

Now consider a hypothetical application in a medical domain that produces the type of output shown in (1) when a physician fails to include a recommended procedure in a plan for treating a specific patient.[1]

(1) a. You seem to have neglected to consider a THORACOSTOMY procedure for this patient.
  b. I propose doing a LEFT thoracostomy.

Using a previous-mention algorithm like the one above will produce the appropriate accentual pattern on the NP *a left thoracostomy* in (1)b because *thoracostomy* is explicitly mentioned in the previous sentence.

Now suppose the physician inadvertently includes the wrong procedure in the treatment plan, say a left *thoracotomy* rather than the intended left *thoracostomy*. Example (2) shows the possible output from the system.

(2) a. You seem to have confused the THORACOTOMY and THORACOSTOMY procedures in your plan for this patient.
  b. I propose doing a left THORACOSTOMY.
  b′. I propose doing a LEFT THORACOSTOMY.
  b″. I propose doing a LEFT thoracostomy.
  b‴. I propose doing a left thoracostomy.

The four accentual possibilities for the NP *a left thoracostomy* in the second sentence are given in (2)b-b‴. Examples (2)b and b′ are both acceptable because they correctly accent the contrastive *thoracostomy*. Based on the the contents of the first sentence, however, the previous-mention strategy would produce the accentual pattern illustrated in (2)b″, which is clearly inappropriate. In fact, such an intonation may cause the hearer to infer that the program's objection was to performing the procedure on the wrong side. Finally, if one considers the terms *left* and *thoracostomy* to be given prior to the utterance because of their inclusion in the physician's plan, the previous-mention strategy would attempt to de-accent both terms as in (2)b‴. Since the NP clearly requires some form of accentuation, alternative strategies are necessary in such a case. Other plausible previous-mention strategies exhibit similar problems for equally simple examples.

We believe that some of the problems associated with the previous-mention strategy in meaning-to-speech systems can be rectified by explicitly modeling contrastive stress. For the example above, the program initially knows that the physician's plan includes a left *thoracotomy* and that the program's plan includes a left *thoracostomy*. Hence, the program can construct an explicit set of alternative procedures from which accentual patterns can be determined. By noting that the alternatives differ not in the side on which they are to be performed, but in the actual type of procedure, the program can easily decide to stress *thoracostomy* rather than *left*. The precise algorithm for contrastive stress assignment is given a more detailed explanation in [18].

We shall also see how the contrastive stress approach can avoid the over-accentuation problem of the previous-mention strategy as well. Consider a patient with two chest wounds: a right lateral wound and a right anterior wound. At some point our hypothetical system may need to address one of these wounds in the following manner.[2]

(3) You need to address the right lateral chest wound in your treatment plan.

Using the previous-mention strategy would lead to the following output if the wound had not been mentioned previously.

(4) You need to address the RIGHT LATERAL CHEST WOUND in your treatment plan.

The contrastive stress algorithm is able to recognize the crucial distinction between the *lateral* and *anterior* properties of the patient's two wounds and assign stress accordingly, producing:

(5) You need to address the right LATERAL chest wound in your treatment plan.

## 3. The Implementation

The present paper describes an implemented system (IBIS) that applies the CCG theory of prosody outlined

[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 ([25]). The morbid nature of the examples, for which we apologize, is due entirely to the special nature of the trauma domain. The lay reader may be interested to know that a *thoracostomy* is the insertion of a tube into the chest, and a *thoracotomy* is a surgical incision of the chest wall. In the examples, accented words are shown in small capitals.

[2] A closely related issue is how the system decides which modifiers are necessary in the description ([20]).

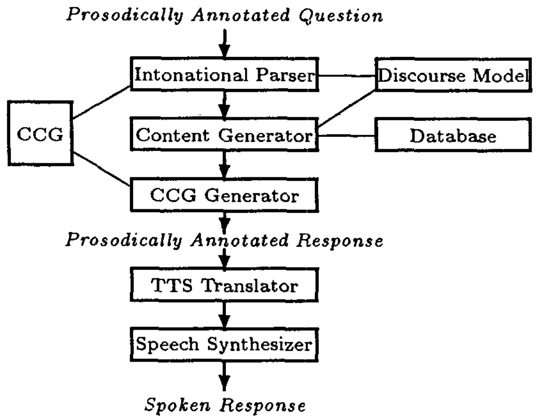

Figure 1: Architecture

in [22, 17, 18] to the the task of specifying contextually appropriate intonation for spoken messages concerning the medical expert system TraumAID, developed independently at Penn (cf. [25]). Our examples below are taken from this domain, in which it is eventually our intention to deploy the generation system in a surgical situation in a critiquing mode, as an output device for the expert system. For the present purpose of illustrating the workings of the generation system, we have chosen a simpler (but sociologically rather unrealistic) database query application.

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 (6).[3]

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

In example (6), capitals indicate stress and brackets informally indicate the intonational phrasing. The intonation contour is indicated more formally using a version of Pierrehumbert's notation ([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 ([2]) distinction between intermediate phrase boundaries, intonational phrase boundaries, and utterance boundaries. 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 since utterance boundaries must be accompanied by a greater degree of lengthening and pausing.

The intonational tunes L+H* LH(%/$) and H* L(L%/$) shown in example (6) convey two distinct kinds of discourse information. First, both H* and L+H* pitch accents mark the word that they occur on (or rather, some element of its interpretation) for *focus*, which in this task 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 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.

We employ a simple bottom-up shift-reduce parser, making direct use of the combinatory prosody theory described in [22, 17, 18], 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, dividing it into theme and rheme, and marking focused items with * as shown in (7). For the moment, unmarked themes are handled by taking the longest unmarked constituent permitted by the syntax.

(7) 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 worked

[3] 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.

out from the grammar rules. The assignment of focus for the rheme of the response (i.e. the instantiated variable) must be worked out from scratch, using techniques for assigning contrastive stress.

The algorithm for assigning contrastive stress works as follows. For a given object $x$ in the rheme of the response, 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 initially restricted by properties or objects explicitly mentioned in the theme of the question. For each property of $x$ in turn, we restrict the set of alternatives to include only those objects having the given property. When imposing the restriction decreases the number of alternatives, we conclude that the given property serves to distinguish $x$ from its alternatives, suggesting that the corresponding linguistic material should be stressed.

For example, for the question given in (6), the content generator produces the following representation, because the theme is "What urinalysis addresses", the rheme is "hematuria", and the context includes alternative conditions and treatments:

(8) 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 (9).[4]

(9) urinalysis@lhstar addresses@lhb 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 [14] as a post-processor to synthesize the speech wave.

## 4. Results

The IBIS system produces distinct intonational differences in minimal pairs of queries like those in examples (10)–(13) below. These minimal pairs 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.

(10) Q: I know that burns induce fever, but
which symptoms do LACERATIONS induce?
L+H* LH% H* LL$
A: LACERATIONS induce BLEEDING.
L+H* LH% H* LL$

(11) Q: I know that burns induce fever, but
which wounds induce BLEEDING?
L+H* LH% H* LL$
A: LACERATIONS induce BLEEDING.
H* L L+H* LH$

(12) Q: I know what CAUSES infection,
but which medications PREVENT infection?
L+H* LH% H* LL$
A: ANTIBIOTICS PREVENT infection.
H* L L+H* LH$

(13) Q: I know what medications prevent NAUSEA,
but which medications prevent INFECTION?
L+H* LH% H* LL$
A: ANTIBIOTICS prevent INFECTION.
H* L L+H* LH$

Examples (10) and (11) illustrate the necessity of the theme/rheme distinction. Although the pitch accent *locations* in the responses in these examples are identical, occurring on *lacerations* and *bleeding*, the alternation in the theme and rheme tunes is necessary to convey the intended propositions in the given contexts. Examples (12) and (13) show that the system makes appropriate distinctions in focus placement within themes and rhemes based on context. More complex examples, like those shown in (14)–(16), illustrate the usefulness of the contrastive stress algorithm for assigning pitch accents in appropriate locations.[5]

## 5. Conclusions

While previous attempts at intonation generation have relied on previous-mention heuristics for assigning accents, the present results show that is is possible to generate synthesized spoken responses with appropriate intonational contours in a database query task using explicit representations of contrastive stress. 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

[4] Full descriptions of the CCG generation algorithm are given in [17].

[5] Further examples of the output of IBIS can be found in [19].

(14) Q: I know which procedure is recommended for the BURN patient,
but which procedure is recommended for the WOUND patient?
L+H* LH% H* LL$
A: A left THORACOTOMY is recommended for the WOUND patient.
H* L L+H* LH$

(15) Q: I know which procedure is recommended for the BURN patient,
but which patient is a left THORACOTOMY recommended for?
L+H* LH% H* LL$
A: A left THORACOTOMY is recommended for the WOUND patient.
L+H* LH% H* LL$

(16) Q: A RIGHT thoracotomy is recommended for the FIRST patient,
but which thoracotomy is recommended for the SECOND patient?
L+H* LH% H* LL$
A: A LEFT thoracotomy is recommended for the SECOND patient.
H* L L+H* LH$

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.

## References


1. Allen, Jonathan, Sharon Hunnicutt, and Dennis Klatt (1987), *From Text to Speech: the MITalk system*, Cambridge, University Press.
2. Beckman, Mary and Janet Pierrehumbert (1986), 'Intonational Structure in Japanese and English', *Phonology Yearbook*, 3, 255-310.
3. Bird, Steven (1991), 'Focus and Phrasing in Unification Categorial Grammar', in S. Bird (ed.), *Declarative Perspectives on Phonology*, Working Papers in Cognitive Science 7, University of Edinburgh, 139-166.
4. Davis, James and Julia Hirschberg (1988), 'Assigning Intonational Features in Synthesized Spoken Directions', *Proceedings of the 26th Annual Conference of the ACL*, Buffalo, 187-193.
5. Gerdeman, Dale and Erhard Hinrichs (1990) 'Functor-driven Natural Language Generation with Categorial Unification Grammars', *Proceedings of COLING 90, Helsinki*, 145-150.
6. Hajičová, Eva, and Petr Sgall (1988), 'Topic and Focus of a Sentence and the Patterning of a Text', in János Petöfi, (ed.), *Text and Discourse Constitution*, De Gruyter, Berlin, 70-96.
7. Halliday, Michael (1970), 'Language Structure and Language Function', in John Lyons (ed.), *New Horizons in Linguistics*, Penguin.
8. 't Hart, J. and A. Cohen (1973), 'Intonation by Rule: a Perceptual Quest', *Journal of Phonetics*, 1, 309-327.
9. 't Hart, J. and R. Collier (1975), 'Integrating Different Levels of Phonetic Analysis', *Journal of Phonetics*, 3, 235-255.
10. Hirschberg, Julia (1990), 'Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech', *Proceedings of AAAI: 1990.*
11. Houghton, George and M. Pearson (1988), 'The Production of Spoken Dialogue,' in M. Zock and G. Sabah (eds), *Advances in Natural Language Generation: An Interdisciplinary Perspective, Vol. 1*, Pinter Publishers, London.
12. Isard, Stephen and M. Pearson (1988), 'A Repertoire of British English Intonation Contours for Synthetic Speech', *Proceedings of Speech '88, 7th FASE Symposium, Edinburgh.*
13. Jackendoff, Ray (1972), *Semantic Interpretation in Generative Grammar*, MIT Press, Cambridge MA.
14. Liberman, Mark and A.L. Buchsbaum (1985), 'Structure and Usage of Current Bell Labs Text to Speech Programs', TM 11225-850731-11, AT&T Bell Laboratories.
15. Monaghan, A.I.C. (1991), *Intonation in a Text-to-Speech Conversion System*, Ph.D dissertation, University of Edinburgh.

16. Pierrehumbert, Janet and Julia Hirschberg (1990), 'The Meaning of Intonational Contours in the Interpretation of Discourse', in Philip Cohen, Jerry Morgan, and Martha Pollack (eds.), *Intentions in Communication*, MIT Press Cambridge MA, 271-312.

17. Prevost, Scott and Mark Steedman (1993), 'Generating Contextually Appropriate Intonation', *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, 332–340.

18. Prevost, Scott and Mark Steedman (1993), 'Using Context to Specify Intonation in Speech Synthesis', *Proceedings of the 3rd European Conference of Speech Communication and Technology (EUROSPEECH)*, Berlin, September 1993, 2103-2106.

19. Prevost, Scott and Mark Steedman (1994), 'Specifying Intonation from Context for Speech Synthesis', unpublished manuscript, University of Pennsylvania.

20. Reiter, Ehud and Robert Dale (1992), 'A Fast Algorithm for the Generation of Referring Expressions', *Proceedings of COLING 92*, 232-238.

21. Rooth, Mats (1985), *Association with Focus*, unpublished PhD dissertation, University of Massachusetts, Amherst.

22. Steedman, Mark (1991), 'Structure and Intonation', *Language*, 68, 260-296.

23. Steedman, Mark (1991), 'Surface Structure, Intonation, and "Focus"', in E. Klein and F. Veltman (eds.), *Natural Language and Speech*, Proceedings of the ESPRIT Symposium, Brussels.

24. Terken, Jacques (1984), 'The Distribution of Accents in Instructions as a Function of Discourse Structure', *Language and Speech*, 27.

25. Webber, Bonnie, R. Rymon and J.R. Clarke (1992), 'Flexible Support for Trauma Management through Goal-directed Reasoning and Planning' *Artificial Intelligence in Medicine* 4(2), April 1992.

26. Young, S. and F. Fallside (1979),'Speech Synthesis from Concept: a Method for Speech Output from Information Systems' *Journal of the Acoustical Society of America*, 66, 685-695.

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