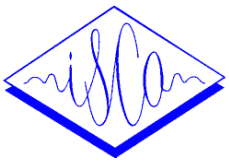


3rd European Conference on
Speech Communication and Technology
EUROSPEECH'93
Berlin, Germany, September 19-23, 1993

USING CONTEXT TO SPECIFY INTONATION IN SPEECH SYNTHESIS*

Scott Prevost & Mark Steedman

Computer and Information Science, University of Pennsylvania
Philadelphia, Pennsylvania USA
(*Internet:* prevost@linc.cis.upenn.edu steedman@cis.upenn.edu)

ABSTRACT

A generator based on Combinatory Categorial Grammar using a simple and domain-independent discourse model can be used to direct synthesis of intonation contours for responses to data-base queries, conveying distinctions of contrast and emphasis determined by the discourse model and the state of the knowledge-base.

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. It has often been noted, however, that the results frequently sound unnatural when taken in context, and may occasionally 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 the area includes an early study by Young and Fallside ([16]), and studies by Terken ([14]), Houghton, Isard and Pearson (cf. [5, 6]), Davis and Hirschberg (cf. [3, 4]), and Ladd et al. ([17]), although the representations of information structure and its relation to syntax employed by these authors are rather different from those we propose.

Consider for example the exchange shown in (1).[1]

(1) I know that BURNS induce FEVER.
but which symptom do STAB wounds induce?
(STAB wounds induce) (BLEEDING).
L+H* LH% H* LL%
Focus *Ground* *Focus*
Theme *Rheme*

Capitals indicate stress, and brackets informally indicate the intonational phrasing. The intonation contour is more formally indicated underneath using Pierrehumbert's notation ([8, 1, 12]). The other annotations indicate that the intonational tunes L+H* LH% and H* LL% convey two distinct kinds of discourse information. First, both pitch accents mark any word that they occur on (or rather, some element(s) of its interpretation) for "focus", which in the context of such simple queries as example (1) usually implies contrast of some kind.[2] Second, the tunes as a whole mark the constituents that bear them (or rather, their interpretations) as having a particular function in the discourse. We have argued at length elsewhere that, at least in this same restricted class of dialogues, one function of the L+H* LH% tune is to mark the "theme" – that is, "what the participants have agreed to talk about". The H* LL% tune (and its relative the H* L tune) mark 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 the correct open proposition as the theme, as shown in example (2).

(2) Q: Which wounds induce BLEEDING?
A: * (STAB wounds induce) (BLEEDING).
L+H* LH% H* LL%

2 COMBINATORY PROSODY

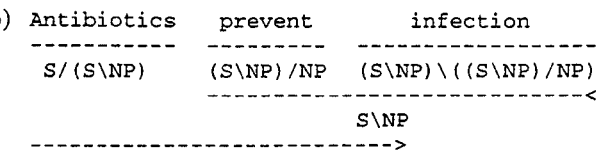
From the examples in the preceding section, it is clear that the unit that we have called the theme is not always a traditional syntactic constituent. Since many problems in the analysis and synthesis of spoken language result from the partial independence of syntactic and intonational phrase boundaries, 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 subconstituencies for sentences, including ones in which the subject and verb of a transitive sentence can exist as a constituent, complete with an interpretation. For example, compare the syntactic deriva-

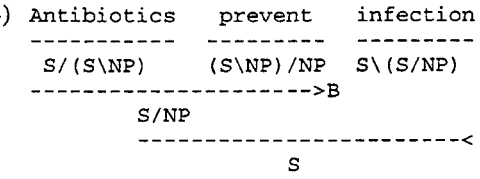
*We are grateful to AT&T Bell Laboratories for allowing us access to the TTS speech synthesiser, and to Mark Beutnagel, Julia Hirschberg, and Richard Sproat for patient advice on its use. The research was supported in part by NSF grant nos. IRI90-18513, IRI90-16592, IRI91-17110 and CISE IIP-CDA-88-22719, DARPA grant no. N00014-90-J-1863, ARO grant no. DAAL03-89-C0031, and grant no. R01-LM05217 from the National Library of Medicine.

[1]The examples used throughout the paper are based on a simple database concerning medical information. This database is in turn modeled on the trauma domain of the TraumAID medical expert system which is under development at Penn ([15]). We ultimately envision constructing a speech output module for this system, which embodies the types of contrastive objects and modifiers that our system handles. The examples given below are artificial, but have been devised on the basis of the system's actual capabilities. The morbid nature of the examples, for which we apologize, is due entirely to the special nature of the trauma domain.

[2]The fraught term "focus" is used throughout this paper strictly in the "narrow" or "phonological" sense in which it refers to the word-level property of bearing a pitch accent.

tions in examples (3) and (4), both of which are licensed by CCG and yield interpretations with identical function-argument structures—*prevent(antibiotics, infection)*.

(3)


(4)


We have argued elsewhere that the notion of constituency espoused by CCG is exactly the same as the notion required to account for prosodic phrasing at all levels, and that the associated information structural categories (like theme, rheme, and focus) are simply the semantic interpretations of surface constituents in this extended sense. We take advantage of this isomorphism between intonational phrasing and CCG constituency by assigning both syntactic and prosodic categories to all lexical items and constituents in the derivation, and then locking the two structural systems together via the following principle.(cf. [12, 13, 10, 11])

(5) 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 enact this condition is to assign functional prosodic categories to constituents bearing pitch accents and argument categories to constituents bearing boundary tones. The theory then allows us to derive a logical form semantics and a representation of information structure for sentences bearing Pierrehumbert-style intonation markings. Although in the interest of brevity we will omit a fuller exposition of the theory, one further point is worth noting. It is often the case that themes are unmarked by any pitch accents or boundary tones. The grammar therefore includes an "Unmarked Theme Promotion Rule" which allows any prosodically unmarked constituent to act as the theme. The ambiguity inherent in such unmarked themes can be resolved only if the actual theme represented in the discourse model can be matched with one of the themes that is non-deterministically proposed by the "Unmarked Theme Promotion Rule."

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.[3]

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 ([3]), and Hirschberg ([4]), 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*:[4]

(6) 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.

For example, in (7) below, the noun *thoracotomy* must remain unstresed while the adjective *left* must be accented in the response, despite having been explicitly mentioned in the text of the question.[5] 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.

(7) 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 (8), in which the adjective is unstressed, is perfectly appropriate:[6]

(8) 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 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

[3] See [10] for an investigation of how much one can get away with on the basis of the question alone.

[4] 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.

[5] 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.

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

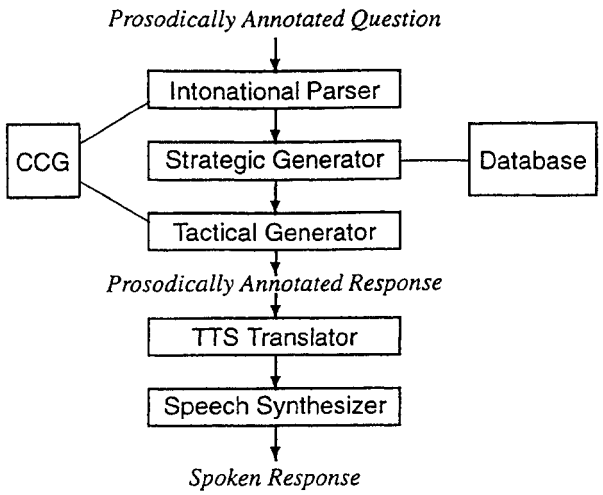


Figure 1: Architecture

situation.

(9) 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.

The algorithm for determining which items are to be stressed for reasons of contrast works as follows.[7] 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.

4 THE IMPLEMENTATION

The present paper is an attempt to apply the theories outlined in the preceding sections to the 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 (10).[8] 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 (11). For the moment, unmarked themes are handled by taking the longest unmarked constituent permitted by the syntax.

(10) I know that burns induce fever,
but WHICH symptoms do LACERATIONS induce?
L+H* LH% H* LL%

(11) prop: $s : \lambda x[symptom(x)\&induce(*lacerations, x)]$
theme: $s : \lambda x[symptom(x)\&induce(*lacerations, x)]/(s : induce(*lacerations, x)/np : x)$
rheme: $s : induce(*lacerations, x)/np : x$

The strategic 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 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 (10), the strategic generator produces the following:

(12) prop: $s : induce(*lacerations, *bleeding)$
theme: $s : induce(*lacerations, x)/np : x$
rheme: $np : *bleeding$

From the output of the strategic generator, the tactical generation module produces a string of words and Pierrehumbert-style markings representing the response, as shown in (13).[9]

(13) lacerations@lhstar induce@lhb bleeding@hstarllb

The final aspect of generation involves translating such a string into a form usable by a suitable speech synthesizer. Currently we use the Bell Laboratories TTS system [7] as a post-processor to synthesise the speech wave. Example (14) shows the translated output for the same example, as it is sent to this synthesiser.

(14) \!*L+H*1 lacerations \!fL1 induce \!pL1 \!bH1
\!*H*3 bleeding \!fL1 \!pL1 \!bL1 \(*[20] \)

5 RESULTS

The system described above produces sharp and natural-sounding distinctions of intonation contour in minimal pairs of queries like those below. Examples (15) and (16) illustrate the system's capability for producing appropriately different into-

[7] 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 [11].

[8] 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.

[9] Full descriptions of the tactical generation algorithm are given in [10] and [11].

nation contours for identical strings 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]

(15) 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%

(16) 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%

Examples (17) and (18) show that the system makes appropriate distinctions in focus placement within themes and rhemes based on context.

(17) 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%

(18) 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%

The issue of focus placement can be crucial in more complex themes and rhemes, as shown below:

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

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

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

6 CONCLUSION

The results show that is 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. 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 entirely 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.

[10]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/eurospeech*.

REFERENCES

[1] Beckman, M. and J. Pierrehumbert: 1986, 'Intonational Structure in Japanese and English', *Phonology Yearbook*, 3, 255-310.

[2] Bird, S.: 1991, 'Focus and phrasing in Unification Categorial Grammar', in S. Bird (ed.), *Declarative Perspectives on Phonology*, Working Papers in Cognitive Science 7, Univ. of Edinburgh. 139-166.

[3] Davis, J. and J. Hirschberg: 1988, 'Assigning Intonational Features in Synthesised Spoken Directions', *Proceedings of the 26th Annual Conference of the ACL*, Buffalo, 187-193.

[4] Hirschberg, J.: 1990, 'Accent and Discourse Context: Assigning Pitch Accent in Synthetic Speech', *Proceedings of AAAI: 1990*.

[5] Houghton, G.: 1986, *The Production of Language in Dialogue: a Computational Model*, unpublished Ph.D dissertation, University of Sussex.

[6] Isard, S. and M. Pearson: 1988 'A Repertoire of British English Intonation Contours for Synthetic Speech', *Proceedings of Speech '88, 7th FASE Symposium, Edinburgh*.

[7] Liberman, M. and A.L. Buchsbaum: 1985, 'Structure and Usage of Current Bell Labs Text to Speech Programs', TM 11225-850731-11, AT&T Bell Laboratories.

[8] Pierrehumbert, J.: 1980, *The Phonology and Phonetics of English Intonation*, Ph.D dissertation, MIT. (Dist. by Indiana University Linguistics Club, Bloomington, IN.)

[9] Pierrehumbert, J. and J. 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.

[10] Prevost, S. and M. Steedman: 1993, 'Generating Contextually Appropriate Intonation', *Proceedings of the Sixth Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, 332–340.

[11] Prevost, S. and M. Steedman: 1993, 'Generating Intonation from Context Using a Combinatory Grammar', manuscript, University of Pennsylvania.

[12] Steedman, M.: 1991, 'Structure and Intonation', *Language*, 68, 260-296.

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

[14] Terken, J.: 1984, 'The Distribution of Accents in Instructions as a Function of Discourse Structure', *Language and Speech*, 27.

[15] Webber, B., 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.

[16] 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.

[17] 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.