

Pragmatic Determinants of Intonation Contours for Dialogue Systems

JUDY DELIN
University of Stirling
jld1@stir.ac.uk

RON ZACHARSKI
IBM
rz@austin.ibm.com

Received April 16, 1995; Accepted September 6, 1996

Abstract. This paper describes an implemented computational model that generates intonation contours for dialogue systems. We concentrate on the relationship between pragmatics and two aspects of intonation: pitch range and pitch accent placement. Pitch range is computed based on the position of an utterance in the discourse structure: utterances that introduce a new topic have an expanded register compared to utterances that continue a topic. Pitch accent placement is based on two pragmatic factors: cognitive status (what the speaker assumes the hearer is attending to) and informativeness (what the speaker assumes to be the interesting or informative component of a phrase). This work suggests that even simple models of discourse topic structure, cognitive status, and informativeness will lead to improved register determination and pitch accent placement in practical conversational systems¹.

Keywords: intonation, pragmatics, concept-to-speech

1. Introduction

The intonation of synthesized speech is often perceived as sounding unnatural. Research suggests that many of the factors that lead to the perception of unnaturalness also affect the intelligibility and comprehensibility of speech (see O'Connell et al., 1968; Luce et al., 1983; Slowiaczek and Nusbaum, 1985; Fowler and Housum, 1987; among others). It is well known that appropriate intonation of an utterance depends on the situation in which that utterance is used. This relationship between situational or pragmatic factors and intonation has been studied from various perspectives in theoretical and computational linguistics (see, for example, Bolinger, 1972, 1986, 1989; Ladd, 1980; Gussenhoven, 1983; Hirschberg and Pierrehumbert, 1986; Steedman, 1991). Although important insights have come out of this work, the exact nature of the pragmatic factors involved in intonation is still a matter of considerable debate.

While implementations of many of these theories have met with some success in text-to-speech systems (most notably, the work of Hirschberg 1992, 1993a), it has been clear for some time that the pragmatic information required for realistic intonation modeling depends on information that is extremely difficult to derive from unrestricted text. However, these theories are easier to implement in CONCEPT-TO-SPEECH systems and these implementations can lead to noticeable improvements in the intonation of such systems². A compelling example of this is illustrated by Davis and Hirschberg (1988) in their "Direction Assistance" system, which produces spoken directions from conceptual representations. Our work represents an extension of this approach to dialogue systems.

This paper describes a system for the production of realistic intonation contours in a dialogue generation/speech synthesis system. We are concentrating on the relationship between pragmatics and two aspects of

intonation: pitch range and pitch accent placement³. For pitch range we compute a starting register for each utterance, based on its position in the discourse structure (represented as a hierarchical organization of super-sentential discourse segments). This position reflects the number of levels pushed or popped since the previous utterance. An additional downstep is included in utterances that (from the speaker's perspective) close the current discourse segment.

The determination of pitch accent location is more complex. Researchers have long suggested that accent placement is determined by givenness—the basic idea is that speakers deaccent expressions that refer to entities that are currently being talked about and accent expressions that refer to entities that are new to the discourse (see, for example, Chafe (1976), Ladd (1980), Bosch (1988), Lambrecht (1985)). We suggest that an additional discourse-pragmatic factor is needed: informativeness. Simply put, speakers, in pursuing a strategy of informativeness, tend to accent constituents that express informative/interesting information and deaccent constituents that express uninformative/uninteresting information. For example, in the exchange

- (1) A: *We're here to deliver the SEven foot concert STEINway.*
 B: *I ordered the NINE foot concert Steinway.*

nine is accented because it is informative in that it distinguishes the piano ordered from the one received, and *foot concert Steinway* is deaccented because it is uninformative—it does not aid in distinguishing the two pianos.

2. Domain

The models of pragmatics and intonation described in this paper are implemented in the BRIDGE speech synthesis system. This system models conversations that occur between pairs of people performing the Map Task (Anderson et al., 1991), a task in which one participant, the route-giver, guides the other, the route-follower, in drawing a route on a map. The route-giver knows the intended route, while the route-follower does not. Neither can see the other's map and there may additionally be mismatches between the landmarks indicated on the map. The BRIDGE system simulates the conversational behavior of both participants in this task, from conceptualizing what is to be said to producing the

required utterance with appropriate intonation⁴. The system as a whole consists of five components:

- a speech act planner (JAM, developed by Carletta (1990)), which determines the propositional content of utterances;
- a pragmatics component based on the notion of a pragmatic file (Heim, 1982; Vallduví, 1990; Zacharski, 1993), which models what the discourse participants are attending to;
- a syntactic planner and generator, which determines how to express a given proposition and creates a syntactic tree for each proposition;
- a contour assignment generator based on a generator developed by Monaghan (1991) for a text-to-speech system, which determines prosodic features such as word stress, pitch accent placement, and pitch range;
- a phoneme-to-speech synthesizer developed by the University of Edinburgh's Centre for Speech Technology Research, which produces synthetic speech.

3. F0 Synthesis

The phoneme-to-speech synthesizer (Campbell et al., 1990) incorporates a model of intonational phonology developed by Ladd (1987). This model is similar to the target and transition approach of Pierrehumbert (1981). The specification of F0 contours occurs in two stages: describing the contour as a sequence of abstract intonational events, and calculating the phonetic values for these events. There are two major types of events: tonal configurations (pitch accents and boundary tones) and register steps. In the current phoneme-to-speech system primary accents are described as H*L, secondaries as H*, and tertiaries as H*H. Final boundaries are specified as L. Register is an F0 band where the pitch accents and boundary tones are realized. Following Ladd (1987) global trends are modeled by changing the register setting for almost every tonal configuration, which eliminates gradience and time dependencies. Register is defined by high, mid, and low lines, which are determined by the following equation:

$$F0 = f_{\min} \cdot f(N) \cdot w^T$$

The speaker-dependent width of the register is represented as w and the high, mid, and low reference lines are specified by T (1, 0, -1, respectively). The

speaker's range is described by the speaker-specific baseline, F_{min} , and the default initial setting, N . The current register setting is represented as $f(N)$, which is defined as

$$f(N) = N \cdot d^i \cdot UR$$

where d is the step size (0.8 in the current model) and i is the number of steps the setting is away from its default value (initially 0). This d^i component handles downstepping and upstepping. UR is the utterance register—the initial register specified for each utterance. The computation of this register is discussed in the following section.

4. Pragmatics and Register

It has often been observed that register (pitch range) is related to discourse structure. Thorsen (1985), for example, has noted that there is a general declination slope across multi-sentential segments and suggests that a hierarchy of textual units is required to account for this overall F0 trend. Brown, Currie et al. (1980) have observed that an utterance that starts a new topic has a larger register than utterances that continue the topic and an utterance that functions to end discussion on the topic has a smaller register than topic-medial utterances. Silverman (1987) has shown that these structure-related variations in pitch range affect the intelligibility of synthetic speech. In BRIDGE, the initial register of each utterance is based on that utterance's relation to the topic structure of the discourse.

BRIDGE, like most dialogue systems, is a planning system. The system starts with a goal—in the case of BRIDGE, the goal is to describe a particular route on a map. It then expands this main goal into a set of subgoals (to describe the first section of the route; to describe the second; and so on). This expansion is recursive—subgoals are expanded into sub-subgoals. This goal structure can be viewed as a tree: the main goal is the root of the tree and the terminal nodes represent the goals associated with individual utterances. The goal tree is built incrementally. At the end of every utterance the system determines what goal to pursue (and whether to expand this goal) based on the current context. BRIDGE performs a simple transformation to convert this tree to a discourse topic tree. For example, the goal tree associated with (2) is shown in Fig. 1 and the discourse topic tree is shown in Fig. 2.

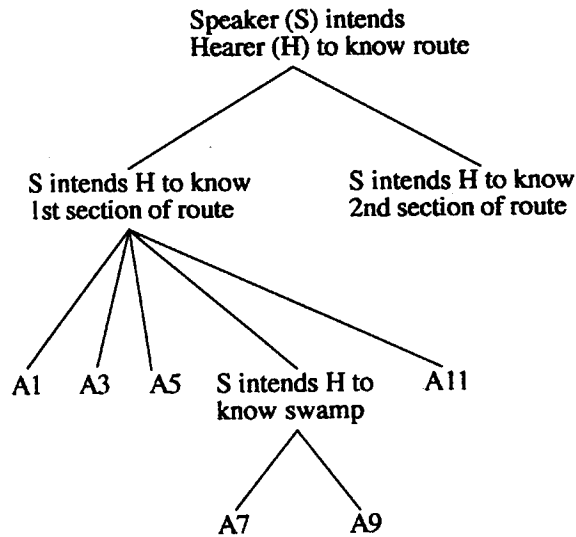


Figure 1. Goal tree associated with Example (2).

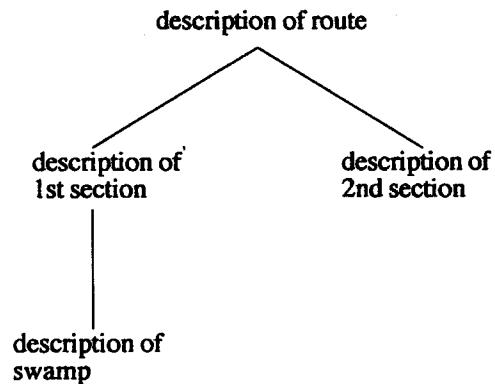


Figure 2. Discourse topic tree associated with Example (2).

- (2) A1: *I want to talk about the first section.*
- B2: *OK.*
- A3: *Do you have the beach?*
- B4: *Yes.*
- A5: *Do you have the swamp?*
- B6: *No.*
- A7: *I want to talk about the swamp.*
- B8: *OK.*
- A9: *The swamp is below the beach.*
- B10: *Right.*
- A11: *The first section goes between the beach and the swamp.*

The algorithm for determining the initial register for an utterance makes use of this discourse topic tree. In a method similar to one proposed by Hirschberg and

Pierrehumbert (1986) register is related to embeddedness of a topic. This embeddedness is represented as a topic level: the main discourse topic ('the description of the route' in the previous example) is a level 0 topic; the immediate subtopics of the main topic ('description of the first section' and 'description of the second section') are Level 1 topics; and so on. The algorithm is as follows:

Algorithm for Utterance Register (UR)

If the current utterance is the initial utterance of a

Level 1 topic

UR = 1.15

otherwise

UR = $0.8^{\text{Level-1}} \cdot \text{FL}$

[where FL (final lowering) is 0.8 if the utterance functions to end discussion on the current topic and is 1.0 otherwise.]

This algorithm encodes several well-known observations about register:

- increasing the register can signal the introduction of a new major topic
- decreasing the register over that of a prior utterance can convey the introduction of a subtopic
- decreasing the register over that of a prior utterance can also convey the closing of a topic

5. Pitch Accent Placement

One fundamental principle embodied in this system is that pitch accent placement is determined by a variety of linguistic factors, from phonological to pragmatic. Some of the factors implemented in BRIDGE include:

- syntactic category (representing a strategy that nouns are more accentable than verbs; verbs more accentable than determiners; etc.)
- linear order (representing a strategy that all things being equal, the last item of a string is more accentable than an inner item)
- cognitive status (representing a strategy that words referring to 'new' discourse entities are more accentable than words referring to 'old' entities)
- informativeness (representing a strategy that words that refer to 'interesting' concepts and properties are more accentable than words referring to uninteresting ones)

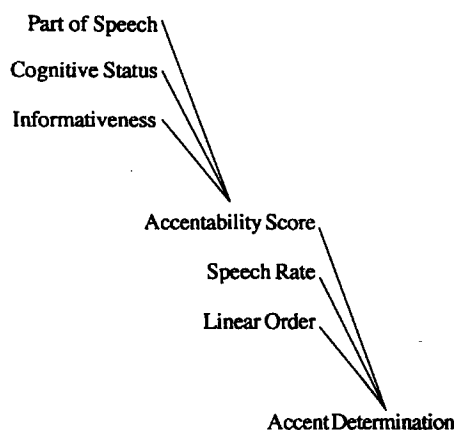


Figure 3. Factors influencing accent determination.

Any one of these factors is not sufficient to uniquely determine the location of pitch accent. For example, knowing the syntactic category of a word (say, a noun) is not sufficient to determine whether that word should receive a pitch accent. Rather, pitch accent location is determined by the interaction of syntactic category, linear order, pragmatic characteristics, as well as other factors. The BRIDGE system utilizes an algorithm developed by Monaghan (1991). This algorithm is roughly illustrated in Fig. 3.

Part of speech, cognitive status, and informativeness are used to calculate an accentability score for each word in the utterance⁵. Speech rate is used to determine prosodic phrases—in fast speech there will be fewer prosodic phrases than there are in slow speech and, as a consequence, there will be fewer accents. A primary accent is computed for each prosodic phrase. Within each phrase, words that share the highest accentability score are considered candidates for accent. The rightmost candidate receives the primary accent (H*L). A rhythm rule places additional accents within the phrase to the left of the primary. This basic procedure illustrated in Fig. 3 is a key thesis of our work. The remainder of the paper describes the two pragmatic factors involved: cognitive status and informativeness.

5.1. Cognitive Status

It is generally accepted in theoretical linguistics that the cognitive status of discourse entities (givenness) plays a major role in pitch accent placement (Faber, 1987; Chafe, 1976; Bosch, 1988). The basic idea is that speakers tend to deaccent expressions that refer to entities that they assume are in the addressee's current

state of attention and accent expressions that refer to entities that are new to the discourse. A particularly clear example of the dominance of this strategy can be seen in (3), in which *the famous springer spaniel* is de-accented because it refers to the dog, Millie, mentioned in the immediately preceding clause:

- (3) *Mr. CLINton appeared to step on Mr. Bush's DOG, MILLie, momentarily, then bent down to PET the famous springer spaniel. So did CHELsea.*
[International Herald Tribune 11 January 1993.47]

In fact, the use of an accent on *the famous springer spaniel* would result in the dispreferred reading that Millie and the springer spaniel are two different dogs:

- (4) *Mr. CLINton appeared to step on Mr. Bush's DOG, MILLie, momentarily, then bent down to pet the famous springer SPANiel.*

Examples such as this suggest that pitch accent does indeed serve as a cue to assist the addressee in determining the cognitive status of discourse entities. In fact, the lack of an accent is more informative than the presence of one, since it indicates the central or salient nature of an entity.

Many computational models of cognitive status are based on a binary distinction—entities are either given or new⁶. The model of cognitive status in BRIDGE is based on the Givenness Hierarchy of Gundel et al. (1993). In this hierarchy there are six linguistically-relevant cognitive statuses which are arranged in a unidirectional entailment relationship. The current BRIDGE model uses four of these statuses: identifiable, familiar, activated, and in-focus⁷. Entities have the status:

identifiable if the speaker can assume that the addressee can determine the class of objects described by that expression or if the speaker can assume that the addressee can either retrieve an existing representation of the speaker's referent or can construct one by the time the utterance has been processed;

familiar if the referent of the expression can be assumed to be already present in the addressee's knowledge store either through recent discourse or through general background information;

activated if the entity is under discussion in the current discourse segment; and

in-focus if the entity can be assumed to be currently at the center of the addressee's attention.

This givenness hierarchy maps to a scale of accentability where expressions referring to in-focus entities are maximally deaccentable and expressions referring to entities which are identifiable but not familiar are maximally accentable:

accent **deaccent**
identifiable > familiar > activated > in-focus

Our computational approach to modelling these statuses is based on the work of Grosz and Sidner (1986)⁸. In their work, the ATTENTIONAL STRUCTURE—that is, the model of what is being attended to at the time, and to what degree—is represented by a stack of focus spaces which contain the discourse entities that are salient at that point in the discourse. The relative salience of particular entities is represented in terms of position in the stack. The INTENTIONAL STRUCTURE—the purpose each segment is intended to achieve, and the hierarchical relationships among these purposes—determine the stack operations of pushes and pops. A pop from the stack would mean that the entities within the focus space were no longer accessible for pronominal or definite reference, requiring explicit reintroduction by a longer description—they are no longer activated. In BRIDGE, the intentional structure is equivalent to the currently active path of the discourse topic tree described in Section 4. Each discourse entity has an activation value which is updated after every utterance. This value is dependent on the entity's past activation value, whether the entity is contained within the current focus space, and the role the entity played in the current utterance. This value is then used to assign the entity to one of the discrete givenness statuses mentioned above.

5.2. Informativeness

A cooperative speaker uses a noun phrase to help the addressee pick out the intended referent. In (5) the speaker uses *it* because she believes the information represented by *it* is sufficient for the addressee to determine what it is that she is talking about.

- (5) *I have been tubed a couple of times and it is uncomfortable going down, but no pain. . .*
[alt.support.eating-disorders]

If she did not believe it was sufficient she would have used a different noun phrase⁹. Many noun phrases describe more than one property of an entity. For example, *the big red ball* describes the entity as having the

properties of being big, red, and a ball. In a given context, some of these properties may be more useful—or informative—in helping the addressee pick out the intended referent. Consider the utterance in (6)

(6) *Give me the red ball.*

Suppose this is uttered in a situation where there are five things on a table: a red sock, a red block, a red apple, a red rose, and a red ball. The noun phrase, *the red ball* describes two properties: being red and being a ball. In this context, the property of being a ball is more informative than being red—the property of being red doesn't distinguish the intended referent from other potential referents and is uninformative¹⁰. The cooperative speaker accents the informative component of the noun phrase and would say:

(7) *Give me the red BALL.*

Consider (6) said in a different context. This time, there are five balls on the table: a black one, a yellow one, a blue one, a green one, and a red one. In this context, the informative property of *the red ball* is the property of being red. Being a ball is uninformative in that it doesn't aid the addressee in distinguishing the intended referent from other potential referents. Once again the speaker accents the word representing the informative description:

(8) *Give me the RED ball.*

This notion of informativeness is intended to capture the idea that, in a given context, some properties may be more useful—or informative—in helping the addressee pick out the intended referent or the intended meaning of a phrase. We suggest that the notion of informativeness is a vital complement to cognitive status.

The idea that informativeness affects pitch accent placement has been suggested by a number of researchers (for example, the notion of 'information' of Hultzén (1956), the notion of 'interest' of Bolinger (1986), the 'news value givenness' of Allerton (1978), the 'pragmatic relations' of Lambrecht (1992), and the 'contrastive properties' of Prevost and Steedman (1994)). In the model described here we develop a definition of informativeness, which incorporates the insights of these researchers and describe how informativeness feeds into the accent placement decision. In the computational model, entities are represented

in a dynamic feature-based taxonomy (a term subsumption language) that represents (1) what properties are normal for each class of entities (NORMS), and (2) what properties distinguish one entity of the class from another (INDICES). To determine accent placement within an expression associated with a particular discourse entity, the system determines the most specific class that contains both this entity and an activated entity, and then marks the norms as uninformative, and the indices as informative. For example, in (10)

- (9) A1: *The first section goes between the rocky beach and the swamp.*
 B2: *I don't understand. Where's the swamp?*
 A3: *The swamp is below the SANDy beach.*

the most specific class that contains the sandy beach and the rocky beach, which has just been activated, is the class of beaches. The property of beachiness is the norm for this class and thus the word *beach* in *sandy beach* is uninformative and deaccented. The property of being sandy distinguishes this particular beach from other activated beaches and thus the word *sandy* is accented. This notion of informativeness plays a crucial role in determining accent placement within an NP. At this point we would like to illustrate how the pragmatic factors we have mentioned affect the output of the BRIDGE system.

6. Example Output

An example of the system's output is shown in (10) (primary accents are described as H*L)¹¹:

- (10) A1: voice(Mary) utterance register(1.15)
 I H*L want to H* talk about the H*L first section LL%
 B2: voice(John) utterance register(1.00)
 H*L alright LL%
 A3: voice(Mary) utterance register(1.00)
 Do you H*H have the rocky H*L beach LL%
 B4: voice(John) utterance register(1.00)
 H*L Yes LL%
 A5: voice(Mary) utterance register(1.00)
 Do you H*H have the H*L swamp LL%
 B6: voice(John) utterance register(1.00)
 H*L No LL%
 A7: voice(Mary) utterance register(0.80)
 I H*L want to H*L talk H*H about the swamp LL%

- B8: voice(John) utterance register(0.80)
H*L OK LL%
- A9: voice(Mary) utterance register(0.64)
The H*L swamp is below the H*L sandy beach LL%
- B10: voice(John) utterance register(0.80)
H*L Right LL%
- A11: voice(Mary) utterance register(0.80)
The H*L first H*H section goes between the swamp and the H*L rocky beach LL%
- B12: voice(John) utterance register(0.80)
H*L Alright LL%
- A13: voice(Mary) utterance register(1.15)
I H*L want to H* talk about the H*L second section LL%

The effects of discourse topic structure on register can be seen throughout this dialogue. For example, the register in A1 and A13 is expanded (“utterance register(1.15)”) because these utterances introduce new major topics. The register in A7 is reduced to indicate the beginning of a subtopic and the register of A9 is reduced to indicate the close of this subtopic. The contribution of cognitive status to deaccenting can be seen in A7, where *swamp* is deaccented because its referent is in-focus (based on its prior mention in A5). Finally, the contribution of informativeness can be seen in A9, where *beach* is deaccented because it does not distinguish the intended referent from the already activated rocky beach. The contribution of informativeness can also be seen in the deaccenting of *section* in A13. In this context, the property of being second is more informative than the property of being a section of a route description¹².

7. Informativeness and Topic-Comment Structure

Following Heim (1982) and Vallduví (1990) an individual’s knowledge store in BRIDGE is modelled as a collection of discourse entities in a file. In the Heim system, a discourse entity corresponds roughly to something that has been introduced by an NP. This decision as to what counts as an entity is well motivated on theoretical grounds. However, we suggest that a more promiscuous ontology will provide a better basis for a model of intonation. In our model, propositions also are discourse entities. For example, the utterance *Kim hugged Lee* is represented by 3 discourse entities: the entity for Kim, the one for Lee and one for the hugging

event itself. The strategy of informativeness applies to all these entities. To appreciate how this expanded ontology relates informativeness to accent placement, it is useful to consult some data presented by Schmerling (1976). Schmerling also takes the view that accentuation does not simply require an assessment of the newness or otherwise of entities, but needs a more subtle appraisal of the role of entities and their semantic relations in utterances. For example, consider (11) and (12):

- (11) *I know who’s standing in front of MARY, but I don’t know who $\left\{ \begin{smallmatrix} \text{MARY's} \\ \text{SHE's} \end{smallmatrix} \right\}$ in front of.*
[Schmerling ex. 142]
- (12) *John insulted MARY, and then SHE insulted HIM.*
[Schmerling ex. 143, Lakoff 1971]

Schmerling correctly notes that even though Mary is referentially given (i.e., ‘in focus’) in the second conjuncts of (11) and (12), the associated referring expressions are accented. Schmerling argues that “it is the newness of the semantic relations that is significant”. In (12) both the predicate, *insulted*, and the participants remain constant in the two conjuncts and do not represent new information. It is the semantic relations that provide new information—Mary changes from being the recipient of an insult to being the giver of one.

This notion of semantic relations is subsumed under the strategy of informativeness. In many cases, a speaker following the strategy of informativeness appears to be answering an implicit question of “What part of the description distinguishes this entity from others?” For example, a speaker chooses to accent *sandy* and deaccent *beach* in the last utterance of

- (13) Ann: *Ben called. You’re to meet him at the rocky beach.*
Clara: *The rocky beach? Where’s that?*
Ann: *North of the SANDy beach.*

because the property of beachiness doesn’t distinguish the sandy beach from other salient entities in the discourse (namely the rocky beach), whereas the property of being sandy does. Similarly, for a discourse entity that represents an eventuality, the speaker determines what distinguishes this eventuality from others. Consider (12), whose discourse entity representation is shown in Fig. 4.

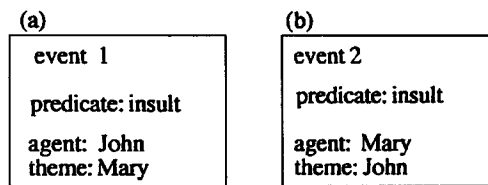


Figure 4. Representation of Example (12).

As Schmerling observes, the change in thematic roles distinguishes these two events. Mary and John are informative in Fig. 4(b) in that they distinguish the event described in (b) from the one described in (a). Thus, their associated referring expressions are accented in (12). It could be argued that this example could be explained by a strategy of deaccenting lexical repetitions—since the verb *insulted* has already been mentioned in the first clause, it is deaccented in the second—and *she* and *him* receive an accent by ‘default’. While this lexical repetition strategy accounts the accent pattern in (11) and (12), it is not a useful strategy for explaining examples such as (14):

- (14) *When I arrived back in Croton I bumped into Laesus, though I had not expected to see him again and HE looked pretty surprised at seeing ME.*
 [Accents marked in original: Lindsey Davis, 1991. *Shadow in bronze*. London: Pan Books. p. 98]

Here *looked pretty surprised at seeing* is deaccented even though it is not a lexical repetition. What is repeated is the meaning—in some sense, *had not expected to see* and *looked pretty surprised at seeing* have a common meaning. The cooperative speaker accents *him* and *me* to help the hearer distinguish the intended event from the one mentioned in the previous clause. In (15) the speaker accents *he* and *her* to assist the hearer in distinguishing the event described from one that the hearer would normally expect:

- (15) *It's not just fans who love her. In a series of sensational concerts in Europe last October, her pianist was none other than the head of the Opéra Bastille in Paris, Myung-Whun Chung, and it was HE who had chosen HER to accompany.*
 [Vanity Fair 56(4).32 April 1993]

Typically, the soloist chooses the accompanist, but the event described goes against this expectation. The speaker helps the hearer make this distinction.

Informativeness correctly accounts for accent placement in all these examples.

7.1. Topic-Comment Structure

In order for an utterance to be relevant to hearers it must ordinarily contain two kinds of information: information that they already possess and on the basis of which a link with existing knowledge can be formed (the ‘topic’), and information that is new, which forms the informative contribution of an utterance (the ‘comment’). This notion of topic-comment structure has been described by a variety of researchers under a number of different names and guises: presupposition-focus of Chomsky (1971) and Jackendoff (1972), open proposition-focus of Ward (1985) and Prince (1986), ground-focus of Vallduví (1990), theme-rheme of Firbas (1972), and topic-comment of Hockett (1958) and Gundel (1974, 1978). While the details of these accounts differ in significant ways, they are all based on intuitions we all share that utterances are ‘about’ something or they ‘link’ up with information the hearer is aware of (the topic) and utterances contain information the speaker is presenting as new to the hearer (the comment)¹³.

It is well known that this distinction is signalled linguistically: in syntax and in prosody (see Gundel, 1988 and chapter 6 of Green, 1989 for reviews of this literature). Several researchers have noted the relationship between topic-comment structure and pitch accent placement. For example, Gundel (1978) argues that “primary stress in a sentence always falls on an element within the comment” and Vallduví (1993) notes that “in English, as in Catalan, (a subset of) the [comment] is marked in situ by intonational prominence”. We support this view that there must necessarily be an accent in that part of the utterance that represents the comment¹⁴.

While the cognitive status of discourse entities (‘givenness’) is solely determined by a speaker’s beliefs about what the addressee is attending to, the determination of topic-comment structure is only partly constrained by these beliefs, since it is also a reflection of what the speaker is interested in (see, for example, Bolinger, 1986). Topic-comment structure is, to a certain extent, a matter of choice. Therefore, the development of a complete and accurate computational model of topic-comment structure would require the daunting task of developing an explicit model of speaker’s interests, which is certainly beyond our current capabilities.

However, significant results in terms of generating appropriate prosodic behavior can be achieved by a simple, less ambitious model. This first approximation uses just a single factor, 'informativeness', to determine topic-comment structure. In our current model, the topic-comment structure of an utterance is defined as the primary division of an utterance into informative and uninformative components. For example, in B's reply in (16)

- (16) A: *Are we learning the fanDANgo or the taran-TELLa in class today?*
 B: *We're learning* [comment *the fanDANgo.*]

the information represented by *the fandango* is the comment because it is informative—it distinguishes the proposition B expressed from A's proposition. The information expressed by the rest of B's utterance (*we're learning*) is uninformative (relative to the comment). (A similar argument can be given for (16B) when it is used as an answer to the *wh*-question *What are we learning in class today?*) Thus, informativeness contributes to the determination of the highest level division of an utterance—topic-comment structure—and the computational approach described in the previous sections can be used to determine this division.

This simple, informativeness-based, model of topic-comment structure works surprisingly well in our speech synthesis system. However, it is just a first approximation of a topic-comment model and subsequent revisions of the model will require additional features. We do not wish to suggest that topic-comment structure can be completely explained by informativeness. While topic-comment structure is to a large extent determined by informativeness, the two notions are distinct. Consider Clara's last response in (17):

- (17) Ann: *What did you get Ben for Christmas?*
 Clara: *I got him a blue SHIRT.*
 Ann: *What did you get Diane?*
 Clara: *I got her* [comment *a RED shirt.*]

The accent on *red* and the lack of an accent on *shirt* are not due to there being a 'narrow comment' (or 'narrow focus') on *red*. The comment is represented by the entire phrase *a red shirt*¹⁵. The relationship between comment and pitch accent is simply that there must be an accent within that part of the utterance that represents the comment. Thus, topic-comment tells us that there must be an accent somewhere within *a red*

shirt. It is informativeness that specifies that *shirt* be deaccented and *red* accented—in this context *red* is informative and *shirt* is not.

Finally, it is interesting to note that the accenting/deaccenting strategy based on informativeness is not available in all languages. Topic-comment structure is the primary division of an utterance and all languages have some way of encoding this structure (see Gundel, 1988). However, the informativeness of individual components within a comment is not always grammatically realized in a language by deaccenting. Whereas English pitch accent placement relies heavily on the strategy of informativeness at the subcomment level, other languages, such as Spanish, Catalan, and Italian, do not. Consider the Spanish example (18), which is similar to (17):

- (18)
 Ana: *Què le regaLAsTe a benjaMÌN por naviDAdes?*
 what i-obj-2s give-2s-past to B. for Xmas?
 "What did you get Benjamin for Christmas?"

Clara: *le regaLÈ una caMIsa NEgra.*
 iobj-2s give-1s-past a-fem shirt black-fem
 "I gave him a black shirt."

Ana: *y a diAna QUÈ le regaLAsTe?*
 And to D. what iobj-2s give-2s-past
 "And what did you get Diane?"

Clara: *Unos pantalones NEgros.*
 a-pl pants black-pl
 "Black pants."

In Clara's last response *Unos pantalones negros* is the grammatical realization of the comment. Of the two properties [being black], and [being pants], [being pants] is the most informative, since blackness does not distinguish this gift from the one given to Ben. Thus, based on the rules presented here for English, we would predict that *negros* would be deaccented and *pantalones* accented¹⁶. However, this is not the case—accent simply goes on the last word in the comment. While English speakers make use of the strategy of informativeness to constrain the location of pitch accents (as illustrated in (17)), Spanish speakers (and Catalan and Italian speakers as well) do not. This offers strong evidence for the position that topic-comment and informativeness are separate notions.

In sum—and returning to practical computational concerns—it is difficult to develop an explicit, comprehensive model of topic-comment. As Bolinger (1986) has noted, topic-comment is ultimately determined by

a speaker's interests. We have described a simple, rudimentary, model of pragmatics and intonation—a model we have implemented in a dialogue generation/speech synthesis system. This work suggests that even simple models of discourse topic structure, cognitive status, and informativeness, will lead to improved register determination and pitch accent placement in conversational systems.

Notes

1. We are deeply grateful to the other members of the BRIDGE project, Bob Ladd and Alex Monaghan. We would like to thank the anonymous reviewers for providing helpful comments on an earlier version of this paper. Special thanks also to Jeanette Gundel, Eric Vallduvi, Ellen Gurman Bard, and Thorstein Freitheim for their critical comments and suggestions at various stages of this project. This research was funded by the UK Economic and Social Research Council, Grant Number R000 23 3460 to Edinburgh University.
2. Text-to-speech systems take written text as input and 'speak' that text. Concept-to-speech systems (or "synthesis from concept" systems) take information expressed in some knowledge representation language and decide both what to say (for example, to decide to express the proposition [ben hit ball]) and how to grammatically realize this utterance (whether to realize [ben hit ball] as *Ben hit the ball*, *He hit it*, *It was Ben who hit the ball*, *The ball was hit*, etc.). Thus, concept-to-speech systems can be viewed in Levelt's (1989) terms as consisting of three components: a conceptualizer, which plans the conceptual representation of the message—the 'preverbal message'; a formulator, which uses lexical knowledge to encode the preverbal message into a grammatical/phonological structure; and an articulator, which realizes this phonological structure as speech.
3. The focus on pitch range and pitch accent placement, and the exclusion of other prosodic features from our study, stemmed from our personal interests and our efforts to keep the research manageable. We recognize that other aspects of intonation, such as tune type (pitch accent type and boundary tones), rhythm, pause duration, and amplitude, are also affected by pragmatic factors. See Hirschberg (1993b) for a review of the relevant research. Also see Brown et al. (1980), Bolinger (1986), Campbell (1992), Hirschberg and Grosz (1992), Hirschberg and Ward (1992), Gundel et al. (1995).
4. This illustrates yet another difference between text-to-speech and concept-to-speech systems. Most commercial text-to-speech systems speak arbitrary text. However, BRIDGE, like other concept-to-speech systems, only talks about one thing—in BRIDGE's case, map routes. The system contains rules specific to constructing good route descriptions and the lexicon is specifically built for this task. The rules and lexicon would need to be altered to port BRIDGE to different applications. However, the pragmatic algorithms described in this paper are general purpose and are applicable to a wide range of domains.
5. The accentability score is the product of two values: syntactic weight and pragmatic weight. The syntactic weight of a word is 4 if that word is a noun, 3 if it is an adjective, verb (excluding auxiliary verbs), or adverb, 2 if it is a preposition, and 0

otherwise. The pragmatic weight of a word is 4 if that word is only identifiable, 3 if it is only familiar, 2 if it is only activated, and 1 if it is either in focus or uninformative (these terms will be explained in Section 5.1).

6. Examples of binary systems include the accessible entities of Youd and House (1991), the notion of salience described in Davis and Hirschberg (1988), and focus set membership described in Hirschberg (1990).
7. The BRIDGE model conflates three of the original six statuses: type identifiable, referential, and identifiable. This has not been done for any theoretical reason (for example, to argue that four statuses are sufficient for explicating the distribution and use of pitch accents in English). Rather, it is for a methodological reason since developing an explicit characterization of the differences between these three statuses is extremely difficult. We have left both this characterization and an examination of its effects on prosody as a subject of future research.
8. See also Linde's (1974) notion of "focus of attention" and Reichman's (1985) seminal work on context spaces.
9. This notion of informativeness follows from Grice's maxim of Quantity (1975):

Q1: "Make your contribution as informative as required (for the current purposes of the exchange)."

Q2: "Do not make your contribution more informative than is required."

Dale (1989) refers to Q1 and Q2 as informational adequacy and informational efficiency—a NP is informational adequate if it enables the hearer to uniquely identify the intended referent and it is sufficient if it is not more informative than necessary (See Passonneau (1995), Dale and Reiter (1995) for more information). As Clark and Wilkes-Gibbs (1990) correctly point out, the informational adequacy condition does not always hold. For example, in spontaneous conversations a speaker may not have enough time to plan and construct a noun phrase that meets this condition.

10. By *uninformative* we do not mean that the information is necessarily redundant and that it can be elided. However, this may be the case as in

i. [at an ice cream shop]

A: *Can I interest you in some ice cream?*

B: *Yes. I'd like a liter of { vaNILla ice cream }
vaNILla*

11. Although formal evaluation of the output of the system has yet to be completed, initial impressions are of an improvement in naturalness and comprehensibility. We have informally compared the output of the system with 'pragmatics on' and 'pragmatics off' by playing two versions of the same dialogues to several audiences, and have received uniformly positive responses to the effect that the 'pragmatics on' condition sounds more natural and coherent.
12. In A1 *section* is also deaccented. In preparing to utter A1 the system knows that it will mention other sections of the route and attempts to distinguish the intended referent from other referents that will be introduced. Bob Ladd has termed this "anticipatory deaccenting". Other examples of this are the deaccenting of *correct* in (i) and the deaccenting of *women* in (ii).
 - i. *A decade later than MOST of my peers, I'd endured STANford University, MALhousie LAW School, and TWO LEgal*

asSOciate jobs—one poLItically correct, one FIScally correct. Maybe I needed POT to help me put UP with the bullshit. [Lia Matera. 1991. *Prior Convictions*. p. 11]

- ii. [Description of conference] Anyway, no DIScourse work. There was one aMERican woman and one Irish woman presenting. [email]

13. This notion of topic is distinct from the notion of discourse topic (van Oosten, 1986) described in Section 4. See Gundel (1988) for more discussion. Topics need not be overtly represented in an utterance.
14. It is not the case that topics are always unaccented. See Gundel (1978) and Gundel et al. (1995).
15. See Ladd (1980) for a similar discussion.
16. To get the full effect of this, consider the following English-like equivalent of (18) (accepting N Adj word order):

Ann: What did you get Ben for Christmas?

Clara: I got him shirt BLACK.

Ann: What did you get Diane?

Clara: Pants BLACK.

Native English speakers would find the accent on *black* in Clara's last response extremely odd. English speakers would say *PANTS black*. Spanish, Catalan, and Italian speakers say *pants BLACK*. (*pantalones NEGROS*, *pantalons NEGRES*, and *pantaloni NERI*, respectively). However, in these languages informativeness may be realized by other prosodic means. See Vallduvfi and Zacharski (1994) for more information.

References

- Allerton, D.J. (1978). The notion of 'givenness' and its relations to presupposition and to theme. *Lingua*, 44:133–168.
- Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., Mcallister, J., Miller, J., Sotillo, C., Thompson H.S., and Weinert, R. (1991). The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366.
- Bolinger, D. (1972). Accent if predictable (if you're a mind reader). *Language*, 48(3):633–644.
- Bolinger, D. (1986). *Intonation and Its Parts: Melody in Spoken English*. Stanford: Stanford University Press.
- Bolinger, D. (1989). *Intonation and Its Uses: Melody in Grammar and Discourse*. Stanford: Stanford University Press.
- Bosch, P. (1988). Representing and accessing focussed referents. *Language and Cognitive Processes*, 3(3):207–231.
- Brown, G., Currie, K.L., and Kenworthy, J. (1980). *Questions of Intonation*. London: Croom Helm.
- Campbell, W.N. (1992). Multi-level timing in speech. Sussex University dissertation.
- Campbell, W.N., Isard, S.D., Monaghan, A.I.C., and Verhoeven, J. (1990). Duration, pitch, and diphones in the CSTR TTS system. *Proceedings of the International Conference on Spoken Language Processing*, pp. 825–828.
- Carletta, J. (1990). Modeling variations in goal-directed dialogue. *Proceedings of the International Conference on Computational Linguistics*, 13:324–326.
- Chafe, W. (1976). Givenness, Contrastiveness, subjects, topics, and point of view. In C.N. Li (Ed.), *Subject and Topic*. New York: Academic Press, pp. 25–55.
- Chomsky, N. (1971). Deep structure, surface structure, and semantic interpretation. In R. Jacobs and P. Rosenbaum (Eds.), *Semantics*. Cambridge: Cambridge University Press, pp. 183–216.
- Clark, H.H. and Wilkes-Gibbs, D. (1990). Referring as a collaborative process. In P.R. Cohen, J. Morgan, and M.E. Pollack (Eds.), *Intentions in Communication*. Boston: MIT Press, pp. 463–493.
- Dale, R. (1989). Cooking up references. *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, vol. 27, pp. 68–75.
- Dale, R. and Reiter, E. (1995). Computational interpretations of the Gricean Maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.
- Davis, J.R. and Hirschberg, J. (1988). Assigning intonational features in synthesized spoken directions. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 26:187–193.
- Faber, D. (1987). Some problems of English nucleus placement. University of Manchester dissertation.
- Firbas, J. (1972). On the interplay of prosodic and non-prosodic means of functional sentence perspective (A theoretical note on the teaching of English intonation). In V. Fried (Ed.), *The Prague School of Linguistics and Language Teaching*. London: Oxford University Press, pp. 77–94.
- Fowler, C.A. and Housum, J. (1987). Talkers' signalling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26:489–504.
- Green, G.M. (1989). *Pragmatics and Natural Language Understanding*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Grice, H.P. (1975). Logic and conversation. In P. Cole and J. Morgan (Eds.), *Speech Acts*. New York: Academic Press, pp. 41–58.
- Grosz, B.J. and Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.
- Gundel, J.K. (1974). The Role of Topic and Comment in Linguistic Theory, University of Texas dissertation.
- Gundel, J.K. (1978). *Stress, Pronominalization and the Given-New Distinction*, University of Hawaii working papers in linguistics NTIS, 10(2):1–13.
- Gundel, J.K. (1988). Universals of topic-comment structure. In M. Hammond, E.A. Moravcsik, and J.R. Wirth (Eds.), *Studies in Syntactic Typology*. Amsterdam: John Benjamins Publishing, pp. 209–239.
- Gundel, J.K., Hedberg, N., and Zacharski, R. (1993). Cognitive status and the form of referring expressions in discourse. *Language*, 69:274–307.
- Gundel, J.K., Hedberg, N., and Zacharski, R. (1995). Prosodic tune and information structure. *Proceedings of the Annual Meeting of the Canadian Linguistic Association*.
- Gussenhoven, C. (1983). Focus, mode, and the nucleus. *Journal of Linguistics*, 19:377–417.
- Heim, I.R. (1982). The Semantics of Definite and Indefinite Noun Phrases. University of Massachusetts dissertation.
- Hirschberg, J. (1990). Accent and discourse context: assigning pitch accent in synthetic speech. *Proceedings of the National Conference on Artificial Intelligence*, 8(2):952–957.
- Hirschberg, J. (1992). Using discourse context to guide pitch accent decisions in synthetic speech. In G. Bailly, C. Benoit and T.R. Sawallis (Eds.), *Talking Machines: Theories, Models, and Designs*. Amsterdam, Elsevier Science Publishers B.V., pp. 367–376.

- Hirschberg, J. (1993a). Pitch accent in context: predicting intonational prominence from text. *Artificial Intelligence*, 63(1):305–340.
- Hirschberg, J. (1993b). Studies of intonation and discourse. *Proceedings of the ESCA Workshop on Prosody*, vol. 41, pp. 90–95.
- Hirschberg, J. and Grosz, B. (1992). Intonational features of local and global discourse structure. *Proceedings of the Speech and Natural Language Workshop*, pp. 441–446.
- Hirschberg, J. and Pierrehumbert, J. (1986). The intonational structuring of discourse. *Proceedings of the Annual Meeting of the Association of Computational Linguistics*, 24:136–144.
- Hirschberg, J. and Ward, G. (1992). The influence of pitch-range, duration, amplitude, and spectral features on the interpretation of rise-fall-rise intonation patterns in English. *Journal of Phonetics*, 20:241–252.
- Hockett, C.A. (1958). *A Course in Modern Linguistics*. New York: Macmillan.
- Hultzén, L.S. (1956). The poet Burns' again. *American Speech*, 31:195–201.
- Jackendoff, R.S. (1972). *Semantic Interpretation in Generative Grammar*. Cambridge: MIT Press.
- Ladd, D.R. (1980). *The Structure of Intonational Meaning*. Bloomington and London: Indiana University Press.
- Ladd, D.R. (1987). A model of intonational phonology for use in speech synthesis by rule. *Proceedings of the European Conference on Speech Technology*, pp. 21–24.
- Lakoff, G. (1971). Presupposition and relative well-formedness. In D.D. Steinberg and L.A. Jakobovits (Eds.), *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology*. Cambridge: Cambridge University Press, pp. 329–340.
- Lambrecht, K. (1992). Sentential-focus structures as grammatical constructions. Paper presented at the Linguistic Society of America Annual Meeting, ms.
- Levelt, W.J.M. (1989). *Speaking: From Intention to Articulation*. Cambridge, MA: MIT Press.
- Linde, C. (1979). Focus of attention and the choice of pronouns in discourse. In T. Givón (Ed.), *Discourse and Syntax*. New York: Academic Press, pp. 337–354.
- Luce, P.A., Fuestel, T.C., and Pisoni, D.B. (1983). Capacity demands in short-term memory for synthetic and natural speech. *Human Factors*, 25(1):17–32.
- Monaghan, A.I.C. (1991). Intonation in a text-to-speech conversion system. University of Edinburgh dissertation.
- O'Connell, D.C., Turner, E.A., and Onuska, L.A. (1968). Intonation, grammatical structure, and contextual association in immediate recall. *Journal of Verbal Learning and Verbal Behavior*, 7:110–116.
- Passonneau, R.J. (1995). Integrating Gricean and Attentional Constraints. *Proceedings of the International Joint Conference on Artificial Intelligence*, 14:1267–1273.
- Pierrehumbert, J.B. (1981). Synthesizing intonation. *Journal of the Acoustic Society of America*, 70(4):985–995.
- Prevost, S. and Steedman, M. (1994). Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.
- Prince, E.F. (1986). On the syntactic marking of presupposed open propositions. *Chicago Linguistic Society*, 22:208–222.
- Reichman, R. (1985). *Getting Computers to Talk Like you and me: Discourse Context, Focus, and Semantics (An ATN Model)*. Cambridge, MA: MIT Press.
- Schmerling, S.F. (1976). *Aspects of English Sentence Stress*. Austin: University of Texas Press.
- Silverman, K. (1987). The structure and processing of fundamental frequency contours. Cambridge University dissertation.
- Slowiaczek, L.A. and Nusbaum, H.C. (1985). Effects of speech rate and pitch contour on the perception of synthetic speech. *Human Factors*, 27(6):701–711.
- Steedman, M. (1991). Structure and intonation. *Language*, 67:260–296.
- Thorsen, N. (1985). Intonation and text in Standard Danish. *Journal of the Acoustical Society of America*, 77:1205–1216.
- Vallduví, E. (1990). The informational component. University of Pennsylvania dissertation.
- Vallduví, E. (1993). Information packaging: A survey, ms.
- Vallduví, E. and R. Zacharski (1994). Accenting phenomena, association with focus, and the recursiveness of focus-ground. *Proceedings of the Amsterdam Colloquium*, 9:683–702.
- van Oosten, J. (1986). The nature of subjects, topics and agents: A cognitive explanation. University of California dissertation.
- Ward, G.L. (1985). The semantics and pragmatics of preposing. University of Pennsylvania dissertation.
- Youd, N. and House, J. (1991). Generating intonation in a voice dialogue system. *European Conference on Speech Technology*, 3:1287–1290.
- Zacharski, R.A. (1993). A discourse pragmatics model of pitch accent in English. University of Minnesota dissertation.