Assigning Intonational Features in Synthesized Spoken Directions *

James Raymond Davis
The Media Laboratory
MIT E15-325
Cambridge MA 02139

Julia Hirschberg
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974

Abstract

Speakers convey much of the information hearers use to interpret discourse by varying prosodic features such as PHRASING, PITCH ACCENT placement, TUNE, and PITCH RANGE. The ability to emulate such variation is crucial to effective (synthetic) speech generation. While text-to-speech synthesis must rely primarily upon structural information to determine appropriate intonational features, speech synthesized from an abstract representation of the message to be conveyed may employ much richer sources. The implementation of an intonation assignment component for Direction Assistance, a program which generates spoken directions, provides a first approximation of how recent models of discourse structure can be used to control intonational variation in ways that build upon recent research in intonational meaning. The implementation further suggests ways in which these discourse models might be augmented to permit the assignment of appropriate intonational features.

Introduction

DIRECTION ASSISTANCE[1] was written to provide spoken directions for driving between any two points in the Boston area[7] over the telephone. Callers specify their origin and destination via touch-tone input. The program finds a route and synthesizes a spoken description of that route. Earlier versions of Direction Assistance exhibited notable deficiencies in prosody when a simple text-to-speech system was used to produce such descriptions[6], because prosody depends in part on discourse-level phenomena such as topic structure and information status which are not generally inferrable from text, and thus cannot be correctly produced by the text to speech system.

To alleviate some of these problems, we modified Direction Assistance to make both attentional and intentional information about the route description available for the assignment of intonational features. With this information, we generate spoken directions using the Bell Laboratories Text-to-Speech System[21] in which pitch range, accent placement, phrasing, and tune can be varied to communicate attentional and intentional structure. The implementation of this intonation assignment component provides a first approximation of how recent models of discourse structure can be used to control intonational variation in ways that build upon recent research in intonational meaning. Additionally, it suggests ways in which these discourse models must be enhanced in order to permit the assignment of appropriate intonational features.

In this paper, we first discuss some previous attempts to synthesize speech from representations other than simple text. We next discuss the work on discourse structure, on English phonology, and on intonational meaning which we assume for this study. We then give a brief overview of Direction Assistance. Next we describe how Direction Assistance represents discourse structures and uses them to generate appropriate prosody.

Previous Studies

Only a few voice interactive systems have attempted to exploit intonation in the interaction. The Telephone Enquiry Service (TES) [19] was designed as a framework for applications such as database inquiries, games, and calculator functions. Application programmers specified text by phonetic symbols and intonation by a code which extended Halliday's[11] intonation scheme. While TES gave programmers a high-level means of varying prosody, it made no attempt to derive prosody automatically from an abstract representation.

Young and Fallside's[20] Speech Synthesis from Concept (SSC) system first demonstrated the gains to be had by providing more than simple text as input to a speech

*The intonational component described here was completed at AT&T Bell Laboratories in the summer of 1987. We thank Janet Pierrehumbert and Gregory Ward for valuable discussions.

[1] Direction Assistance was originally developed by Jim Davis and Tom Trobaugh in 1985 at the Thinking Machines Corporation of Cambridge.

synthesizer. SSC passed a network representation of syntactic structure to the synthesizer. Syntactic information could thus inform accenting and phrasing decisions. However, structural information alone is insufficient to determine intonational features[10], and SSC does *not* use semantic or pragmatic/discourse information.

Discourse and Intonation

The theoretical foundations of the current work are three: Grosz and Sidner's theory of discourse structure, Pierrehumbert's theory of English intonation, and Hirschberg and Pierrehumbert's studies of intonation and discourse.

Modeling Discourse Structure

Grosz and Sidner[9] propose that discourse be understood in terms of the purposes that underly it (INTENTIONAL STRUCTURE) and the entities and attributes which are salient during it (ATTENTIONAL STRUCTURE). In this account, discourses are analyzed as hierarchies of segments, each of which has an underlying **Discourse Segment Purpose** (DSP) intended by the speaker. All DSPs contribute to the overall **Discourse Purpose** (DP) of the discourse. For example, a discourse might have as its DP something like 'intend that Hearer put together an air compressor', while individual segments might have as contributing DSP's 'intend that Hearer remove the flywheel' or 'intend that Hearer attach the conduit to the motor'. Such DSP's may in turn be represented as hierarchies of intentions, such as 'intend that a hearer loosen the allen-head screws', and 'intend that Hearer locate the wheel-puller'. DSPs **a** and **b** may be related to one another in two ways: **a** may DOMINATE **b** if the DSP of **a** is partially fulfilled by the DSP of **b** (equivalently, **b** CONTRIBUTES TO **a**). So, 'intend that Hearer remove the flywheel' dominates 'intend that Hearer loosen the allen-head screws', and the latter contributes to the former. Segment **a** SATISFACTION-PRECEDES **b** if the DSP of **a** must be achieved in order for the DSP of **b** to be successful. 'Intend that Hearer locate the wheel-puller' satisfaction-precedes 'intend that Hearer use the wheel-puller', and so on. Such intentional structure has been studied most extensively in task-oriented domains, such as instruction in assembling machinery, where speaker intentions appear to follow the structure of the task to some extent. In Grosz and Sidner's model, part of understanding a discourse is reconstructing the DP, DSPs and relations among them.

Attentional structure in this model is an abstraction of 'focus of attention', in which the set of salient entities changes as the discourse unfolds.[2] A given discourse's attentional structure is represented as a stack of FOCUS SPACES, which contain representations of entities referenced in a given DS, such as 'flywheel' or 'allen-head screws', as well as the DS's DSP. The accessibility of an entity — as, for pronominal reference — depends upon the depth of its containing focus space. Deeper spaces are less accessible. Entities may be made inaccessible if their focus space is popped from the stack.

Intonational Features and their Interpretation

This model of discourse is employed for expository purposes by Hirschberg and Pierrehumbert[12] in their work on the relationship between intonational and discourse features. In Pierrehumbert's theory of English phonology[16], intonational contours are represented as sequences of high (**H**) and low (**L**) tones (local maxima and minima) in the FUNDAMENTAL FREQUENCY (f0). Pitch accents fall on the stressed syllables of some lexical items, and may be simple **H** or **L** tones or complex tones. The four bitonal accents in English (**H*+L, H+L*, L*+H, L+H***) differ in the order of tones and in which tone is aligned with the stressed syllable of the accented item — the asterisk indicates alignment with stress. Pitch accents mark items as intonationally prominent and convey the relative 'newness' or 'salience' of items in the discourse. For example, in (1a), *right* is accented (as 'new'), while in (1b) it is deaccented (as 'old').

(1) a. Take a right, onto Concord Avenue.
    b. Take another right, onto Magazine Street.

Different pitch accents convey different meanings: For example, a **L+H*** on *right* in (1a) may convey 'contrastiveness', as after the query *So, you take a left onto Concord?*. A simple **H*** is more likely when the direction of the turn has *not* been questioned. A **L*+H**, however, can convey incredulity or uncertainty about the direction.

INTERMEDIATE PHRASES are composed of one or more pitch accents, plus an additional PHRASE ACCENT (**H** or **L**), which controls the pitch from the last pitch accent to the end of the phrase. INTONATIONAL PHRASES consist of one or more intermediate phrases, plus a BOUNDARY TONE, also **H** or **L**, which falls at the edge of the phrase;

[2]See [1] and [3] for earlier AI work on global and local focus.

we indicate boundary tones with an '%', as **H%**. Phrase boundaries are marked by lengthened final syllables and (perhaps) a pause — as well as by tones. Variations in phrasing may convey structural relationships among elements of a phrase. For example, (2) uttered as two phrases favors a non-restrictive reading in which the first right happens to be onto Central Park.

(2) Take the first right [,] onto Central Park.

Uttered as a single phrase, (2) favors the restrictive reading, instructing the driver to find the first right which goes onto Central Park.

TUNES, or intonational contours, have as their domain the intonational phrase. While the meaning of tunes appears to be compositional — from the meanings of their pitch accents, phrase accents, and boundary tones[15], certain broad generalizations may be made about particular tunes in English. Phrases ending in **L H%** appear to convey some sense that the phrase is to be completed by another phrase. Phrases ending in **L L%** appear more 'declarative' than 'interrogative' phrases ending in **H H%**. Phrases composed of sequences of **H*+L** accents are often used didactically.

The PITCH RANGE of a phrase is (roughly) the distance between the maximum f0 value in the phrase (modulo segmental effects and FINAL LOWERING effects) and the speaker's BASELINE, defined for each speaker as the lowest point reached in normal speech over all utterances. Variation in pitch range can communicate the topic structure of a discourse[12, 18]; increasing the pitch range of a phrase over prior phrases can convey the introduction of a new topic, and decreasing the pitch range over a prior phrase can convey the continuation of a subtopic. After any bitonal pitch accent pitch range is compressed. This compression, called **catathesis**, or **downstep**, extends to the nearest phrase boundary. Another process, called FINAL LOWERING, involves a compression of the pitch range during the last half second or so of a 'declarative' utterances. The amount of final lowering present for utterance appears to correlate with the amount of 'finality' to be conveyed by the utterance. That is, utterances that end topics appear to exhibit more final lowering, while utterances within a topic segment may have little or none.

Intonation in Direction-Giving

To identify potential genre-specific intonational characteristics of direction-giving, we performed informal production studies, with speakers reading sample texts of directions similar to those generated by Direction Assistance. From acoustic analysis of this data, we noted first that speakers tended to use **H*+L** accents quite frequently, in utterances like that whose pitch track appears in Figure 1. The use of such contours has been associated in the literature with 'didactic' or 'pedantic' contexts. Hence, the propensity for using this contour in giving directions seems not inappropriate to emulate.

We also noted tendencies for subjects to vary pitch range in ways similar to proposals mentioned above — that is, to indicate large topic shifts by increasing pitch range and to use smaller pitch ranges where utterances appeared to 'continue' a previous topic. And we noted variation in pausal duration which was consistent with the notion that speakers produce longer pauses at major topic boundaries than before an utterance that continues a topic. However, these informal studies were simply intended to produce guidelines.

In the intonation assignment component we added to Direction Assistance, pitch accent placement, phrasing, tune, and pitch range and final lowering are varied as noted above to convey information status, structural information, relationships among utterances, and topic structure. We will now describe how Direction Assistance works in general, and, in particular, how it uses this component in generating spoken directions.

Direction Assistance

Direction Assistance has four major components. The **Location Finder** queries the user to obtain the origin and destination of the route. The **Route Finder** then finds a 'best' route, in terms of drivability and describability. Once a route is determined, the **Describer** generates a text describing the route, which the **Narrator** reads to the user. In the work reported here, we modified the Describer to generate an abstract representation of the route description and replaced the Narrator with a new component, the **Talker**, which computes prosodic values from these structures and passes text augmented with commands controlling prosodic variation to the speech synthesizer.

Generating text and discourse structures

The Describer's representation of a route is called a **tour**. A tour is a sequence of **acts** to be taken in following the

Figure 1: Pitch Track of Subject Reading Directions

route. Acts represent something the driver must do in following the route. Act types include **start** and **stop**, for the beginning and ending of the tour, and various kinds of turns. A rich classification of turns is required in order to generate natural text. A 'fork' should be described differently from a 'T' and from a highway exit. Turning acts include **enter** and **exit** from a limited access road, **merge**, **fork**, **u-turn**, and **rotary**.

For each act type, there is a corresponding descriptive schema to produce text describing that act. Text generation also involves selecting an appropriate cue for the act. There are four types of cues: *Action* cues signal when to perform an act, such as "When you reach the end of the road, do x". *Confirmatory* cues are indicators that one is successfully following the route, such as "You'll cross x" or "You'll see y". *Warning* cues caution the driver about possible mistakes. *Failure* cues to describe the consequences of mistakes (e.g. "If you see x, you have gone too far") have not yet been implemented. In general, there will be several different items potentially useful as action or confirmatory cues. The Describer selects the one which is most easily recognized (e.g. a bridge crossing) and which is close to the act for which it is a cue.

Descriptive schemas are internally organized into syntactic constituents. Some constituents are constant, and others, e.g. street names and direction of turns, are slots to be filled by the Describer from the tour. Constituents are further grouped into one or more (potential) intonational phrases. Each phrase will have a pitch range, a preceding pause duration, a phrase accent, and a boundary tone assigned by the Talker. Phrases that end utterances will also have a final lowering percentage. Where schemas include more than one intonational phrase, relationships among these phrases are documented in the schema template so that they may be preserved when intonational features are assigned.

Intentional structure is also represented at the level of the intonational phrase. Unlike in Grosz and Sidner's model, a single phrase may represent a discourse segment. This departure stems from our belief that, following [12, 15], certain intonational contours can communicate relationships among DSP's.[3] Certain relationships among DSP's are specified within schemas; others are determined from the general task structure indicated by the domain and the particular task structure indicated by the current path.

[3]It is possible that the intermediate phrase may prove an even better unit for discourse segmentation.

Constituents may be annotated with semantic information to be used in determining information status. Semantic annotations include the type of the object and a pointer (to the internal representation for the object designated). For each type of object, there is a predicate which can test two objects of that type for co-designation. For example, for purposes of reference or accenting we may want to treat 'street' and 'avenue' as similar.

Each DS has associated with it a focus space. Following [2], a focus space consists of a set of FORWARD-LOOKING CENTERS, potentially salient discourse entities and modifiers. Focus spaces are pushed and popped from the FOCUS STACK as the description is generated, according to the relationships among their associated DS's.

As an example, the generator for the rotary act appears in figure 2. This schema generates two sentences, second of which is a conjunction. One slot in this schema is taken by an NP constituent for the rotary. The `make-np-constituent` routine handles agreement between the article and the noun. A second slot is filled with an expression giving the approximate angular distance traveled around the rotary. The actual value depends upon the specifics of the act. A third slot in this schema is filled by the name of the street reached after taking the rotary. The choice of referring expression for the street name depends upon the type of street. No cues are generated here, on the grounds that a rotary is unmistakable.

```
(defun disc-seg-rotary (act)
  (list
    (make-sentence
      "You'll" "come" "to"
      (make-np-constituent '("rotary")
         :article :indefinite))
   (make-conjunction-sentence
    (make-sentence
      "Go" (rotary-angle-amount
             (get-info act 'rotary-angle))
      "way" "around" (make-anaphora nil "it"))
    (make-sentence
      "turn" "onto"
       (make-street-constituent
         (move-to-segment act) act)))))
```

Figure 2: Generator for Rotary Act Type

Assigning Intonational Features

The **Talker** employes variation in pitch range, pausal duration, and final lowering ratio to reflect the topic structure of the description, or, the relationship among DS's as reflected in the relationship among DSP's. Following the proposals of [12], we implement this variation by assigned each DS an embeddedness level, which is just the depth of the DS within the discourse tree. Pitch range decreases with embeddedness. In Grosz and Sidner's terms, for example, for DS_1 and DS_2, with DSP_1 dominating DSP_2, we assign DS_1 a larger pitch range than DS_2. Similarly, if DSP_2 dominates DSP_3, DS_3 will have a still smaller pitch range than DS_2. Sibling DS's will thus share a common pitch range. Pitch variation is perceived logarithmically, so pitch range decreases as a constant fraction (.9) at each level, but never falls below a minimum value above the baseline. Also following [12], we vary final lowering to indicate the level of embeddedness of the segment completed by the current utterance. We largely suspend final lowering for the current utterance when it is followed by an utterance with greater embedding, to produce a sense of topic continuity. Where the subsequent utterance has a lesser degree of embedding than the current utterance, we increase final lowering proportionally. So, for example, if the current utterance were followed by an utterance with embedding level 0 (i.e., no embedding, indicating a major topic shift), we would give the current utterance maximal final lowering (here, .87). Pausal duration is greatest (here, 800 msec) between segments at the least embedded level, and decreases by 200 msec for each level of embedding, to a minimum of 100 msec between phrases. Of course, the actual values assigned in the current application are somewhat arbitrary. In assigning final lowering, as pitch range and intervening pausal duration, it is the relative differences that are important.

Accent placement is determined according to relative salience and 'newness' of the mentioned item.[12, 14, 5] (We employ Prince's[17] $Given_s$, or given-salient notion here to distinguish 'given' from 'new' information. However, it would be possible to extend this to include hierarchically related items evoked in a discourse as also given, or 'Chafe-given'[17], were such possibilities present in our domain.) Certain object types and modifier types in the domain have been declared to be potentially salient. When such an item is to be mentioned in the path description, it is first sought in the current focus space and its ancestors. In general, if it is found, it is deaccented; otherwise it receives a pitch accent. If the object is not a potentially salient type, then, if it is a function word, it is deaccented, otherwise it is taken to be a miscellaneous content word and receives an accent by default. In some cases, we found that — contra current theories of focus — items should remain deaccentable even when the focus spaces containing them have been popped from the focus stack. In particular, items in the current focus space's preceding sibling appear to retain their 'givenness'. Reanalysis to place both occurrences in the same segment or to ensure that the first is in a parent segment seemed to lack independent justification. So, we decided to allow items to remain 'given' across sibling segment boundaries, and extended our deaccenting possibilities accordingly.

We vary phrasing primarily to convey structural information. Structural distinctions such as those presented by example (2) are accomplished in this way.

Intentional structure is conveyed by varying intonational contour as well as pitch range, final lowering, and pausal duration. A phrase which required 'completion' by another phrase is assigned a low phrase accent and a high boundary tone (this combination is commonly known as CONTINUATION RISE).[15] For example, since we generate VP conjunctions primarily to indicate temporal or causal relationship (e.g *Stay on Main Street for about ninety yards, and cross the Longfellow Bridge.*), we use continuation rise in such cases on the first phrase.

The sample text in Figure 3 is generated by the system. Note that commands to the speech synthesizer have been simplified for readability as follows: '**T**' indicates the topline of the current intonational phrase; '**F**' indicates the amount of final lowering; '**D**' corresponds to the duration of pause between phrases; '**N***' indicates a pitch accent of type **N**; other words are not accented. Phrase accents are represented by simple **H** or **L**, and boundary tones are indicated by %. The topic structure of the text is indicated by indentation.

Note that pitch range, final lowering, and pauses between phrases are manipulated to enforce the desired topic structure of the text. Pitch range is decreased to reflect the beginning of a subtopic; phrases that continue a topic retain the pitch range of the preceding phrase. Final lowering is increased to mark the end of topics; for example, the large amount of final lowering produced on the last phrase conveys the end of the discourse, while lesser amounts of lowering within the text enhance the sense of connection between its parts. Pauses between clauses are also manipulated so that lesser amounts of silence separate pieces of text which are intended to be semantically more related to one another. For example, the segment beginning with *You'll come to a rotary...* is separated from the previous discourse by a pause of 600 msec, but phrases within this segment describing the procedure to follow once in the rotary are separated by pauses of only 400 msec.

```
T[170] H*+L If your H*+L car is on the H*+L
       same H*+L side of the H*+L street as
       H*+L 7 H*+L Broadway Street L H\% D[600]
   T[153] H*+L turn H*+L around L H\%
   T[153] F[.90] and H*+L start H*+L driving
          L L\%  D[600]
   T[153] F[.90] H*+L Merge with H*+L Main
          Street L L\% D[600]
   T[153] H*+L Stay on Main Street for about
          H*+L one H*+L quarter of a H*+L mile
          L H\% D[600]
   T[153] F[.90] and H*+L cross the Longfellow
          H*+L Bridge L L\% D[600]
   T[153] F[.96] You'll H*+L come to a
          H*+L rotary L L\% D[400]
      T[137] H*+L Go about a H*+L quarter
             H*+L way H*+L around it
             L H\% D[400]
      T[137] F[.90] and H*+L turn onto
             H*+L Charles Street L L\% D[600]
   T[153] H*+L Number H*+L 130 is about H*+L
          one H*+L eighth of a H*+L mile
          H*+L down L H\% D[400]
      T[137] F[.87] on your L+H* right
             H* side L L\%
```

Figure 3: A Sample Route Description from Direction Assistance

Summary

We have described how structural, semantic, and discourse information can be represented to permit the principled assignment of pitch range, accent placement and type, phrasing, and pause in order to generate spoken directions with appropriate intonational features. We have tested these ideas by modifying the text generation component of Direction Assistance to produce an abstract representation of the information to be conveyed. This 'message-to-speech' approach to speech synthesis has clear advantages over simple text-to-speech synthesis, since the generator 'knows' the meanings to be conveyed. This application, while over-simplifying the relationship between discourse information and intonational features to some extent, nonetheless demonstrates that it should be possible to assign more appropriate prosodic features automatically from an abstract representation of the meaning of a text. Further research in intonational meaning and in the relationship of that meaning to aspects of discourse structure should facilitate progress toward this goal.

References

[1] Barbara Grosz. The Representation and Use of Focus in Dialogue Understanding. Phd thesis, University of California at Berkeley, 1976.

[2] B. Grosz, A. K. Joshi, and S. Weinstein. Providing a Unified Account of Definite Noun Phrases in Discourse. *Proceedings of the Association for Computational Linguistics*, pages 44–50, June 1983.

[3] Candace Sidner. Towards a computational theory of definite anaphora comprehension in English discourse. PhD thesis, MIT, 1979.

[4] M. Anderson, J. Pierrehumbert, and M. Liberman. Synthesis by rule of English intonation patterns. *Proceedings of the conference on Acoustics, Speech, and Signal Processing*, page 2.8.1 to 2.8.4, 1984.

[5] Gillian Brown. Prosodic structure and the given/new distinction. In Cutler and Ladd, editors, *Prosody: Models and Measurements*, chapter 6, Springer Verlag, 1983.

[6] James R. Davis. Giving directions: a voice interface to an urban navigation program. In *American Voice I/O Society*, pages 77–84, Sept 1986.

[7] James R. Davis and Thomas F. Trobaugh. *Direction Assistance*. Technical Report, MIT Media Technology Lab, Dec 1987.

[8] Marcia A. Derr and Kathleen R. McKeown. Using focus to generate complex and simple sentences. *Proceedings of the Tenth International Conference on Computational Linguistics*, pages 319–325, 1984.

[9] Barbara J. Grosz and Candace L. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

[10] Dwight Bolinger. Accent is predictable (if you're a mind-reader). *Language*, 48:633-644, 1972.

[11] M. A. K. Halliday. *Intonation and Grammar in British English*. Mouton, 1967.

[12] J. Hirschberg and J. Pierrehumbert. The intonational structure of discourse. *Proceedings of the Association for Computational Linguistics*, pages 136–144, July 1986.

[13] Kathleen R. McKeown. Discourse strategies for generating natural-language text. *Artificial Intelligence*, 27(1):1–41, 85.

[14] S. G. Nooteboom and J. M. B. Terken. What makes speakers omit pitch accents? an experiment. *Phonetica*, 39:317–336, 1982.

[15] J. Pierrehumbert and J. Hirschberg. The meaning of intonation contours in the interpretation of discourse. In *Plans and Intentions in Communication*, SDF Benchmark Series in Computational Linguistics, MIT Press, forthcoming.

[16] Janet B. Pierrehumbert. *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT, Dept of Linguistics, 1980.

[17] Ellen F. Prince. Toward a taxonomy of given - new information. In Peter Cole, editor, *Radical Pragmatics*, pages 223–256, Academic Press, 1981.

[18] Kim E. A. Silverman. *Natural prosody for synthetic speech*. PhD thesis, Cambridge Universtity, 1987.

[19] L. Witten and P. Madams. The telephone inquiry service: a man-machine system using synthetic speech. *International Journal of Man-Machine Studies*, 9:449–464, 1977.

[20] S. J. Young and F. Fallside. Speech synthesis from concept: a method for speech output from information systems. *Journal of the Acoustic Society of America*, 66(3):685–695, Sept 1979.

[21] J. P. Olive and M. Y. Liberman. Text to speech – An overview. *Journal of the Acoustic Society of America, Suppl. 1*, 78(3):s6, Fall 1985.