Modeling Contrast in the Generation and Synthesis of Spoken Language

Scott Prevost

The Media Laboratory
Massachusetts Institute of Technology
20 Ames Street
Cambridge, MA 02139-4307 USA

ABSTRACT

This paper presents an implemented model of spoken language processing that accounts for intonational phenomena associated with semantic contrasts. The model determines accentual patterns based on sets of alternative properties from a knowledge base and a contrastive stress algorithm. The results of applying the model to a natural language generation program illustrate the advantages over previous models based on lexical "givenness."

1. INTRODUCTION

One of the key obstacles in gaining acceptance of synthetic speech output for computer applications is the inability in many instances for such programs to produce natural sounding intonation. In English, the selection of a given intonational pattern for an utterance can affect the relationship the utterance bears to previous utterances, and in extreme cases can completely alter its meaning. For example, consider the utterances below, where capitalization marks the words bearing pitch accents.

(1) Speaking of BILL,
JOHN thought he would WIN, but he DIDN'T.

(2) Speaking of BILL,
JOHN thought he would WIN, but HE didn't.

In the first case, the final clause of the utterance can be paraphrased as "Bill didn't win." In the second case, however, the final clause must be paraphrased as "Bill didn't think that he (Bill) would win." Examples such as these illustrate that algorithmic approaches for assigning intonational parameters to synthesized speech must rely not only on orthographic and syntactic clues, but also on the semantics of the intended speech.

In other cases, the choice of accentual pattern for a given utterance may depend on prior utterances, as shown in the examples below. While the final sentences in these examples may be considered to have the same meaning, their accentual patterns are clearly distinct and cannot be interchanged without sounding markedly unnatural.

(3) Yesterday, we drove to the beach. The weather was rainy and windy for most of the trip, so we didn't make very good time. Fortunately, when we ARRIVED at the beach, the weather turned BEAUTIFUL.

(4) Last week we went on vacation. During the inland portion of our trip, the weather was dreadful. Fortunately, when we arrived at the BEACH, the weather turned BEAUTIFUL.

In order to capture such contextual effects in intonation, text-to-speech (TTS) and meaning-to-speech (MTS) systems have employed a number of useful heuristics which cover a wide array of examples (Hirschberg 1990; Monaghan 1991). The limitations of these heuristics, particularly with respect to the phenomenon of *contrastive stress*, are explored in the remainder of this paper. In the following sections, we present a model for determining intonational patterns in an MTS system and briefly discuss the relevant aspects of an implementation designed to produce spoken descriptions of objects.

2. ACCENTUATION PATTERNS

Given the semantic nature of intonational patterns and the obvious contextual effects, predicting the distribution of pitch accents for a given utterance is indeed difficult. In broad terms, Bolinger (1972) defined the problem as *semantic highlighting*, whereby lexical items are stressed based on their contextual interest, or roughly speaking, how much they contribute to the hearer's model of the conversation. While Bolinger's definition is left intentionally vague, it nonetheless forms the basis for accent prediction strategies in state-of-the-art TTS systems.

2.1. The Previous Mention Strategy

The stress patterns in examples such as (3) and (4) can be predicted by a set of heuristics based on textual *givenness* (cf. Hirschberg 1990; Monaghan 1991). We refer to this technique, which has been widely used in text-to-speech applications, as the *previous mention strategy*. In its most simple form, the strategy works as follows:

1. Assign accents to open-class lexical items (e.g. nouns, verbs, other content words).
2. De-accent all closed-class lexical items (e.g. function words).
3. De-accent any lexical items that were already mentioned in the local discourse segment.

2.2. Contrastive Stress

The effect of the accentuation strategy described above is to equate Bolinger's notion of what constitutes semantic "interest" with the given/new distinction. That is, items bearing semantic content are accented on first mention only. While this application of semantic highlighting covers many cases, it fails to account for the cases where information is highlighted for reasons other than its "new" status in the discourse. Consider the following example:

(5) YESTERDAY, we drove from the MOUNTAINS to the BEACH. The WEATHER at the BEACH was PERFECT, but the MOUNTAIN weather was HORRENDOUS.

In this example, both *beach* and *mountain* are accented in the second sentence despite their having been introduced in the first. Consequently, the decision to accent these items cannot be based on the previous mention strategy.

The accentual pattern in (5) can be considered to be an instance of *contrastive stress*. That is, *beach* is accented because it stands in direct contrast to some other salient item in the discourse, namely the mountain. This phenomenon is clearly evident in cases where pronouns receive stress, such as (6). Since pronominalized items are generally considered "given" by the presence of their antecedent, the previous mention strategy cannot possibly account for their accentuation.

(6) Bill and I went to a new restaurant last night. I HATED it, but HE LIKED it.

2.3. Contrastive Stress in Natural Discourse

While it is generally quite easy to concoct examples for which the previous mention strategy is inadequate, the data described in this section verifies that contrastive accentuation occurs on contextually "given" items quite frequently in naturally occurring speech. The data was extracted from the Switchboard corpus, a collection of over 2000 digitized telephone conversations collected at Texas Instruments. Since the corpus is a general tool for studying numerous aspects of speech data, the subjects were aware of neither the nature of the present research nor the intonational theories espoused by this writer.

Since the notions of both "givenness" and "contrastiveness" are somewhat vague in the literature on spoken discourse, we examined utterances of the form "but he..." to determine how often the subject pronoun received stress. Because of the explicitly contrastive nature of the "but" construction, we could be reasonably certain that any accentuation applied to these pronouns was likely to be attributable to a semantic contrast among competing items from the discourse. Moreover, since pronouns are generally "given" by previous occurrences of their antecedent, we avoided the difficult subjective task of assigning a given/new status to items.

In total, 162 occurrences of "but he ..." were extracted from 1022 conversations in the Switchboard corpus. Of those, 33 exhibited some degree of accentuation on "he," as determined by a combination of subjective judgments (by the author) and pitch track analyses. Of the 33 occurrences of stressed "he," two were discounted as examples of contrastive stress because the immediately preceding discourse (approximately ten utterances) did not support such an interpretation. In the remaining cases, an antecedent for the stressed pronoun was clearly established in the previous utterances.

The results of the experiment, presented in Table 1, show that 31 (19.14%) of the 162 subject pronouns in explicitly marked contrastive constructions of the form "but he ..." received some degree of stress that can reasonably be attributed to their contrastive status. Although it is impossible to extrapolate the data to cases where contrastive constructions are not marked explicitly by coordinators like "but" or "however," the data does provide clear evidence that pronouns, despite their status as "given," are eligible to receive stress.

Disc	# He	# But he	# Contrastive	% Contrastive
1	157	4	0	0%
2	212	8	0	0%
3	265	6	1	16.67%
4	325	12	2	16.67%
5	245	14	0	0%
6	298	10	1	10.00%
7	326	14	4	28.57%
8	249	6	0	0%
9	277	9	2	22.22%
10	276	13	4	30.77%
11	274	13	2	15.38%
12	334	10	3	30.00%
13	427	21	8	38.10%
14	315	10	1	10.00%
15	312	12	3	25.00%
Totals	4292	162	31	19.14%

Table 1: Contrastive occurrences of "but he ..." in the Switchboard Corpus (phase 1)

2.4. Contrastive Stress Algorithm

In the previous examples, items were shown to be accented for contrastive purposes when some other salient alternative had been clearly established in the discourse. In each of those examples, however, the items in question were represented by relatively simple referring expressions. For more complex referring expressions, such as those containing adjectives or relative clauses, the distribution of accents among the lexical items comprising the referring expression marks the contrast. Consider the following examples:

(7) The patient broke her LEFT leg, NOT her RIGHT leg.

(8) The patient broke her left LEG, NOT her left ARM.

When some discourse entity needs to be contrasted with some other salient entity, the choice of accentual pattern is

dependent on the set of features that discriminates between those items. Given an utterance which includes a referring expression for some entity *x*, and a set of alternative entities for *x* as determined by the prior discourse and an associated knowledge base, the set of features to be accented can be decided by the algorithm described below.

- Let RSET include *x* and its alternatives.
- Let PROPS be a list of all properties (features) of *x*, ordered so that nominal properties take precedence over adjectival properties
- Let CSET be the (initially empty) set of properties of *x* that must be accented for contrastive purposes.

For each property *p* in PROPS, do the following. If *p* is not a property of each member of RSET, eliminate those entities from RSET for which *p* does not hold and include *p* in CSET. Stop when RSET contains only *x*.

3. INTONATIONAL TUNES

The previous section outlined the importance of modeling accentuation patterns that serve to contrast between competing discourse items. Being able to predict the distribution of accents, however, is not sufficient for producing contextually appropriate and natural-sounding intonation since there are several different types of accents to choose from. Furthermore, intonational tunes are comprised not only of pitch accents, but also of phrasal and boundary tones that delimit intermediate and intonational phrases respectively (Pierrehumbert 1980). Consider the following examples in which intonational tunes are shown with a variant of Pierrehumbert's notation.

(9) Q: I know which leg the OLD patient broke, but which leg did the YOUNG patient break?
A: The YOUNG patient broke her LEFT leg.
L+H* L H* L L$

(10) Q: I know which patient broke her RIGHT leg, but which patient broke her LEFT leg?
A: The YOUNG patient broke her LEFT leg.
H* L L+H* L H$

In these examples, H* and L+H* represent two different types of pitch accents in Pierrehumbert's intonational classification scheme, each characterized by a high pitch on the accented syllable of the word on which it occurs. The latter is intoned as a distinct rise from a lower pitch at the beginning of the word to the high pitch on the accented syllable. The occurrences of L represent phrasal tones that demarcate intermediate phrases and control the pitch from the most recent pitch accent to the end of the phrase. Finally, L$ and H$ represent rising and falling intonational boundaries respectively. In the present classification scheme, L$ and H$ differ from Pierrehumbert's (1980) L% and H% in the degree of pausing associated with them, the former being slightly longer than the latter. This differentiation is necessary for the purposes of generation so that full sentence boundaries can be distinguished from intra-sentential boundaries.

In previous work, Prevost (1995) and Prevost and Steedman (1994) have argued that the two basic intonational tunes in examples (9) and (10), H* L (L%/L$) and L+H* L (H%/H$), can be directly mapped onto the *information structure* of such simple declarative utterances. Following Halliday (1972) and others, the information structure, which refers to the packaging of information within an utterance, is divided into two parts: the *theme* and the *rheme*. The theme (or *topic*) of an utterance, which is often intonationally marked by the tune L+H* L (H%/H$), denotes that part of the utterance which links it to prior utterances. So, for example, the theme of the answer in example (9) might be propositionally represented as $\lambda x.broke(young\text{-}patient,x)$ since the phrase "the young patient broke" links the answer to the previous utterance. The rheme (or *comment*) of an utterance, which is often intonationally marked by the H* L (L%/L$) tune, denotes that part of the utterance which forms the core contribution to the discourse (i.e. the new or particularly salient information). In example (9), the rheme might be propositionally represented simply as *left-leg* or more abstractly as $\lambda P.P(left\text{-}leg)$.

Given the division of a simple declarative utterance into theme and rheme, the mapping described above dictates which intonational tunes are associated with phrases within the utterance. While the mapping controls the placement of phrasal and boundary tones, it does not dictate the locations of pitch accents. For this, we rely on the previous mention strategy and the contrastive stress algorithm described in Section 2.4. Elements of the theme and rheme that receive stress on the basis of newness or contrastiveness are said to be in *focus*. Based on the discussion above, focused elements of the theme receive L+H* accents, while rhematic focused elements receive H* accents. While there is no clear evidence that the shape of contrastive accents differs from non-contrastive accents (Bolinger 1972), the amplitude of contrastive accents often overshadows other accents in an utterance. We denote this by marking contrastive accents with the subscript *c* (as in H^*_c) and realizing them with slightly high pitch than other accents in the utterance.

4. IMPLEMENTATION

The intonational theory and algorithms presented above are implemented in a spoken language generation system that produces spoken descriptions of objects from a small knowledge base. The details of the natural language generation scheme, which are beyond the scope of this paper, are provided in Prevost (1995). The present section briefly describes how the model of intonation under discussion is embodied by the implementation.

The natural language generator is divided into three phases: high-level content planning, sentence planning and surface generation. During the high-level content planning stage, propositions which satisfy the given communicative goal are

selected from the knowledge base and sorted based on their relevancy, pre-compiled templates for object descriptions (cf. McKeown 1985) and a number of rhetorical constraints (cf. Hovy 1993). Principle among these constraints is the notion that consecutive utterances share semantic material. The sharing of material in effect dictates the division of utterances into theme and rheme and consequently determines the corresponding intonational tunes.

During the sentence planning phase, high-level propositions are converted into representations that more fully constrain the possible sentential realizations. This phase, which forms the bridge between language-independent propositions and language-specific syntactic constructs, determines the choice of referring expressions for discourse entities. Since the contrastive stress algorithm described in Section 2 relies on such specifications, it is during this sentence planning phase that the algorithm is invoked and locations of pitch accents within theme and rheme phrases are determined.

In the final stage of speech production, a surface generator (Prevost 1995) based on Combinatory Categorial Grammar (Steedman 1991) converts the output of the sentence planner into sentences with intonational annotations. These sentences are then synthesized to produce speech with contextually-appropriate intonation.[1] Examples (11) and (12) show the result of invoking the generator twice with the goal of describing two objects from the knowledge base. Note that although the information conveyed about the two items is quite similar, the intonational patterns for example (12) are clearly based on the context provided by example (11).

(11) The X4 is a SOLID-state AMPLIFIER.
L+H* L H* H* L L$
It COSTS EIGHT HUNDRED DOLLARS,
H* H* H* H* L H%
and PRODUCES ONE hundred watts-per-CHANNEL.
H* H* H* L L$
It was PRAISED by STEREOFOOL, an AUDIO JOURNAL
H^*_c $!H^*_c$ L H% H* H* L H%
but was REVILED by AUDIOFAD, ANOTHER audio journal.
H^*_c $!H^*_c$ L H% H* L L$

(12) The X5 is a TUBE amplifier.
$L+H^*_c$ L H^*_c L L$
IT costs NINE hundred dollars,
$L+H^*_c$ L H^*_c L H%
and produces TWO hundred watts-per-channel.
H^*_c L L$
IT was praised by Stereofool AND Audiofad.
$L+H^*_c$ L H^*_c L L$

[1] We currently use the AT&T TTS system to synthesize the program's output, but other synthesizers may also be employed.

5. CONCLUSIONS

The results show that it is possible to produce spoken output in meaning-to-speech systems that intonationally conveys important contrastive distinctions. Examples (11) and (12) clearly illustrate that even items that are contextually "given" (i.e. previously mentioned) are eligible to receive contrastive stress under the current intonational model.

6. ACKNOWLEDGMENTS

The author is grateful for the advice and helpful suggestions of Mark Steedman, Justine Cassell, Matthew Stone and Beryl Hoffman. Without the AT&T Bell Laboratories TTS system, and Julia Hirschberg's patient advice on its use, this work would not have been possible. This research was funded by NSF grants IRI91-17110 and IRI95-04372 and the generous sponsors of the MIT Media Laboratory.

7. REFERENCES

1. Bolinger, D. (1972). Accent is predictable (if you're a mind reader). *Language*, 48:633–644.

2. Halliday, M. (1970). Language structure and language function. In Lyons, J., editor, *New Horizons in Linguistics*, pages 140–165. Penguin.

3. Hirschberg, J. (1990). Accent and discourse context: assigning pitch accent in synthetic speech. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, pages 952–957.

4. Hovy, E. (1993). Automated discourse generation using discourse structure relations. *Artificial Intelligence*, 63:341–385.

5. McKeown, K. (1985). *Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text*. Cambridge University Press, Cambridge.

6. Monaghan, A. (1991). *Intonation in a Text-to-Speech Conversion System*. Ph.D. thesis, University of Edinburgh.

7. Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, Massachusetts Institute of Technology.

8. Prevost, S. (1995). *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation.* Ph.D. thesis, University of Pennsylvania.

9. Prevost, S. and Steedman, M. (1994). Specifying intonation from context for speech synthesis. *Speech Communication*, 15:139–153.

10. Steedman, M. (1991). Structure and intonation. *Language*, pages 260–296.

Sound File References:

[a703s01.wav]

[a703s02.wav]