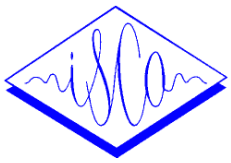

The ESCA Workshop
on Speech Synthesis
Autrans, France
September 25-28, 1990

Using Discourse Context to Guide Pitch Accent Decisions in Synthetic Speech

Julia Hirschberg
AT&T Bell Laboratories — 2D-450
Murray Hill NJ 07974 USA

ABSTRACT

The paper describes the assignment of PITCH ACCENT in **NewSpeak**, an interface to the Bell Laboratories Text-to-Speech System, which infers limited discourse-level information, including GIVEN/NEW distinctions, and some information on FOCUS, TOPIC and CONTRAST, along with improved part-of-speech distinctions, to assign intonational features for unrestricted text.[1]

1. INTRODUCTION

How speakers decide what to emphasize and de-emphasize in natural speech, has long been the subject of scholarly debate. In natural speech, words that appear more intonationally prominent than others are said to bear PITCH ACCENTS.[2] Although pitch accent is a perceptual phenomenon, words hearers typically identify as accented tend to differ from their DEACCENTED versions (those not bearing pitch accents) in pitch, duration, amplitude, and spectral characteristics. Accented words are usually identifiable in the FUNDAMENTAL FREQUENCY CONTOUR (f0) as local maxima or minima, aligned with the word's stressed syllable; their duration and amplitude tend to be greater than their deaccented counterparts. The vowel in the stressed syllable of a deaccented word is often reduced from the full vowel of the accented version.

While formerly it was believed that syntactic information determines speakers' accent decisions, it is now recognized that many kinds of information contribute [2]. Experimental studies (e.g. [4]) have shown that speakers associate accent with a word's GIVEN/NEW STATUS — whether the item represents 'old' information, which a speaker is entitled to believe is shared with his/her hearer, or 'new' information in a discourse [8]. And analyses of large corpora of recorded speech (e.g. [1]) provide evidence that part-of-speech information or constituent structure alone are insufficient for modeling pitch accent assignment in natural speech. However, such findings have yet yet to be reflected in algorithms for accent assignment in speech synthesis. Commonly, speech synthesizers use only a simple distinction between functions words (e.g. prepositions, pronouns) and content words (e.g. nouns, verbs) — possibly with minimal information about surface position — to assign pitch accent. Such approaches tend to accent too many words in synthesis of longer stretches of text; in isolated sentences, they predict only about 75-80% 'acceptable' accent assignment

This paper presents work incorporating various types of INFORMATION STATUS, including a given/new distinction, limited information on FOCUS, TOPIC and CONTRAST, and more sophisticated part-of-speech distinctions, to assign pitch accent for unrestricted text. The algorithm described has been trained on prosodically labeled corpora of read speech. This algorithm is currently used in **NewSpeak**, an interface to the Bell Laboratories Text-to-Speech System (TTS) [7] which takes as input unrestricted text and outputs text interspersed with escape sequences which control intonational features in TTS. It is also being used to hypothesize accentuation information for large corpora of prosodically unlabeled speech, so that accent can be used as a variable in statistical analysis; and to provide a prosodic hypothesis used to speed prosodic labeling of other speech corpora.

2. ACCENT ASSIGNMENT IN NEWSPEAK

NEWSPEAK generates accent assignments based upon more sophisticated use of syntactic information as well as higher-level discourse information. NewSpeak's text analysis component takes as input text annotated

[1] Thanks to Ken Church, Richard Omanson, and Richard Sproat for helpful comments and discussion.
[2] While, in English, each word has a characteristic (lexical) stress pattern, not every word is accented.

with part-of-speech information from Church's tagger [5]. Some errors in accent prediction occur because an item's part-of-speech assignment is ambiguous between function and content word, such as preposition vs. verbal preposition/particle (e.g. *John left* **in** *the limo* vs. *John left* **in** *the typo*) or conjunction vs. discourse marker (e.g. *They left after lunch* **and** *landed in France in time for dinner.* vs. *The left after lunch.* **And,** *they landed in France in time for dinner.*) However, not all function words that are correctly classified are in fact deaccented, as illustrated by a sample paragraph from the FM Radio Newscasting Database[3] (where accented words are represented in upper-case and deaccented in lower-case):

(1) a. Newsreader: In NINETEEN SEVENTY-SIX, DEMOCRATIC GOVERNOR MICHAEL DUKAKIS FULFILLED a CAMPAIGN promise to DE-POLITICIZE JUDICIAL APPOINTMENTS.

b. He NAMED REPUBLICAN Edward HENNESSY, to HEAD the STATE SUPREME JUDICIAL COURT,

c. For HENNESSY, it was ANOTHER STEP along a DISTINGUISHED CAREER THAT BEGAN as a TRIAL LAWYER, and LED to an APPOINTMENT AS ASSOCIATE Supreme Court Justice in nineteen seventy-ONE.

d. THAT YEAR THOMAS MAFFY, NOW PRESIDENT of the MASSACHUSETTS BAR Association, was HENNESSY'S LAW clerk.

The relative pronoun *that* in (1b), the determiner *that* in (1d), and the preposition *as* in (1c) are all accented. A function-content distinction operating even on correctly tagged data will successfully predict only 85% of pitch accents in (1).[4]

If function words are not always deaccented, content words are not always accented. Consider, for example, *campaign promise* in (1a), MASSACHUSETTS BAR *Association* and LAW *clerk* in (1d). These COMPLEX NOMINALS are sequences of nouns whose semantico-syntactic structure maps to differences in stress assignment. Some, like CAMPAIGN *promise* are stressed on the left, with consequent deaccenting of the right member of the nominal; others, like *nineteen seventy-six* and *State Supreme Judidical Court*, are stressed on the right; and still others, like *judicial appointments*, may be stressed either on the right or left. NewSpeak gets citation-form stress assignment for complex nominal's from Sproat's NP parser [10].

However, while there are regularities in the citation form of stress assignment for such phrases, these explain only part of the accenting of complex nominals. Note, for example, that while *nineteen seventy-six* in (1a) is entirely accented, the subsequent *nineteen seventy*-ONE in (1c) exhibits a different pattern; and only *associate* is accented in the nominal ASSOCIATE *Supreme Court Justice* in (1c), although in citation form, all components would be accented. We can explain this behavior in terms of the sensitivity of accent decisions to context. In simple terms, *nineteen seventy* and *Supreme Court Justice* are deaccented in (1c) because they represent GIVEN information in their context of utterance.

In NewSpeak, the accenting of complex nominals and other content words is mediated by by the inference of discourse-level information on the information status of mentioned entities. Drawing upon work in Artificial Intelligence[6], NewSpeak infers such features from an analysis of orthographic features of unrestricted text, including paragraphing and punctuation as well as DISCOURSE MARKERS.[5] These indicators are used to build up a model of the text's ATTENTIONAL STRUCTURE, a hierarchical structuring of concepts mentioned or evoked in the discourse, which can be taken to represent information that is 'given' in the text. In NewSpeak, this structure is implemented as a stack of FOCUS SPACES, each a set of roots of open-class items mentioned within a portion of the discourse. Focus spaces are updated depending upon the topic structure inferred from the text. The first focus space is hypothesized to represent the GLOBAL FOCUS for the text, intuitively, the set of general concepts characterizing the text. While the set of items in local focus is constantly subject to change, items in global focus remain so throughout the discourse. These focus spaces can be understood as similar to the nested contexts of a block-structured programming language.

In NewSpeak's original accent assignment algorithm, items in either global or local focus were treated as given information. Subsequent mention of given items was deaccented, in line with empirical results

[3] Being collected by SRI International (Patti Price), Boston University (Mari Ostendorf), and MIT (Stefanie Shattuck-Hufnagel).

[4] This relatively high success rate is in part to the tendency of news readers to accent content words at the end of phrases, even where other speakers normally would not (e.g. *trial lawyer* in (1c) [3]. And overall, the function-content distinction predicts 80% of accent decisions correctly in the 5-minute text from which this sample was taken.

[5] Words such as *now* and *well*, which convey explicit information about discourse structure.

[4] suggesting that listeners associate accented items with new information and deaccented items with old information. While questions such as the domain over which items remain given — and the process by which they lose their givenness — are open research questions, it is also clear that not every item which is given — under any reasonable definition — will be deaccented. Consider, for example, the accenting of the clearly 'given' HENNESSY in (1b)-(1d). To address this phenomenon, NewSpeak does limited inference of additional discourse characteristics, such as topic and focus, which also influence human accent assignment. For example, the accenting of HENNESSY discussed above can be explained in terms of change in topic, from Dukakis to Hennessy to Maffy and then back to Hennessy. Accent here is used to indicate these shifts. To infer likely topic and focus behavior, NewSpeak currently uses surface order and part-of-speech together with local and global focus information. This represents only a start, however, on the approximation of these discourse features.

3. PRELIMINARY RESULTS

Thus far, results from testing variations of NewSpeak's algorithm on samples of recorded (read) speech suggest certain tentative conclusions. (It should be stressed that analysis of a much larger amount of labeled speech will be necessary to demonstrate their usefulness — and will also permit the analysis of interactions among the structural and discourse features described below.) Clearly, the simple association of function word with deaccenting employed in most text-to-speech systems must be modified. In NewSpeak's current accent assignment algorithm, closed-class items are divided into three categories. Possessive pronouns (including *wh*-pronouns), definite and indefinite articles, copulas, coordinating and subordinating conjunctions, existential *there*, *have*, accusative pronouns and *wh*-adverbials, most prepositions, positive modals, positive *do*, as well as some particular adverbials like *ago*, nominative and accusative *it* and nominative *they*, and some nominal pronouns (e.g. *something*) are identified as 'closed, deaccented'. And certain of these classes are marked for further reduction in synthesis via CLITICIZATION, involving the removal of adjacent word boundaries and vowel reduction. 'Closed, accented' items, on the other hand, include the negative article, negative modals, negative *do*, most nominal pronouns, most nominative and all reflexive pronouns, pre- and post-qualifiers (e.g. *quite*), pre-quantifiers (e.g. *all*), post-determiners (e.g. *next*), nominal adverbials (e.g. *here*), interjections, particles, most *wh*-words, plus some prepositions (e.g. *despite, unlike*). Other word classes (adjectives, adverbials,common and proper nouns, verbs) are deemed 'open'. For purposes of acquiring given/new information, only open-class items are considered, although how much of this category to consider is also subject to variation.

The collection and manipulation of the attentional state representation has been varied experimentally in the following ways: Both global and local focus representations have been manipulated independently such that the global focus space may be set, and the local focus spaces updated, by the orthographic phrase, the sentence, or the paragraph. So, for example, the global space may be set after the first phrase, sentence or paragraph of a text. The local stack can be updated independently at the end of each phrase, sentence or paragraph — although discourse markers will push or pop the stack as well. For the current experiments, the best results to date have come when the global space is defined to be the first full sentence of the text and the local attentional stack is updated by paragraph. The content of both global and local focus spaces have also been varied systematically by word class, so that all open-class words, nouns only, or nouns plus some combination of verbs and modifiers are allowed to affect — and be affected by — the attentional state representation. Preliminary results, which again should be taken as suggestive only, indicate that focal spaces defined in terms of roots of all content words, rather than nominals only, or even all nonverb roots, provide the best accent prediction.

Finally, some experimentation has been done to relate the accenting of items currently in local focus with structural and discourse-based indicators of contrastiveness. For example, the referential strategy of PROPER-NAMING [9], in which the use of proper names was found to focus attention, helps to explain behavior such as the accenting of HENNESSY, described above. It is conjectured that such referential behavior might indicate the speaker's attempt to focus attention upon persons recently mentioned, when other focii have intervened since their introduction. This strategy, together with others which can be inferred from surface and syntactic features of the text, such as the preposing of adverbials and of prepositional phrases (e.g., THAT YEAR in (1d)) and the reintroduction of items in global but not in local focus, have been tested as predictors of accent with some success.

4. DISCUSSION

This paper has described the pitch accent assignment strategy employed in NewSpeak, an interface to the Bell Labs Text-to-Speech System, which employs a hierarchical representation of the attentional structure of the discourse, together with more traditional syntactic information, to assign intonational features in the synthesis of unrestricted text. It has also sketched experiments currently being performed to refine that algorithm, by modifying traditional uses of word class, key word, and surface position information, and by varying the construction of and interaction between the components of a model of attentional state. The testing of various discourse models against pitch accent placement in actual speech, should also add to our set of evaluation criteria for such models.

From a theoretical point of view, such analysis should bring us closer to understanding how to model pitch accent in human speech. However, real progress will depend upon the availability of large amounts of prosodically labeled data. An immediate application for NewSpeak's accent assignment algorithm, in addition to its use in pre-processing text for TTS, is to provide prosodic labeling hypotheses to speed the subsequent hand labeling of prosodic features for large corpora; the post-editing of prosodic labels appears to be considerably faster and the labeled speech can then be used in further training of the algorithm. Lacking such labeled corpora, the algorithm can be used as a rough substitute for accent information. Currently, it is being used to predict accent assignment for a large corpus of read sentences being analyzed for durational characteristics, with remarkable accuracy.[6]

While the use of higher level discourse information to inform algorithms for pitch accent assignment appears to be a useful strategy for modeling accent assignment in natural speech, it may indeed turn out not to be desirable to emulate natural speech in synthetic speech. However, clearly, whatever variation eventually emerges as desirable between synthetic speech and human speech should clearly be intentional rather than chance. Demonstrating that one speech synthesizer is preferable to another, or that one prosodic strategy is to be favored over another, in terms of simple human preference, is notoriously difficult to accomplish. Comparison of the output of algorithms used to assign intonational features in synthetic speech with prosodic features in natural speech would thus appear a useful alternative.

References

[1] B. Altenberg. *Prosodic Patterns in Spoken English: Studies in the Correlation between Prosody and Grammar for Text-to-Speech Conversion*, volume 76 of *Lund Studies in English*. Lund University Press, Lund, 1987.

[2] D. Bolinger. Accent is predictable (if you're a mindreader). *Language*, 48:633–644, 1972.

[3] D. Bolinger. *Intonation and Its Uses: Melody in Grammar and Discourse*. Edward Arnold, London, 1989.

[4] G. Brown. Prosodic structure and the given/new distinction. In D. R. Ladd and A. Cutler, editors, *Prosody: Models and Measurements*. Springer Verlag, Berlin, 1983.

[5] K. W. Church. A stochastic parts program and noun phrase parser for unrestricted text. In *Proceedings of the Second Conference on Applied Natural Language Processing*, pages 136–143, Austin, 1988. Association for Computational Linguistics.

[6] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204, 1986.

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

[8] E. Prince. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics*. Academic Press, New York, 1981.

[9] A. J. Sanford, S. C. Garrod, K. Moar, and H. Al-Ahmar. Naming, role-descriptions, and main and secondary characters in discourse comprehension. Reported in A. J. Sanford. Aspects of pronoun interpretation: Evaluation of search formulations of inference. In G. Rickheit and H. Strohner, editors, *Inferences in Text Processing*, pages 183–204. North-Holland, Amsterdam, 1985.

[10] R. Sproat. Stress assignment in complex nominals for English text-to-speech. In *Proceedings of the Tutorial and Research Workshop on Speech Synthesis*, Autrans, France, September 1990. European Speech Communication Association.

[6] Checking 300 sentences from a corpus of 2775 uncovered only thirty errors in accent assignment (accented, deaccented, cliticized).