APPLICATION OF THE LIBERMAN-PRINCE STRESS RULES TO COMPUTER SYNTHESIZED SPEECH

David L. McPeters* and Alan L. Tharp
Computer Science Department
North Carolina State University
Raleigh, North Carolina 27650 USA

ABSTRACT

Computer synthesized speech is and will continue to be an important feature of many artificially intelligent systems. Although current computer synthesized speech is intelligible, it cannot yet pass a Turing test. One avenue for improving the intelligibility of computer synthesized speech and for making it more human-like is to incorporate stress patterns on words. But to achieve this improvement, a set of stress prediction rules amenable to computer implementation is needed.

This paper evaluates one such theory for predicting stress, that of Liberman and Prince. It first gives an overview of the theory and then discusses modifications which were necessary for computer implementation. It then describes an experiment which was performed to determine the model's strengths and shortcomings. The paper concludes with the results of that study.

I INTRODUCTION

Since speech is such an important component of human activities, it is essential that it be included in computer systems simulating human behavior or performing human tasks. Advantages of interacting with a computer system capable of speech include that

a) special equipment (e.g. a terminal) is unnecessary for receiving output from the device.

b) the output may be communicated to several people simultaneously.

c) it may be used to gain someone's attention.

d) it is useful in communicating information in an emergency.

The primary methods for generating computer synthesized speech are 1) to use a lexicon of word pronunciations and then assemble a message from these stored words or 2) to use a letter-to-sound translator. A shortcoming common to both methods, and of interest to linguists and more recently computer scientists, is the inclusion of English prosody in computer synthesized speech e.g. Klatt [6], Lehiste [8], Witten et al [11] and Hill [5]. Of the three primary components of English prosody, this paper considers only stress (the other two are intonation and pause). It applies the theory for stress prediction proposed by linguists Mark Liberman and Alan Prince [9] to computer synthesized speech. Their theory was chosen primarily as a result of it having received widespread attention since its introduction (see Paradis [10], Yip [12], Fujimura [3 and 4] and Basboll [2]).

II THE LIBERMAN-PRINCE MODEL

In addition to the attention it received, the Liberman-Prince model [9] (hereafter referred to as the LP model) is attractive for computer application for two other reasons. First, the majority of its rules can be applied without knowledge of the lexical category (part-of-speech) of the word being processed since the rules are based only on the sequences and attributes of letters in a word. This feature is especially important in an unrestricted text-to-speech translation system. Secondly, since the metrical trees that define the prominence relations are a common data structure, a computer model may be designed which remains very close to the foundations and intentions of the theoretical model.

This section will summarize the LP theory as presented in [9]. The LP method of predicting stress focuses on two attributes of vowels: + or - long and + or - low. The e of be is +long while the e of pet is -long. Each of the vowels has both a + and - long pronunciation. For example: state, sat, pint, pin, snow, pot, cute, and cup. The attribute + or - low is named for the height of the tongue in the mouth during articulation of the sound (see Figure 1). During production of a +low vowel, the tongue is low in the mouth while it is high for a -low vowel. Speaking aloud the words in the figure demonstrates this difference.

*Current address: Bell Laboratories, Indianapolis, Indiana 46219.

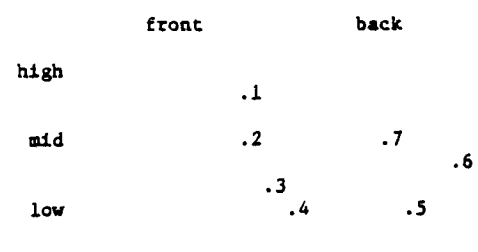

Figure 1. Tongue vowel positions. The relative position of the highest points of the tongue in vowels in 1 heed, 2 hid, 3 head, 4 had, 5 father, 6 good, 7 food. [7].

Stress is not inherent to vowels in isolation but is present only within words. Stress of a vowel phoneme within a word is a relative quality that is noticeable only by contrast with surrounding phonemes. Consonant phonemes may also be defined in terms of several different attributes, but within this theory their main purpose is to combine with vowels to complete the syllable structure of the words.

In English, each syllable of a word must contain at least one vowel. A syllable can be a single vowel, rode-o; it may be an open syllable with the vowel at a syllable boundary, po-lice, ar-ticulate or it may be a closed syllable with the vowel surrounded by consonants, Mon-tana. The term 'vowel' in this context means vowel phoneme and not orthographic vowel; the same is true for consonants. The th in thing is considered a single consonant phoneme.

The LP model defines context sensitive rules that can be used to predict which vowels within a word should be stressed. The three rule types are:

1) English Stress Rule and the Stress Retraction Rule - ESR and SRR,

2) English Destressing Rule - EDR, and

3) Exceptionless Vowel Lengthening Rule - EVL.

As the names imply, the first and second rules deal with assignment of + or - stress, while the third predicts which vowels should be long. All three rules operate within a word from right to left.

In the first stage, the shape of the penultimate (next-to-last) syllable determines the assignment of the + stress attribute using the ESR rule. "If the penultimate vowel is short and followed by (at most) one consonant, then stress falls on the preceding syllable," [9] as in Table 1(a). "If the penultimate vowel is long [Table 1(b)] or followed by two or more consonants [Table 1(c)] then it must bear stress itself." [9] Each of the previous statements assumes the final vowel is short. The fourth case of the ESR says that if the final vowel is long then it must bear stress, Table 1(d). (See [9] for exceptions to this first stage.)

TABLE 1. Examples of the ESR.

a.	b.	c.	d.
América	aróma	defóctive	negáte
canónical	Cardóna	referéndum	repúte
Éverest	hormónal	amálgam	eróde
asparagus	horízon	eréctor	balloón
polýgamous	desírous	anárthrous	bállyhóo
élephant	adjácent	Charýbdis	éxplóit

In the second stage, the +stress attribute is assigned based on the position of the leftmost +stress vowel in the word. Since the rule retracts stress across the word it is called the Stress Retraction Rule (SRR).

The ESR and SRR mark certain vowels to be stressed; this however does not imply that when the word is spoken, each of the vowels will be stressed. There are instances, depending on the characteristics of the word, where vowels will lose their stress through the application of the English Destressing Rule (EDR).

The EDR depends on the notion of metrical trees whose purpose it is to give an alternating rhythm to the syllables of a word and define the relative prominence of each syllable within the word. Rhythm is reflected by the assignment of the attribute s, strong, to stressed syllables and w, weak, to unstressed syllables. For the words labor, caprice, and Pamela the trees are simple (see Figure 2). The first rule in building the tree is if the vowel is -stress then its attribute is w, if the vowel is +stress then it may be s or w. The root node of any independent subtree or the root node of the final tree is not labeled. The s w labeling defines a contrast between two adjacent components of a word; therefore, a solitary s or w would have no meaning.

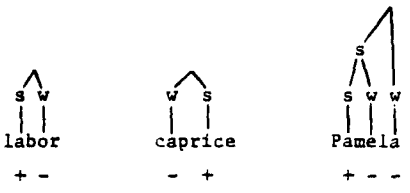

Figure 2. Assignment of s and w.

Each time a +stress is assigned by either the ESR or the SRR an attempt is made to add to the tree. As in the word labor a node is added to the tree and the vowels are marked s or w according to their stress markings, + or -. Next, any unattached vowels to the right of the new node are added, as with Pamela. This builds a series of binary subtrees that are necessarily left branching

(see Figure 3). There are some situations where nothing can be added to the tree after the assignment of +stress. Such words cause a rephrasing of the second step above to become: next attach any vowels to the right of the present vowel that have not been attached during the operation of a previous rule.

These two steps allow trees such as those in Figure 4 to be formed. Two questions remain. How is the tree completed? How are the s, w relations defined above the vowel level?

To answer the first question; after all unattached vowels to the right have been attached into a left branching subtree, this subtree is joined to the highest node of the subtree immediately to the right, if it exists (see Figure 5).

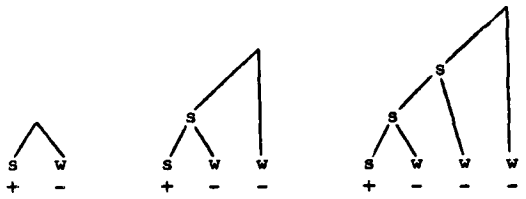

Figure 3. Leftbranching binary subtrees.

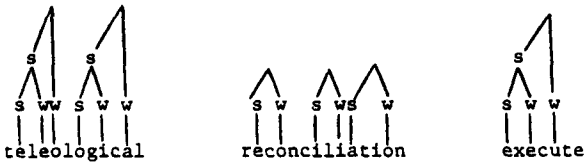

Figure 4. Connection of unattached nodes.

The s, w assignment is made by the Lexical Category Prominence Rule (LCPR). In its simplest form it states: In the configuration [N1,N2] within a lexical category, N2 is s if and only if it branches. The LCPR has already been used in the stress assignments of teleological, Pamela, and execute, to connect unattached vowels to the right of the + - sequences. The LCPR also follows the convention that no -stress vowel is assigned s.

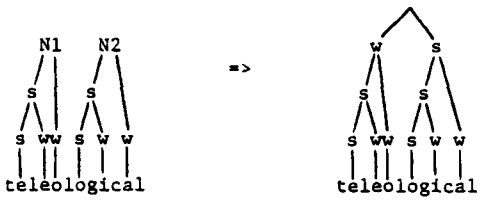

Figure 5. LCPR example.

To insure that all vowels are included in the tree, one final step is necessary as illustrated by the word Monongahela. Following the rules as previously outlined will generate a stress assignment and tree such as that in Figure 6(a). The first vowel must be included in the tree to produce Figure 6(b). This is done as the last stage of tree building. The LCPR is used in this case to join the vowel and the tree structure and to assign s, w values.

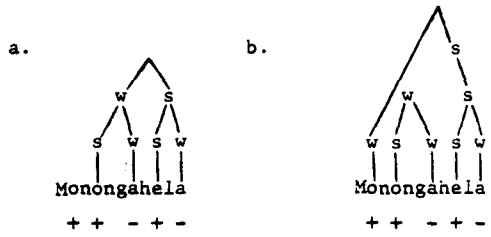

Figure 6. Final step in treebuilding.

The English Destressing Rule (EDR) is used to determine which vowels should be reduced. Generally two things happen when a vowel is reduced. First, it will lose its +stress attribute and secondly, the vowel sound will be reduced to a schwa (an indeterminate sound in many unstressed syllables, e.g. the leading a in America). The rule is based on the tree prominance relations of the metrical trees, and is restricted to operating on only those vowels that have been marked +stress by either the ESR or SRR (see [9]).

Finally the Exceptionless Vowel Lengthening Rule (see [9]) is applied to handle apparent exceptions in the operation of the ESR, e.g. words such as alien, simultaneous, radium and labia which contain a vowel sequence preceding the vowel to be stressed.

III IMPLEMENTATION

Converting a theoretical model such as that proposed by LP into a computerized implementation poses problems. One concern is whether the rules and definitions of the theory are well suited to a computer implementation, or if not, must they be transformed to such an extent that they no longer resemble the originals? Fortunately the LP theory is expressed in rules and definitions that easily lend themselves to an implementation.

Overcoming other problems while remaining close to the LP theory involves a careful combination of three factors. First, certain modifications must be made with the application of the rules for locating the +stress attribute and building metrical trees. Second, several assumptions must be made about the exact definitions of the terms such as VOWEL and CONSONANT. Third, some of the rules which are too general must be restricted. None of these modifications causes a drastic reshaping of the model.

Three outcomes exist for a word being processed by such a system. One, the stress pattern of the word will be correctly predicted. Two, the stress pattern of the word will be incorrectly predicted. Three, the word will drop through without the system being able to predict any stress. Any modifications, assumptions or restrictions imposed should be done with the primary intent of reducing the number of words for which an incorrect stress pattern is predicted, even if this means increasing the number of words which drop through.

One modification was to use a phonetic translation of the word instead of its standard spelling. This meant working from an underlying representation rather than the surface representation. By working from the underlying representation, the attributes +-stress, and +-low could be differentiated from the phonetic alphabet character directly because a +long vowel and a -long vowel would be represented by two different characters in the phonetic alphabet. Four immediate results occur from making this modification. First, single consonant sounds such as the th in thing are represented by a single character. However, the same is not true for dipthongs. Both IPA symbols and VOTRAX codes (a VOTRAX ML-I speech synthesizer was used to output the results of the stress prediction) for dipthongs are multiple character codes. Second, in a phonetic translation all reduced vowels are already reduced. Therefore for the most part the EDR is of little value. It only retains its usefulness for initial syllables that are not stressed but whose vowel is not schwa. This syllable will draw stress by the SRR creating a situation for the EDR to apply. Third, the ESR and SRR also operate less freely because they will not apply stress to a schwa. Fourth, a new rule is required to operate in conjunction with the EVL. This rule must give a final +long vowel, such as the e in story, the -long attribute so that the ESR can correctly assign stress.

A second change was that the SRR could be applied in accordance with the principle of disjunctive ordering. This situation results from the fact that a translator system has no lexicon. Although the words therefore cannot be marked for a particular type of stress retraction (SRR), it does not cause a major problem.

One implication of these modifications is the sequential ordering of the rules which group words into classes based solely on the characteristics of their phonetic translation. Therefore any set of stress rules should be organized in terms of a 'best fit' mode of application. Secondly, the stress rules cannot be defined in a way that can differentiate syllable boundaries, so no rule can be based on the concept of a 'light' or 'heavy' syllable. Although the stress rule input form does allow an affix option, it should be kept in mind that the en of enforce is considered a prefix as well as the en of English. Finally, there can be no distinction between words based on the word stem or the word origin, except, in the case of word origin, if it can be defined in terms of a distinct affix. For example the Greek prefix hetero in: heterodox, heteronyn, or heterosexual is a candidate for long retraction by the SRR.

Although the application model is a modified version of the LP model, it still operates in the manner of their original intent.

IV EVALUATION

An experiment was conducted to evaluate stress placement using the computerized version of the LP model. A random sample of unique English words and their correct phonetic translations used for the experiment was selected from the American Heritage Dictionary [1]. Five hundred pairs of random numbers were generated; the first number in the pair was a random number between one and the page number of the last page in the dictionary and the second one was a random number between one and sixty. For each pair, the first number was the page on which the random word was to be found and the second number, n, determined the word to be the n'th on the page. If n was larger than the actual number of words on the page, then n modulo the number of words on the page was used. If the selected word was not polysyllabic, it was rejected. Using this technique, 357 unique random words were selected. Each word was translated into ASCII codes for the VOTRAX according to the phonetic translation in the dictionary. These translations were then given as input to the stress system.

Because the words in the random sample contain combinations of primary, secondary, and tertiary stress, several methods arise for evaluating the results (listed in the order of importance):

1) The number of words completely correct, the number of words incorrect, and the number of words which dropped through.

2) The number of times primary, secondary, and tertiary stress were each individually predicted correctly regardless of the other two.

3) The number of times when secondary or tertiary stress was incorrectly predicted.

4) The number of times secondary or tertiary stress was predicted but the word did not require it.

5) The number of times secondary or tertiary stress was needed but not predicted.

The figures for the first evaluation are shown in Table 2. The totally correct words are slightly under two thirds of the entire sample. However, when the words with correct stress and the words which fell through are combined, the total is slightly over 70%.

TABLE 2. Words: correct, incorrect, unmodified.

	Correct	Incorrect	Unmodified
#	226	101	30
%	63.3	28.29	8.4

The results of the second evaluation are shown in Table 3. While primary stress is predicted correctly in 75% of the cases, secondary stress is only 53% and tertiary stress occurs too infrequently to make any observations. The number in parentheses in Table 3 indicates the total number of the particular stress level required.

TABLE 3. Individual stress levels correct.

	Primary	Secondary	Tertiary
#	270(357)	68(128)	3(4)
%	75.63	53.12	75.

The third evaluation results are shown in Table 4. The 19% in which secondary stress was placed on the wrong syllable is small but still significant. Again tertiary stress occurrences were too few to make observations.

TABLE 4. Incorrect prediction of secondary and tertiary stress.

	Secondary	Tertiary
#	25(128)	1(4)
%	19.53	25.

The results of the fourth test are given in Table 5. Considering that there were 357 words in the sample, this is a relatively small number of erroneous predictions.

TABLE 5. Stress that should not have been predicted.

	Secondary	Tertiary
#	3	1

Finally the fifth evaluation leads to Table 6. This table shows the number of times secondary or tertiary stress was required but not predicted. An interpretation of this table suggests that for 35 words which needed both primary and secondary stress, only primary stress was predicted. These words are also included in the incorrectly stressed words of Table 2. The importance of this fact appears when one considers that the stress pattern is partially correct, but is not distorted by incorrect stressing. Therefore even though partial, this stress pattern would be an improvement. If these words are now combined with the totally correct words and those which dropped through, they equal 291 words or 81.51%, i.e. almost 82% of the words can be stressed totally, partially, or left unchanged.

TABLE 6. Secondary and tertiary stress which was not predicted.

	Secondary	Tertiary
#	35(128)	0
%	27.34	0

With 63.3% of the sample words completely correct, 73.10% of the sample words completely or partially correct, 8.4% unmodified and 18.49% in error, this test has demonstrated that the stress model defined by the stress system and its input rules does work in a substantial percentage of cases.

Of the 66 words that were incorrectly stressed, most fall into one of four categories.

1) Two syllable words where the vowel pattern is -long -long or +long +long and the last syllable is stressed. In these cases the stress system incorrectly assigns stress to the first vowel: e.g., transact, mistrust.

2) Words in which the ESR or SRR skips over syllables that should be stressed, e.g. isodynamic, epoxy, comprehend, remitter, inopportune.

3) When in a two syllable word, the word stem vowel is short and the prefix or suffix vowel is long, the long vowel is marked for stress, e.g. fancied.

4) The LCPR does not correctly assign nodes s, w, values, e.g. contumacy, gastight.

Each of these groups is an exception to a larger group whose stress patterns fit the predicted patterns.

A final question is: How well does this system predict stress in the most common English words? Of the 200 most common, 162 have a single vowel in their phonetic translation and therefore would drop through the system without being modified. Of the 38 remaining words, 33 are correctly stressed by the stress system, leaving 5 incorrectly stressed. However, since these are the most common of words of English, it would seem reasonable to include these words as special rules

in the rule system of the translator and not allow the stress system to operate on them.

V SUMMARY

Computer synthesized speech and linguistic theories for predicting stress can interact with one another to mutual benefit. Computer synthesized speech techniques can be used to evaluate the linguistic theory. Just as computers have been used so often to evaluate theories in other disciplines, so too can they be used in linguistics. The organization, speed, accuracy and unbiasedness of the computer makes it superior to a person in many respects for judging a hypothesis.

On the other hand, the linguistic theories can provide a substantial base on which to build language components of artificially intelligent systems. The intelligibility of computer synthesized speech can be improved with the application of linguistic theories for predicting stress such as that proposed by Liberman and Prince.

Evaluations such as that presented in this paper will be of value not only in comparing competing theories but will also be helpful in determining whether the accuracy of a theory's predictions is acceptable for a particular application and where improvements may be made to the theory.

VI REFERENCES

1. American Heritage Dictionary, 1980.

2. Basboll, H., Phonology, Language and Speech, 23: 91-111, 1980.

3. Fujimura, O., Perception of Stop Consonants with Conflicting Transitional Cues: A Cross-Linguistic Study, Language and Speech, 21, 337-346, 1978.

4. Fujimura, O., Modern Methods of Investigation in Speech Production, Phonetica, 37: 38-54.

5. Hill, D. R., A program structure for event-based speech synthesis by rules within a flexible segmental framework, Int. Journal of Man-Machine Studies, 10: 285-294, 1978.

6. Klatt, D. H., Linguistic uses of segmental duration in English: Acoustic and perceptual evidence, Journal of the Acoustical Society of America, 1976.

7. Ladefoged, P., A Course in Phonetics, Harcourt Brace Jovanovich, Inc., 1975.

8. Lehiste, I., Suprasegmentals, The M.I.T. Press, 1970.

9. Liberman, M. and Prince, A., On Stress and Linguistic Rhythm, Linguistic Inquiry, 8(2): 249-336, 1977.

10. Paradis, C., The Role of Canadian Raising and Analysis in Syllabic Structure, Canadian Journal of Linguistics, 25, 35-45, 1980.

11. Witten, I. H. and Abbess, J., A microcomputer-based speech synthesis-by-rule system, Int. Journal of Man-Machine Studies, 11: 585-620, 1977.

12. Yip, M., The Metrical Structure of Regulated Verse, Journal of Chinese Linguistics, 8: 107-125, 1980.