THE SIMULATION OF STRESS PATTERNS IN SYNTHETIC SPEECH - A TWO-LEVEL PROBLEM

Timothy J Gillott

[1]Department of Artificial Intelligence
Hope Park Square
University of Edinburgh
Edinburgh EH9 2NH
Scotland

ABSTRACT

This paper is part of an MSc. report on a program called GENIE (Generator of Inflected English), written in CProlog, that acts as a front end to an existing speech synthesis program. It allows the user to type a sentence in English text, and then processes it so that the synthesiser will output it with natural-sounding inflection; that is, as well as transcribing text to a phonemic form that can be read by the system, it assigns this text an f0 contour. The assigning of this stress is described in this paper, and it is asserted that the problem can be solved with reference to two main levels, the sentential and the syllabic.

0. General

The paper is divided into three main sections Firstly, Section 1 deals with the problem of stress, its various components and their relative importance. It also discusses (briefly) the two-level nature of the problem.

Part II examines the problems that the model must face in dealing with stress assignment, and further develops the contention that these problems must be dealt with at the sentential and the syllabic levels. It proposes a phonological solution to the problem of syllabic stress, based on the Dependency Phonology framework, and suggests a modified function and content word algorithm to deal with sentential stress assignment.

Part III deals with the actual algorithms developed to deal with the problems. A fair amount of familiarity with Prolog is assumed, but the code itself is not examined too deeply.

In addition, possible improvements are discussed, briefly, at the end of the paper. As this program is a prototype, there will be many such improvements, although there are no plans to produce an enhanced model at the present date. It should also be borne in mind that as this paper is primarily a report on a piece of software the linguistic bases behind some of the algorithms are by no means dealt with as comprehensively as they might be.

1. The Role of Stress in Utterances

This is by no means intended to be a comprehensive analysis of stress assignment in English, rather it is a brief review of some of the most important acoustic factors which together go to make up the perceptual phenomenon of stress, and in particular those factors most relevant to the text-to-speech program.

Stress is the name given to the group of acoustic phenomena that result in the perception of some words in utterances as being more important than others. There is no one-to-one correspondence of the acoustic level with the perceptual one, but all the members of the above group contribute to some extent, some with more effect than others. The three most important, pitch, intensity and duration, will be briefly reviewed.

1.1 Pitch

Intelligibility of English utterances is to a large extent dependent on contrasting pitch. No lexical distinction is made on the basis of pitch as in a tone language such as Mandarin, but pitch does have the property of radically altering the semantics of a sentence. Very often, pitch change is the only way to disambiguate sentences that are otherwise syntacticaly and lexically identical.

For example, consider the two examples below. They are both syntactically (and lexically) identical, but the differing intonation patterns cause the semantic interpretation of the two to differ considerably:

The elephants are charging.

The elephants are charging.

The first sentence conveys the information that a group of elephants happen to be performing a certain action, that of charging, whereas the important information contained in the second is that it is elephants that are doing the charging, as opposed to rhinos or white mice. This is what is meant by saying that the movement of pitch is closely connected with semantic content.

[1]Now at: British Telecom Research Laboratories, Martlesham Heath, Ipswich, Suffolk IP5 7RE, UK

An important point arises here; this is that although the meaning of the whole sentence is changed by the different intonation pattern, the actual words themselves retain the same meaning in both examples. That is, there are two levels of semantic information contained within a sentence; morphological (word level) and sentential (utterance level). This distinction is important and runs through the whole problem of synthetic stress assignment, and will be considered in more detail later in the paper.

Although sentential stress often varies, morphological stress does so much less frequently. For instance, the stressed syllable is the first one in the word "elephant". To put it on the second syllable would destroy the semantic message conveyed by the word "elephant". When morphological stress does differ within the same word, it invariably accompanies a radical difference in the semantics of a verb, and is usually syntactically defined; viz próject (the noun) as opposed to projéct (the verb).

It is obvious to say that pitch varies to indicate stress within both words and utterances. But how does it vary? It would be tempting to say that a stressed syllable is always signalled by a rise in pitch, as in the examples above. This is indeed true in a great number of cases, but by no means all, as pointed out by Bolinger (Bolinger 1958). For instance, consider the following phrase (taken to mean "do continue"):

Go on.

Clearly in this common utterance, it is the "on" that is emphasised, and it can easily be seen that pitch is lower for this word. Bolinger determined that pitch movement, rather than pitch rise only, is the important factor and that the point in the sentence where intonation is perceived to rise or fall serves as an important indicator of stress.

1.2 Intensity

The subjective impression often gained from a stressed word in an utterance is that it is somehow "louder" than the non-stressed words. If this were so, it would be reasonable to assume that there would be some physical evidence for this in terms of effort made by the speaker, and in terms of measurable intensity. Until fairly recently, no method existed to prove satisfactorily that effort increased when a word was stressed, but experiments by Ladefoged (Ladefoged 1967) to obtain myographs of intercostal muscle movement have revealed a heightened tension in these muscles when articulating stressed syllables. The same set of experiments also revealed a small increase in subglottal pressure when a speaker emphasised a syllable. So physiological evidence does point to increased effort expelling the airstream when stressed syllables are produced. This should have some correlate in measured intensity.

1.3 Duration

Duration is recognised as being connected with the perception of stress, even if people tend not to recognise it as such. This holds for synthetic speech as well as for natural speech. Experiments carried out with an early version of the stress assignment program indicated that duration is useful, if not essential, to produce a natural-sounding stress pattern, particularly sentence-finally. A sentence with natural f0 movement and durational increase on the stressed syllables was contrasted with the same sentence with just f0 movement. The result was perceptively more natural-sounding with both pitch movement and durational increase, although it was perfectly intelligible without the durational increases. This ties in with observed phenomena in natural speech and will be discussed below.

1.4 Relative Importance of Pitch, Intensity and Duration

Experiments conducted by Dennis Fry (Fry 1955) indicated that the three contributive factors discussed above are by no means equally important in stress perception. A minimal pair list was taken, and stressed syllables were presented with two out of the three factors present, to see what effect this would have on perception. This is to say that the words would be introduced with pitch movement and durational increase, but no change in intensity: or intensity and pitch change would be varied normally, but duration of all syllables would be kept constant. The results showed that pitch was by far the most significant factor in stress perception, followed by duration. Intensity was relatively unimportant even to the point of being mistaken for another parameter (Bolinger, op. cit).

Bolinger found that an increase in intensity with no corresponding pitch increase was nevertheless heard as a pitch raise. Interestingly enough, a drop in intensity was not heard as a drop in pitch, merely as a form of interference, as if the speaker's words were being carried away by the wind.

Similar experiments carried out with an early version of this program indicated that the same could be observed in synthetic speech. Intonation clearly had the greatest effect on intelligibility; duration was seen to be important but not vital to intelligibility; and intenisty was seen to be relatively unimportant.

It was therefore decided to represent stress in the program as a combination of intonation movement and durational change. Intensity was not included because the software that drove the synthesiser had no facility for user alteration of this parameter. Taking into account the relative unimportance of intensity as a cue for stress, it was not though worthwhile to introduce such a facility to the driver software.

2. Problems Facing the Model: Types of Stress

It can be seen from the brief outline given above that GENIE must deal with a complex problem in assigning stress to utterance. The program must take the whole utterance, assess it in order to see where stress peaks should occur, and assign them dynamically. A complex phenomenon has to be represented using very sparse information.

2.1 Types of Stress

Stress assignment is a complex issue at at least two linguistic levels. As seen in 1.1 above, there is a notion of stress both at the syllabic and the sentential level. Even if the stressed words were predicted correctly within the sentence by the program (and this is a far from trivial problem) there still remains the problem of correctly predicting the stressed syllable(s) within the words themselves. Many theories have been advance, both syntactical (eg Chomsky and Halle 1968) and metrical (eg Liberman 1979) to propose a solution to this problem in natural speech. Whilst acknowledging these hypotheses, a phonological solution will be proposed which seems to handle at least as many cases as do the foregoing. This is the theory that has been implemented in GENIE, and although at present it is in a prototype stage only, it works well.

This solution takes as its base the Dependency model of vowel space, and proposes that it is possible, at least for English and possibly for other stress languages, to predict syllabic stress on the position of the syllabic nucleus within a "sonance hierarchy". This is a central notion of the Dependency Phonology model (Anderson 1980), and a brief outline of the model follows for those unfamiliar with it.

2.1.1 A Brief Outline of the Dependency Model of Vowel Space

Various phonological theories have argued for a non-discrete vowel space, as opposed to a discrete scale as evidenced in Chomsky and Halle's system of assigning vowels fixed heights, eg +low etc. Among the models arguing for such a non-discrete space is Dependency Phonology (Anderson, 1980), which takes as its position that there exists a linear "scale of sonance" from which continuum points can be chosen. These points are recognised as vowels. In fact the model goes further than this in postulating a scale of sonance for all sounds, as will be seen below.

The notion "scale of sonance" needs some clarification. Sonance, or sonority as it is also known, is best defined acoustically. A highly sonant sound is characterised by having a high energy content and strong formant banding when examined on a broad-band spectrogram. These qualities are those possessed by vowels, and in fact the model equates sonance with "vowelness", the degree to which a given sound is like a vowel. Thus on the "sonance hierarchy", vowels have the most sonant position, and the continuum goes from this point via liquids, nasals, voiced fricatives and voiceless fricatives to voiceless plosives, the least sonant of all. Thus the points of the scale are distinguished from each other in that their acoustic makeup possesses an amount of "vowelness" that can be compared with that of their neighbours on the scale. This system is the exact opposite in concept to the Chomsky and Halle type stepped scale; it is a stepless scale.

The part of the sonance hierarchy that interests us most is the more vocalic end. However, the scrutiny will extend to cover all sounds.

2.1.2 Using the Model

This is all very well in theory, but it must be applied. As was said before, the central idea is that words can be assigned stress on the basis of the positions occupied by their component segments on the sonance hierarchy. Taking vowels only for a moment, let us see how this works. The vocalic end of the scale can be seen as shown below, always bearing in mind that labels such as "V" or "VSon" are only points along a continuum:

VVSon	+
VVC	Sonance
VSon	
VC	-

Thus a word like "proposal" can be seen to have three syllabic nuclei, one of VC, one of VVC, and one of VC. Following the notion of sonance as the guiding principle, it can be seen that the primary stress should be awarded to the diphthong. And this is indeed true.

But what about words whose syllabic nuclei both appear to share the same point on the scale, eg "rabbit", "object"? To attempt to explain this, the notion of the sonance of individual vowels must be considered.

Vowels themselves can be ranked on a scale of sonance. Some vowels are more sonant than others. Examples of this would be [a] as opposed to [i] or [u]. The theory of Natural Phonology (Donegan and Stampe) express this concept in terms of colour. [a] is more sonant and less "coloured", in this model, than [i] or [u]. In Dependency Theory, the difference is expressed in terms of "vowelness" or sonance. This notion equates to acoustic values, where [a] is seen to have more energy than [i] or [u] due to the wider exit shape of the vocal tract for the former. Experiments carried out by Lehiste (Lehiste 1970) show that this is also borne out perceptually. When speakers were asked to produce [a] and [u] at what they considered to be the same "loudness", the dB reading for [a] was in fact considerably lower than that for [u]. This showed that [a] was perceived as being in some way "louder" and requiring some compensation in order to pronounce it at the same subjective level as [u].

Thus it seems reasonable to propose a scale of sonance for vowels as well as more generally for all speech sounds. When a word like "rabbit" is examined, it can be seen that [ae] wins the stress assignment as it is much more sonant than [ɪ].

Counter examples do exist, and will be briefly outlined. As it is not the main purpose of this paper to expound a linguistic theory, the outline will not be as rigorous as it might otherwise have been. These counter examples divide roughly into three groups.

(1) Two forms of the same word can have different stress assignment depending on their syntactic category. Thus:

Noun	object
Verb	object

The only explanation that can be advanced for this in terms of the theory proposed above is that the two VC groups are close to each other in terms of sonance. [ɒ] and [ɛ] are both reasonably near the centre of the tongue height space. Pairs that exhibit similar behaviour seem to share this characteristic:

NOUN	VERB
próject	projéct

It is suggested that only such pairs of words that have VC groups whose sonance levels are sufficiently close can exhibit this behaviour, and even then no explanation can be advanced as to why this should be so. It seems likely that the only explanation is a syntactic one.

(2) Words such as "balance", "valance", etc present a problem as it is not immediately apparent as to why the stress should be assigned to the first VSon group; both the vowels are the same. However, it should be remembered that nasals possess less overall energy than do liquids, albeit not much less. It is suggested that a VNasal group is marginally less sonant than a VLiquid group.

(3) Words with suffixes also tend to present a problem, viz:

plástic but plastícity.

It is suggested that the only answer to this is a syntactic one.

Many words were examined in this way, and although there was never anything like one hundred percent correctness, it was seen that such a notion could form the basis for a robust, compact algorithm for syllabic stress assignment, without the need for many production-type rules as seen in the systems that use MIT-type syntactic stress assignment rules. It can also be seen from the above that a syntactic component will probably be needed to supplement the purely phonological solution in a developed system. However, it is submitted that an algorithm based on this system will be considerably less cumbersome than those currently used, and should also produce a compact, natural solution to the problem.

2.2 Sentential Stress

The problem of stress, as stated above, is a two-level problem. As well as being assigned to syllables within the word, stress is also assigned to the whole sentence. The problem is that no one seems to have produced a definitive set of rules from which an algorithm for sentential stress assignment can be evolved. Most text-to-speech systems use the notion of "function" and "content" words. While by no means claiming to solve this problem, an algorithm will be suggested for sentential stress assignment which works somewhat better than those in present systems.

3. Algorithms Developed

This selection will explain how GENIE deals with the two-level problem of stress assignment. It must be emphasised that the solution proposed is little more than a prototype, and does not present a complete solution to this complex problem. The operation of the Prolog will be examined in principle, but without going too deeply into the code.

3.1 Sentence Processing

Firstly, the user types in a sentence in normal English text, with word boundaries indicated in the normal way by spaces. Each word is read in and instantiated to an item in a Prolog list. Element separations are indicated by commas. Now the program has to convert the English list elements to a phonetic transcription. The approach taken was not to use grapheme-to-phoneme for this prototype system. Instead, the words were looked up in a dictionary and the relevant list changed element by element. An example will clarify the stages up to this point:

English text: this is a tricky project.

List form: this,is,a,tricky,project,.

"Phonetic" form: [dh,qq,i,s,i,z,qq,zz,a,
ch,ci,rr,i,k,ky,kz,i,
p,py,pz,rr,o,j,jy,e,k,
k,ky,kz,t,.]

This sentence now has to be classified using two criteria; firstly the punctuation (giving the overall sentence type) and the syntactic structure. The last element in the list is a full stop. This tells the program that the sentence is a declarative. If it had had a question mark, further processing would have been done to determine what type of question, ie WH-question, reverse-NP question etc. When this has been done, the relevant intonation pattern is selected. Notice that the sentence is not parsed in any recognised way to determine the type of intonation pattern. There are merely a series of informal questions ie "Is sentence a question? If it is, is this question a WH-question?" These informal checks seem to be all that is necessary.

3.2 Assignment of Intonation

The two level problem of intonation assignment is dealt with in this program by first assigning an intonation contour to the sentence, and then modifying the words that the program selects as stressed. The following general scheme was adopted:

(1) Assign a general intonation slope to the sentence.

(2) Fit it to the length of the sentence

(3) Find the stressed word(s) in the sentence

(4) Assign stress peaks to them

(5) Interpolate values either side of these peaks to form a slope

Note that this description is really too vague to be called an algorithm. Each section contains algorithms, however, and they will be explained in turn.

3.2.1 Assignment of General f0 Contours

The classification of the sentence was done in order that the program should select the correct intonation slope, peak values etc for the type of sentence typed in. These slopes are simply Prolog lists of small integers, eventually intended to be read by the program as f0 values. The values used were obtained from analysis of recorded sentences spoken by the author. For instance, the "skeleton slope" for a declarative sentence was found, when the relevant Hz values had been translated into values suitable for the program, to descend from an initial value of 12 to a final value of 6. The slope was expressed thus:

[12,11,10,9,8,7,6]

It can be seen that as all sentences are different lengths, this general slope must somehow be "fitted" to the sentence. "Length" in this context refers to the length of a Prolog list; thus the list above would have a length of seven elements, each element being delimited by a comma.

The transcribed list above is rather longer; it has 30 elements. Obviously each sentence is going to differ in length. The algorithm eventually adopted was as follows:

(1) Find the length of the phonetic list

(2) Find the length of the selected skeleton slope

(3) Perform an integer division on the length of the phonetic list by the length of the slope

(4) Use the result as a sentinel. The head of the skeleton slope is assigned to a third list until the sentinal number is exceeded. In this way, a list is built up which has repeated occurences of the skeleton slope values to allow a slope of the same length as the phonetic list to be built up, although the original skeleton remains the same length.

(5) When the slope is empty, any remaining elements in the sentence list are assigned to the last non-null value in the slope.

Parts (1) to (3) of the algorithm were easy. The built-in predicate length/2 found the lengths of the relevant lists. Part (4) was a recursive routine that built up a list of integers, doing one of two things as conditions in the algorithm dictated:

(a) If the element in the phonetic list is a phone and the value of the sentinel variable has not been exceeded, then assign the present value of the head of the skeleton slope to the list being built up. Then recurse down the phonetic list but not the slope, so as to assign the same value to the next element in the phonetic list.

(b) If the sentinel value has been exceeded, then recurse down both the phonetic list and the slope so as to assign the next value in the slope to the phonetic list.

Part (5) is self-explanatory; the sentence is always longer than the slope by a few elements, so a "filler" element was necessary. This was the end pitch of the slope list, which for a surprisingly large number of sentence types was 6.

3.2.3 Finding the Stressable Words

The system used by most text-to-speech systems to select stressable words is that of content and function word, and this system is no exception. However, it was mentioned that the algorithm used was a slight improvement on existing ones. The algorithms that exist tend to use a strategy of stressing the last content word in a sentence. While this is reasonable as stress in English tends to occur climactically, it results in a rather monotonous rendition of sentences if more than one is spoken in succession.

The algorithm that was developed carries its improvement in the way it controls which content words are to be stressed in any given sentence, and works as follows:

(1) If the sentence is a declarative, an emphatic or a WH-question, then select for stressing any content words that occur AFTER the verb.

(2) If the sentence is an NP-AUX inversion question and there are content words after the verb, stress the content words, but not the verb. The main verb is taken as the marker, not the auxiliary.

(3) If, in either of the above types, there are no content words after the verb, then stress the verb.

This covers a substantial subset of the commonly occurring stress patterns in English, but by no means all. One major improvement to this program lies in increasing the subset dealt with. This algorithm is readily admitted to be the most unsatisfactory area of the program. The notion of the verb as a marker is linguistically suspect, and only acts as a convenient marker for the program to recognise. Stress can occur both before and after the verb, and in the present implementation there is as yet no means of dealing with this.

3.2.4 Assigning Stress Peaks

The procedure that finds the stressable words uses the original English text in Prolog-list form. The list is searched according to the following algorithm:

(1) Go through the list recursively, checking each word for membership of the "verb" list. When one is found, go to (2).

(2) Search the remaining part of the list recursively until a content word is found. Find out what position this element is in the list, and then assign its phonetic counterpart a syllabic stress pattern. If no word is found, keep searching until the end of the list is found, in which case go back to the verb and assign it a syllabic stress pattern.

(3) If neither verb nor content words are found, report an error.

3.2.4.1 Assigning Syllabic Stress

Before the word(s) chosen by the foregoing algorithm can be assigned to the list, the correct syllable within that word must be stressed. This is where the principle of sonance hierarchy comes in. It was mentioned that there is a notion of a scale of sonance. This notion was implemented quite simply. Each member of the scale is given a weighted valued dpending on its sonance, ranging from 1 for a voiceless plosive to 11 for a diphthong followed by a sonant. The list used for this is the phonetic version of the English text word. For example, suppose the word "program" had been chosen to be assigned the stress peak. This word would be represented in the system's phonetic alphabet as

[p,py,pz,rr,oa,ob,g,gy,gz,rr,aa,m]

This list, when the syllabic stress assignment routine had performed its function, would have a companion list that looked like this:

[1,-1,-1,1,9,-1,1,-1,-1,1,8,1]

The -1 values are dummy values given to elements such as "PY" and "PZ" which are needed by the system in order to produce the various acoustic components of plosives and have no relevance to stress assignment. Hence they are given very low values to preclude their ever being chosen to act as a stress peak.

Another routine takes the maximum integer value in the list and marks its position. A copy of this list has a special symbol substituted for the relevant element, thus:

[1,-1,-1,1,$,-1,1,-1,-1,8,1]

and this symbol is inserted into the main list. This can be done by virtue of the fact that the phonetic list is in face made up of smaller lists of the individual phonetic representations of the English words. There is a straight forward substitution of the special symbol in the list seen above for the phoneme that occupies the same position in the phonetic representation that has just had syllabic stress assigned to it. This list is then integrated into the main list.

The result of all this is a list very similar to the original phonetic rendition of the English text, but with a special symbol substituted at the point that has been chosen to have stress assigned to it.

The next step is to transfer all this to the intonation slope that was created earlier. For this process, the list with the special symbol and the list representing the general intonation trend for the required sentence are both searched down recursively; if the symbol is found at the head of the phonetic list, the relevant stress peak value (an f0 value obtained from recorded speech) is inserted in its place in a third list. Otherwise, the values of the slope are transferred to this third list.

3.2.5 Interpolation

This process ensures that there is a smooth rise and fall towards and away from the selected peak so as to give a natural effect. It takes advantage of the interpolation procedures already existing in the synthesis program. The stress peak is again found by searching down the list in a similar manner to that described above. When it is found, the following algorithm is followed.

(1) Obtain the value of the stress peak

(2) Obtain the value of the element on the left hand side of the peak

(3) Average the values obtained above

(4) Assign the result to the element on the left of the peak

(5) Do the same for the value on the right of the peak

The basic assignment of intonation to the sentence is now complete. There are, however, two additional modifications to be performed. One is invoked if there is more than one content word after the verb. Initially, both of these are assigned the same stress value, but before the interpolation is assigned, the second peak is reduced by a fixed amount that depends on the type of sentence.

The second is performed if the final word is stressed on the final syllable. It was found that a normal slope after a word-final stress peak was not steep enough to produce a convincing pitch fall. This was countered by inhibiting the normal interpolation routine to the right of any such peaks.

3.3 Durational Assignment

The synthesis program to which GENIE acts as a front-end has a set of standard durations that are assigned to phonemes. To assign duration the following algorithm was adopted:

Search down the phonetic list after stress peak assignment, doing:

(1) If the head of the list is the special symbol, increase the standard duration of the element by one.

(2) Plosive subelements (the PY, PZ etc. phones referred to earlier) have their durations doubled to increase plosive frication. Similar elements at the end of sentences have their durations tripled.

(3) Non-stressed elements with a duration above a certain level have their durations reduced by a fixed proportion.

The default is "assign -1 in all other cases". This signals to the system that a default duration should be assigned to the element.

The outcome of all this are three lists; the phonetic list, a list of f0 values and a list of durations, the last two simulating the stress patterns found in a similar sentence in natural speech.

The durational alterations were found on a "suck it and see" basis. Initially it was how to deal with durational assignment, other than lengthening duration in stressed positions. Successive values were put in in all strategic positions in the program, and the resulsts were tested by ear.

4. Improvements

As mentioned before, this program is only a prototype. The main stress assignment algorithms need to be refined; more syntactic types need to be included so that a larger corpus of English syntactic types can be included. In particular, the syllabic stress assignment program should perhaps contain some syntactic information to help the basic algorithm where phonology is inadequate.

REFERENCES

ANDERSON, J M and EWEN, C eds.
Studies in Dependency Phonology
Ludwigsburg Studies in Language and Linguistics
1980

BOLINGER, D L
A Theory of Pirch Accent in English Word 1958

CHOMSKY, N and HALLE, M
The Sound Pattern of English
New York Harper & Row 1968

FRY, D
Duration and Intensity as Physical Correlates of Linguistic Stress
Journal of the Acoustic Society of America No.27, pp 765-8 1955

LADEFOGED, P
Three Areas of Experimental Phonetics
Chapter I: Stress and Respiratory Action pp 1-49
Oxford University Press 1967

LEHISTE, I
Suprasegmentals
MIT Press 1970

LIBERMAN, L
The Intonational System of English
New York 1979