INTONATION SYNTHESIS BASED ON METRICAL GRIDS

J. Pierrehumbert, Bell Laboratories, Murray Hill, NJ 07974

A System for computer synthesis of fundamental frequency contours from a linguistically motivated abstract representation is described. The program is demonstrated using LPC speech whose Fo contours have been replaced by synthetic Fo contours.

The intonation synthesis program described here represents a preliminary step towards a computer model of English intonation patterns. In attempting to construct such a model, one of the most basic questions is what the abstract representation of different patterns should be. Here, the intonation pattern is described by local tonal specifications, which are taken to be level tones, and an envelope which defines the overall shape of the Fo contour and controls the realization of tones in the fundamental frequency domain. Tones are assigned to the text rather sparsely, and rules of interpolation govern the Fo between tonal assignments. Special implementation rules are used to create an intonation break at a phrase boundary. To synthesize neutral declarative Fo patterns, tonal specifications are made on the basis of a metrical grid, a representation of stress relationships which can be constructed from the syntax of an utterance [Liberman and Prince (1977)]. A number of optional variations on this pattern of tonal assignment may also be synthesized using the program.

The input to the program is a phoneme string annotated with phoneme durations, word and phrase boundaries, and four tones: L (low), LM (low-mid), HM (high-mid) and H (high). Formula 1 shows the input to the program to generate the Fo contour displayed in the figure for the sentence "In November, the region's weather was unusually dry.".

1) * i 4 n 3 * (H) n 9 o 12 v 6 e & 13 m 7 b 2 er 16 % dh 4 uh 4 *
(LM) r 6 ii & 12 d 4 zh 3 i 5 n 5 z 3 * (HM) w 6 e & 13 dh 3 er 11 *
w 4 uh 9 z 4 * (HM) uh 8 n 8 y 3 uu & 12 zh 6 uu 7 uh 4 l 4 ii 9 *
(H) d 9 r 3 ai & 43 $

"*" represents a word boundary; "%" and "$" are phrase boundaries. Tones assigned to a word are located on stressed syllables, indicated in the input by "&". At present, information about phoneme durations is used only to locate the tones at the correct point within the syllable. A version of the program which is being developed, however, will also simulate some of the segmental effects on Fo which are apparent in part a) of the figure.

The major phrase boundary "$" triggers the creation of a declining envelope for the Fo contour, shown by dashed lines in the figure. The baseline, which is the lowest level the Fo is allowed to reach and is also the level at which a L is implemented, declines linearly. The topline, which defines how a H would be implemented at any point in time, declines more steeply than the baseline and also declines more steeply early in the phrase than later. Data on declarative sentences in Sorensen and Cooper (in press) confirm that the topline has these characteristics in production. LM and HM are realized respectively one third and two thirds of the way up from the baseline to the topline, on a linear scale.

Each tone is realized as a 6 centisecond level section in the Fo contour. The Fo between tones is computed by treating LM, HM, and H as defining a peak in the contour, and L as defining a valley. When two peaks are sufficiently far apart, the Fo falls to the baseline and tracks the baseline before starting a rise to the next peak; when two peaks are closer, the fall from the first intersects the rise to the second, and so the Fo does not reach the baseline. For an input like 1) which contains no L's, the height and location in time of all local minima are thus determined by the height and separation in time of adjacent peaks. The finished piecewise linear contour is smoothed by running means of 9

centiseconds to create the contour shown in the figure.

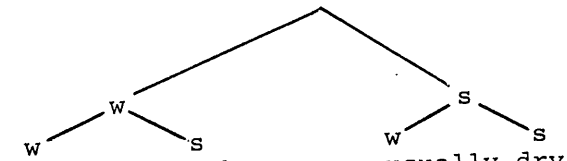
Details of this general scheme for implementing tones depends on the distinction between nuclear tones and prenuclear tones. It is well-established that the nuclear tone, which is the last tone before the minor phrase boundary "%" and falls on the main stress of the phrase, has distinctive phonetic characteristics [Ashby (1978), O'Connor and Arnold (1961), Crystal (1969)]. The program makes the following differences between nuclear and prenuclear tones: Prenuclear tones are placed at the end of the stressed vowel to which they are assigned. Nuclear tones, by contrast, occur earlier in the vowel. In accordance with Ashby's results, the peak is placed a fixed distance, here 6 centiseconds, into the stressed vowel. This means that it is near the beginning of a stressed vowel in a phrase-final syllable, which is subject to phrase-final lengthening, but near the middle of the vowel in a non-phrase-final stressed syllable, which is typically shorter. Secondly, the fall after a prenuclear tone is less steep than the fall after a nuclear tone. The program assigns slope on either side of a prenuclear tone by formula 2).

2) $$s = 4 * (1.5 - (70/(x + 43))),$$

where "x" is the distance of the tone above the baseline, in Hz, "s" is the slope to the left of the tone, and "-s" is the slope to the right. The slope function increases with peak height, but saturates at approximately 4 Hz per centisecond for extremely high peaks. For a L tone, x is Ø and the computed "s" is negative; the Fo therefore falls rather than rises to the left of the tone, and rises rather than falls to the right. The fall after a nuclear tone is assigned twice the slope computed by 2). The nuclear peak is thus asymmetric, since the slope of the rise up to the tone is computed by 2). The third difference between nuclear and prenuclear tones is that the fall after a nuclear tone reaches a lower value than one would expect from extrapolating minima earlier in the phrase. This feature of the nuclear accent is handled by lowering a trapdoor in the baseline after the nucleus, as can be seen in the figure.

Informal experimentation suggested that peak height in neutral declarative intonation is related to phrasal stress subordination, and that maintaining this relationship is an important ingredient in a satisfactory synthesis of this class of intonation patterns. An algorithm for tonal assignment in neutral declarative intonation was worked out on the basis of this hypothesis. The algorithm makes use of the trees and metrical grids developed in Liberman (1975) and Liberman and Prince (1977); we go over a single example here, and refer the reader to the original sources for a more general account. To compute a stress contour for a phrase, we begin by constructing a syntactic tree for it. 3) represents such a tree for the phrase "the region's weather was unusually dry" ("the" and "was" are assumed to be cliticized and so are not entered as words in the tree).

3)



Syntactic node labeling is omitted to make room for node labeling referring to stress relationships. Stronger stress is assigned to the right of two sister nodes by Chomsky and Halle's Nuclear Stress Rule, and this relationship is indicated by labeling the right node "s" (strong) and the left node "w" (weak). A focused element attracts stress, overriding the Nuclear Stress Rule; if "unusually " were focused in 3), for example, it would be labelled with "s", and "dry" would be labelled with "w". Our aim now is to assign a measure of absolute prominence to each stressed syllable in a way which reflects the stress relationships in the tree in 3). A numbering scheme which accomplishes this, in fact the flattest numbering scheme which gives each "w" lower prominence than its sister "s", is to count up the total number of "w" nodes dominating the syllable. Adding 1 to this number yields a standard stress transcription in which 1 is the highest stress and higher numbers represent lower stress. The

outcome in this case is 3 2 2 1. It should be noted that this transcription differs from what Chomsky and Halle (1968) would generate for the same phrase; using their rules, "unusually high" would have the contour 2 1 in isolation, but the stress on "unusually" is downgraded to 3 when the phrase is embedded as in 3) (p. 84). The Liberman and Prince transcription is adopted for the present purposes because informal experimentation suggested that Fo peaks should not be downgraded in this fashion. Following Liberman (1975), a condition requiring alternation of prominence is also imposed in constructing metrical grids for a phrase (p. 280). In the case of an extended right-branching construction, such as "organized on the model of eleven gallons of worms", this condition rules out the contour 2 2 2 2 1 but permits 2 3 2 3 1. Stress levels as determined in this way may then be translated directly into tonal specifications: 1 is H, 2 is HM, and 3 is LM. Lower levels of stress are not assigned tones. No tones are assigned after the main stress of the phrase; for example, if "unusually" is focused in "the region's weather is unusually dry", the nuclear tone falls on "unusually", and "dry", having no tone of its own, is implemented on the base-line after the steep fall from the nuclear tone.

In addition to the neutral declarative intonation pattern illustrated in the figure, the program can be used to synthesize a number of variant intonation patterns. Tones may be assigned not only to stressed syllables, but also to phrase boundaries. A LM assigned to a phrase boundary creates a continuation rise; an utterance initial H results in a high Fo onset, which, as O'Connor and Arnold point out (p. 71) adds a note of vivacity or vehemence to the utterance. Further discussion of boundary tones may be found in Liberman (1975). Secondly, English has a number of intonation patterns in which a stressed syllable has a low pitch. One such pattern is the "contradiction contour" discussed in Liberman and Sag (1974). The program currently permits the assignment of a L to a prenuclear syllable; work is underway on modeling the behavior of low tones in nuclear position.

A number of features of the approach to intonation taken in developing this synthesis program have been confirmed by recent experiments. Wales and Toner (in press) studied what kinds of ambiguous sentences may be disambiguated using intonation. The only successful disambiguations in their study involved sentences with two possible surface structure bracketings; homonyms could not be disambiguated using intonation, nor could deep structure ambiguities which were not reflected in surface structure bracketing. This result is what the approach taken here would predict, since the distinctive phonetic treatment of the nuclear tone provides a mechanism for marking phrase boundaries, but no mechanisms are provided for marking deep structure phrase boundaries or for distinguishing readings of homonyms. Experiments done by Streeter (1978) also confirm that Fo can be used to disambiguate phrasing. Nakatani and Schaffer (1978) report experiments on the perception of word boundaries in reiterant speech. They found that Fo is not a cue for word boundary location when the stress contour is fixed. This is again what the present approach would predict: the synthesis program provides no way of marking word boundaries using Fo since any sequence of words with the same stress contour will be assigned Fo patterns in the same way.

References

Ashby, Michael (1978) "A Study of Two English Nuclear Tones", Language and Speech 21, No. 4, 326-336.

Chomsky and Halle (1968) The Sound Pattern of English, Harper and Row, New York.

Crystal, D. (1969) Prosodic Systems and Intonation in English, Cambridge University Press, London.

Liberman and Sag (1974) "Prosodic Form and Discourse Function" in Papers from the 10th Regional Meeting of the Chicago Linguistic Society, Chicago.

Liberman, M. (1975) The Intonation System of English, MIT Ph.D. Dissertation, reproduced by the University of Indiana Linguistics Club, Bloomington.

Liberman and Prince (1977) "On Stress and Linguistic Rhythm", Linguistic Inquiry 8 No. 2, 249-336.

Nakatani and Schaffer (1978) "Hearing 'words' without words: Prosodic cues for word perception", J. Acoust. Soc. Am., 63 No. 1, 234-244.

O'Connor and Arnold (1961) Intonation of Colloquial English, Longmans, London (new edition 1963).
Sorensen and Cooper (in press) "Syntactic Coding of Fundamental Frequency in Speech Production", in R. A. Cole (ed.), Perception and Production of Fluent Speech, Erlbaum, Hillsdale, NJ.
Streeter, L. (1978) "Acoustic Determinants of Phrase Boundary Perception", J. Acoust. Soc Am., 64, No. 6, 1582-1592.
Wales and Toner (in press) "Intonation and Ambiguity" in Cooper and Walker (eds.), Sentence Processing.

Figure Caption

a) is an Fo from a natural utterance of the sentence "In November, the regions weather was unusually dry". b) shows how an Fo contour for this sentence is computed by the synthesis program. The dashed lines represent the envelope of the contour and the solid curve is the computed Fo contour.