S Y M P O S I U M P R O S O D I E

TORONTO MAY 29-30 1981

Synthesising English intonation

J. PIERREHUMBERT
M.I.T.
Cambridge MA

Summary

This paper describes a computer program for the synthesis of F_o contours in English. Although it can also generate expressive contours, we concentrate here on neutral declarative contours. The F_o contour is a result of prosody (or choice of stress, phrasing and intonation pattern) and local perturbations caused by the speech segments We are mainly concerned here with the prosody of utterances. Intonation is viewed as a string of high (H) and low (L) target values linked together by transitions. We describe how these target values, the pitch accents, the phrase accents, the boundary tones and the transitions are determined. These elements have been found to generate a very satisfactory approximation to normal speech.

Résumé

Nous présentons un programme pour la synthèse des contours mélodiques de l'anglais. Le contour de F_o résulte du jeu de la prosodie (accent, découpage en groupes de sens, intonation) et de la microprosodie. Nous nous limitons ici à la prosodie de la phrase déclarative neutre. Le patron mélodique étant considéré comme une suite d'éléments discrets haut (H) et bas (L) reliés par des transitions, nous décrivons une méthode pour déterminer les accents de mélodie ("pitch accents"), les accents de groupe ("phrase accents"), les tons de fin de groupe ("boundary tones") et les transitions. Cette méthode a déjà donné des résultats très satisfaisants.

SYNTHESIZING ENGLISH INTONATION

Janet Pierrehumbert

Bell Laboratories
Murray Hill, NJ

This paper describes a computer program for synthesizing English fundamental frequency contours, and sketches the theory which underlies it. The program was developed as a tool for investigating the different intonation patterns of English, and it takes as input an intonational transcription aligned with a segmental transcription. When the input is supplied from the computer terminal, the program will generate not only neutral declarative intonation, but also a number of more expressive patterns. A program which generates a transcription for the FØ synthesis program has also been developed, so that it can be used to supply intonation automatically in the Olive and Liberman text-to-speech system. Since a text-to-speech system does not have access to the pragmatic factors which would govern a person's choice among expressive intonation patterns, a neutral declarative pattern is provided. Here, we will concentrate on the synthesis of such neutral declarative patterns. A more complete description of the FØ synthesis program, including comparisons to previous proposals, can be found in Pierrehumbert (forthcoming). Pierrehumbert (1980) develops a phonological account of the English intonation system.

We would like to begin by introducing some background assumptions which are illustrated in Figure 1. Following Lehiste and Peterson (1961), Lea (1972), and others, we distinguish two contributions to this FØ contour. The prosody, or choice of stress, phrasing, and intonation pattern, is responsible for its overall shape. In addition, the speech segments introduce local perturbations: High vowels raise the FØ; unvoiced obstruents raise the FØ at the onset of a following vowel; voiced obstruents and glottal stops are associated with a dip in FØ. The work described here is concerned only with prosodic effects on FØ, and does not attempt to model the segmental effects. Next, we would like to distinguish the pitch range from the other features of the intonation. Authors since Trager and Smith (1951) have recognized that it is possible to produce the same "tune" in different ranges. For example, a calling intonation may be produced in a small range if the person addressed is close by, and in a larger range if he is far away. A more subtle point is that the pitch range becomes smaller and lower during the course of an utterance.

This effect, known as the declination effect, is documented for English in Maeda (1976) and O'Shaughnessy (1976). It has also been reported in many other languages. In keeping with this distinction, the intonation pattern is described in the synthesis program through the interaction of two levels: a pitch range, which varies in time, and a specification of the tune. The pitch range is described by a baseline, or current bottom of the speaker's range, and a topline, or current top of the range [1]. The tune is described in terms of fractions from the bottom to the top of the current range. In Figure 2, the dotted lines delimit the pitch range. The implementation of the tune by the FØ synthesis program is shown as a solid line.

Just as it is useful to decompose the continuously changing vocal tract resonances into a string of speech segments, it is also useful to decompose the tune into a string of discrete elements. There have been three general types of proposals about what these elements are. Ohman analyzes Swedish FØ contours in terms of impulses fed to a linear filter; Fujisaki et al. present a related model for Japanese. Bolinger (1951), O'Shaughnessy (1976) and Clark (1978) view FØ contours as a series of instructions to raise or lower the FØ. A third tradition, represented by Pike (1945), Trager and Smith (1951), Liberman (1975), and Bruce (1977), views tunes as a series of target values which are connected by transitions. The work presented here is in this third tradition. Clearly, any given FØ contour can be described using any of these approaches, so arguments for one approach over another must be based on regularities in the entire system of intonation. Such arguments are presented in Pierrehumbert (1980). An important feature of the present target model is that the transition between two targets is not always monotonic. In Figure 2, the transition between one target and the next dips down. This makes it possible to use a sparser specification of targets than would otherwise be necessary for a good synthesis.

The location of the targets with respect to the speech segments is determined by the stress pattern and phrasing. Specifically, a stressed syllable may carry a pitch accent, consisting of either a single target or a sequence of two targets. In Figure 3, for example, the first stressed syllable is marked with a low target while the second is marked with a high target. In Figure 4, "fath" in "father" has associated with it a low target followed by a high one [2]. A stressed syllable in a word which is presupposed in the discourse ordinarily lacks a pitch accent. In this case, its FØ contour is determined by the transition between targets on either side, just as if it were unstressed.

Additional targets are found at the beginning and end of the phrase. Boundary tones, found right at the phrase boundary, determine the FØ at the onset and the offset of the phrase. In Figure 3, we see a high initial boundary tone. A high final boundary tone in Figure 4 is responsible for the rise at the end of the phrase. The phrase accent follows the nuclear pitch accent (the pitch accent on the main stress of the phrase), and controls the FØ from there to the phrase boundary. As many authors have noted, stressed syllables after the nuclear accent cannot carry a pitch accent. (O'Connor and Arnold, 1961; Crystal, 1969; Vanderslice and Ladefoged, 1972; Ashby, 1978) The phrase accent can be seen clearly in FØ contours like Figure 4 where the nuclear stress is early in the phrase. Here the phrase accent, T_3, is low.

Much as the possible continuum from low to high front vowels is divided into the front vowels of English, target values are also divided into linguistic categories. The traditional number of categories for target level in English is four. However, Pierrehumbert (1980) showed that it is possible to formulate allophonic rules which reduce this number to two, High and Low (hereafter H and L). The American English neutral declarative intonation pattern has H pitch accents and a L phrase accent. The boundary tone is typically H if the phrase is nonterminal, and L if it is terminal. Figure 1 is an example of this type of intonation [3].

Two properties of the H pitch accent play a crucial role in our synthesis of this type of intonation. First, the target value corresponding to the H depends on emphasis or prominence; the greater the prominence, the higher the target value. (L pitch accents, in contrast, are lower under greater prominence.) For this reason, the target values in the phrase bob up and down in accordance with the phrasal stress subordination. It is our impression that the speaker has a great deal of latitude in deciding what difference in target level to use in marking a given stress relationship. We have found, however, that using target levels of 1.0, 0.7, and 0.4 to mark the nuclear accent and two weaker levels of prenuclear accent results in a good quality synthesis of the intonation. A neutral phrasal stress pattern (that is, a phrasal stress pattern which does not focus any particular part of the sentence) may be determined from its syntactic structure using the Nuclear Stress Rule and the metrical grid theory proposed in Liberman (1975). The Olive and Liberman text-to-speech system does not have access to a complete syntactic structure, and so heuristics worked out by Mark Liberman are used to arrive at a phrasal stress pattern. The nuclear stress (which receives a target value of 1.0) is placed on the last content word of the phrase; alternating

values of 0.4 and 0.7 are placed on preceding word stresses. This rough approximation to the facts of normal speech has been found to give very satisfactory results. In contrast, trial syntheses of intonation which lacked alternation in peak height were found to produce a sing-song effect, which became objectionable on extended texts. The present rules occasionally give words an anomalous prominence in the phrase. It is our feeling that this problem cannot be overcome without a model of the discourse factors which control focus in real speech.

H accents have a second property which is crucial to the FØ synthesis program; they are connected by dipping transitional functions, so that they come out as peaks in the contour. The transition between L and another target, in contrast, is monotonic, so that L's tend to come out as nonpeaks. The present system takes target values under .2 to be L and the rest to be H [4]. When two successive targets are both H, a local minimum between the two peaks is computed as a function of their separation in time and frequency. This minimum analytically determines a parabola which is fit between the two targets.

The rules for computing the minimum were worked out on the basis of the strengths and weaknesses of a previous target-based synthesis program, described in Pierrehumbert (1979). It was found that the dip should be greater between more separated targets than between less separated ones. Furthermore, the FØ contour should anticipate the level of the next target from the moment it leaves the previous one. This second point becomes more critical as the two targets are closer in time. In computing the minimum, the simplest case is that of two successive targets at the same level above the baseline. In this case, the minimum is a function of their separation in time alone. It is computed as a fraction of the distance from the baseline to the target level by the following formula:

$$F = 1 - (T * .005) \qquad \text{for } T < 20$$

$$F = .9 - ((T-20) * .015) \qquad \text{for } 20 \leq T < 80$$

Here, F is the computed fraction and T is the separation in time of the targets in centiseconds. The parameters are suitable for a normal careful speech rate, and would have to be changed in a synthesis-by-rule system which spoke slower than a real person would, as many do. In the rare cases when T exceeds 80 cs., the targets are implemented separately; the fall from the first takes up 40 cs., and then the FØ tracks

the baseline until it begins to rise 40 cs. before the second. When the two targets are at different levels, F' is computed by a procedure which scales F on the basis of the relation between the lower and higher targets. The final F is then the geometric mean of F and F'. A check is made that F does not exceed the lower target value, and if so it is reset to that value. Thus the dipping disappears in the limiting case where the two targets are very separated in frequency relative to their separation in time. Figure 5 shows the transitions fit between two targets of various levels separated by 40 cs.

The interaction of the phrase accent and the boundary tone with the nuclear pitch accent generates effects which must be replicated for the synthesis to sound natural. These effects have been most thoroughly studied in neutral declarative intonation, where L follows a nuclear H accent. It has been found that the H often occurs earlier in the stressed syllable than a prenuclear H would. The fall from H to L must be steeper than falls elsewhere for the pitch accent to sound nuclear. Lastly, when the boundary tone is L, it is lower than an extrapolation of low points earlier in the phrase would predict. (Olive, 1974; Maeda, 1976; Mattingly, 1968). In order to generate these phonetic details, the program treats the nuclear accent as a special case. First, the H target is placed early in the syllable when it is on the last syllable in the phrase, and otherwise at the end of the syllable. An earlier version of the program followed Ashby's (1978) suggestion that the peak on a nuclear stressed syllable is a fixed distance from the beginning; however, this rule placed the peak too early in words like "rival" where the stressed syllable is long but not phrase final. In order to guarantee a steep nuclear fall, the program still incorporates Ashby's observation that the fall takes a fixed amount of time, here 20 cs. The program also makes the nuclear fall go below the baseline for the phrase as a whole by lowering a trapdoor in the baseline after the nuclear stress. This can be seen in Figure 2. What the program does, then, is to generate for each nuclear H accent a following L phrase accent which has a particular location in time and pitch. Although we interpret the fall theoretically as H L, only the H has to be specified in the input. A target value for a H boundary tone may be specified optionally in the input. If it is missing, the program generates a L boundary tone at the end of the trapdoor.

A number of features of the approach to intonation taken here have been confirmed by work on the production and perception of intonation. Wales and Toner (1980) investigated what

kinds of ambiguities can be resolved by intonation. They found that intonation can resolve surface structure phrasing ambiguities, but not lexical or deep structure ambiguities. Streeter (1978) also reports that intonation can disambiguate phrasing. The approach taken here predicts these results: The boundary tones and the distinctive phonetic treatment of the nuclear tone provide ways of marking phrasing, but no mechanism is provided for marking particular readings of sentences with other types of ambiguities. Nakatani and Schaffer (1978) report experiments on the perception of word boundaries in reiterant speech. They found that FØ is not a cue for word boundary location when the stress pattern is fixed. FØ was helpful in locating the word boundary when it marked a stress pattern which was compatible with only one location of the boundary, given the constraints of the experiment. These results also conform to the approach taken here. The pitch accents provide a way of marking stress, and given the stress pattern, listeners would in some cases be able to infer the location of the word boundary. However, for any given stress pattern, the transition between targets is insensitive to the location of the boundary, and provides no additional information about where it is. The framework outlined here also makes some predictions about how FØ can be used as a cue for stress. A word with a given stress pattern can have many different FØ contours, depending on the type of pitch accent and its location in the phrase. Some of these FØ contours represent well-formed English intonation patterns on more than one assumption about where the stress is. For others, only one interpretation of the stress pattern is likely or even possible. It is only in the second kind of case that FØ would be a cue for stress. This general picture has been supported by experimental work. Fry (1958) and Morton and Jassem (1965) found that more than one FØ configuration is interpreted as a cue for stress. On the other hand, some FØ patterns Fry investigated were poor cues for stress. On our interpretation, this was because the contours were intonationally ambiguous. Nakatani and Aston (forthcoming) report that FØ is not a cue for stress on a noun following a focused adjective. This result follows from the fact that post-nuclear stressed syllables cannot be marked with pitch accents.

Footnotes

[1] An experiment reported in Pierrehumbert (1980) and Liberman and Pierrehumbert (1979) suggests that this description affords too many degrees of freedom; the slope of declination higher in the range appears to be constrained by the slope of the baseline.

[2] Due to the slow rate at which FØ changes are carried out, the high target actually falls on the post-stress syllable.

[3] According to Crystal, the most common declarative intonation pattern in British English is a terraced pattern instead of a pattern like Figure 1. The terraced pattern also occurs with a more marked meaning in American English. It is analyzed in Pierrehumbert (1980) as arising from H+L pitch accents.

[4] This rule is a serious simplification. A more complete model of English intonation would have to represent tone type directly.

References

Ashby, M. (1978) "A Study of Two English Nuclear Tones", Lang. Speech 21, 326-336.

Bolinger, D. (1951) "Intonation: Levels versus configurations", Word 7, 199-210.

Bruce, G. (1977) Swedish Word Accents in Sentence Perspective Travaux de l'Institut de Linguistique de Lund, CWK Gleerup, LiberLaromedel, Lund.

Clark, M. (1978) "A Dynamic Treatment of Tone, with Special Attention to the Tonal System of Igbo", Ph.D dissertation, U. Mass. Amherst. Distributed by Indiana University Linguistics Club, Bloomington.

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

Fry, D. B. (1958) "Experiments in the perception of stress", Lang. Speech 1, 125-152.

Fujisaki, H., H. Hirose, and K. Ohta (1979) "Acoustic Features of the Fundamental Frequency Contours of Declarative Sentences in Japanese", Research Institute of Logopedics and Phoniatrics Annual Bulletin No. 13, University of Tokyo, 163-173.

Lea, W. A. (1972) "Intonation Cues to the Constituent Structure and Phonemics of Spoken English", Ph. D dissertation, Purdue University.

Lehiste, I. and G. E. Peterson (1961) "Some basic considerations in the analysis of intonation", J. Acoust. Soc. Am. 33, 419-425.

Liberman, M (1975) "The Intonation System of English", Ph.D dissertation, M.I.T. Published by Garland Press, New York in 1979.

Liberman, M. and J. Pierrehumbert (1979) "A Metric for the Height of Certain Pitch Peaks in English", paper presented at the 98th meeting of the Acoustical

Society of America. Abstract in J. Acoust. Soc. Am., S64.

Maeda, S. (1976) "A Characterization of American English Intonation", Ph.D dissertation, M.I.T.

Mattingly, I. (1968) "Synthesis by Rule of General American English", Supplement to Status Report on Speech Research April 1968, Haskins Laboratories, New Haven.

Morton, J. and W. Jassem (1965) "Acoustic Correlates of Stress", Lang. Speech 8, 159-181.

Nakatani, L. and C. Aston (forthcoming) "Perceiving the stress pattern of words in sentences", in J. Acoust. Soc. Am.

Nakatani, L. and J. Schaffer (1978) "Hearing 'words' without words: prosodic cues for word perception", J. Acoust. Soc. Am. 63, 234-244.

O'Connor, J. O. and G. F. Arnold (1961) Intonation of Colloquial English, Longmans, London.

Öhman, S. (1967) "Word and Sentence Intonation: a Quantitative Model", Speech Transmission Laboratory Quarterly Progress and Status Report 2-3, Royal Institute of Technology, Stockholm, 20-54.

Olive, J. (1974) "Speech Synthesis by Rule", Speech Communication Seminar, Stockholm, Aug 1-3, 1974.

Olive, J. and M. Liberman (1979) "A Set of Concatenative Units for Speech Synthesis", in Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America, Acoustical Society of America, New York, 515-518.

O'Shaughnessy, D. (1976) "Modelling Fundamental Frequency and its Relationship to Syntax, Semantics, and Phonetics", Ph.D dissertation, M.I.T.

Pierrehumbert, J. (1979) "Intonation Synthesis Based on Metrical Grids", in Speech Communication Papers Presented at the 97th Meeting of the Acoustical Society of America, Acoustical Society of America, New York, 523-526.

---- (1980) "The Phonology and Phonetics of English Intonation", Ph.D dissertation, M.I.T. To be published by M.I.T. Press.

-----(forthcoming) "Synthesizing Intonation", in J. Acoust. Soc. Am.

Pike, K. L. (1945) The Intonation of American English, University of Michigan, Ann Arbor.

Streeter, L. (1978) "Acoustic determinants of phrase boundary perception", J. Acoust. Soc. Am. 64, 1582-1592.

Trager, G. L. and H. L. Smith (1951) Outline of English Structure, Battemburg Press, Norman OK.

Vanderslice, R. and P. Ladefoged (1972) "Binary Suprasegmental Features and Transformational Word-Accentuation Rules", Language 48(4), 819-839.
Wales, R. and H. Toner (1980) "Intonation and Ambiguity", Sentence Processing, edited by W. Cooper and E. Walker, Erlbaum, Hillsdale NJ, 135-158.

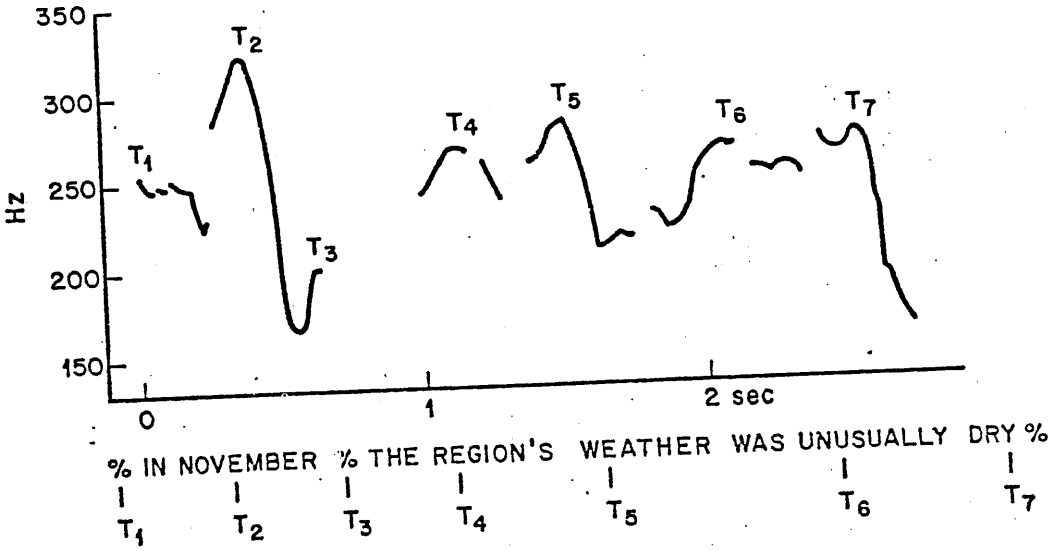

Figure 1 A neutral declarative intonation pattern. Targets marked are those used as input to the FØ synthesis program.

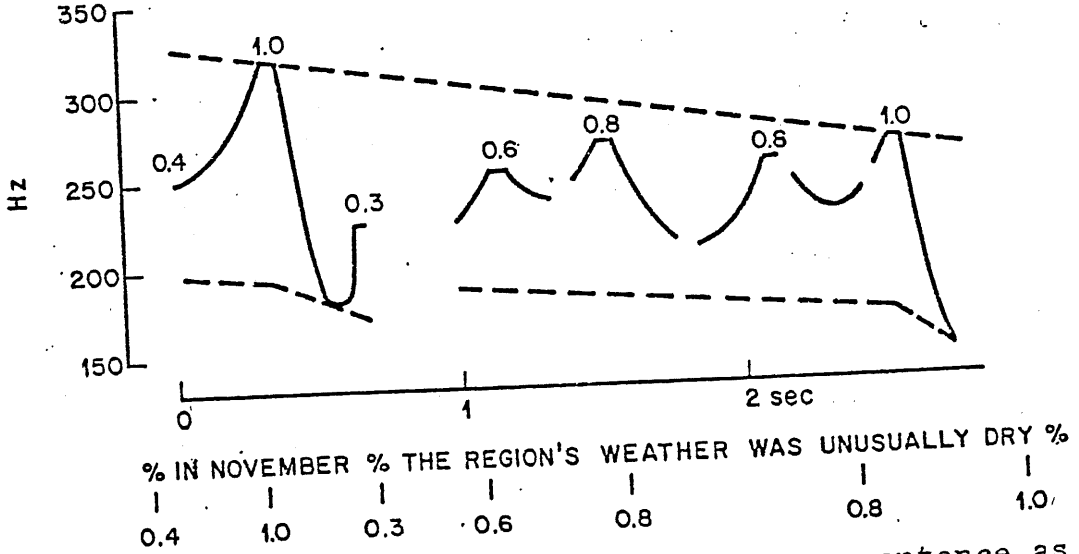

Figure 2 An intonation pattern for the same sentence as it is synthesized by the FØ synthesis program.

Figure 3 An example of the "surprise/redundancy" intonation pattern discussed in Liberman (1975).

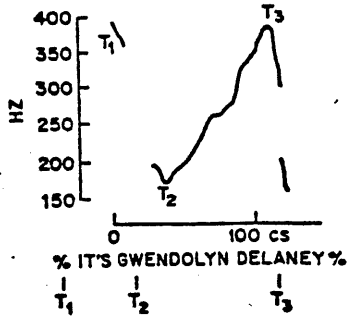

Figure 4

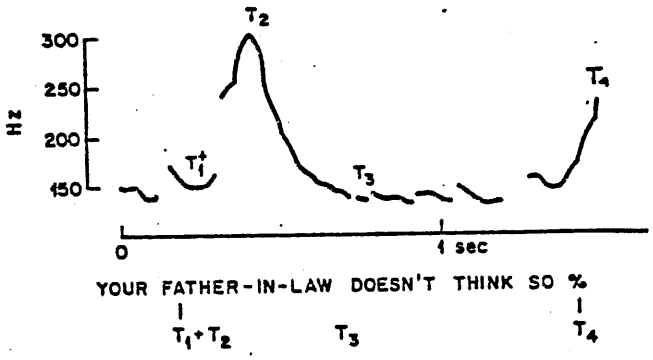
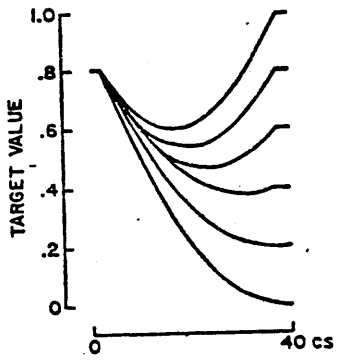
Figure 5 Superimposed FØ contours showing transitions computed between two targets which are 40 cs. apart.