Automatic Recognition of Intonation Patterns

Janet B. Pierrehumbert

Bell Laboratories
Murray Hill, New Jersey 07974

1. Introduction

This paper is a progress report on a project in linguistically based automatic speech recognition, The domain of this project is English intonation. The system I will describe analyzes fundamental frequency contours (F0 contours) of speech in terms of the theory of melody laid out in Pierrehumbert (1980). Experiments discussed in Liberman and Pierrehumbert (1983) support the assumptions made about intonational phonetics, and an F0 synthesis program based on a precursor to the present theory is described in Pierrehumbert (1981).

One aim of the project is to investigate the descriptive adequacy of this theory of English melody. A second motivation is to characterize cases where F0 may provide useful information about stress and phrasing. The third, and to my mind the most important, motivation depends on the observation that English intonation is in itself a small language, complete with a syntax and phonetics. Building a recognizer for this small language is a relatively tractable problem which still presents some of the interesting features of the general speech recognition problem. In particular, the F0 contour, like other measurements of speech, is a continuously varying time function without overt segmentation. Its transcription is in terms of a sequence of discrete elements whose relation to the quantitative level of description is not transparent. An analysis of a contour thus relates heterogeneous levels of description, one quantitative and one symbolic. In developing speech recognizers, we wish to exploit achievements in symbolic computation. At the same time, we wish to avoid forcing into a symbolic framework properties which could more insightfully or simply be treated as quantitative. In the case of intonation, our experimental results suggest both a division of labor between these two levels of description, and principles for their interaction.

The next section of this paper sketches the theory of English intonation on which the recognizer is based. Comparisons to other proposals in the literature are not made here, but can be found in the papers just cited. The third section describes a preliminary implementation. The fourth contains discussion and conclusions.

2. Background on intonation

2.1 Phonology

The primitives in the theory are two tones, low (L) and high (H). The distinction between L and H is paradigmatic; that is, L is lower than H would be in the same context. It can easily be treated as a distinction in a single binary valued feature. Utterances consist of one or more intonation phrases. The melody of an intonation phrase is decomposed into a sequence of elements, each made up of either one or two tones. Some are associated with stressed syllables, and others with the beginning and end of the phrase. Superficially global characteristics of phrasal F0 contours are explicated in terms of the concatenation of these local elements.

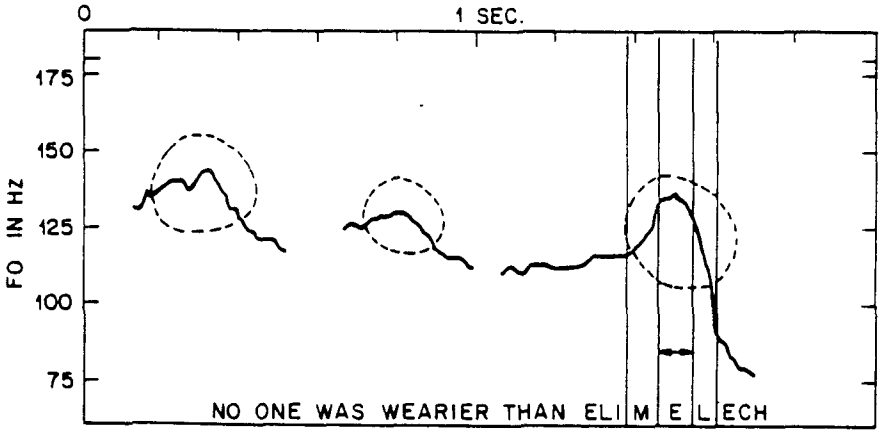

Figure 1: An F0 contour with three H* pitch accents, which come out as peaks. The alignment of "Elimelech" is indicated.

*This work was done at MIT under NSF Grant No. IST-8012248.

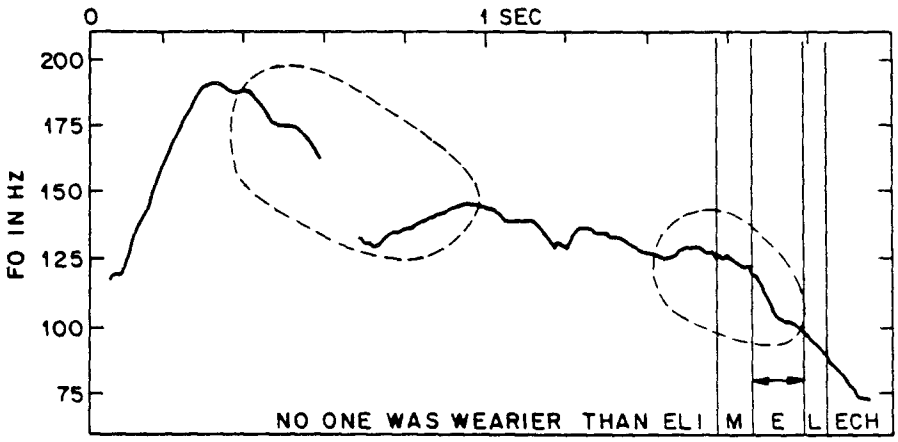

Figure 2: The two H+L* accents in this contour are circled. Compare the F0 contour on the stressed syllable in "Elimelech" to that in Figure 1.

The characteristic F0 configurations on stressed syllables are due to pitch accents, which consist of either a single tone or a sequence of two tones. For example, each peak circled in Figure 1 is attributed to a H pitch accent. The steps circled in Figure 2 are analyzed as H+L, because they have a relatively higher level just before the stress and a relatively lower level on the stress. In two tone accents, which tone falls on the stress is distinctive, and will be transcribed with a *. In this notation, the circled accents in Figure 2 are H+L*. Seven different accents are posited altogether. Some possible accents do not occur because they would be neutralized in every context by the realization rules. Different types of pitch accents can be mixed in one phrase. Also, material which is presupposed in the discourse may be unaccented. In this case, the surrounding tonal elements control its F0 contour.

The tonal correlates of phrasing are the boundary tones, which control the F0 at the onset and offset of the phrase, and an additional element, the phrase accent, which controls the F0 between the last pitch accent and the final phrase boundary. The boundary tones and the phrase accent are all single tones, either L or H. In what follows, a "%" will be used as the diacritic for a boundary tone. Figure 3 shows two pitch contours in which a L phrase accent is followed by a H% boundary tone. When the last pitch accent is early in the phrase, as in 3A, the level of the phrase accent is maintained over a fairly long segmental string ("doesn't think"). In 3B, on the other hand, the pitch accent, phrase accent, and boundary tone have all been compressed onto a single syllable.

As far as is known, different pitch accents, phrase accents, and boundary tones combine freely with each other. This means that

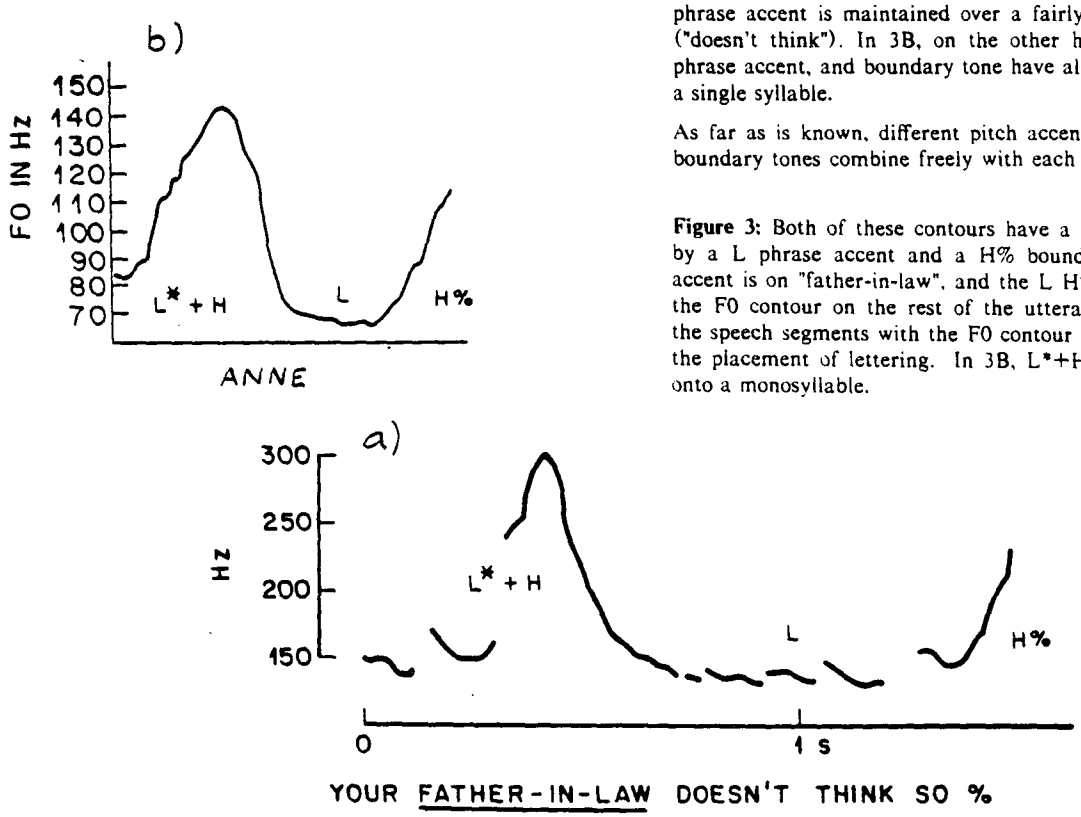

Figure 3: Both of these contours have a L*+H accent followed by a L phrase accent and a H% boundary tone. In 3A, the accent is on "father-in-law", and the L H% sequence determines the F0 contour on the rest of the utterance. The alignment of the speech segments with the F0 contour is roughly indicated by the placement of lettering. In 3B, L*+H L H% is compressed onto a monosyllable.

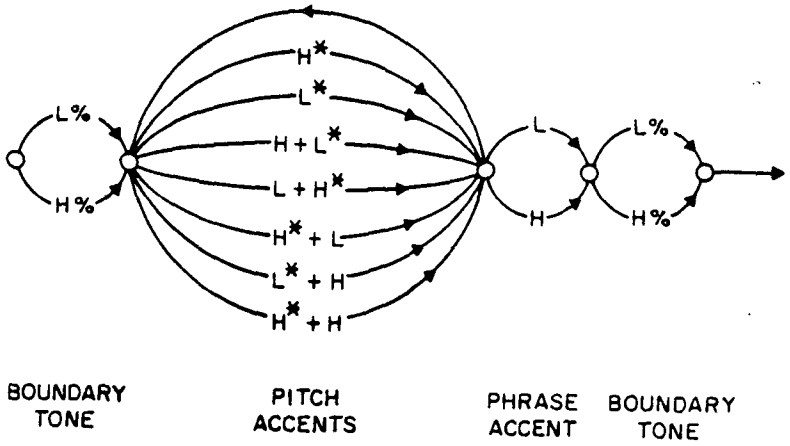

Figure 4: The grammar of the phrasal tunes of English given in Pierrehumbert (1980).

the grammar of English phrasal melodies can be represented by a transition network, as shown in Figure 4. This grammar defines the level of description that the recognizer attempts to recover. There is no effort to characterize the meaning of the transcriptions established, since our focus is on the sound structure of speech rather than its meaning. In production, the choice of loci for pitch accents depends on the focus structure of the sentence. The choice among different melodic elements appears to be controlled by the attitudes of the speaker and the relation of his phrase to others in the discourse. Meanings suggested in the literature for such elements include surprise, contradiction, elicitation, and judiciousness.

2.2 Phonetics

Two types of rules have a part in relating the tonal level of description to the continuously varying F0 contour. One set of rules maps tones into crucial points, or targets, in the F0 contour. Both the small tonal inventory and the sequential decomposition proposed depend on these rules being nontrivial. Specifically, a rule of downstep lowers a H tone in the contexts H+L _ and H L+_. The value of the downstepped H is a fixed fraction of the value for the preceding H, once a phrasal constant reflecting pitch range is subtracted. Iterative application of this rule in a sequence of accents which meet its structural description generates an exponential decay to a nonzero asymptote. A related rule, upstep, raises a H% after a H phrase accent. This means that the L* H H% melody often used in yes/no questions (and illustrated in Figure 5 below) takes the form of a rise--plateau--rise in the F0 contour. Differences in the relative level of accent tones can also result from differences in the emphasis on the material they are associated with. This is why the middle H* in Figure 1 is lower than the other two, for example.

A second class of rules computes transitions between one target and the next. These fill in the F0 contour, and are responsible for the F0 on syllables which carry no tone. Transitions are not always monotonic; in Figure 1, for example, the F0 dips between each pair of H accents. Such dipping can be found between two targets which are above the low part of the range. Its extent seems to depend on the time-frequency separation of the targets.

In certain circumstances, a single tone gives rise to a flat stretch in the F0 contour. For example, the phrase accent in Figure 3A has spread over two words. This phenomenon could be treated either at a phonological level, by linking the tone to a large number of syllables, or at a phonetic level, by positing a sustained style of transition. There are some interesting theoretical points here, but they do not seem to affect the design of an intonation recognizer.

Note that the rules just described all operate in a small window, as defined on the sequence of tonal units. To a good approximation, the realization of a given tonal element can be computed without look-ahead, and looking back no further than the previous one. Of course, the window size could never be stated so simply with respect to the segmental string; two pitch accents could, for example, be squeezed onto adjacent syllables or separated by many syllables. One of the crucial assumptions of the work, taken from autosegmental and metrical phonology, is that the tonal string can be projected off the segmental string. The recognition system will make strong use of the locality constraint that this projection makes possible.

2.3 Summary

The major theoretical innovations of the description just sketched have important computational consequences. The theory has only two tones, L and H, whereas earlier tone-level theories had four. In combination with expressive variation in pitch range, a four tone system has too many degrees of freedom for a transcription to be recoverable, in general, from the F0 contour. Reducing the inventory to two tones raises the hope of reducing the level of ambiguity to that ordinarily found in natural language. The claim that implementation rules for tonal elements are local mean that the quantitative evidence for the occurrence of a particular element is confined to a particular area of the F0 contour. This constraint will be used to simplify the control structure. A third claim, that phrasal tunes are constructed syntactically from a small number of elements, means that standard parsing methods are applicable to the recognition problem.

3. A recognition system

The recognition system as currently implemented has three components, described in the next three sections. First, the F0 contour is preprocessed with a view to removing pitch tracking

errors and minimizing the effects of the speech segments. Then, a schematization in terms of events is established, by finding crucial features of the smoothed contour through analysis of the derivatives. Events are the interface between the quantitative and symbolic levels of description; they are discrete and relatively sparse with respect to the original contour, but carry with them relevant quantitative information. Parsing of events is carried out top down, with the aid of rules for matching the tonal elements to event sequences. Tonal elements may account for variable numbers of events, and different analyses of an ambiguous contour may divide up the event stream in different ways. Steps in the analysis of an example F0 contour are shown in Figure 5.

3.1 Preprocessing

The input to the system is an F0 contour computed by the Gold Rabiner algorithm (Gold and Rabiner, 1969). Two difficulties with this input make it unsuitable for immediate prosodic analysis. First, the pitch tracker in some cases returns values which are related to the true values by an integer multiplier or divisor. These stray values are fatal to any prosodic analysis if they survive in the input to the smoothing of the contour. This problem is addressed by imposing continuity constraints on the F0 contour. When a stray value is located, an attempt to find a multiplier or divisor which will bring it into line is made, and if this attempt fails, the stray value is deleted. In our experience, such continuity constraints are necessary to eliminate sporadic errors; without them, no amount of parameter tweaking suffices.

A second problem arises because the speech segments perturb the F0 contour; here, consonantal effects are of particular concern. There are no F0 values during voiceless segments. Glottal stops and voiced obstruents depress the F0 on both sides. In addition, voiceless obstruents raise the F0 at the beginning of a following vowel. Because of these effects, a attempt was made

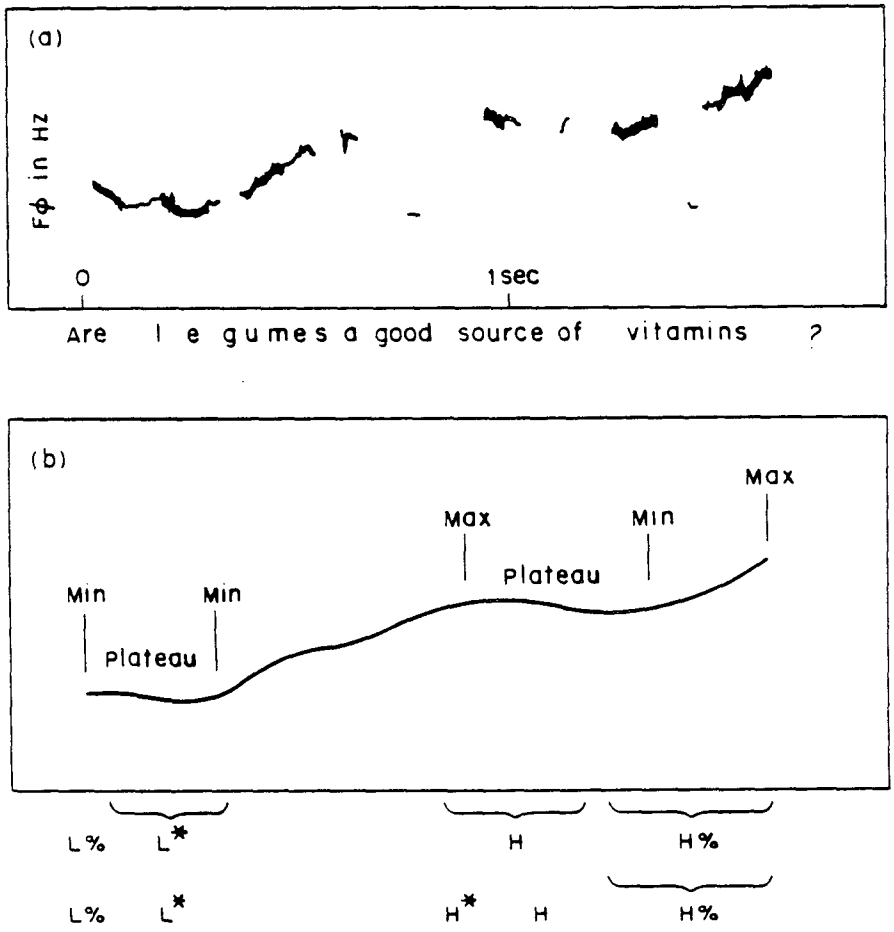

Figure 5: Panel A shows an unprocessed F0 contour. The placement of lettering indicates roughly the alignment of tune and text. Parts of the F0 contour which survive the continuity constraints and the clipping are drawn with a heavier line.

Panel B shows the connected and smoothed F0 contour, together with its event characterization. The two transcriptions of the contour are shown underneath. The alignment of tonal elements indicates what events each covers.

to remove F0 values in the immediate vicinity of obstruents. An adapted version of the Fleck and Liberman (1982) syllable peak finder controlled this clipping. Our modification worked outward from the syllabic peaks to find sonorant regions, and then retained the F0 values found there. In Figure 5A, the portions of the F0 contour remaining after this procedure are indicated by a heavier line. The retained portions of the contour are connected by linear interpolation. Following Hildreth and Marr's work on vision, the connected contour is smoothed by convolution with a Gaussian in order to permit analysis of the derivatives. The smoothed contour for the example is shown in Figure 5B.

3.2 Schematization

Events in the contour are found by analysis of the first and second derivatives. The events of ultimate interest are maxima, minima, plateaus, and points of inflection. Roughly speaking, peaks correspond to H tones, some valleys are L tones, and points of inflection can arise through downstep, upstep, or a disparity in prominence between adjacent H accents. Plateaus, or level parts of the contour, can arise from tone spreading or from a sequence of two like tones. Events are implemented as structures which store quantitative information, such as location, F0 value, and derivative values.

Maxima and minima can be located as zeroes in the first derivative. Those which exhibit insufficient contrast with their local environment are suppressed; in regions of little change, such as that covered by the phrase accent in Figure 3A, this threshholding prevents minor fluctuations from being treated as prosodic. Plateaus are significant stretches of the contour which are as good as level. A plateau is created from a sequence of low contrast maxima and minima, or from a very broad peak or valley. In either case, the boundaries of the plateau are marked with events, whose type is relevant to the ultimate tonal analysis. These events are not located at absolute maxima or minima, which in nearly level stretches may fall a fair distance from points of prosodic significance. Instead, they are pushed outward to a near-maximum, or a near-minimum. The event locations in Figure 5B reflect this adjustment. Minima in the absolute slope, (which form a subset of zero crossings in the second derivative) are retained as points of inflection if they contrast sufficiently in slope with the slope maxima on either side. In some cases, such points were engendered by smoothing from places where the original contour had a shelf. In many others, however, the shoulder in the original contour is a slope minimum, although a more prototypical realization of the same prosodic pattern would have a shelf. Presumably, this fact is due to the low pass characteristics of the articulatory system itself.

3.3 Parsing

Tonal analysis of the event stream is carried out by a topdown nondeterministic finite state parser, assisted by a set of verification rules. The grammar is a close relative of the transition network in Figure 1. (There is no effort to make distinctions which would require independent information about stress location, and provision is made for the case where the phrase accent and boundary tone collapse phonetically.) The verification rules relate tonal elements to sequences of events in the F0 contour. As each tonal element is hypothesized, it is checked against the event stream to see whether it plausibly extends the analysis hypothesized so far. The integration of successful local hypotheses into complete analyses is handled conventionally (see Woods 1973).

The ontology of the verification rules is based on our understanding of the phonetic realization rules for tonal elements. Each rule characterizes the realization of a particular element or class of elements, given the immediate left context. Wider contexts are unnecessary, because the realization rules are claimed to be local. Correct management of chained computations, such as iterative downsteps, falls out automatically from the control structure. The verification rules refer both to the event types (e.g. "maximum", "inflection,") and to values of a small vocabulary of predicates describing quantitative characteristics. The present system has five predicates, though a more detailed accounting of the F0 contour would require a few more. One returns a verdict on whether an event is in the correct relation to a preceding event to be considered downstepped. Another determines whether a minimum might be explained by a non-monotonic F0 transition, like that pointed out in Figure 1. In general, relations between crucial points are considered, rather than their absolute values. Even for a single speaker, absolute values are not very relevant to melodic analysis, because of expressive variation in pitch range. Our experiments showed that local relations, when stated correctly, are much more stable.

Timing differences result in multiple realizations for some tonal sequences. For example, the L* H H% sequence in Figure 5A comes out as a rise--plateau--rise. If the same sequence were compressed onto less segmental material, one would see a rise--inflection--rise, or even a single large rise. For this reason, the rules OR several ways of accepting a given tonal hypothesis. As just indicated, these can involve different numbers of events.

The transcription under figure 5B indicates the two analyses returned by the system. Note that they differ in the total number of tonal elements, and in the number of events covered by the H phrase accent. The first analysis correctly reflects the speaker's intention. The second is consistent with the shape of the F0 contour, but would require a different phrasal stress pattern. Thus the location of the phrasal stress cannot be uniquely recovered from the F0 contour, although analysis of the F0 does constrain the possibilities.

4. Discussion and conclusions

4.1 Intellectual antecedents

The work described here has been greatly influenced by the work of Marr and his collaborators on vision. The schematization of the F0 contour has a family resemblance to their primal sketch, and I follow their suggestion that analysis of the derivatives is a useful step in making such a schematization.

Lea (1979) argues that stressed syllables and phrase boundaries can be located by setting a threshhold on F0 changes. This procedure uses no representation of different melodic types, which are the main object of interest here. Its assumptions are commonly met, but break down in many perfectly well-formed English intonation patterns.

Vives et al. (1977) use F0 in French to screen lexical hypotheses, by placing restrictions on the location of word boundaries. This procedure is motivated by the observation that the F0 contour constrains but does not uniquely determine the boundary locations. In English, F0 does not mark word boundaries, but there are somewhat comparable situations in which it constrains but does not determine an analysis of how the utterance is organized. However, the English prosodic

system is much more complex than that of French, and so an implementation of this idea is accordingly more difficult.

4.2 Segmentation and labelling

The approach to segmentation used here contrasts strongly with that used in the past in phonemic analysis. Whereas the HWIM system, for example, proposed segmental boundaries bottom up (Woods et al., 1976), the system described here never establishes boundaries. For example, there is no point on the rise between a L* and a H* which is ever designated as the boundary between the two pitch accents. Whereas phonetic segments ordinarily carry only categorical information, the events found here are hybrids, with both categorical and quantitative information. A kind of soft segmentation comes out, in the sense that a particular tonal element accounts for some particular sequence of events. Study of ambiguous contours indicates that this grouping of events cannot be carried out separately from labelling. Thus, there is no stage of analysis where the contour is segmented, even in this soft sense, but not labelled.

It is not hard to find examples suggesting that the approach taken here is also relevant for phonemic analysis. Consider the word "joy", shown in Figure 6. Here, the second formant falls from the palatal locus to a back vowel position, and then rises again for the off-glide. A different transcription involving two syllables might also be hypothesized; the second formant could be falling through a rather nondistinct vowel into a vocalized /l/, and then rising for a front vowel. Thus, we can only establish the correct segment count for this word by evaluating the hypothesis of a medial /l/. Even having done so, there is no argument for boundary locations. The multiple pass strategy used in the HWIM system appears to have been aimed at such problems, but does not really get at their root.

4.3 Problems

A number of defects in the current implementation have become apparent. In the example, the amount of clipping and smoothing needed to suppress segmental effects enough for parsing results in poor time alignment of the second transcription. The H* in this analysis is assigned to "source", whereas the researcher looking at the raw F0 contour would be inclined to put it on "gumes". In general, curves which are too smooth may still be insufficiently smooth to parse. An alternative approach based on Hildreth's suggestions about integration of different scale channels in vision was also investigated. (Hildreth, 1980.) Most of the obstacles she mentions were actually encountered, and no way was found to surmount them. Thus, I view the separation of segmental and prosodic effects on F0 as an open problem. Adding verification rules for segmental effects appears to be the most promising course.

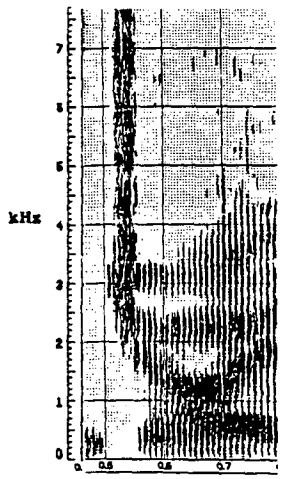

Figure 6: A spectrogram of the word "joy", cut out of the sentence "We find joy in the simplest things." The example is taken from Zue et al. (1982).

Two classes of extraneous analyses generated by the system merit discussion. Some analyses, such as the second in Figure 5, violate the stress pattern. These are of interest, because they inform us about how much F0 by itself constrains the interpretation of stress. A second group, namely analyses which have too many tonal elements for the syllable count, is of less interest. A future implementation should eliminate these by referring to syllable peak locations.

Acknowledgements

I would like to thank Mitch Marcus and Dave Shipman for helpful discussions.

References

Fleck, M. and M. Y. Liberman (1982). "Test of an automatic syllable peak finder," J. Acoust. Soc. Am. 72, Suppl.1 S78.

Gold, B. and L. Rabiner (1969). "Parallel Processing Techniques for Estimating Pitch Periods of Speech in the Time Domain." J. Acoust. Soc. Am. 46, 442-448.

Hildreth, E. (1980). "Implementation of a Theory of Edge Detection," Artificial Intelligience Laboratory Report AI-TR-579, MIT.

Lea, W. A. (1979). "Prosodic Aids to Speech Recognition." in W. A. Lea, ed. Trends in Speech Recognition. Prentice Hall, Englewood Cliffs N.J. 166-205.

Liberman, M. Y. and J. Pierrehumbert (forthcoming in 1983). "Intonational Invariance under Changes in Pitch Range and Length." Currently available as a Bell Labs Technical Memorandum.

Marr, D. (1982). Vision. W. H. Freeman and Co., San Francisco.

Pierrehumbert, J. (1980). "The Phonology and Phonetics of English Intonation." PhD dissertation, MIT. (forthcoming from MIT Press).

Pierrehumbert, J. (1981). "Synthesizing intonation." J. Acoust. Soc. Am. 70, 985-995.

Vives, R., C. Le Corre, G. Mercier, and J Vaissiere (1977). "Utilisation, pour la reconnaissance de la parole continue, de marqueurs prosodiques extraits de la frequence du fondamental." 7iemes Journees d'Etudes sur la Parole, Groupement des Acousticiens de Langue Francais, 353-363.

Woods, W.A. (1973). "An Experimental Parsing System for Transition Network Grammars." in R. Rustin, ed., Natural Language Processing. Algorithmics Press, Inc., New York.

Woods, W.A, M. Bates, G. Brown, B. Bruce, C. Cook, J. Klovstad, J. Makhoul, B. Nash-Webber, R. Schwartz, J. Wolf, and V. Zue (1976). "Speech Understanding Systems Final Report Volume II." BBN Report No. 3438.

Zue, V., F. Chen, and L. Lamel (1982). Speech Spectrogram Reading: Special Summer Course. MIT.