JAPANESE PROSODIC PHRASING AND INTONATION SYNTHESIS

Mary E. Beckman[1] and Janet B. Pierrehumbert
Linguistics and Artificial Intelligence Research
AT&T Bell Laboratories,
600 Mountain Ave, Murray Hill, NJ 07974

ABSTRACT

A computer program for synthesizing Japanese fundamental frequency contours implements our theory of Japanese intonation. This theory provides a complete qualitative description of the known characteristics of Japanese intonation, as well as a quantitative model of tone-scaling and timing precise enough to translate straightforwardly into a computational algorithm. An important aspect of the description is that various features of the intonation pattern are designated to be phonological properties of different types of phrasal units in a hierarchical organization. This phrasal organization is known to play an important role in parsing speech. Our research shows it also to be one reflex of intonational prominence, and hence of focus and other discourse structures. The qualitative features of each phrasal level and their implementation in the synthesis program are described.

1. INTRODUCTION

In this paper, we will present a computer program for synthesizing fundamental frequency contours for standard Japanese. Fundamental frequency (*f0*) is the paramount physical correlate of the sensation of pitch, and, in many languages, the time course of f0 is one of the primary phonetic manifestations of intonation. This is especially true in Japanese, where duration and amplitude do not have the consequential role in communicating intonational structure that they do in English (Beckman, 1986). Accordingly, a program for synthesizing Japanese f0 contours is tantamount to a computational implementation of a theory of Japanese intonation.

The theory that we have implemented in our synthesis program is based on a review of the literature in English and Japanese, and on the results of an extensive series of experiments in which we examined and made f0 measurements of about 2500 intonation contours in order to resolve some of the many problems not answered in the literature. These experiments have uncovered important facts about the hierarchical structure underlying Japanese prosody and about the manifestations of focus in Japanese. We have incorporated these discoveries in our synthesis program, which, we believe, covers all known qualitative characteristics of Japanese intonational melody. Informal listening tests by Japanese speakers indicate that the f0 contours which the program produces sound quite natural. In some cases, the synthesized contours were even preferred to the genuine human intonation contours on which they are modeled.

Although the main concern of our research was to provide an accurate phonological and phonetic characterization of Japanese intonational structure that could be used in the automatic computation of f0 contours, our description of Japanese prosodic phrasing and intonation synthesis is also of direct relevance to issues in several other areas, including the role of prosodic phrasing in the parsing of speech, the relationship between intonational patterns and discourse phenomena such as focus, and the development of a more accurate understanding of the phonological mechanisms of intonation as a universal component of human speech. The computer implementation of the theory in turn should provide a practical tool for further research in these areas. These other background issues are discussed in Sections 1.1-1.3. Section 2 then summarizes the characteristics of Japanese intonation that we have incorporated in our synthesis program, and Section 3 gives a detailed account of the program itself.

1.1 Prosodic Phrasing and Syntactic Parsing

Prosodic organization of the sort that we discovered for Japanese bears strongly on current issues in syntactic parsing. It is well known that intonational phrase boundaries can play a crucial role in parsing speech. For example, if the sentence in (1) is said without any internal phrase boundaries, it produces a garden path; the human parser interprets *several bugs* as the object of *left*, and then is unable to arrive at a syntactic role for the final verb phrase.

(1) When we left several bugs in the program still hadn't been corrected.

On the other hand, if the sentence is produced with the intonation break indicated by the comma in (2), *several bugs* is readily interpreted as the subject of the main clause.

(2) When we left, several bugs in the program still hadn't been corrected.

Intonation breaks can also be used to disambiguate sentences with ambiguous scope of negation or conjunction. Thus in example (3), the break represented by the comma forces the reading in which the scope of negation is the main verb clause (*Because they were mad, they didn't leave*), as opposed to the reading in which the scope of negation is the subordinate clause (*It was not because they were mad that they left*).

(3) They didn't leave, because they were mad.

Similarly in (4), the break after *mnemonic rhyme* prevents *sublime* from modifying *free meter*, whereas under the alternative phrasing in (5), *sublime* is taken to modify both conjuncts.

(4) Sublime mnemonic rhyme, and free meter.

(5) Sublime, mnemonic rhyme and free meter.

In reviewing these examples, we have spoken as if there were only one type of intonational phrase boundary. And the most substantial current proposal about the role of intonational phrasing in the parsing of Japanese (Marcus and Hindle, 1985) takes into account only a single level of phrasing. In actuality, however, Japanese and English both have several different types of intonational phrase, which are related to each other hierarchically.[2] As Marcus and Hindle point out to us, major modifications to their proposal will be necessary to accommodate the role of the complete hierarchical intonational structure in parsing.

1. Present address: Ohio State University, Department of Linguistics, 1841 Millikin Rd, Columbus, OH 43210.

1.2 Focus and Discourse Structure

Another major result of our experiments was to be able to describe the manifestations of focus in terms of the phonological structures we discovered. We use the word *focus* here in the sense of Chomsky (1971), to characterize words or phrases which are intonationally marked as prominent. This contrasts with usage in the AI literature, where the *focus space* is used to describe entities which are assumed to be salient with respect to a given discourse segment. However, the concepts are related to each other via the broader concept of the *attentional structure*, as described in Grosz and Sidner (1985).

Broadly speaking, intonational prominence is used to modify the attentional state. A word or phrase that is marked by intonational prominence is made phonetically more salient; its prosodic coloring is more attention-demanding than it otherwise would be. One reason for a word or phrase to receive intonational prominence is that it refers to something which is being added to the focus space. Or, if the entity referred to is already in the focus space, the word or phrase may be made intonationally prominent because the referent is under contrast or in some other way plays a marked role in the utterance. The presence or absence of intonational prominence is thus very much analogous to the use of full referring expressions versus pronominal forms.

The analogy breaks down, however, when the range of possible use is considered. Pronominal forms and other sorts of anaphora can be used in place of full referring expressions only in some syntactic categories and positions. Intonational prominence, by contrast, can be absent or present on any word. Therefore, the study of how intonational prominence is used promises to make crucial contributions to developing a theory of attentional structure. But an accurate controlled study of the use of intonational prominence is impossible without an exact characterization of the form of intonational prominence. A precise phonological and phonetic description of intonational structure is thus an important prerequisite to the development of theories of discourse structure.

We also note that it is crucial to take focus, in the linguistic sense, into account in addressing the role of intonational phrasing in parsing. One of the main results of our experiments was the discovery that focus systematically affects prosodic phrasing in Japanese. Any parser intended for use with real speech must be able to accommodate the way in which focus and syntactic structure interact to determine the observed phrasing.

1.3 Japanese and English Intonation

A final motivation for our description of Japanese was to contribute to a more universal understanding of intonational structure. Our work is in some sense an extension of work on an earlier model of English intonation (Pierrehumbert, 1980, 1981; Liberman and Pierrehumbert, 1984; Anderson, Pierrehumbert, and Liberman, 1984). We first became interested in synthesizing f0 contours in Japanese because there are known to be formal differences between Japanese and English prosody. We wished to discover what aspects of a theory developed for English prosody would carry over to a language which differed in many ways, and how such shared principles would interact with language-specific principles.

1.3.1 Basic Principles — One principle that can be assumed to be universal is the notion that intonation is separable from the text of an utterance not just physically but also linguistically. When a speaker produces an utterance with a give intonation pattern, he is implementing two separate strings of phonological elements in parallel. The textual string of distinctive segmental events that is realized in the spectral patterns of the utterance is conceptually distinct from the string of distinctive melodic events that is realized in the f0 contour. The physical implementations of these two representational strings are coordinated by a phonological specification of the alignment between the textual events (phonemic segments and phrasal groups of segments) and the melodic events (tones and tone configurations).

1.3.2 English Tone Configurations — In English, as is well known, there are two types of basic tone configurations. Some tone configurations, which are called pitch accents, are placed on especially prominent syllables in a phrase. If the placement of the special prominences shifts because of emphasis or focus, the pitch accents move along with them. Other tone configurations are placed at the edges of phrases without regard for the locations of the prominent syllables within the phrases. If the phrasing changes, these tones must also move. For both types of tone configuration, the speaker can select among several different patterns. His choice appears to convey a message about propositional attitude. For example, one pattern might suggest that the speaker is impatiently repeating what he feels should be obvious to the listener while another would imply that he is uncertain about the relevance of what he is saying, as illustrated in Figure 1.

1.3.3 Stress — Japanese phrasal prosody differs from English in several crucial ways. First, Japanese does not have lexical stress as English does. The prominent syllables that carry pitch accents in English are marked also by a rhythmic salience — an extra duration and loudness that adds another sort of prosodic

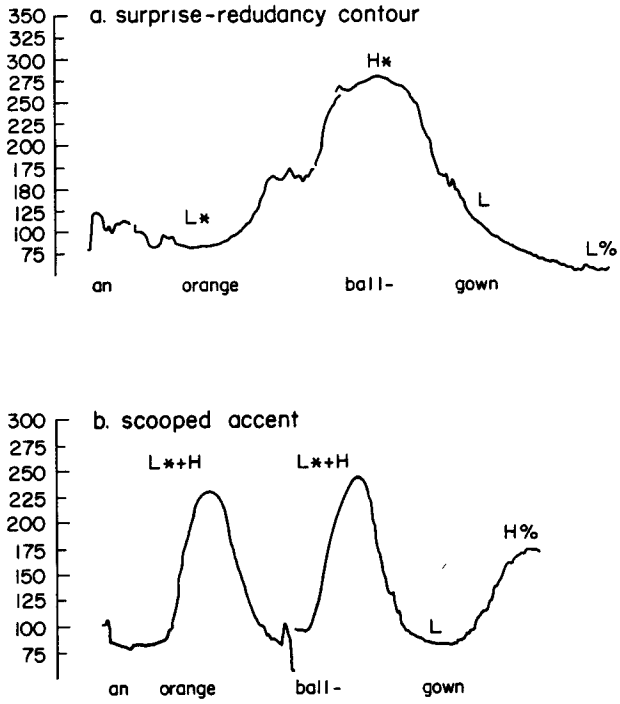

Figure 1. Fundamental frequency contours for two intonation patterns for the utterance *An orange ballgown*. The tones in the melody are transcribed using the notation of Pierrehumbert (1980, 1981), with "*" for the tone in a pitch accent that associates to the stressed syllable and "%" for a boundary tone. Version (a) is a "surprise-redundancy contour" with a L* pitch accent on the stressed syllable in *another*, a H* pitch accent on *orange*, and a L% boundary tone. Version (b) implies uncertainty, with a scooped rising accent (L*+H) on each word followed by a L H% phrase-final boundary sequence

2. Section 2 summarizes our results on the levels of phrasing found in Japanese. Beckman and Pierrehumbert (forthcoming) give a detailed comparison to the analogous levels of phrasing in English.

prominence to the intonational prominence of the pitch accent. Especially prominent elements in a Japanese utterance can also be longer and louder, but unlike in English, this rhythmic prominence is not a lexical feature. That is, words in Japanese do not have the lexical markings of stress that in English give a rhythmic prominence to the first syllable in *seven* and the second syllable in *eleven* even in the absence of a pitch accent. Instead, Japanese has a lexical distinction between *accented* and *unaccented* words.

1.3.4 Japanese Lexical Accent — Accented words have a fundamental frequency fall at some designated syllable; around the lexically designated location there is a sharp descent from a relatively higher pitch level to a relatively lower one. We represent this fall as a sequence of a high tone and a low tone, or H L, as illustrated in the following schematization of the accented word *yamaza'kura*:

(6) ya maza' ku ra
HL

Here the line coming up from the H indicates that the high tone is *associated to* the designated syllable *za'*. That is, the realization of the H tone in the resulting f0 contour must occur concurrently with the production of the syllable's segments. The relatively lower pitch level of the L immediately following the associated H results in the pitch fall of the accent.

Unaccented words differ from accented words in having no syllable designated to carry the H of the accent fall, and hence no lexically associated tone, as in:

(7) mu ra sa ki i ro

Since the presence or absence of an accent HL sequence is a property of the component lexical items, an entire sentence may have no accents; this contrasts with the situation in English, where it is impossible to utter a sentence without placing a pitch accent on at least one syllable.

1.3.5 Choice of Tune and Phrasing — Another important difference is that, utterance-internally in Japanese, there is no paradigmatic choice among different tone patterns to express differences in meaning such as uncertainty or impatient rejoinder. In other words, the shape of the accent HL contour is a property of the lexical feature *accented*, and there is nothing corresponding to the choice of tone pattern for the pitch accent in English. At the end of the phrase, however, there is a distinction between rising and falling contours, which can convey the sort of meanings expressed by the choice of tone patterns at the edges of phrases in English. Because of the lexical origin of the phrase-internal tone features in Japanese, the system of phrasal intonation is relatively impoverished compared to English. Other than the limited choice of pattern type at the end of the phrase, the only dimensions of variation seem to be different choices of phrasing and of pitch range. Our experiments were designed to explore how phrasing is conveyed and what the consequences of local manipulations of pitch range are.

2. THE HIERARCHY OF PHRASE LEVELS

In our data, we have found evidence for three levels of phrasing marked by f0 features. We call these three types of phrases the *accentual phrase*, the *intermediate phrase*, and the *utterance*.

2.1 The Accentual Phrase

The lowest level, the accentual phrase, is a phrasal unit containing at most one accent. This unit may be a single word. However, when words are combined into sentences, it is quite usual for some to lose their status as separate accentual phrases. Noun-noun compounds typically form a single accentual phrase, as do adjective-noun sequences or sequences of direct object and governing verb.

Apart from the possible occurrence of an accent, the hallmark of an accentual phrase is an f0 rise at its beginning. We account for this rise by positing a L% tone (the *boundary L%*)[3] marking the phrase boundary, and a H tone (the *phrasal H*) associated with a designated syllable near the beginning of the phrase. If the sample accented and unaccented words shown above in (6) and (7) were produced as complete accentual phrases, they might be represented as in (8):

(8) ya maza' ku ra mu ra sa ki i ro
(L%) H HL L% (L%) H L%

The tones that we have represented here are the only ones we posit for the accentual phrase.[4] We interpret f0 patterns at places not occupied by the indicated tones as arising from a phonetic process which interpolates between the assigned target values for these tones.

This notion of phonetic interpolation differs radically from more traditional representations of the accentual phrase. Studies of Japanese in the school of modern generative phonology have asserted that the accentual phrase is the domain of a process called *tone spreading*, whereby tones are copied from their originally specified places to associate to every syllable in the phrase. Thus in accented phrases, the L tone of the accent is made to associate with all syllables following the accent in the phrase. The H tone, conversely, is made to associate with all syllables preceding the accent, except possibly for the first, which might be associated instead to a L tone (corresponding to the L% that we take as marking the preceding phrase boundary). In unaccented phrases, similarly, the phrasal H tone is thought to be associated to all the syllables after the first. These assumptions give rise to representations like those in (9). The phonetic prediction of such a representation is that a spread tone will be realized as a sustained pitch level over the syllables to which it is copied.

(9) ya maza' ku ra mu ra sa ki i ro
L H L L H

Our data, however, demonstrate that Japanese actually has no such rules of tone spreading. For example, in an utterance-medial unaccented phrase, there is a smooth fall from the phrasal H tone near the beginning of the phrase to the L% at the boundary before the next accentual phrase. The slope of this fall varies inversely with the separation of the two tones, as would be expected if a simple linear interpolation between fixed end point values were stretched to occupy a larger and larger distance. This generalization is illustrated in Figure 2, which shows f0 contours for segmentally matched unaccented sentence-medial phrases with 1, 2, 3, 5, and 6 syllables intervening between the phrasal H and the boundary L% for the next accentual phrase. Slopes of regression lines fit over the H-L% transition are indicated. The inverse correlation between these slopes and the number of syllables in the phrase is not compatible with the notion that the phrasal H tone has spread to associate with all following syllables up to the boundary L%. It would arise naturally, however, by a

3. Here we use the % notation used by Pierrehumbert (1980) to designate a boundary tone.

4. Note that we put the first L% tone in each phrase in parentheses, because we consider it to be an edge feature of the preceding accentual phrase rather than of the accentual phrase being represented.

phonetic process which interpolates linearly between the values of the H on the second syllable and the L%.

The finding that Japanese has no tone spreading is particularly significant, since most modern theories of phonology assume that surface phonological representations (those which are interpreted phonetically) are fully specified, meaning that a specific feature value must be assigned wherever a feature of some sort could be assigned a value. There has been considerable controversy about what phonological rules are necessary to generate the correct fully specified representations. Our results show, however, that at least for tone, the surface representations are only partially specified. That is, only some of the syllables that could in theory be assigned tonal values actually have associated tones. This is consistent with a view in which the surface representations are merely descriptions of the phonetic form, in a spirit similar to what Marcus et al. (1983) have proposed for surface syntactic representation.

2.2 The Intermediate Phrase

The partially specified tone patterns at the accentual phrase level are grouped together prosodically into units at the next higher level of phrasing, that of the intermediate phrase. An intermediate phrase consists of one or more accentual phrases (only rarely more than three). An intermediate phrase boundary is often marked by a pause or *pseudo-pause* (a pre-pausal "winding down" of production speeds unaccompanied by any actual momentary cessation of production). Also, the L% boundary tone for the last accentual phrase in an intermediate phrase is markedly lower than at a medial accentual phrase boundary. Perhaps the most salient and systematic characteristic of the intermediate phrase, however, is that it is the domain of a process known as *catathesis*. Catathesis compresses the pitch range following an accent. This compression affects all tones up to the intermediate phrase boundary, but it does not propagate to the tones belonging to the following intermediate phrase.[5] If an intermediate phrase contains more than one accent, the multiple applications of catathesis cumulate, so that the pitch range can be extremely compressed by the end of the phrase.

An important finding of our experiments is that phrasing at this level is a fairly reliable indicator of focus. Even in syntactic structures where no phrase break is normally expected in neutral renditions, focus will introduce an intermediate phrase boundary right before the focused word or phrase. For example, in one of our experiments, subjects consistently introduced an intermediate phrase boundary between the words in an adjective-noun sequence when the discourse context gave the noun a contrastive emphasis. Often this striking use of phrasing was accompanied by local expansion of the pitch range on the focused item, affecting the f0 values of its phrasal H, accent tones, and boundary L%. In a sizeable number of utterances, however, the change in phrasing was the only consequence of focus.

We suspect that this relationship between phrasing and focus reveals something about the prominence structure internal to the intermediate phrase. In English, the last accented item in a phrase is generally agreed to be the strongest one. If, in Japanese, the strongest item in a phrase is instead in first position, one strategy for marking intonational prominence would be to structure the phrasing of the utterance so as to place the focused item at the beginning of an intermediate phrase. In English, focused items are sometimes set off by phrase boundaries in this way, but this use of phrasing is not nearly as characteristic as the manipulation of local pitch range and of syllable duration and amplitude to put a stronger rhythmic "beat" on the lexically stressed syllable. We believe that this contrast between English and Japanese is related to a difference in prosodic structure. The focused item in Japanese cannot be made more prominent by manipulating the rhythmic prominence of the stressed syllable, because Japanese does not have stress in the sense that English does.

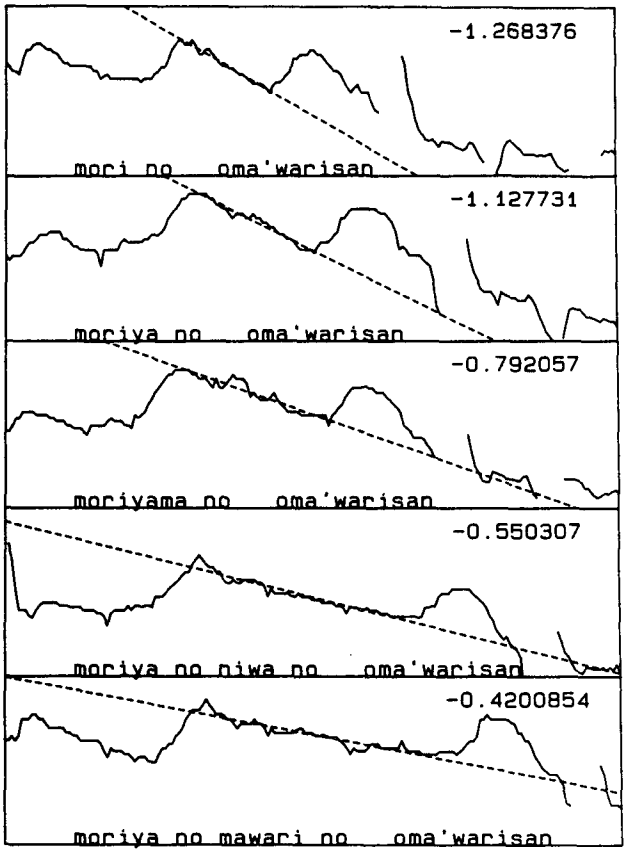

Figure 2. Fundamental frequency contours for five segmentally matched unaccented phrases with varying numbers of syllables between the phrasal H and the boundary L%. The dashed line in each panel is a regression curve fit to the f0 values between the two tones, and the number in the upper right is the slope of the regression curve.

2.3 The Utterance

Our third level of phrasing is the utterance. The phonological mark of an utterance is that it has an initial L% boundary tone. It is also the type of phrase which can be ended with a question rise, a pattern which we account for by the insertion of a H% boundary following the L% ending the last accentual phrase.

In our experiments, the utterance also seemed to be the domain for two phonetic processes affecting the pitch range. One is *declination*, which gradually lowers the pitch range as a function of distance from the beginning of the utterance. Unlike catathesis, it operates without regard to what tones are present. The other is *final lowering*, which further lowers the pitch range in anticipation of the end of the utterance. Questions exhibit declination but not final lowering. There is some reason to suppose that they are subject to *final raising*, which expands the pitch range at the end of the utterance. In particular, the H% boundary tone ending a question is considerably higher than H tones elsewhere in the sentence.

Final lowering is seen in English as well as in Japanese, and was originally supposed to define a comparable utterance level there. More recently, Hirschberg and Pierrehumbert (1986) have

5. Catathesis does affect the L% at the boundary between two intermediate phrases. This is why we consider the L% to be a property of the end of the preceding accentual phrase rather than of the beginning of the next accentual phrase, as shown above in representation (8).

proposed that final lowering is not a prosodic property specific to a particular phonological phrase level in English, but rather is a more direct phonetic expression of discourse structure. We now suspect that final lowering in Japanese is similar, and in Beckman and Pierrehumbert (forthcoming), we suggest that declination also is such a paralinguistic discourse phenomenon. In the current implementation of the intonation synthesizer we treat final lowering and declination as utterance-level properties. On the other hand, we do make the amount of lowering in each utterance a user-controllable variable, so that it should not be difficult to test these more recent suggestions.

2.4 Other Miscellaneous Effects

In addition to the various phrase-specific f0 features discussed so far, there are certain other qualitative differences among tones. For example, our experiments showed that the H tone of the lexical accent is generally higher than the phrasal H of the accentual phrase. We account for this difference by giving the accent H intrinsically more *tonal prominence*. That is, we automatically assign it a higher target value within the local pitch range.

Another important effect is that when the initial syllable in the following accentual phrase is lexically long or accented, the preceding boundary L% is *weak*. That is, it undergoes a phonetic lenition that causes the tone to be realized in the f0 contour with only a very short duration and with a target f0 value that is relatively higher than it otherwise would be. (As in English, low tones are made more tonally prominent by lowering.)

Finally, the tonal prominence of a boundary L% reflects the boundary strength; the L% boundary tone is more tonally prominent (lower) at an intermediate phrase than at a mere accentual phrase boundary, and still more prominent at an utterance boundary.

3. THE F0 SYNTHESIZER

The phrasal f0 features outlined thus far are generated automatically by our synthesis program from a user-provided script that identifies the locations of the appropriate phrase boundaries and lexically determined accents in the time pattern of speech segments for an utterance. Thus at the accentual phrase level, the synthesizer inserts the phrasal H and boundary L% at the appropriate places relative to the phrase ends, and assigns the H of the accent to the designated syllable along with the accent L at the appropriate time delay. At the intermediate phrase level, the program triggers a compression of the pitch range at each accent, lowering the values of all subsequent tones until the end of the phrase. And at the utterance level, it sequentially lowers the f0 values of the tones to generate the rule-prescribed time courses of declination and final lowering. The techniques used to implement these effects are quite similar to those used in the English synthesizer developed earlier by Anderson, Pierrehumbert, and Liberman (1984), and are applied in the same order.

3.1 The Schematized f0 Contour

First, the input routines parse the user-provided script, filling in system defaults for unspecified values to produce a set of values for speaker variables and phrasal structures. Once the the script has been interpreted, the next step is to construct a schematic version of the f0 contour in which tones appear as level stretches. The values that must be computed in constructing the schematic are the temporal location of each stretch and its duration and f0 value.

3.1.1 Timing — The location and duration of each tone is determined by the time pattern of the speech segments, and by our theory of the rules which align tones with segments. For example, the stretch for a medial L% begins at the end of the last segment before the relevant phrasal boundary. The difference in timing between a weak L% and a strong L% (see Section 2.4) is accomplished by giving a weak L% only a point duration and a strong L% the "standard tone duration" (a speaker- and rate-specific value roughly the length of a short syllable). The beginning of the following phrasal H can then be located immediately after the end of the L%.

In the present version of the synthesizer, the "standard tone duration" is the only possible duration for a tone that is not a point. The user can specify its actual millisecond value in his script for the utterance, or he can include it in a file of user-defined defaults for the speaker, or, if the system-provided default is appropriate for the speaker and rate, he can leave the value unspecified.[6] The locations and types of the various phrase boundaries and the location of the accent, on the other hand, are specific to an utterance, and must be specified by the user in the utterance script.

3.1.2 Rules for the f0 Value — The f0 value of each tone is determined by the interaction of relationships such as the following:

High versus Low: A low tone is lower than a high tone in the same local pitch range setting.

Intrinsic prominence of accents: The H in an accent is higher than the phrasal H tone.

Boundary tone weakening: The L% boundary tone is higher if the first syllable of the upcoming phrase is long or accented.

Boundary strength: The L% boundary tone is lower at an intermediate phrase boundary than at an accentual phrase boundary, and lower yet at an utterance boundary.

In the synthesizer, all of these qualitative differences have been made precise, with numerical values for the various relations estimated from the results of our experiments. Obviously, several rules interact to control the value for any single tone. For instance, a boundary tone might be raised because the following phrase begins with a long syllable, but lowered because it is at an intermediate phrase boundary.

3.1.3 The Tone-Scaling Domain — The tone-scaling domain within which these rules operate is a normalized transformed hertz domain, which reflects the overall choice of pitch range and the intonational prominence of each accentual phrase. The lower bound of the tone-scaling domain is defined by a *reference line* (r), which is set to the lowest value in the speaker's range. The upper bound of the overall pitch range is a *high-tone line* for the intermediate phrase (h), which is set to the highest possible H tone value in that phrase. The size of the overall pitch range is thus $h - r$. By raising h, this overall pitch range is expanded for "speaking up" (as it would be in natural speech if the speaker is excited or projecting his voice).

Various uses of this tone-scaling domain are illustrated in Figure 3. For example, catathesis is realized as a proportional compression of the overall pitch range that reduces the value of h at each accent according to the formula:

$$(10)\ h_{new} = c * (h_{old} - r) + r \quad [c < 1]$$

Note that in this equation the proportional reduction of h is normalized to the overall pitch range, so that it can be expressed as a constant value c.

The prominences of different accentual phrases relative to the strongest element in the intermediate phrase are also normalized to this overall pitch range, so as to be readily interpretable and easily specified by the user. A *local tone-scaling domain* is calculated for each accentual phrase on the basis of its relative prominence. (This can be thought of as setting a local accentual-phrase value for the high-tone line h_{α}, as illustrated in Figure 3.)

6. These three options are available also for other underived variables such as the position relative to the end of the utterance where final lowering should begin.

The relations among tones described above are then similarly expressed as prominence values normalized to this local tone-scaling domain. In this way the relationships can be expressed as speaker-specific constants despite changes in overall pitch range and local focus, and interactions among them can be multiplicative within the tone-scaling domain. Within the local tone-scaling domain, H tones are scaled upward and L tones are scaled downwards. That is, prominence values for H tones increase from 0 to 1 as f0 goes up from r to h, whereas those for L tones decrease from 1 to 0, as indicated by the different prominence scales to the right of the transformed hertz domain in Figure 3.

Our use of this transformed hertz domain follows broadly the conceptual structure for English tonal scaling developed in Liberman and Pierrehumbert (1984). Differences between the two models appear to reflect differences between Japanese and English. For example, many English L tones appear below the reference line whereas Japanese L tones are all realized above it, in the same overall region as H tones.

Of the various quantitative values used in tone scaling, those of the reference line, of the high-tone line, of the catathesis ratio constant, and of the other constants for the relations among tones are all speaker variables like the "standard tone duration" for timing. Therefore, they are implemented in the synthesizer as variables that can be specified in the utterance script or in a separately provided defaults file, and which revert to the system default value if left unspecified by the user. The prominence

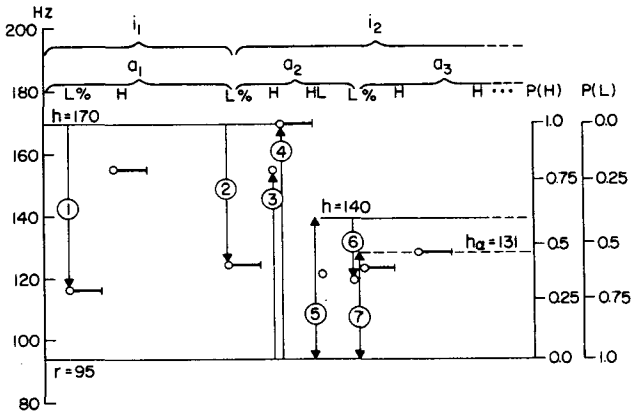

Figure 3. Tone-scaling domain with f0 values computed for the first nine tones in the utterance *mayumi-wa ANA'TA-ni aima'sita ka?* ('Did Mayumi meet YOU?'). Braces at top show the accentual phrase and intermediate phrase grouping. The reference line is 95 Hz and the high-tone line is 170 until reduced by the catathesis at the accent in *ana'ta*. Values for the *y*-axis are hertz on scale to left, and H-tone and L-tone prominences (as scaled in the initial pitch range) on scales to right. Labeled arrows illustrate the application of representative tone scaling rules. (1) Boundary strength at utterance-initial boundary: L%(u)=0.7. (2) Boundary strength at intermediate-phrase boundary: L%(i)=0.6. (3-4) Relationship between phrasal H and accent H: accent H=1.0, phrasal H=0.8. (5) Catathesis constant is 0.6 and reduces high-tone line to 140 Hz. (6) Boundary strength at accentual-phrase boundary with weak L% tone because of long initial syllable in *aima'sita*: L%(a)=0.5, weak L%=0.85; weak L%(a)=0.5*0.85=0.425. (7) Accentual phrase *aima'sita* is subordinated to the focused accentual phrase *ana'ta-ni* by P=0.8, which locally compresses the tone-scaling domain by making a reduced local high-tone line: h_α=131 Hz.

value of each accentual phrase, on the other hand, is specific to its particular degree of subordination to the head of its intermediate phrase, and must be specified in the utterance script.

3.2 The Finished f0 Contour

When the tones have been located in time and frequency, several adjustments are made to produce a finished natural intonation contour from the schematized f0 contour. First, the tones are connected by linear interpolation, as shown in Figure 4a. Declination now applies, as well as final lowering in declaratives (Figure 4b). The resulting contour is then smoothed by convolution with a square window of roughly syllable width.[7] Step functions in f0 now appear more realistically as gradual rises (Figure 4c). Finally, a small amount of random jitter is added to prevent the occurrence of unnaturally flat sections and unnaturally smooth ramps, and the f0 value is set to zero during portions corresponding to voiceless segments (Figure 4d). In order to listen to the results, the computed f0 contour is then substituted for the natural contour in an LPC-coded version of the utterance, and the speech is resynthesized.

CONCLUSION

The model of Japanese intonation implemented in the synthesis program accounts for all of the characteristics of Japanese intonational structure that we have been able to document in our experiments. Some future modifications to the model will probably be necessary as we learn more about how the highest level of phrasing behaves in long connected passages. For example, as noted above, we suspect on the basis of recent work on English (Hirschberg and Pierrehumbert, 1986) that some of the characteristics that we have identified with the utterance in the present model are actually reflections of discourse structure rather than features specific to a well-defined type of unit within the hierarchy of prosodic phrases.

Constructing the f0 synthesizer has been useful in confirming our phonological and phonetic model of Japanese intonation. We believe that the synthesizer will also be useful in generating controlled materials for investigating the use of intonational prominence and the role of phrasing in parsing speech.

ACKNOWLEDGEMENTS

Ken Church, Julia Hirschberg, and Mitch Marcus gave useful comments on earlier drafts of this paper.

APPENDIX: GLOSSARY

catathesis. A sudden compression of pitch range that is triggered by a particular tonal configuration, and that lowers all tones following the trigger within some phrasal unit. In Japanese, catathesis is triggered by every accent, and in English, by every bitonal pitch accent.

declination. A gradual lowering of the pitch range that is effected as some function of time from the beginning of an utterance without regard to the tonal structure.

final lowering. A gradual lowering of the pitch range starting at some distance from the end of the utterance.

fundamental frequency. The reciprocal of the period in a periodic signal, and the main physical correlate of pitch. Fundamental frequency is abbreviated *f0* and is measured in periods per second (unit *hertz*). In speech, f0 corresponds to the frequency of vibration of the vocal cords during voiced segments.

H. A high tone.

7. The rates of the declination and of the final lowering and the size of the smoothing window are speaker- and rate-specific variables like the reference line, and are treated in the same way in the synthesis program.

high-tone line. In Japanese tone-scaling, the upper bound of the pitch range. Its f0 value corresponds to that of a hypothetical highest possible H tone in that range.

intonational phrase. A prosodic unit delimited phonologically by some sort of intonational feature such as a boundary tone.

L. A low tone.

LPC coding. A specification of the spectral characteristics of a signal in terms of sets of *linear predictor coefficients* at fixed intervals. An *n*th-order analysis of the signal is obtained by a least squares estimation of successive samples within an analysis frame from the linear combination of the last *n* samples. The set of predictor coefficients for each analysis frame can then be used as a filter for an input pulse train to synthesize a new signal with the same spectral pattern and an arbitrarily different f0 pattern.

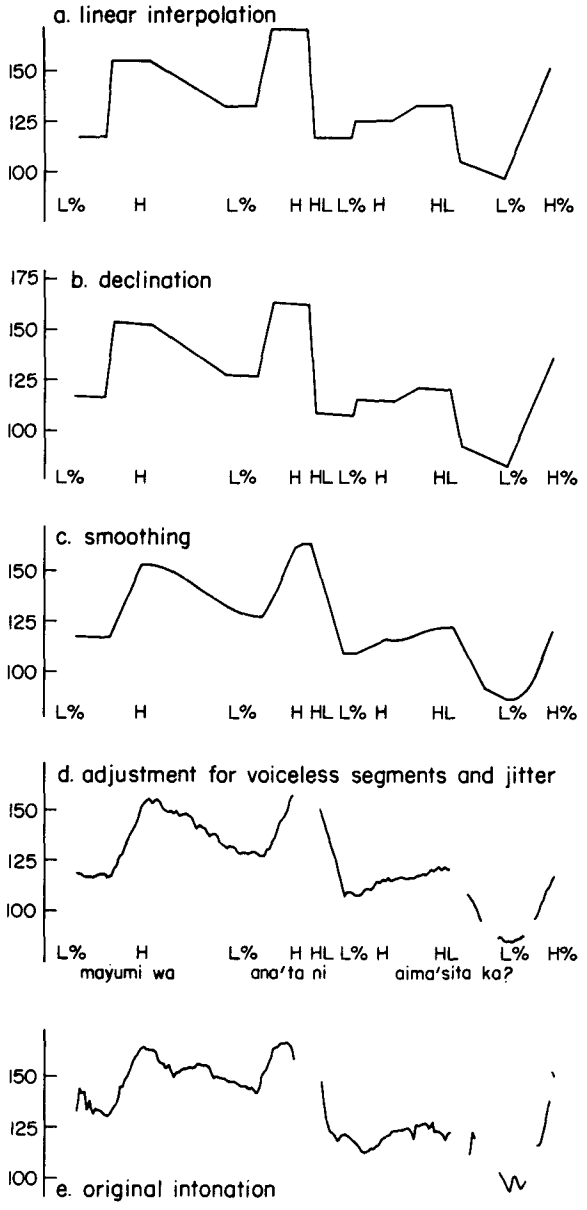

Figure 4. Adjustments for making a finished f0 contour from schematic tone level stretches for utterance shown in Figure 3. (1) Linear interpolation fills in unspecified values between tones. (2) Declination applies, but not final lowering, because the utterance is a question ending in a H% boundary tone. (3) The contour is smoothed by convolution with a syllable-sized square window. (4) Jitter is added and f0 values excised during voiceless segments [t], [ʃ], and [k]. (5) The f0 contour of the original utterance is shown for comparison with (4).

pitch accent. A tonal configuration that is associated to a designated syllable in an utterance, and that marks the syllable (or the word containing the syllable) as *accented* or intonationally prominent. In Japanese, accent consists of a pitch fall from H tone to L at a lexically designated syllable in a word. In English, an accent is any one of six tonal patterns (H*, L*, H*+L, L*+H, H+L*, L+H*) that can be associated to a lexically designated syllable.

pitch range. The spread of fundamental frequency between the "floor" of a speaker's voice and the highest f0 appropriate to the occasion. Linguistic factors such as prominence or intonational focus (see Section 1.2) can locally affect pitch range, but it is determined overall by paralinguistic factors such as degree of animation and projection; the overall pitch range is raised or expanded when the speaker "speaks up" to project his voice, or when he is excited.

prosody. The rhythm and melody of speech as specified phonologically in the representation of its phrasal organization and intonational structure, and as realized phonetically in duration and loudness and pitch patterns.

reference line. In Japanese tone-scaling, the bottom of the pitch range, corresponding to the lowest possible f0 value for a tone in a speaker's pitch range.

standard Japanese. The speech of educated Tokyo speakers, as prescribed by the Japanese Broadcasting Corporation.

stress. A local non-tonal prominence on a lexically designated syllable in an English word, which is realized phonetically in the rhythmic pattern of relative lengths and loudnesses, and also by certain segmental patterns such as vowel and consonant lenition.

tone. The basic phonological element representing distinctive events in the melody — i.e., the melodic counterpart of a phonemic segment in the text string. We believe that these melodic segments are target pitch level specifications such as "high" and "low" rather than specifications of pitch change such as "rise" and "fall". (See Pierrehumbert and Beckman (forthcoming) for detailed arguments on this point.) In both English and Japanese, there are two tone types — H and L — and the type of each tone in an utterance, and its temporal location and f0 value reflect the prosodic phrasing and intonational focus structure of the utterance.

REFERENCES

Anderson, Mark D., Janet B. Pierrehumbert, and Mark Y. Liberman. 1984. "Synthesis by Rule of English Intonation Patterns." *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing.*

Beckman, Mary E. 1986. *Towards Phonetic Criteria for a Typology of Lexical Accent.* Netherlands Phonetic Archives No 7, Foris Publications.

Beckman, Mary E., and Janet B. Pierrehumbert. forthcoming. "Intonational Structure in Japanese and English." *Phonology Yearbook*, Vol. 3.

Chomsky, N. 1971. "Deep structure, surface structure, and semantic interpretation." In D.D. Steinberg and L.A. Jakobovits, eds., *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics, and Psychology.* Cambridge

University Press, Cambridge, 183-216.

Grosz, B. and C. L. Sidner. 1985. "The Structures of Discourse Structure." 6097, BBN Laboratories and Technical Note #369 from the AI Center, SRI International. To appear in *Computational Linguistics.*

Hirschberg, Julia, and Janet Pierrehumbert. 1986. "The Intonational Structuring of Discourse." This volume.

Liberman, Mark, and Janet Pierrehumbert. 1984. "Intonational Invariance under Changes in Pitch Range and Length." In M. Aronoff and R.T. Oehrle, eds., *Language Sound Structure.* MIT Press.

Marcus, M., D. Hindle, and M. Fleck. 1983. "D-Theory: Talking about Talking about Trees." *Proceedings of the 21st Annual Meeting of the Association for Computational Linguistics,* 129-136.

Marcus, M., and D. Hindle. 1985. "A computational account of extra-categorial elements in Japanese." Paper distributed at the SDF Japanese Syntax Workshop, UCSD, San Diego, March 1985.

Pierrehumbert, Janet B. 1980. *The Phonology and Phonetics of English Intonation.* MIT dissertation.

——. 1981. "Synthesizing Intonation." *Journal of the Acoustical Society of America,* **70**: 985-995.

Pierrehumbert, Janet B., and Mary E. Beckman. forthcoming. "Japanese Tone Structure." Paper submitted to *Linguistic Inquiry.*