

JAPANESE INTONATION SYNTHESIS USING SUPERPOSITION AND LINEAR ALIGNMENT MODELS

Jennifer J. Venditti* and Jan P. H. van Santen†

* Bell Laboratories and Ohio State University

† Oregon Graduate Institute Center for Spoken Language Understanding

ABSTRACT

This paper outlines a new approach to Tokyo Japanese intonation synthesis, in which the F0 contour of an utterance is generated using the superposition of multi-level phrase curves and lexical accent curves, coupled with linear alignment models which determine the precise alignment of the curves with the segmental material. We first discuss the construction of a phrase curve used to model the prosodic domain termed the 'UA-group' (defined below), and describe the alignment of this curve with the syllabic structure of the utterance. Then, we describe a separate accent curve, carrying independent prominence specification, which is added to this UA-curve in the case of accented phrases. The alignment of the accent curve with the segments is determined by linear alignment models.

1. BACKGROUND

There are a number of different ways that Tokyo Japanese intonation has been described for purposes of speech synthesis. Two well-known approaches to modeling Japanese intonation are the J_ToBI tone-sequence model [1, 3, 7, *inter alia*] and the Fujisaki superpositional model [2, *inter alia*]. In the tone-sequence model, high and low tones are phonologically associated to linguistically relevant morae or boundaries in the text input. These tones are realized in a phonetic F0 space determined by a topline and reference line, and can be scaled within this space using a small set of parameters. The F0 contour is generated by connecting the tone levels by linear interpolation and smoothing. In contrast, the superpositional model proposed by Fujisaki uses the addition of accent and phrase components (curves), as well as a variable F0 baseline, to generate the F0 contour. The phrase curves result from impulse commands located at intonation phrase edges, while the more local accent curves result from rectangular commands placed at key points in the speech stream. The height (aka. amplitude) of these commands, and also the duration of accent commands, can be freely manipulated based on linguistic and other properties of the text. The two types of commands are then smoothed by separate filters having specific invariant mathematical properties, and added (in the log domain) to produce the final F0 contour.

The tone-sequence and Fujisaki models have generated good-quality intonation in various Japanese text-to-speech synthesis systems. However, both approaches are too constrained in the F0 shapes they can produce, and in the precise alignment of these shapes with the actual speech signal. In the tone-sequence model, each tone is assigned the duration of its associated mora (boundary tones have zero duration), and a simple smoothing algorithm applied to the constructed contour is what determines the align-

ment of the F0 with the segments. In addition, current implementations use a small set of tone scaling parameters, which improperly restricts the shape of the output contour. In the Fujisaki model, in contrast, the accent/phrase commands can be located at any point in the speech stream, and accent commands may take any duration. However, the shape of each curve is constrained by the fixed properties of the smoothing filter itself, and thus lacks the variability needed to allow for influence of segmental structure and local prominence on the precise F0 alignment.

In this paper we present an alternative approach to Japanese intonation modeling which allows for a wider variation in shape of generated contours, and also provides a principled quantitative description of F0 contour alignment. Our model is seated in the superposition approach to intonation modeling, though it differs from Fujisaki's model in some important respects, as we describe in detail in the next section.

2. SUPERPOSITION-ALIGNMENT MODEL FOR JAPANESE

Our new approach to Japanese intonation modeling is partially motivated by the observation that the local prominence of an accent can vary continuously and independently from the prominence of the overall phrase in which it occurs. We have observed this in monomorphemic words, polymorphemic compound words, and multi-word phrases. Figure 1 shows an example of four repetitions of the noun phrase *sakihodo-no ma'ngô* "the mango from before" uttered by the same speaker during different readings of a cooking recipe. In the different repetitions, the local prominence of the accent on *ma'ngô* varies continuously, resulting in an F0 contour in which the accent resembles anything from a 'shoulder' (panel A) to a local 'peak' (panel D). The height of the accent under differences in prominence varies independently from the height of the phrase-initial rise (left side of each panel). The figure shows a marked concavity in the contour as the accentual prominence increases. We model this patterning in a superposition-based approach by isolating two component curves, the *UA-group phrase curve* and the *accent curve*, each of which may carry an independent prominence specification.

The UA-group curve is associated to the prosodic domain we term the *UA-group*: a sequence of zero or more lexically unaccented words followed by a lexically accented word or higher-level prosodic phrase boundary (such as the *Iphrase*, *minor phrase*, or *major phrase*).¹ That is, a UA-group may be composed of a se-

¹ This paper discusses the modeling of only the lower-level UA-group. The *Iphrase* is a higher-level unit similar to the Japanese ToBI *intonation phrase*, and is parsed from the input text based on syntactic and lexical factors. The minor and major phrases are defined by orthographic con-

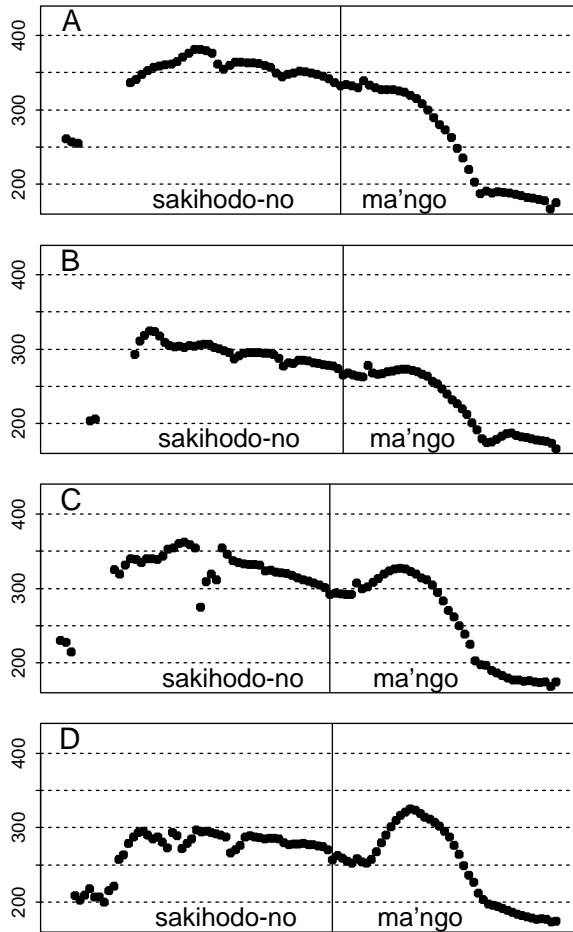


Figure 1: Four repetitions of the noun phrase *sakhodo-no ma'ngô* “the mango from before”. Vertical lines mark the start of the accented mora /ma/ in *ma'ngô*.

quence of unaccented words (U^+), unaccented words followed by an accented word (U^+A), or a single accented word (A). This prosodic unit is similar to the *accent phrase* defined by Fujisaki, and is similar in size though not identical to the Japanese ToBI *accentual phrase*. The UA-group curve characterizes the overall shape of the phrase, including the initial rise and plateau portions. In addition to this phrase curve, accented UA-groups (such as those shown in Figure 1) are modeled by the addition of a separate accent curve. By separating out these two component curves of the UA-group, it is possible to independently vary the heights of each curve, based on syntactic, pragmatic, or discourse properties of the text.

This superposition-based approach differs significantly from Fujisaki’s superpositional model in some key respects. First, the definition and domain of the *phrase* and *accent* curves is markedly different. The domain which Fujisaki terms the *accent phrase* is modeled in his approach by an *accent curve*, whose shape characterizes both the initial rise as well as the accentual peak-fall. In our model, in contrast, this domain is modeled by the addition of two separate curves, as described above. This allows for the

ventions, such as the comma and period, respectively. See [4] for further discussion of these levels.

much needed independent variation of local prominences within the phrase. In the Fujisaki model, since the shape of the accent curve is determined by the invariant mathematical properties of the smoothing filter, the resulting F0 shape is severely restricted. Specifically, the F0 height of the initial rise can only be higher or equal to that of the accentual peak-fall, and any local concavity between these two points cannot be modeled. Panel A of Figure 1 shows a contour which could be adequately modeled using Fujisaki’s approach. However, panels B-D, in which the continuously increasing prominence of the accent produces a local peak, cannot be modeled by a single accent curve in Fujisaki’s approach. In these cases, the contour would have to be modeled using two separate accent curves: one for the initial rise and plateau, and one for the accent rise-fall. The tone-sequence model also runs into similar problems. The contour in panel A could be modeled as one single *accentual phrase*, while those in panels B-D would need to be represented categorically different: as two separate phrases. In addition, the tone scaling parameters currently used in implementations of this approach do not allow the accent peak to be lower than the initial rise, which poses a problem for modeling the contour in panel A. In contrast to the Fujisaki and tone-sequence models, our approach is able to describe the continuous variation in local accent prominence produced by this speaker using a single mechanism: an *accent curve* of variable height added on top of the *UA-group phrase curve*, also of variable height. In the resulting contour, the height of the accent portion may vary continuously, and is only indirectly dependent on the overall height of the phrase curve on which it rides.

2.1. COMPONENTS OF THE UA-GROUP

Our implementation models the F0 contour of a UA-group using a constructed *UA-group curve* (including initial rise and plateau regions), with the optional superposition of an *accent curve* in the case of an accented UA-group. We will describe the shape and alignment characteristics of each of these curves below.

INITIAL RISE It is well-known that in Tokyo Japanese, the moraic structure of the first syllable of a phrase has an effect on the timing of the phrase-initial F0 rise. The rise ends around the offset of the 2nd mora when the initial syllable is light, while it ends near the offset of the 1st mora when the initial syllable is heavy. In addition, the F0 contour starts from a higher value in the heavy syllable condition, in comparison with light syllable condition [3, *inter alia*]. Figure 2 shows an example of these differences. Our own experimental data show that, while the alignment and F0 starting height do differ (as shown in the figure), the F0 height at the end of the rise does not differ significantly among the two syllable types (start F0: $t(258)=-6.64$, $p<.01$; end-of-rise F0: $t(258)=-2.19$, $p>.01$). In addition, the slope of the rise also does not significantly differ among the two conditions, as shown in Figure 3.

Based on this observation, we have chosen to model the phrase-initial rise using an ascending sigmoid with invariant slope which shifts its alignment depending on the structural property of the initial syllable, as schematized in Figure 4. In the light syllable condition, the inflection at the end of the rise is aligned with the end of the 2nd mora, while in the heavy syllable condition, it is aligned with the end of the 1st mora. Modeling the initial rise

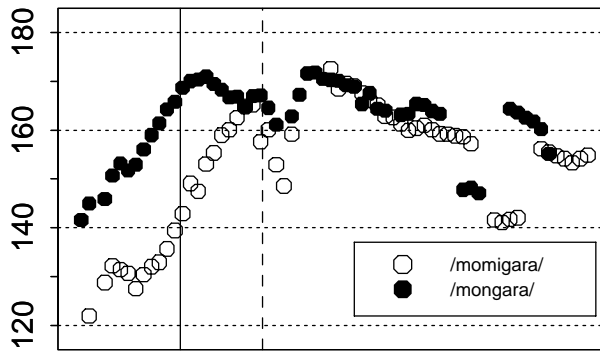


Figure 2: An example of the initial rise in phrases starting with a light vs. heavy syllable, uttered by the same speaker. Light: *momigara desu* “It’s a rice husk” (hollow) vs. heavy: *mongara desu* “It’s a folded pattern” (filled). Solid and dashed vertical lines mark the end of the 1st and 2nd morae in the heavy and light syllable words, respectively. The contours are aligned by the start of the file.

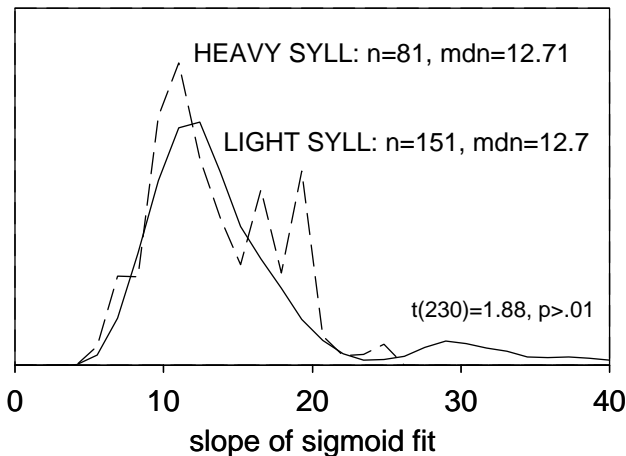


Figure 3: Histograms of slope values of sigmoids fit to initial rise in phrases starting with a light (solid line) vs. heavy (dashed line) syllable.

in this way causes the observed differences in start F0 height to fall out from differences in rise alignment, as schematized in the figure.²

PLATEAU Our experimental data indicate that the F0 in the plateau region of the UA-group curve falls linearly in totally unaccented phrases (e.g. Figure 2) and late-accented phrases (e.g. Figure 1). The F0 fall is significantly correlated with the number of mora in the plateau (corr=0.57, $t=14.46$, $p<.01$).³ Therefore, we model the plateau, the portion of the UA-group curve stretching from the end of the rise to the phrase end (in unaccented

²Currently, we use a sigmoid with invariant slope and duration. However, in future implementations, it would be desirable to use linear alignment models to tie the F0 rise to the durations of the segments which comprise the first two morae, much like we do for the accent curve (see below).

³Note that the height of the high F0 of the rise does not significantly vary according to phrase length.

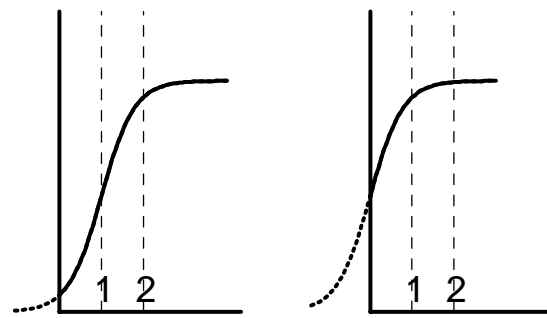


Figure 4: The alignment of a rising sigmoid to model phrases with initial light vs. heavy syllables. ‘1’ and ‘2’ mark the end of the 1st and 2nd mora in the phrase, respectively.

UA-groups) or to the start of the accented mora (in accented UA-groups), as a linear descent which is attributed to an underlying *minor phrase curve* upon which the UA-group curve rides (see [5, 4] for more details of this higher-level curve).

END In unaccented UA-groups, the F0 contour is modeled by the UA-group curve which is made up of the initial rise and plateau portions only. In accented UA-groups, there is an additional ‘end’ component to the UA-group curve. The end portion stretches from the end of the plateau to the end of the UA-group. This is currently modeled by a descending sigmoid: the upper inflection point is aligned with the end of the plateau, and the lower inflection is predicted by a linear alignment model which takes into consideration the durations of the accented and post-accentual regions, as described below. Therefore, in accented phrases, the UA-group curve consists of the rise-plateau-end portions.

ACCENT In addition, accented UA-groups have an additional accent curve which is added to generate the output contour, as outlined in Section 2. Figure 5 shows a monomorphemic accented word in which the linearly descending portion of the UA-group curve (the ‘end’, shown by the sloped line) is subtracted from the surface contour, resulting in a curve which is an estimate of the accent curve (shown in the bottom of the figure).

The peak of the isolated accent curve (dotted vertical line) most often corresponds to the local peak (or ‘shoulder’) in the original F0 contour. The accent peak has been generally thought to occur around the end of the accented mora (here, the first mora of the long syllable /nyU/). However, our experimental data suggest that its location can vary considerably, and is highly dependent on the durations of the segments in the accented mora and the duration of the remainder of the UA-group. Figure 6 plots data from an experiment in which verbs (e.g. *no’mu* “to drink”) were uttered with following verbal affixes ranging from 0–5 morae in length.⁴ In the figure, the short vertical lines mark the locations of the start (peak/shoulder) and end (lower inflection of ‘end’ portion) of the accentual fall, respectively. The figure shows that the start of the fall occurs at the end of the accented mora only in the case of *no’mu* with no following affixes. In the other conditions with more post-accentual morae, the start of the fall oc-

⁴1: *no’mu*, 2: *no’mu-no*, 3: *no’mu-no-da*, 4: *no’mu-no-da-ga*, 5: *no’mu-no-da-kedo*, 6: *no’mu-no-da-keredo*.

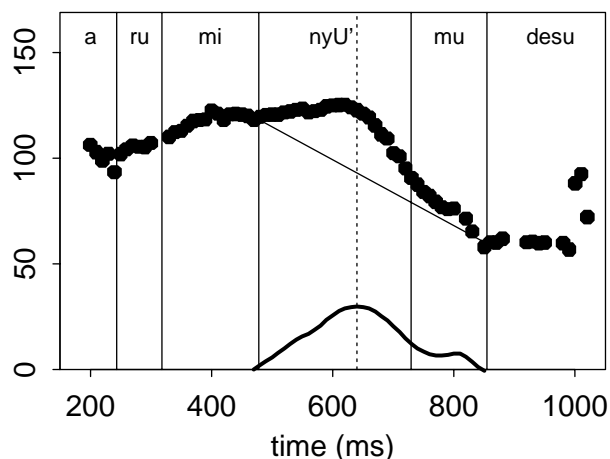


Figure 5: An example of a UA-group F0 contour with accent curve subtracted: *aruminyûmu desu* “It’s aluminum”. Solid vertical lines mark syllable boundaries.

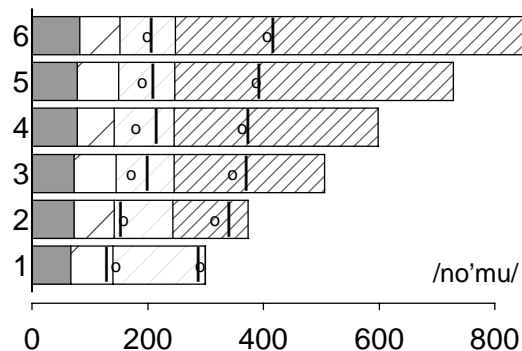


Figure 6: Mean durations of accented mora onset /n/, vowel /o/, following /mu/ and remaining portion of the phrase. Mean locations of start and end of F0 fall are marked with vertical lines, and predicted values of both points are marked with small circles.

curs (well-)within the following mora /mu/ (while still being perceived as falling on the lexically-accented mora /no/). However, even though the location of these points can vary considerably, we have found that they can be predicted by a linear alignment model which takes into consideration the duration of the accented mora onset, vowel, and (a transform of) the remaining portion of the UA-group. The resulting predictions are shown by the small circles in Figure 6.

But what points on the accent curve are important? So far, we have mentioned only two points, but there is no reason to a priori restrict analysis to just these. Rather, our approach broadens the scope to predict the location of *many* points on the curve. First, we sample the accent curve at locations corresponding to percentages of the maximal height (100% is the peak location, 50% pre-peak location is the time point where the accent curve is half of the maximal height, etc.). We call these time points *anchor points*. Then, the timing of each of these anchor points is predicted by an *alignment model*: the weighted linear sum of the durations of ‘parts’ of the associated sequence of phonemes, such as durations of the onset, vowel, and post-accentual region. Data from many

accent curves with varying segmental make-up and durations are used to estimate the regression weights, which we call *alignment parameters*. Finally, we take a roughly bell-shaped normalized accent curve and use it as a common template from which accent curves of varying durations can be generated using these parameters (see [4, 5, 6] for more details).

3. SUMMARY

This paper has outlined a new approach to Tokyo Japanese intonation synthesis, in which the F0 contour is generated by the superposition of UA-group (and other) phrase curves and lexical accent curves. The approach uses linear alignment models which determine the precise alignment of the curves with the segmental material. The result is a quantitative model of Japanese intonation contours which is also true to the underlying phonological contrasts in the intonation system. Our model maintains the important distinction between lexically unaccented vs. accented phrases, modeled using the UA-group curve vs. UA-group curve plus an accent curve, respectively. It accounts for the reported ‘strong/weak’ low F0 contrast in phrase-initial position, modeled by an ascending sigmoid with invariant slope which shifts its alignment. It also accounts for the reported contrast in the scaling of high tones (H- vs. H*): since the accent curve rides on top of the UA-group curve, this results in a higher accent F0 peak in early-accent phrases in comparison with unaccented phrases (this relative scaling need not be the case in late-accent phrases, as shown in Figure 1, panel A). In addition, modeling accented phrases as a superposition of an accent curve and a UA-group curve allows for continuous variation of the accent height due to differences in intonational prominence. This results in seemingly distinct output contours in which the accent can vary from a local ‘peak’ (high prominence) to a ‘shoulder’ (low prominence), but which are produced by the same underlying mechanism.

4. REFERENCES

1. Beckman, M. E., and Pierrehumbert, J. B. 1986. Intonational structure in Japanese and English. *Phonology Yearbook* 3:255–309.
2. Fujisaki, H., and Hirose, K. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journ. of the Acoustical Society of Japan* 5:233–242.
3. Pierrehumbert, J. B., and Beckman, M. E. 1988. *Japanese Tone Structure*. MIT Press.
4. van Santen, J. P. H. 1998. Intonation: The superpositional approach. (chapt. 6.3) In *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, R. Sproat, Ed. Kluwer Academic Publishers.
5. van Santen, J. P. H., and Möbius, B. 1997. A model of fundamental frequency contour alignment. In *Proc. of the ESCA Workshop on Intonation* (Athens, Greece), pp. 321–324.
6. van Santen, J. P. H., Möbius, B., Venditti, J. J., and Shih, C. 1998. Description of the Bell Labs intonation system. In *Proc. of the 3rd ESCA Workshop on Speech Synthesis* (Jenolan Caves, Australia), pp. 293–298.
7. Venditti, J. J. Japanese ToBI labelling guidelines. 1995. [http://ling.ohio-state.edu/Phonetics/J_ToBI/jtobi_home_page.html].