# MODELLING JAPANESE INTONATION USING PENTATRAINER2

Albert Lee, Yi Xu

Speech Hearing and Phonetic Sciences, UCL (University College London)
kwing.lee.10@ucl.ac.uk, yi.xu@ucl.ac.uk

## ABSTRACT

This paper presents results from Japanese intonation modelling using PENTAtrainer2, an articulatory synthesiser. Our first aim is to show that PENTA, on which PENTAtrainer2 is based, can achieve high accuracy in predictive synthesis of varying intonation contours. We trained the synthesiser on a 6251-sentence functionally annotated corpus and generated $F_0$ contours for each communicative condition. The accuracy of speaker-dependent and independent synthesis, together with naturalness ratings, show that PENTA is effective in modelling Japanese intonation. This suggests that once contextual variability is incorporated into a model, multi-functional targets alone would suffice as the prosodic representation even in a sizeable corpus.

**Keywords**: Focus, Japanese, Sentence type

## 1. INTRODUCTION

Analysis-by-synthesis is a robust way of showing how capable a model is in capturing variability in intonation. Through the accuracy of synthesis, one can compare various models using the same dataset. This approach, especially when the synthesis is predictive, is a big step toward solving the 'lack of reference problem' [1] in prosodic research.

The modeling of Japanese sentential prosody dates back at least to 1960s. To date, many models have been introduced, of which AM [2–4] and the Fujisaki Model [5] are among the most influential. AM models intonation by interpolating sparsely distributed tones (H and L), whereas Fujisaki Model superimposes the output of two second-order linear filters with a base frequency value. Here we test a third approach - Parallel Encoding and Target Approximation (PENTA) model [6, 7], and assess whether it can synthesize Japanese intonation with satisfactory accuracy, like it can for other languages [7, 8], or in Japanese word prosody [9].

PENTA takes a different set of assumptions from both AM and Fujisaki Model. It differs from the former in treating variation of $F_0$ alignment as contextual variation rather than as a part of phonology or phonetic implementation rules; and from the latter in that it has no stipulation on how many tiers a language can use at once to encode communicative information. A detailed explanation of PENTA, as well as its comparison with [2, 4], can be found in [10]. See also Figure 1 for an example of PENTA-style functional annotation.

Two issues will be addressed in this paper: (1) whether PENTA can predictively synthesize accurate $F_0$ contours for a speaker who is not part of the training corpus; and (2) whether the accuracy metrics employed by PENTA reflect its effectiveness in reality. These issues will be elaborated in the methodology below.

## 2. METHODS

### 2.1. The corpus

**Table 1**: *Corpus used in the present study. 'A' stands for 'accented' and 'U' for 'unaccented'.*

| | | Word I | | Word II | | Word III | | |
|---|---|---|---|---|---|---|---|---|
| Short | A | 'mei-ga<br>May が<br>May-NOM | × | 'momo<br>腿<br>thigh | × | -o 'mita<br>を見た<br>-ACC saw | × | ? |
| | U | mei-ga<br>姪が<br>Niece-NOM | | momo<br>桃<br>peach | | -ni nita<br>に似た<br>-DAT resembled | | 。 |
| Long | A | 'muumin-ga<br>ムーミンが<br>Moomin-NOM | × | 'budou<br>武道<br>martial arts | × | -o 'mita<br>を見た<br>-ACC saw | × | ? |
| | U | noumin-ga<br>農民が<br>Farmer-NOM | | budou<br>葡萄<br>grapes | | -ni nita<br>に似た<br>-DAT resembled | | 。 |

We collected a corpus of Japanese sentences for this study. There are 6,400 utterances (2 sentence lengths × 8 accented conditions × 2 sentence types × 4 focus conditions × 5 repetitions × 10 speakers). For each target sentence there are four possible focus conditions, namely, initial, medial, final, and neutral. The sentence types are yes/no questions vs. statements. Each sentence is either eight or 11 morae in length. Focus was elicited by having the speaker produce the question and the (corrective) statement in pair. Of the 6,400 utterances collected, 149 had to be discarded due to mis-production of the accent condition. A total of 6251 sentences were retained.

### 2.2. PENTAtrainer2

We used PENTAtrainer2 [7] to obtain pitch targets which were then used to synthesise $F_0$ contours. PENTAtrainer2 is a software package for semi-automatic analysis and synthesis of speech melody based on PENTA [6, 11]. It was written in Java

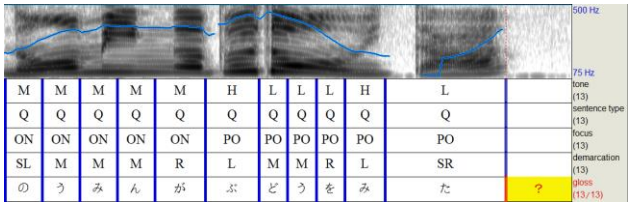
controlled by a set of Praat [12] scripts. The basic idea of PENTAtrainer2 is to extract the underlying pitch targets defined in terms of height (*b*), slope (*m*), and strength ($\lambda$) by means of stochastic analysis-by-synthesis based on quantitative Target Approximation (qTA) [13].

The analysis-by-synthesis in PENTAtrainer2 was controlled by simulated annealing, a machine learning algorithm [14]. To apply PENTAtrainer2, users first use the Annotation script to divide into layers the functions to be modelled, and label all the function-internal intervals, as illustrated in Figure 1. They then apply the learning script to extract globally optimal values of *b*, *m* and $\lambda$ for each of the functional combinations. The performance of the modelling is assessed numerically by root-mean-square error (RMSE) and Pearson's *r*. RMSE indicates the average mismatch of the synthetic and original contours while correlation indicates the mismatch between the shape and the alignment of the contours [7]. Other things being equal, an accurate synthesis would yield small RMSE and large *r* values.

### 2.2. Annotation

In this study, the raw sound data were first chunked into individual utterances, and then alternatively segmented by mora and by syllable in Praat. Under moraic segmentation, a heavy syllable is segmented into two intervals equal in duration. Then, the segmented data were functionally labelled like in Figure 1.

**Figure 1**. *Functional annotation in PENTAtrainer2. The labeled functions are* Tone, Sentence Type, Demarcation *and* Focus.




On the Tone tier each interval is marked H, M, or L, following [9]. In the case of syllabic segmentation an accented heavy syllable is labelled F. H represents the high target in an accented word (cf. H* in [15]), whereas M stands for the high target elsewhere (cf. H- in [15]). The low target in an accented word is marked L. Under syllable segmentation, pitch accent as hosted in a heavy syllable is hypothesised as bearing a falling target, thus the F label. Sentence Type is either Q (uestions) or S (tatements). Note that these labels provided no phonetic guidance to PENTAtrainer2, as they are treated simply as category identifiers. On the Focus tier, intervals in a focussed sentence are labeled as on-focus, pre-focus, or post-focus [16], and those in a neutral sentence are all labelled N (neutral). The The Demarcative tier contains information of the position of an interval in the sentence, comprising five categories – left/right edge of word, middle of word, and left/right edge of sentence. These four tiers combined give rise to 72 unique communicative conditions for the corpus, which means that the entire corpus will be synthesised using 72 sets of qTA parameters (*b, m,* $\lambda$).

During learning, globally optimal parametric values were obtained after 1,000 reiterations of target optimization. Section 3.1 reports the accuracy of speaker-dependent synthesis – synthesis of the $F_0$ contours of a given speaker using the global parameters learned from his/her own utterances. In Section 3.2, the results of predictive synthesis accuracy is presented. Here we adopted the Jackknife procedure [17], where the global parametric values of all speakers save one are averaged and used to predict the $F_0$ contours of the speaker being left out. The procedure is repeated ten times such that all ten speakers' data are assessed.

### 2.4. Naturalness judgment

The synthesis quality was also assessed perceptually in a naturalness judgment test. Sixteen monolingual native Japanese listeners (3 male) were recruited as subjects. They were all born and raised in the Greater Tokyo area (Tokyo, Saitama, Kanagawa, and Chiba), and aged between 23 and 37 years old (mean age = 27.9). Most subjects had arrived in the UK for less than a year, except one who had arrived for 12 months, and another who had spent two years in the USA. None reported any (history of) speech or hearing impairment.

The listening test took place in a quiet room in University College London. Subjects were seated in front of a laptop computer, which displayed the Praat [12] ExperimentMFC interface, and wearing circumaural headphones. They listened to each stimulus and rated the naturalness on a 1~5 scale, with 5 being the most natural. Each stimulus could be replayed up to three times.

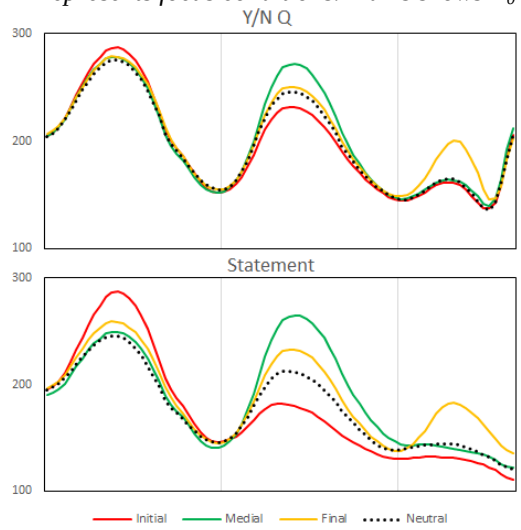
## 3. RESULTS

### 3.1 Partial acoustic analysis

Figure 2 shows averaged $F_0$ contours of an accented sentence spoken in different sentence types and focus conditions. Visual inspection suggests that there is on-focus raising of $F_0$ peak as well as post-focus compression of $F_0$ range. There is also a sentence-final rise, which is typical of questions in Japanese [3, 4]. We then compared each focus

condition*sentence type with its neutral focus counterpart, and ran repeated-measures ANOVA on each subset (N=6) of the results. For statements under initial focus, for example, there is significant on-focus raising of maximum $F_0$ ($F(1,9)=61.9$ $p<0.001$), echoing with [18], as well as significant interaction between focus and the accent condition of the focused item on post-focus mean $F_0$ ($F(1,9)=32.9$ $p<0.001$), among other effects. All these focus markers are in line with those reported in other studies on Japanese prosodic focus [4], [19].

**Figure 2**. *$F_0$ contours each averaged from 50 repetitions (mei-ga momo-o mita). The left panel is yes-no questions, and the right panel statements. Colour of the curves represents focus conditions. Y axis shows $F_0$ in Hz.*



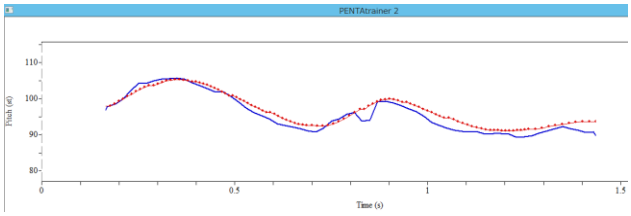

### 3.2. Speaker-dependent synthesis accuracy

**Table 2**: *Accuracy of speaker-dependent synthesis.*

| Sentence type | Focus | Mora | | Syllable | |
|---|---|---|---|---|---|
| | | RMSE | *r* | RMSE | *r* |
| Question | Final | 1.495 | 0.915 | 1.514 | 0.913 |
| | Initial | 1.662 | 0.917 | 1.776 | 0.900 |
| | Medial | 1.555 | 0.921 | 1.644 | 0.902 |
| | Neutral | 1.603 | 0.908 | 1.711 | 0.890 |
| | **Sub-avg** | **1.579** | **0.915** | **1.662** | **0.901** |
| Statement | Final | 1.416 | 0.911 | 1.391 | 0.910 |
| | Initial | 1.725 | 0.928 | 1.823 | 0.915 |
| | Medial | 1.513 | 0.923 | 1.625 | 0.906 |
| | Neutral | 1.657 | 0.891 | 1.678 | 0.876 |
| | **Sub-avg** | **1.578** | **0.913** | **1.630** | **0.901** |
| **Grand average** | | **1.578** | **0.914** | **1.646** | **0.901** |

Table 2 shows the respective mean synthesis accuracy under moraic and syllabic segmentation in terms of RMSE and *r*. Here the articulatory parameters used to synthesise the $F_0$ contours of a given speaker is obtained through training on the utterances of the same speaker. Across sentence types and focus conditions, synthesis accuracy is high with $r > 0.9$ and RMSE < 1.8 in most cases. Although *r* appears to be greater in certain contexts, those cases do not see a smaller RMSE at the same time, suggesting that no particular condition is more accurately modelled than the others. We also carried out visual inspections as illustrated in Figure 3, and found that the resynthesized contours were highly similar to their original counterpart.

**Figure 3**. *An interface of PENTAtrainer2 for visual inspection of synthesis accuracy. The blue curve is the $F_0$ contour of a natural utterance whereas the red dotted curve represents the corresponding resynthesis. The target sentence is muumin-ga budou-o mita* 'Moomin watched martial arts', *with focus on the first word.*




### 3.3. Speaker-independent synthesis accuracy

Table 3 shows mean predictive synthesis accuracy under the Jackknife procedure, where the speaker being modelled was excluded from training. The resynthesis deviated more from natural utterances (cf. 3.2 above). The overall accuracy is RMSE = 2.733 and $r = 0.8$ for moraic segmentation, and RMSE = 3.024 and $r = 0.702$ for syllabic segmentation. The advantage of the mora segmentation is greater here than in 3.2.

**Table 3**. *Synthesis accuracy of PENTAtrainer2 under Jackknife procedure by sentence type and focus condition.*

| Sentence type | Focus | Mora | | Syllable | |
|---|---|---|---|---|---|
| | | RMSE | *r* | RMSE | *r* |
| Question | Final | 2.468 | 0.849 | 2.658 | 0.769 |
| | Initial | 2.940 | 0.787 | 3.130 | 0.736 |
| | Medial | 2.639 | 0.804 | 2.947 | 0.721 |
| | Neutral | 2.554 | 0.842 | 2.979 | 0.732 |
| | **Sub avg** | **2.650** | **0.821** | **2.929** | **0.739** |
| Statement | Final | 2.766 | 0.739 | 3.018 | 0.663 |
| | Initial | 3.242 | 0.793 | 3.327 | 0.751 |
| | Medial | 2.687 | 0.800 | 2.975 | 0.662 |
| | Neutral | 2.578 | 0.785 | 3.154 | 0.581 |
| | **Sub avg** | **2.817** | **0.779** | **3.118** | **0.664** |
| **Grand average** | | **2.733** | **0.800** | **3.024** | **0.702** |

### 3.4. Naturalness judgment results

Results of the naturalness judgment test are found in Table 4. We are interested in whether Type of stimuli (original vs. synthesised) affects how a listener rates the naturalness of stimuli, or whether its interaction with other effects reaches statistical significance. Result of a repeated measures ANOVA shows that Type of stimuli has no significant main effect on naturalness judgment rating. This suggests that the two types of stimuli sounded equally natural to the native listeners. The grand mean rating of natural stimuli is 3.688, which is close to that of synthesised stimuli (3.658, out of a 1~5 scale).

Incidentally, on the whole, statements (mean=3.817) were judged to sound more natural than questions (mean=3.528).

**Table 4**: *Mean naturalness ratings by focus condition and sentence type.*

| Sentence type | Focus | Original | Synthesis | Average |
|---|---|---|---|---|
| Question | Final | 3.397 | 3.441 | 3.419 |
| | Initial | 3.588 | 3.691 | 3.64 |
| | Medial | 3.566 | 3.456 | 3.511 |
| | Neutral | 3.507 | 3.581 | 3.544 |
| | **Sub-avg** | **3.515** | **3.542** | **3.528** |
| Statement | Final | 3.816 | 3.809 | 3.813 |
| | Initial | 3.772 | 3.669 | 3.721 |
| | Medial | 3.801 | 3.743 | 3.772 |
| | Neutral | 4.051 | 3.875 | 3.963 |
| | **Sub-avg** | **3.860** | **3.774** | **3.817** |
| **Grand average** | | **3.688** | **3.658** | **3.673** |

## 4. DISCUSSION

The present study has shown that Japanese sentential prosody can be modelled with parametric representations based on PENTA, an articulatory-functional model. Compared to a previous study on lexical prosody [9] (Speaker dependent [mora] RMSE = 1.088, $r$ = 0.914; [syllable] RMSE = 1.092 $r$ = 0.896), results in Table 2 are very similar. On the other hand, synthesis accuracy is much lower under Jackknife procedure (Table 3), compared to [9] where [mora] RMSE = 1.739, $r$ = 0.853; [syllable] RMSE = 2.227, $r$ = 0.796, suggesting that there is more cross-speaker variability in sentential prosody than in lexical prosody. This observation echoes with [18] where some focus cues like pre-focus $F_0$ lowering was found to be optionally used by some speakers, whereas other cues like post-focus compression were consistently used by all; for word prosody, such freedom is less common owing to the need to mark lexical contrasts.

Our results agree with [9] where moraic segmentation yielded better synthesis accuracy. This may seem to suggest that the mora is the true tone-bearing unit in Japanese for tonal target approximation, contra other languages like Mandarin and English where tonal targets are hosted in the syllable. However, we are hesitant to come to such a conclusion because a heavy syllable under moraic segmentation comprises two intervals, but one interval under syllabic segmentation. This means that by nature the former involves more degrees of freedom, leading to better ability to capture variability. Thus these results cannot be taken as an answer to what the domain of target approximation of Japanese is; the question needs to be tackled through better controlled experiments, which take into account confounds from degrees of freedom. A follow-up study is under way to address this issue.

The high synthesis accuracy is supported by naturalness judgment ratings by native listeners. This means that the synthetic stimuli do not sound different from the natural ones to our participants. By extension, the remaining errors not captured by PENTAtrainer2 do not make the resynthesis any less natural-sounding. This means that the key information has been successfully encoded in the learned parameters. Therefore, PENTAtrainer2 can offer, for purposes like perception tests, natural sounding stimuli which are free of cross-repetition inconsistency common in natural stimuli.

A further implication of our results is that the PENTA model as well as its prosodic representation are well suited for Japanese. Syllable-by-syllable target specification, as we have shown, is adequate for a corpus with numerous non-contextual (i.e. functional) variations. The encoding schemes of all functions jointly determine a unique articulatory target of each syllable. Then, by incorporating articulatory factors [20], there is no need to specify temporal alignment of tone. Whether our approach is superior to other frameworks is still an open question, but we have shown that PENTA representation is at least as suitable for Japanese as for other languages like Mandarin and English [7, 8].

On a side note, the training process of PENTAtrainer2 is also reminiscent of child language acquisition. The development of infant speech relies on audition – deaf children cannot learn to speak by themselves [22, 23]. Over the course of repetitions PENTAtrainer2 refines its articulatory parameters in order to generate $F_0$ contours that are more similar to the original, just as infants gradually refine their articulation over time by listening to themselves during practice.

The present study is but a first step. To fully understand the nature of Japanese prosody, future research could compare different theories through their modelling performances using the present data set. A number of tools are being developed to compare PENTA and other models in a fair manner.

## 5. CONCLUSION

Compared to a previous study on Japanese lexical prosody, the synthesis accuracy of PENTAtrainer2 was highly comparable. Our naturalness judgment test showed that resynthesis did not sound different from the natural stimuli to the native listeners, confirming that the accuracy measurements were effective. These results pave the way for future efforts on model comparison, which is necessary for a thorough understanding of Japanese prosody.

## 6. REFERENCE


[1] Xu, Y. 2011. Speech prosody: A methodological review. *J. Speech Sci*. 1, 85–115.

[2] Pierrehumbert, J. 1980. *The phonology and phonetics of English intonation*. PhD thesis, MIT.

[3] Beckman, M., Pierrehumbert, J. 1986. Intonational structure in Japanese and English. *Phonol. Yearb.* 3, 255–309.

[4] Pierrehumbert, J., Beckman, M. 1988. *Japanese Tone Structure*. Cambridge: MIT.

[5] Hirose, K., Fujisaki, H., Yamaguchi, M. 1984. Synthesis by rule of voice fundamental frequency contours of spoken Japanese from linguistic information. *Proc. 1984 IEEE Int. Conf. Acoust. Speech, Signal Process*. San Diego, 2.13.1–2.13.4.

[6] Xu, Y. 2005. Speech melody as articulatorily implemented communicative functions. *Speech Commun.* 46, 220–251.

[7] Xu, Y., Prom-on, S. 2014. Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Commun.* 57, 181–208.

[8] Liu, F., Xu, Y., Prom-on, S., Yu, A. 2013. Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modelling. *J. Speech Sci.* 3, 85–140.

[9] Lee, A., Xu, Y., Prom-on, S. 2014. Modeling Japanese F0 contours using the PENTAtrainers and AMtrainer. *Proc TAL 2014* Nijmegen, 164–167.

[10] Xu, Y., Lee, A., Prom-on, S., Liu, F. in press. Explaining the PENTA model: A reply to Arvaniti & Ladd (2009). *Phonology*.

[11] Xu, Y., Wang, Q. 2001. Pitch targets and their realization: Evidence from Mandarin Chinese. *Speech Commun.* 33, 319–337.

[12] Boersma, P., Weenink, D. 2015. *Praat: doing phonetics by computer* [Computer program]. Retrieved 24 March 2014 from http://www.praat.org/

[13] Prom-on, S., Xu, Y., Thipakorn, B. 2009. Modeling tone and intonation in Mandarin and English as a process of target approximation. *J. Acoust. Soc. Am.* 125, 405–424.

[14] Kirkpatrick, S., Gelatt, C., Vecchi, M. 1983. Optimization by simulated annealing. *Science* 220(4598), 671–680.

[15] Venditti, J. 2005. The J_ToBI model of Japanese intonation. In: Jun, S.-A. (ed) *Prosodic typology: The phonology of intonation and phrasing*. New York: OUP, 172–200.

[16] Xu, Y., Xu, C., Sun, X. 2004. On the temporal domain of focus. *Proc Speech Prosody 2004* Nara, 81–84.

[17] Quenouille, M. 1986. Notes on bias in estimation. *Biometrika* 43, 353–360.

[18] Lee, A., Xu, Y. 2012. Revisiting focus prosody in Japanese. *Proc Speech Prosody 2012* Shanghai, 274–277.

[19] Sugahara, M. 2003. *Downtrends and post-focus intonation in Tokyo Japanese*. PhD thesis, University of Massachusetts, Amherst.

[20] Xu, Y., Sun, X. 2002. Maximum speed of pitch change and how it may relate to speech. *J. Acoust. Soc. Am.* 111, 1399–1413.

[21] Raphael, L., Borden, G., Harris, K. 2011. *Speech science primer: Physiology, acoustics, and perception of speech*, 6th ed. Baltimore: Wolters Kluwer.

[22] Oller, D., Eilers, R. 1988. The role of audition in infant babbling. *Child Dev.* 59, 441–449.