ISCA Archive
http://www.isca-speech.org/archive

4th International Symposium on
Tonal Aspects of Languages (TAL-2014)
Nijmegen, The Netherlands
May 13-16 2014

Modeling Japanese F0 contours using the PENTAtrainers and AMtrainer

Albert Lee[1], Yi Xu[1], Santitham Prom-on[2]

[1]Dept. of Speech, Hearing and Phonetic Sciences, University College London, United Kingdom.
[2]Dept. of Computer Engineering, King Mongkut's University of Technology Thonburi, Thailand.

kwing.lee.10@ucl.ac.uk, yi.xu@ucl.ac.uk, santitham@cpe.kmutt.ac.th

Abstract

This study is a user report on modeling and predicting the F_0 contour of Japanese utterances with PENTA Model and AM Theory. Through a 2640-word corpus of Japanese, the two models were compared using two versions of PENTAtrainer, as well as newly developed AMtrainer, respectively. The respective modelling and synthesis accuracies of different models are discussed with reference to the type of input they require. The satisfactory results achieved by these tools demonstrate their potential in offering a fair and direct comparison between the two models. We call for more collaborative effort in comparative modeling to achieve this goal.

Index Terms: AM Theory, PENTA Model, Pitch accent

1. Introduction

One challenge faced by researchers working on speech prosody is the 'lack of reference problem' [1]. That is, the study of prosody does not have a reference like word identity in the case of studying segments (except in the case of lexical tone). The resulting inadequate understanding of the meaning communicated through prosody has led to rival theories of intonation coexisting over the past decades.

Computational modeling is a rigorous way of testing our predictive knowledge of prosody. However, little work has been done on comparing the effectiveness of different models, using comparable data and evaluation protocols, with a few exceptions [2], [3]. For Japanese intonation, one of the prevailing models is the Autosegmental-Metrical (AM) Theory [4], [5], which we will compare with the Parallel Encoding and Target Approximation (PENTA) model here in terms of their learning accuracy.

1.1. PENTA Model

Proposed by Xu [6], PENTA assumes that speech prosody conveys multiple communicative functions simultaneously, each with a unique encoding scheme. Unlike other functional (as opposed to formal) theories, PENTA assumes that functional encoding is done through an articulatory mechanism of target approximation [7], and that such an articulatory mechanism is responsible for the final production of both lexical tones and intonation.

A number of theoretical assumptions set PENTA apart from other theories – first, it leaves no tone-bearing unit (e.g. syllable) unspecified for pitch target (contra the sparse tones assumption in AM); second, phrasing is but a communicative function e.g. [8] encoded in parallel with others like sentence type and lexical tone, rather than being part of an abstract superordinate hierarchy; third, F_0 turning points such as 'peaks' and 'valleys' are not targets themselves, but are merely byproducts of target approximation.

1.2. AM Theory

AM is a theory of intonational phonology, based mainly on Pierrehumbert [5] and subsequent work. In AM, intonation is viewed as a linear series of H and L tones, which correspond to prominent F_0 maxima and minima. Alignment (timing in relation to segments) and scaling (height) are the two principle dimensions characterizing the phonetic realization of tones.

AM differs from PENTA in many ways. Notably, AM assumes sparse tones specification, i.e. not all syllables need to be specified for tone targets. Unlike target approximation in PENTA, AM assumes linear or non-linear interpolation between surface F_0 turning points to be the core mechanism of generating continuous F_0 contours. Moreover, other factors being held constant, AM assumes temporal alignment of tones relative to segments to be phonologically specified; whereas PENTA has no 'alignment' specification other than full synchrony of pitch target with the syllable. Most importantly, whereas pitch targets in PENTA are articulatory-based parametric representation of communicative functions, in AM pitch accents and boundary tones are symbolic representation of autonomous phonological elements.

2. Methodology

2.1. The corpus

The corpus used in the present study was previously reported in Lee et al. [9]. A total of 33 Japanese words were chosen as stimuli (see Table 1). The target words varied in length (1-4 morae), accent condition (unaccented and initial/medial/penultimate/final accent), and syllable structure (CVCV, CVn, CVV). From eight speakers, altogether 2,640 utterances (33 target words × 8 speakers × 5 repetitions × 2 speech rates) were collected. The target words are framed in the carrier sentence *Jiten-ni X-mo nottemasu* 'The word X too is found in the dictionary'.

1-mora	CV		
UA (L-H)	ne		
1 (H*-L)	`ne		
2-mora	CV	CVV	CVN
UA (LH-H)	mane	mai	
1 (H*L-L)	`memo	`mei	`men
2 (LH*-L)	mu`ne		
3-mora	CV	CVV	CVN
UA (LHH-H)	mimono	mimei, neimo	momen
1 (H*LL-L)	`menami	`meimu, `nimei	`ninmu
2 (LH*L-L)	na`name	me`mai	ni`man
3 (LHH*-L)	mimo`no	nui`me	
4-mora	CVCV	CVV	CVN
UA (LHHH-H)	monomane	meimei	nennen
1 (H*LLL-L)		`muumin	`nannen
2 (LH*LL-L)	mi`namina		
3 (LHH*L-L)	nama`nama	mei`mei	men`men
4 (LHHH*-L)	anoma`ma	nimai`me	ninen`me

Table 1. *List of stimuli used in the present study*

2.2. PENTAtrainer

PENTAtrainers are two semi-automatic software packages for analysis and synthesis of speech melody based on communicative functions and Target Approximation model [6], [7]. They are both in the form of Praat [10] scripts. The basic idea of PENTAtrainers is to extract the underlying pitch targets defined in height (*b*), slope (*m*), and strength (λ) by means of automatic analysis-by-synthesis based on the quantitative Target Approximation (qTA) [11].

PENTAtrainer1 extracts target parameters locally unit by unit through exhaustive search. For each target interval (typically the syllable), PENTAtrainer1 compares all possible combinations of *b*, *m*, and λ within the search ranges and finds the parameter combination that generates F_0 contours with the least difference from the original. It also records learning accuracy in terms of Root-Mean-Square Error (RMSE) and Pearson's *r* for each labeled interval, as well as the mean RMSE and global *r* for all the labeled intervals.

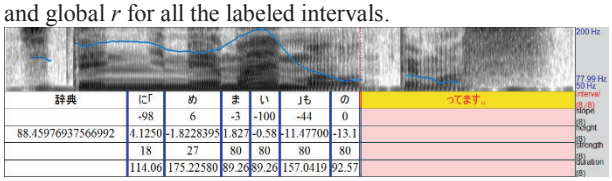

Figure 1. *Extraction of model parameters for each labeled interval by PENTAtrainer1 (*Jiten-ni me`**ma**i-mo nottemasu *'The word "memai" too is found in the dictionary'). In order, the second to the fifth tiers show slope, height, strength, and duration of the labeled intervals. The parameter numbers in Tiers 2-5 are extracted rather than manually entered.*

PENTAtrainer2 extracts qTA targets globally from an entire corpus by means of analysis-by-synthesis based on simulated annealing [12]. To apply it, users need to annotate each interval with labels for the functions being modeled, as illustrated in Figure 2. The result of the target extraction will be globally optimal values of *b, m*, and λ for each of the functional combinations. Like PENTAtrainer1, PENTAtrainer2 records RMSE and *r* values as indicators of modeling performance.

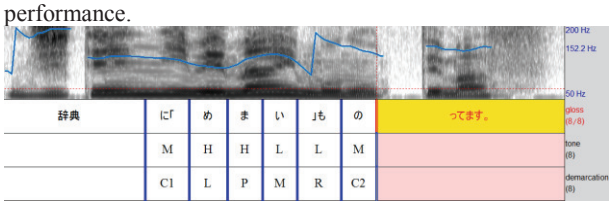

Figure 2. *Functional annotation in PENTAtrainer2* (same sentence as in Fig. 1). *The labeled functions are* Tone *and* Demarcation.

With both PENTAtrainers, predictive F_0 contour generation can be performed with categorical target parameters. With PENTAtrainer2, the categorical parameters are extracted directly. With PENTAtrainer1, the categorical parameters are the mean parameters of all the individual tokens of the same category. F_0 contours generated with the categorical parameters can then be compared to those of the natural utterances.

2.3. AMtrainer

AMtrainer is a Praat-based training model newly developed by the second author. It provides a similar user interface as PENTAtrainer1, but the parameters extracted are *location* and *height*. Built upon algorithms proposed in [13], AMtrainer take as input point tier labels (see Figure 3), which correspond to specific F_0 turning points on the surface; the rest of the F_0 contours are assumed to result from linear or sagging interpolation between the turning points.

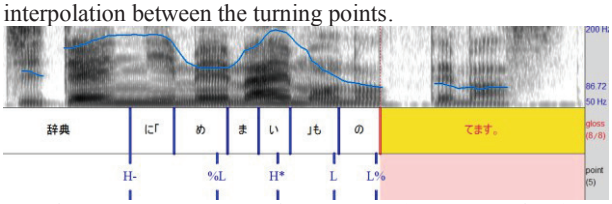

Figure 3. *ToBI annotation for AMtrainer (*Sentence as Fig. 1*)*

The present analysis follows the standard J-ToBI annotation convention [14] for Japanese lexical prosody. The annotation of an unaccented word consists of boundary tone (L%), and phrasal tone (H-), whereas that of an accented word contains also pitch accent (H*+L). Note that for simplicity's sake here the phrasal tone (H-) is omitted in cases where pitch accent occurs in the first or second mora of the word. See Venditti [14] for more information regarding J-ToBI.

Annotation for AMtrainer was performed in three steps. First, continuous F_0 contours were obtained from ProsodyPro [15]. Then, from these data the F_0 turning points corresponding to %L, H-, H*+L, and L% were identified for each utterance; and subsequently, converted to Praat TextGrid files to be used as the input for AMtrainer. The criteria for identifying F_0 turning points were as follows:

Tone	Definition
H- (#1)	This tone corresponds to the beginning of the case marker *–ni*, which is part of the carrier sentence that precedes the target word. The inclusion of this tone is to allow for interpolation with the following %L.
%L	This tone corresponds to minimum F_0 in the first mora of the target word.
H- (#2)	This tone corresponds to the maximum F_0 velocity value in the second and third morae of the target word.
H*	This tone corresponds to maximum F_0 in the accent host mora and the ensuing one.
L	This tone corresponds to the minimum F_0 velocity value in the first two post-accent morae.
L%	This tone corresponds to minimum F_0 in the mora after the target word (i.e. *no-*).

Table 2. *Label extraction criteria for AMtrainer*

Annotation files generated under Table 2 were then fed into AMtrainer for further analyses, of which results will contain the height and the location of each tone, as well as RMSE and Pearson's *r* between the original and model-generated F_0 contours for each utterance.

2.4. Analyses

The accuracy of PENTAtrainers depends on both the target approximation algorithm and how well the annotation/categorization scheme represents the sources of variation of the data. Here we consider two schemes *–Mora,* and *Syllable*.

The analysis will be presented over three subsections below. PENTAtrainers assess the goodness of fit between the synthesized and original F_0 contours using two measurements, namely, learning accuracy and synthesis accuracy. Although the two measurements are highly similar in nature, the design of AMtrainer renders it only possible to yield the former, in terms of which, in Section 3.1, we will compare AMtrainer and the PENTAtrainers, before proceeding to our discussion of the *synthesis accuracy* results of PENTAtrainer1 and PENTAtrainer2. In Section 3.2, we consider 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 3.3, the results of predictive

synthesis accuracy is presented. Here we adopt the Jackknife procedure [16], where the global parametric values of all speakers save one are averaged and used to predict F_0 contours of the speaker being left out. The procedure is repeated eight times such that all eight speakers' data are assessed.

	Mora	Syllable
Tone	H,M,L	H,M,L,F
Demarcation	C1,L,M,R, P,LP,C2	C1,L,M,R, P,LP,C2
TBU	Mora	Syllable

Table 3. *Functional labels used in the three annotation schemes for PENTAtrainer1 and PENTAtrainer2.*

3. Results

3.1. Learning accuracy

Table 4 shows the learning accuracy of AMtrainer, PENTAtrainer1, and PENTAtrainer2. Both annotation schemes under PENTAtrainer1 (second and third groups from top) yielded higher Pearson's *r* (0.998 and 0.994) and lower RMSE (0.101 and 0.122) than the other groups, suggesting that synthesized F_0 contours from PENTAtrainer1 differed less from the original. AMtrainer reached similar learning accuracy to PENTAtrainr2.

	Segmentation	Accented		Unaccented		Overall	
		RMSE	*r*	RMSE	*r*	RMSE	*r*
AMTrainer		0.623	0.972	0.727	0.765	0.654	0.909
P1	Mora	**0.112**	**0.998**	**0.075**	**0.992**	**0.101**	**0.996**
	Syllable	0.136	0.997	0.09	0.985	0.122	0.994
P2	Mora	1.117	0.960	1.021	0.804	1.088	0.913
	Syllable	1.129	0.958	1.006	0.743	1.092	0.893

Table 4. *Learning accuracies of AMtrainer, PENTAtrainer1 and PENTAtrainer2.*

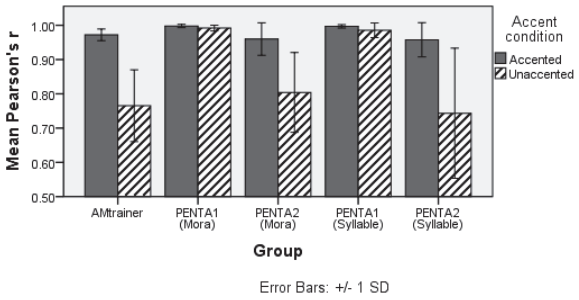

Figure 4. *Mean Pearson's r of by all three trainers.*

An interesting pattern emerged after subsetting data according to accent conditions (accented vs. unaccented). As is obvious in Table 4, learning accuracy was considerably lower in unaccented words than in accented words for AMtrainer and PENTAtrainer2; for PENTAtrainer1 accent condition did not improve learning accuracy much. Moreover, for PENTAtrainer2, learning accuracy of unaccented words was higher with mora being the tone-bearing unit than otherwise.

3.2. Speaker-dependent predictive synthesis

In this subsection synthesis accuracy results are reported. Assessment of accuracy is based on all the original F_0 contours of a given speaker compared with those generated from the global parametric values learned from all the utterances of the same speaker. Note that since PENTAtrainer1 only extracts local parametric values of individual utterances, here the global values used for PENTAtrainer1 are the result of averaging over individual local values.

Table 5 shows that PENTAtrainer2 has an advantage over PENTAtrainer1 in synthesis accuracy. This difference is to be solely attributed to the sources of global parametric values used for generating F_0 contours – for PENTAtriner1, the global parametric values are the averages of local *b, m*, and λ, whereas for PENTAtrainer2, the global values are directly obtained through optimizations over an entire corpus. The present results thus show the effectiveness of global optimization for predictive synthesis.

	Segmentation	Accented		Unaccented		Overall	
		RMSE	*r*	RMSE	*r*	RMSE	*r*
P1	Mora	1.786	0.924	1.796	0.674	1.789	0.849
	Syllable	1.746	0.941	2.054	0.628	1.839	0.846
P2	Mora	**1.117**	**0.962**	1.021	**0.804**	**1.088**	**0.914**
	Syllable	1.129	0.960	**1.006**	0.748	1.092	0.896

Table 5. *Accuracies of speaker-dependent synthesis by both PENTAtrainers.*

Figure 5 shows the synthesis accuracy of the PENTAtrainers under different accent conditions. Similar to what was observed in Table 4, unaccented words consistently achieved weaker Pearson's *r* than their accented counterparts. Note that PENTAtrainer1 achieved much weaker *r* than it did in Table 4, because here F_0 contours were synthesized using averaged global values, whereas in Table 4 synthesis was based on local parametric values, and did not have to capture cross-repetition variations.

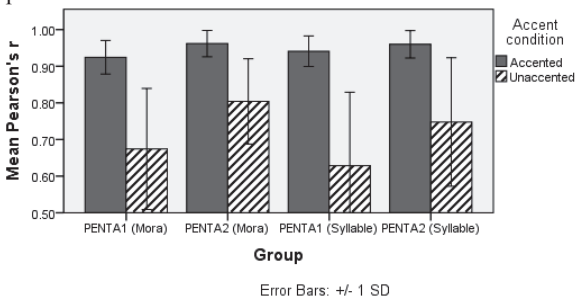

Figure 5. *Mean Pearson's r of speaker-dependent synthesis by both PENTAtrainers.*

Finally, in an attempt to improve the synthesis accuracy of unaccented words, we tested an additional function 'Word Length' (i.e. 1-4 morae), alongside 'Tone' and 'Demarcation'. Despite using more predictors (11→32 for moraic segmentation, 15→41 for syllabic segmentation), and the known effect of word length on F_0 in Japanese [17], we did not see considerable improvement in synthesis accuracy for unaccented words, with RMSE=0.952, *r*=0.797 (down from 0.804) for moraic segmentation, and RMSE=0.958, *r*=0.793 (from 0.748) for syllabic segmentation. This suggests that 'Word Length' was not effective in capturing the remaining variation in the data.

3.3. Speaker-independent Predictive synthesis

	Segmentation	Accented		Unaccented		Overall	
		RMSE	*r*	RMSE	*r*	RMSE	*r*
P1	Mora	1.761	0.925	1.807	0.668	1.775	0.847
	Syllable	1.938	**0.932**	2.307	0.585	2.050	0.826
P2	Mora	**1.726**	0.921	**1.767**	**0.696**	**1.739**	**0.853**
	Syllable	2.088	0.877	2.547	0.608	2.227	0.796

Table 6. *Accuracies of speaker-independent predictive synthesis*

Using the Jackknife procedure, the predictive power of the global articulatory parameters of PENTAtrainers was assessed for each speaker in the corpus. As can be seen from Table 6,

once recordings of the speaker being assessed is excluded from the training corpus, PENTAtrainer2 no longer showed absolute advantage over PENTAtrainer1. This is especially the case with unaccented words.

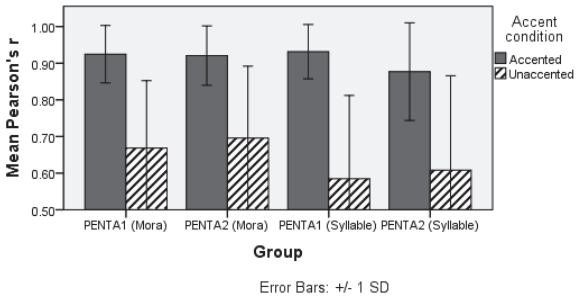

Figure 6. *Mean Pearson's r of speaker-independent predictive synthesis (Jackknife procedure)*

4. Discussion

AM and PENTA differ in terms of several theoretical assumptions, which have implications on their predictive power. Notably, in the former a tone target is a point in the surface contour but an underlying linear trajectory in the latter. In addition, for AM the temporal alignment of a tone in relation to segments is flexible and has to be specified. For PENTA pitch targets and the tone-bearing units are synchronized and so no further alignment specification is needed.

We are unable to assess whether AM or PENTA is superior in the present study because tone labelling for AMtrainer was performed post-hoc based on the actual location of acoustical landmarks whereas the categories used in PENTAtrainers were pre-defined. The goodness of fit by AMtrainer only reflects the effectiveness of linear and sagging interpolation as an F_0 contour generation mechanism; in contrast, both PENTAtrainers perform predictive synthesis based on functional categories. That AM labels were extracted from the actual location of acoustical landmarks limits the comparability between the tools here. To render AM more comparable to models that take categorical input like PENTA, the height value and temporal alignment of its labels need to be predicted by an algorithm rather than added post-hoc. To do so, one may first find out the segmental anchoring behavior of each tone [18] and then calculate the temporal alignment of the tones in each utterance; whereas for scaling there is no simple way of prediction yet. Once the issue of post-hoc annotation for AMtrainer is overcome, it would be ideal to test the assumptions of temporal alignment vs. target approximation using a dataset that controls for speech rate, like the one used in the present study.

Nonetheless, this paper has shown that both AM and PENTA can fit Japanese word prosody non-predictively with satisfactory accuracy. The fact that both models do almost equally well means that AMtrainer and PENTAtrainer can serve as a platform for direct comparisons between the two theories if used properly. We call for collaborative efforts to reach this goal by investigating more types of speech data and devising a protocol of annotation for unbiased contrast of the models.

5. Conclusions

This paper has presented a user report of AMtrainer and the PENTAtrainers. We aim to set a fair platform for further comparison between PENTA Model and AM Theory in modeling and predicting the F_0 contours. Both PENTAtrainer1 and PENTAtrainer2 reached an accuracy of predictive synthesis as high as r>0.9, showing that PENTA is an effective tool in F_0 modeling. The high accuracy achieved by AMtrainer reflects the effectiveness of sagging and linear interpolation as a means of contour generation. Meanwhile, the similarity between the results from AMtrainer and PENTAtrainers suggests that there is a potential for AMtrainer to predictively generate F_0 contours with functional and categorical input. But a more objective way of generating the input for AMtrainer is needed before a full comparison between the two theories is possible.

6. Acknowledgements

Prof. Carlos Gussenhoven kindly provided invaluable comments on an earlier draft of this paper. A part of Sections 3.2 and 3.3 was presented at the 3rd NINJAL International Conference on Phonetics and Phonology (ICPP3), Tokyo, December 2013. All errors and inaccuracies remain our own.

7. References

[1] Y. Xu, "Speech prosody: A methodological review," *Journal of Speech Sciences*, vol. 1, no. 1, pp. 85–115, 2011.

[2] S. Raidt, G. Bailly, B. Holm, and H. Mixdorff, "Automatic generation of prosody: Comparing two superpositional systems," in *Proc. of Speech Prosody 2004*, 2004.

[3] X. Sun, "The determination, analysis, and synthesis of fundamental frequency," PhD thesis, Northwestern University, 2002.

[4] J. B. Pierrehumbert and M. E. Beckman, Japanese Tone Structure. Cambridge, MA: MIT, 1988.

[5] J. B. Pierrehumbert, "The phonology and phonetics of English intonation," PhD thesis, MIT, 1980.

[6] Y. Xu, "Speech melody as articulatorily implemented communicative functions," *Speech Commun*, vol. 46, no. 3–4, pp. 220–251, Jul. 2005.

[7] Y. Xu and Q. E. Wang, "Pitch targets and their realization: Evidence from Mandarin Chinese," *Speech Commun*, vol. 33, no. 4, pp. 319–337, Mar. 2001.

[8] Y. Xu and M. Wang, "Organizing syllables into groups: Evidence from F0 and duration patterns in Mandarin," *J Phon*, vol. 37, no. 4, pp. 502–520, 2009.

[9] A. Lee, Y. Xu, and S. Prom-on, "Mora-based pre-low raising in Japanese pitch accent," in *Proc. of Interspeech 2013*, 2013, pp. 3532–3536.

[10] P. P. G. Boersma and D. J. M. Weenink, "Praat: Doing phonetics by computer." [Computer program]. Retrieved 18 Dec 2013 from http://www.praat.org/.

[11] S. Prom-on, Y. Xu, and B. Thipakorn, "Modeling tone and intonation in Mandarin and English as a process of target approximation," *JASA*, vol. 125, no. 1, pp. 405–424, 2009.

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

[13] J. B. Pierrehumbert, "Synthesizing intonation," *JASA*, vol. 70, no. 4, pp. 985–995, 1981.

[14] J. J. Venditti, "The J_ToBI model of Japanese intonation," in in *Prosodic typology: The phonology of intonation and phrasing*, S.-A. Jun, Ed. New York, NY: Oxford University Press, 2005, pp. 172–200.

[15] Y. Xu, "ProsodyPro: A tool for large-scale systematic prosody analysis," in *Proc. of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 2013, pp. O1–1.

[16] M. H. Quenouille, "Notes on bias in estimation," *Biometrika*, vol. 43, no. 3/4, pp. 353–360, Dec. 1956.

[17] E. O. Selkirk, T. Shinya, and S. Kawahara, "Phonological and phonetic effects of Minor Phrase length on F0 in Japanese," in *Proc. of Speech Prosody 2004*, 2004.

[18] T. Ishihara, "Tonal alignment in Tokyo Japanese," PhD thesis, University of Edinburgh, 2006.